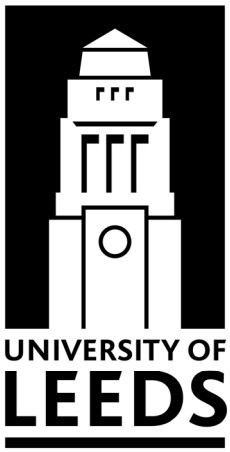
**Institute for Transport Studies**
**FACULTY OF THE ENVIRONMENT**




# DTA2006

First International Symposium
on Dynamic Traffic Assignment

JUNE 21ST-23RD 2006


Engineering and Physical Sciences
Research Council

Many transport measures envisaged for the future have a fundamentally dynamic element, ranging from real-time driver information to congestion charges affecting the temporal distribution of demand.

This symposium will be the first event to gather leading international researchers to evaluate progress with the formidable problem of modelling network dynamics to support the design and evaluation of such policies, and to identify the gaps and major challenges for future research efforts.

DTA2006 will bring together leading researchers in the field of Dynamic Traffic Assignment (DTA) over transport networks. The aim is to provide a stimulating but informal atmosphere in which to debate and identify key developments in the field and to identify future research possibilities. Perhaps, it can be hoped, it can also be a way of generating new collaborative research arrangements between attendees.

The hope is, of course, that this is the first in a series of such symposia; indeed, interest has already been expressed in hosting a follow-on event. However, no formal organising or decision-making body has been formed, and it is hoped that at the meeting we can debate the most effective way to continue the communication and collaboration, including the need for a standing organising committee. This first symposium is organised by a group of researchers from the University of Leeds, UK (see below). The impetus for our decision to organise DTA2006 has drawn from our involvement in several smaller scale initiatives on this topic, specifically a series of workshops involving UK and Japanese universities, a series of UK/Italian workshops, and a two-day workshop in Belfast in 2004 attracting a variety of participants. The International Scientific Committee for DTA2006 draws substantially on those that took a strong part in such events, and we acknowledge their efforts in laying the groundwork for a wider symposium. The impetus has also drawn from the communication we have with other leading researchers from around the world, and the ISC has also drawn on these contacts in its membership.

Financial sponsorship for the event is being provided by the UK Engineering and Physical Sciences Research Council.

**Organising Committee**
David Watling (chair; Leeds, UK)
Chandra Balijepalli (Leeds, UK)
Richard Connors (Leeds, UK)
Agachai Sumalee (Leeds, UK)

**International Scientific Advisory Committee**
David Boyce (Northwestern, USA)
Guilio Cantarella (Salerno, Italy)
Malachy Carey (Belfast, UK)
Terry Friesz (Penn State, USA)
Ben Heydecker (London, UK)
Masao Kuwahara (Tokyo, Japan)
Hong Lo (Hong Kong, China)
Hani Mahmassani (Maryland, USA)
Srinivas Peeta (Purdue, USA)
Mike Smith (York, UK)
Thanasis Ziliaskopoulos (Northwestern, USA; Thessaly, Greece)

# Delegate List

| Surname | First name | Affiliation | Email |
|---|---|---|---|
| Addison | Puff | Centre for Transport Studies, UCL | puff@transport.ucl.ac.uk |
| Balakrishna | Ramachandran | MIT | rama@mit.edu |
| Balijepalli | Chandra | ITS, University of Leeds | cbalijep@its.leeds.ac.uk |
| Bar-Gera | Hillel | Ben-Gurion University | bargera@bgu.ac.il |
| Bell | Michael | Imperial College London | m.g.h.bell@ic.ac.uk |
| Bie | Jing | Hong Kong Univ of Sci and Tech | jbie@ust.hk |
| Bliemer | Michiel | Delft University of Technology | m.c.j.bliemer@tudelft.nl |
| Blumberg | Michal | Ben-Gurion University | michal.nitzani@gmail.com |
| Cantarella | Giulio E. | University of Salerno, Italy | g.cantarella@unisa.it |
| Carey | Malachy | Queens University, Belfast | m.carey@qub.ac.uk |
| Chow | Andy | Centre for Transport Studies, UCL | andy@transport.ucl.ac.uk |
| Clark | Anna | ITS, University of Leeds | traacl@leeds.ac.uk |
| Clegg | Richard | University of York | richard@richardclegg.org |
| Connors | Richard | ITS, University of Leeds | R.D.Connors@its.leeds.ac.uk |
| Durlin | Thomas | ENTPE/INRETS, France | thomas.durlin@entpe.fr |
| Evans | Suzanne | BirkBeck College | s.evans@bbk.ac.uk |
| Florian | Michael | Universite de Montreal | mike@crt.umontreal.ca |
| Fox | Ken | Halcrow | foxk@halcrow.com |
| Friesz | Terry | The Pennsylvania State University | tfriesz@psu.edu |
| Gao | Song | Caliper Corporation | songgao@alum.mit.edu |
| Gbadamosi | Kolawole | Centre for Transport Studies, Olabisi Onabanjo | kt_bad@yahoo.com |
| Gentile | Guido | University "La Sapienza", Rome | guido.gentile@uniroma1.it |
| Golanski | Yann | NNDG, Mathematics, University of York. | yg2@york.ac.uk |
| Grebenc | Andrej | University of Ljubljana | agrebenc@s5.net |
| Heydecker | Benjamin | Centre for Transport Studies, UCL | ben@transport.ucl.ac.uk |
| Hicks | Jim | PB Consult | jhicks@pbconsult.com |
| Iryo | Takamasa | Kobe University | iryo@kobe-u.ac.jp |
| Juran | Ido | Technion, Israel | idoj@pgl.co.il |
| Kalafatas | Georgios | Purdue University | gkalafatas@gmail.com |
| Koh | Andrew | ITS, University of Leeds | a.koh@its.leeds.ac.uk |
| Kurauchi | Fumitaka | Kyoto University | kurauchi@urbanfac.kuciv.kyoto-u.ac.jp |
| Kuwahara | Masao | Tokyo University | kuwahara@iis.u-tokyo.ac.jp |
| Li | Minwei | Delft University of Technology, Netherlands | m.li@tudelft.nl |
| Liu | Henry | University of Minnesota | henryliu@umn.edu |
| Liu | Ronghui | ITS, University of Leeds | r.liu@its.leeds.ac.uk |
| Lo | Hong | Hong Kong Univ of Sci and Tech | cehklo@ust.hk |
| Maher | Mike | Napier University | m.maher@napier.ac.uk |
| Meschini | Lorenzo | University "La Sapienza", Rome | lorenzo.meschini@uniroma1.it |
| Mitsakis | Evangelos | Hellenic Inst. Of Transport,CRTH | emit@certh.gr |
| Mounce | Richard | Abu Dhabi University / University of York | richardmounce@hotmail.com |
| Nakayama | Shoichiro | Kanazawa University, Japan | snakayama@t.kanazawa-u.ac.jp |
| Noekel | Klaus | PTV AG | klaus.noekel@ptv.de |
| ozbay | kaan | Rutgers University, Dept.of Civil Envir Engr. | kaan@rci.rutgers.edu |
| Peeta | Srinivas | Purdue University | peeta@purdue.edu |
| Polak | John | Imperial College London | j.polak@imperial.ac.uk |
| Prashker | Joseph N. | Transportation Research Inst., Technion | prashker@netvision.net.il |
| Ramadurai | Gitakrishnan | Rensselaer Polytechnic Institute, Troy, NY, USA | ramadg@rpi.edu |
| Rosa | Andrea | Napier University | a.rosa@napier.ac.uk |
| Rydergren | Clas | Linköpings university | clryd@itn.liu.se |
| Schmoecker | Jan-Dirk | Imperial College London | j-d.schmoecker@ic.ac.uk |
| Shepherd | simon | ITS University of Leeds | s.p.shepherd@its.leeds.ac.uk |
| Smith | Mike | University of York | mike@st-pauls-square.demon.co.uk |
| SONG | Ziqi | The University of Hong Kong | ziqi@hkusua.hku.hk |
| Sumalee | Agachai | ITS, University of Leeds | A.Sumalee@its.leeds.ac.uk |
| Szeto | Wai Yuen | Transport Engineering, Trinity College Dublin | ceszeto@yahoo.com.hk |
| Taale | Henk | Delft University of Technology | h.taale@tudelft.nl |
| Tampère | Chris | Katholieke Universiteit Leuven | chris.tampere@bwk.kuleuven.be |
| Teklu | Fitsum | ITS, University of Leeds | tra2ft@leeds.ac.uk |
| Tuydes | Hediye | Middle East Technical University, Ankara, Turkey | htuydes@metu.edu.tr |
| Wainaina | Simon | JMP Consulting | simon.wainaina@jmp.co.uk |
| Waller | S. Travis | The University of Texas-Austin | stw@mail.utexas.edu |
| Watling | David | ITS, University of Leeds | d.p.watling@its.leeds.ac.uk |
| Xiang | Yanling | Atkins | yanling.xiang@atkinsglobal.com |
| Yanmaz-Tuzel | Ozlem | Rutgers University, Dept.of Civil Envir Engr. | ozlem.yanmaz@gmail.com |
| Yperman | Isaak | Katholieke Universiteit Leuven | isaak.yperman@bwk.kuleuven.be |
| Ziliaskopoulos | Athanasios | University of Thessaly and Northwestern University | ziliasko@uth.gr |

# DTA 2006: First International Symposium on Dynamic Traffic Assignment

**Centenary Gallery, Parkinson Building, University of Leeds, UK**
**June 21st–23rd 2006**

## *Programme*

### Wednesday 21st June

09:00 – 10:00 Registration

10:00 Welcome to DTA2006
Watling, Balijepalli, Connors & Sumalee

*Session 1: Properties of dynamic network loading models*

10:15 Analysis Of Dynamic Traffic Assignment
Heydecker & Addison

10:45 Analysis and comparison of macroscopic approaches to Dynamic Traffic Assignment
Cantarella, Carteni, de Luca & Punzo

11:15 Coffee break

*Session 2: Dynamic link models*

11:45 Dynamic Queuing & Spillback in Analytical Multiclass Dynamic Traffic Assignment Model
Bliemer

12:15 Multi-Commodity Dynamic Network Loading with Kinematic Waves and Intersection Delays
Tampere & Yperman

12:45 A Dynamic Network Loading Model Based on the Variations Solution Procedure
Blumberg & Bar Gera

1:15 Lunch

*1:45-2:45 Session 3: Poster session on Dynamic network loading*

Application of a New Dynamic Traffic Assignment Model for Assessment of Moving Bottlenecks
Juran, Prashker, Bekhor & Ishai

A Graph-Based Formulation For The Single Destination Dynamic Traffic Assignment Problem
Kalafatas & Peeta

The Enhanced Lagged Cell-Transmission Model
Szeto

Route Generation and Dynamic Traffic Assignment for Large Networks
Taale & Bliemer

Time and Space Discretization in Dynamic Traffic Assignment Models
Gentile, Nöekel & Meschini

*Session 4: Dynamic network loading*

2:45 – 3:45 Panel Discussion 1. Dynamic Network Loading: Which Models and What Level of Space/Time Discretisation are Appropriate?
Peeta (chair), Juran, Szeto, Taale, Gentile

3:45 Coffee break

*Session 5: Dynamic network assignment models*

4:15 A Simulation Based Dynamic Traffic Assignment Model: Dynameq
Florian, Mahut & Tremblay

4:45 Traffic Assignment On Networks With Time-Varying Flows, While Approximating Continuum Flows On Links
Carey

5:15 Close for the day. Participants make their own dinner arrangements.

≈≈≈≈≈≈≈≈≈≈≈≈≈≈
≈≈≈≈≈≈≈≈≈≈≈≈≈≈

## Thursday 22nd June

*Session 6: Formulation and Properties of DUE Models*

09:00 The Development of a Probit-Based Stochastic Dynamic Traffic Assignment Model
Maher & Rosa

09:30 Analysis of Dynamic System Optimum and Externalities with Departure Time Choice
Chow

10:00 Existence, Uniqueness, Stability and Bilevel Optimisation Of Dynamic Traffic Equilibria
Golanski, Mounce & Smith

10:30 Dynamic Non-cooperative Games as a Foundation for Modeling Dynamic User Equilibrium
Friesz, Mookerherjee & Kwon

11:00 Coffee break

*Session 7: DUE solution algorithms*

11:30 A Link-Node Complementarity Model and Solution Algorithm for Dynamic User Equilibria with Exact Flow Propagation
Ban, Liu, Ferris & Ran

12:00 Inner Approximation Algorithm for the Discrete Time Varying User Equilibrium Problem
Chang & Ziliaskopoulos

12:30 Descent Direction Based Solution Algorithm for DUE Assignment
Sumalee & Kuwahara

1:00 Lunch

*1:30-2:30 Session 8: Poster session on Deployment of DTA models*

Obstacles to Deployment of Dynamic Traffic Assignment Models
Ziliaskopoulos & Barratt
Observability In Estimating Time Dependent Origin Destination Flows From Traffic Counts
Balakrishna, Ben-Akiva & Wen
Dynamic Origin-Destination Matrix Estimation From Link Counts: An Approach Coherent With Dynamic Traffic Assignment
Durlin & Henn
Behavior-Consistent Deployable Traffic Routing under Information Provision
Paz & Peeta

*Session 9: Deployment of DTA models*

2:30 – 3:30 Panel discussion 2. Challenges to the Real-Life Deployment of DTA Models.
Florian (chair), Ziliaskopoulos, Balakrishna, Taale, Peeta, Waller

3:30 Coffee break

3:45 Coaches depart to take participants to Yorkshire Sculpture Park for …

- Philosophical session: Future of DTA? (details to follow)
- Discussion of, and vote on, future conference plans, dates, venues
- Conference Dinner

10:00 Coaches return to Leeds

≈≈≈≈≈≈≈≈≈≈≈≈≈≈
≈≈≈≈≈≈≈≈≈≈≈≈≈≈

## Friday 23rd June

*Session 10: Day-to-day dynamic models*

09:00 Modeling of Commuters' Day-to-Day Learning Behavior
Ozbay & Yanmaz-Tuzel

09:30 Doubly Dynamic Simulation Model for Traffic Assignment
Balijepalli, Watling & Liu

10:00 The Dynamic Assignment of Tours in Congested Networks with Pricing
Polak & Heydecker

10:30 Coffee break

*Session 11: Dynamic Congestion Pricing*

11:00 Day-to-day Congestion Pricing Policies towards System Optimal
Yang & Szeto

11:30 A Computable Theory of Dynamic Congestion Pricing
Friesz, Kwon & Mookerherjee

12:00 Temporal Externality in Dynamic User Equilibrium with Heterogeneous Travellers – Who Makes Congestion Worse?
Iryo

12:30 Lunch

*1:00 – 2:00 Session 12: Poster session on Alternative models and problems for DTA*

Stability Domains of Traffic Equilibrium: Directing Traffic System Evolution to Equilibrium
Lo & Bie
Equilibrium Dynamic Traffic Assignment with Adaptive Routing Choices
Gao
Stability of Network Flows with Bounded Rational Route Choice
Nakayama
Dynamic Simulation-Based Model of Urban Taxi Services
Song & Tong
Combining DTA Approaches for Studying Road Network Robustness
Li, Taale & Van Zuylen

*Session 13: Equilibrium versus non-equilibrium DTA*

2:00 – 3:00 Panel discussion 3. Non-Equilibrium And Other Alternatives To Dynamic User Equilibrium: What Potential Do They Offer?
Lo (chair), Cantarella, Gao, Li, Nakayama, Ramadurai

3:00 Coffee break

*Session 14: Public transport*

3:30 A First Approach to Dynamic Frequency-Based Transit Assignment
Schmöcker, Bell & Kurauchi

4:00 Towards a Holistic Frequency-Based Transit Assignment Model– The Stochastic Process Approach
Teklu

4:30 Closing session: Discussion of publication plans, future symposia

5:00 Conference closes

≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈
≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈

Updated 12 June 2006.

# ANALYSIS OF DYNAMIC TRAFFIC ASSIGNMENT

*BG Heydecker: University College London, England ben@transport.ucl.ac.uk*
*JD Addison: University College London, England puff@transport.ucl.ac.uk*

## Abstract
Dynamic road traffic assignment models can be used explicitly to represent and analyse route choice and travel times in networks for which the demand for travel varies over time and congestion arises. Static models cannot represent this phenomenon in a satisfactory manner. Despite the long-term research activity and the several approaches that have been introduced, no dominant approach has emerged to dynamic analysis. This seems to be due, at least in part, to the multiplicity of interrelated requirements on dynamic models for them to perform in a satisfactory. Here, we consider analytical approaches to dynamic traffic assignment, and in particular the extent to which they address the issues of realism that arise. This leads to restrictive requirements on traffic models for satisfactory dynamic behaviour. The wider role of assignment modelling is considered in the context of activity analysis, transport planning, and network design. We show how dynamic assignment can be developed together with departure-time choice to serve in these functions. We present these results as a contribution to the literature on dynamic assignment and to support further analysis and model development.

## 1 Introduction
Road traffic assignment models are used to provide estimates of flows in a network and the associated costs of travel. The main components of a model of this kind are the demand for travel, the network that is available for their use in travelling, and a modelling principle that describes the way in which they choose between the possibilities that are available to them. The purpose of a models of this kind is to estimate the likely state of a network under specified demands (either in the form of values or functions of costs), and to help evaluate changes to this demand, to the network, and to the way in which travellers exercise the choices that are available to them. For practical use, models of this kind should be self-consistent and solvable conveniently, yielding results that are reliable and repeatable, and that stand scrutiny.

When Wardrop (1952) proposed his two principles (user equilibrium and system optimal) of route choice, they were presented in verbal form. Within 4 years, Beckmann (1956) had established an equivalent convex mathematical programme for the case in which demand is constant over time. The solution for this is in the form of constant flows on routes through the network, with associated constant costs of travel. Satisfactory methods to calculate solutions for general networks based upon this took a further 10 to 20 years longer: after some 40 years, Patriksson (1994) presented a substantial review of this static case, showing that considerable interest and activity remained.

In networks where the demand for travel exceeds the capacity of the network, transient congestion will arise. Because this arises from the accumulation of traffic during the period of overload, static models cannot represent this phenomenon in a satisfactory manner. The topic of dynamic road traffic assignment has a history that is almost as long as that of its static counterpart. However, despite this long-term research activity and the several approaches that have been introduced, no dominant approach has emerged for dynamic analysis. This seems to be due, at least in part, to the multiplicity of interrelated requirements on dynamic models for them to perform in a satisfactory manner. By comparison, there are limited, if any, counterpart requirements in the static case.

In his review of 10 years ago, Patriksson wrote

> "So far, no well-founded dynamic models free from any serious anomaly such as instant propagation of some travellers, infinite cycling, failure to recognize the first-in-first-out principle, etc., have appeared, and their numerical solution most often rely on a time-discretization which brings the dynamic model into a (typically very large) static one."

In the present paper, we consider analytical approaches to dynamic traffic assignment that have been developed during the time since then, and in particular the extent to which they address the issues that Patriksson and others have raised. By analysis of the equilibrium condition in the dynamic context, this identifies necessary conditions on traffic flows for an assignment to remain in equilibrium over time. Central to this discussion is exploration of ways in which formulation of dynamic traffic assignment goes beyond a sequence of static assignments.

We consider dynamic traffic assignment as a process in continuous time, and perform much of the present analysis on this basis. This enables us to call on the calculus in our analysis. However, as a practical matter, numerical solution methods usually work in discrete time. Managing the discrete time solution of a continuous time process brings with it technical interest, and leads to certain insights into the dynamic formulation. This analytical approach can be used to inform model specification and formulation, solution approaches, and certain aspects of simulation approaches.

Further analysis of dynamic traffic assignment leads to restrictive requirements on traffic models for satisfactory dynamic behaviour, which can be used to eliminate all but a few candidates. This analysis has been used to support the development of various solution approaches that can be used to calculate dynamic traffic equilibria in networks of practical size. The wider role of assignment modelling is considered in the context of activity analysis, transport planning, and network design. We show how dynamic assignment can be developed together with departure-time choice to serve in these functions. We present these results as a contribution to the literature on dynamic assignment and to support further analysis and model development.

# 2 Traffic performance models

## 2.1 Introduction

Dynamic traffic assignment represents the response of travellers to variation over time in the travel conditions between which they choose. Some of the variations in travel time involved in this result from congestion, thus forming an interaction between choice and its consequences to be equilibrated in some way. In view of the importance of the time variation in travel time in dynamic assignment, we consider first the influences on travel times, most important amongst which is the effect of variations in traffic on travel times. These traffic models play a fundamental role in dynamic assignment, and can be identified with some of the issues that Patriksson raised with early approaches to analysis of this kind.

## 2.2 Link representation

We consider first a single section of road in the form of a unidirectional link that connects two nodes of a network. We consider elementary requirements on the model relationship that is used for the dynamics of traffic on links of this kind, and show that several problematic issues are associated with this.

We suppose that there is a capacity $Q$ for the link, which is the least upper bound on the rate at which traffic can leave it, and also that there is a greatest lower bound $\phi$ on travel time, which obtains in free-flow conditions. We denote as $e(t)$ the rate at which traffic enters the link at time $t$, and as $g(t)$ the rate at which traffic leaves it at that time. The amount $x(t)$ of traffic present on the link at time $t$ can be used to represent the state of the link. The requirement of conservation of traffic on each link $a$ at time $t$ can be expressed as

$$\dot{x}(t) = e(t) - g(t) \qquad \forall t \tag{1}$$

where $\dot{x}$ denotes the derivative of $x$ with respect to time. Clearly we require that each of the primary quantities non-negative, so that

$$\left.\begin{aligned} e(t) &\geq 0 \\ g(t) &\geq 0 \\ x(t) &\geq 0 \end{aligned}\right\} \qquad \forall t \tag{2}$$

Although there is no requirement on the sign of $\dot{x}$, the requirement in (2) that $x$ be positive is necessary and is independent of the other requirements.

### 2.3 Travel time and outflow

The role of a link traffic model is to calculate, for a specified temporal inflow profile, what the resulting travel time and temporal outflow profile will be. Here, we denote as $\tau(s)$ the time of egress associated with entry to the link at time $s$, so that the travel time is $\tau(s) - s$. Because of conservation of traffic, the outflow $g(t)$ is closely associated with this travel time.

We adopt at this point the first-in first-out (FIFO) condition, which specifies that the time $\tau(s)$ at which traffic leaves a link increases with the time $s$ of entry to it, so that

$$t \geq s \quad \Rightarrow \quad \tau(t) \geq \tau(s). \tag{3}$$

When $\tau(s)$ is differentiable, this leads to

$$\dot{\tau}_a(s) \geq 0. \tag{4}$$

Where this FIFO condition is interpreted as applying to variations in the expected time of egress of an individual vehicle according to its time of entry, it seems to be entirely reasonable. However, if it is applied to the times of egress of different vehicles, then it precludes overtaking which can be unduly restrictive, especially where vehicles of distinct kinds are present. The importance of the FIFO condition is that it provides a clear reference for the analysis of traffic models that corresponds to a clearly defined, if special, case. This can be used as a point of comparison for more elaborate analysis where that is required.

The FIFO condition can be used, following Astarita (1996), and Heydecker and Addison (1996) to establish an important interrelationship between the time $\tau(s)$ of egress, and the entry and exit flows associated with a link. The FIFO condition (3) means that the amount of traffic that has exited from a link before time $\tau(s)$ at which a vehicle that entered at time $s$ does is exactly equal to the amount of traffic that had entered the link before time $s$. Thus we have

$$\int_{s'=-\infty}^{s} e(s') \quad ds' \quad = \quad \int_{t'=-\infty}^{\tau(s)} g(t') \quad dt'. \tag{5}$$

Differentiating this with respect to time of entry $s$, and using the fundamental theorem of calculus and the chain rule of differentiation gives

$$e(s) \quad = \quad g\left[\tau(s)\right] \quad \dot{\tau}(s) \tag{6}$$

This expresses the relationship between variations in travel time and those in the rate of flow at fixed points along the trajectory that is followed by a vehicle: it therefore describes the way in which flow propagates along the links of a network. According to this relationship, any function that specifies the time of egress from a link can be used to determine the rate of outflow at that time, and conversely a function that specifies the rate of outflow determines the time of egress implicitly. Furthermore, this flow propagation condition can be used together with the FIFO condition (4) to recover one or other of the positivity conditions in (2) on the entry and exit flows. However, if the FIFO condition fails, then one or other of the flow propagation condition (6) and the positivity requirement on outflow $g$ must fail.

A distinction arises between the state equation (1), which expresses conservation of traffic on a link, and the flow propagation equation (6), which expresses the relationship between variations in travel time and the outflow. The state equation is definitional and applies at an instant, whereas the flow propagation equation depends on the FIFO principle and applies along a vehicle trajectory.

### 2.4 Requirements on traffic models

Several requirements arise on traffic models for use in dynamic traffic assignment. Most of the issues raised by Patriksson, as well as various others, are associated with this component, and can be solved by choosing between possibilities accordingly. These requirements can be summarised as follows.

| | |
|---|---|
| Positivity: | $e(s) \geq 0 \; \forall s \quad \Rightarrow \quad x(t) \geq 0, g(t) \geq 0 \quad \forall t$ |
| Conservation: | $\dot{x}(t) = e(t) - g(t) \qquad \forall t$ |
| FIFO: | $t \geq s \quad \Rightarrow \quad \tau(t) \geq \tau(s) \quad \forall s, \forall t$ |
| Minimum travel time: | $\exists \phi > 0 \; \forall s \; \tau(s) - s \geq \phi$ |
| Finite clearing time: | $\exists S \; \forall s \; s \leq S \Rightarrow \; \tau(s) \leq S + \phi$ |
| Capacity: | $\exists Q \; \forall t \;\; g(t) \leq Q$ |
| Causality: | $\forall s, \forall t \;\; t \geq s \quad \Rightarrow \quad \tau(s)$ is not affected by $e(t)$. |

We note that correct flow propagation, and the consequent interrelationship between travel time and link outflow according to (6) follows from these conditions. The minimum travel time ensures that no flows propagate instantaneously and that infinite cycling is not possible in equilibrium, whilst the finite clearing time ensures that no travellers remain on the network indefinitely and that it returns to free-flow conditions after the study period. The causality requirement ensures that response follows stimulus. Thus many of the shortcomings that have been encountered in dynamic traffic assignment models can be ascribed to the traffic models that have been adopted, and can be avoided by choosing traffic models appropriately.

### 2.5 Choice of link traffic model

The requirements on models of traffic behaviour expressed in section 2.4 can be used to eliminate most candidates from use in dynamic traffic assignment. Hurdle (1986) Daganzo (1995), Astarita (1996), Heydecker and Addison (1998) and Mun (2001) have all contributed to this process. Thus, for example, the outflow functions proposed by Merchant and Nemhauser (1978) in their pioneering work specify that $g(t)$ is determined by $x(t)$, but this violates the requirement of causality, as noted by Hurdle (1986) and Daganzo (1995) amongst others. Daganzo (1995) showed that exit time functions should not depend on inflow or outflow at that time, though Carey, Ge and McCartney (2003) have developed a satisfactory formulation in which exit time depends on the outflow at the time of egress.

Friesz, Bernstein, Smith, Tobin and Wie (1993) considered a model form in which the exit time depends on the state at the instant of entry, so that

$$\tau(s) = s + \phi + d\left[x(s)\right] \tag{7}$$

where $d$ is the delay incurred. Mun (2001), and Nie and Zhang (2005) have shown that the only satisfactory form for the delay function $d(.)$ in respect of the requirements listed above (notably FIFO) is linear, in which case $d(x) = x/Q$, which gives a model that behaves satisfactorily.

Mun extended this affine travel time concept by subdividing each link into a freely flowing section followed by a capacitated one, with free-flow travel times given respectively by $\phi - \Delta t$ and $\Delta t$ respectively: Mun's canonical choice for the free-flow time $\Delta t$ of the capacitated part of the link is the time increment of solution. In the case that $\Delta t = \phi$ this corresponds to Friesz's model (7) whilst in the limit as $\Delta t \to 0$, this corresponds to the deterministic queue.

Other traffic models that are known to have satisfactory properties include Lighthill and Whitham's (1955) kinematic wave model. However, solution of this is computationally demanding, though convenient solution methods have been developed for simplified versions of this (see, for example, Daganzo, 1994).

From this analysis, we see that acceptable traffic models for use in dynamic traffic assignment are scarce. This contrasts with the case of static assignment, in which satisfactory properties can be established for broad classes of cost functions – Smith (1979) showed that monotonicity is adequate for this.

# 3 Network Loading

## 3.1 Introduction

The next topic for discussion is that of how the temporal profile of flows and travel times throughout a network can be calculated from specified inflows onto routes. This process is known as dynamic network loading, and is typically undertaken many times during the solution of a dynamic assignment. Analysis of this process calls upon that of link traffic models introduced in Section 2. Here, we show how relevant quantities can be calculated for the network using the link model, and derive certain network properties.

## 3.2 Multi-commodity flows

Each link of a network can appear on several distinct routes and will in general be used by traffic of several different kinds (in particular, traffic travelling to distinct destinations); we classify traffic in this sense as separate commodities. We suppose that the travel time associated with use of each link is calculated according to the whole state of the link, possibly making allowance for differing characteristics of various kinds of traffic, and then applies equally to all traffic entering it at the same time. Thus let $e_a^c(s)$ be the rate at which traffic of commodity $c$ enters link $a$ at time $s$, and let $g_a^c(t)$ be the corresponding commodity-specific link outflow at time $t$. By definition, the total link inflow $e_a(s)$ at time $s$ can be found as

$$e_a(s) = \sum_c e_a^c(s)$$

The associated time $\tau_a(s)$ of egress from that link can then be calculated according to the condition (5) from the traffic model for that link using the temporal profile of this total inflow.

A direct consequence of this is that according to this, FIFO behaviour will be respected between different traffic commodities using the link. Multi-commodity FIFO requires that the resulting time of egress, and hence also its derivative, apply equally to all traffic irrespective of commodity $c$. Using (5), the commodity-specific link outflows $g_a^c(t)$ are then given by

$$g_a^c\left[\tau_a(s)\right] = \frac{e_a^c(s)}{\dot{\tau}_a(s)} \qquad \forall c, \quad \forall a, \quad \forall s. \tag{8}$$

This result, which was noted by Papageorgiou (1990) shows how the traffic model, which applies to all the traffic on a link irrespective of its origin, destination or any other classification, can be used directly to calculate the correct traffic propagation and hence link outflows for the different traffic commodities in the network. If travel times and flow propagation are calculated according to this, traffic of one commodity cannot then overtake that of another. This avoids by construction problems of the kind identified previously by Carey (1992) that arise when the value of an objective function is improved by promoting one segment of traffic and holding back another.

## 3.3 Flows at Nodes

Analysis of flows at node in a network can be undertaken by appealing to conservation of flows. Here, we treat zone centroid connectors in the same way as other links, so do not distinguish them. First, we note that the flow of each flow commodity $c$ is conserved at each node $n$, so that

$$\sum_{b \in B(n)} g_b^c(t) = \sum_{a \in A(n)} e_a^c(t)$$

where $B(n)$ is the set of all links that lead into node $n$, and $A(n)$ is the set of all links that lead from node $n$. Because this applies to each commodity separately, it also applies to the total flows entering and leaving a node.

### 3.4 Calculation of costs

The cost $C_p(s)$ of travel that is associated with use of path $p$ starting at time $s$ can be calculated from the cost of using each link on the path at the time it will be reached by following a vehicle trajectory. This can be expressed using the time $\tau_{ap}(s)$ at which the entry to link $a$ will be reached by a traveller who sets out on path $p$ at time $s$. Thus

$$C_p(s) = \sum_{a \in p} c_a\left[\tau_{ap}(s)\right]. \tag{9}$$

This process of calculating costs according to the time of entry to each link corresponds to an ideal model of travel cost: it represents the cost that would be incurred by a traveller following that route. Because the partial travel times along the paths correspond to components of the cost, it is also known as a *nested cost operator* (Wie, Tobin, Friesz and Bernstein, 1995).

## 3 Models of demand for travel

### 4.1 Introduction

We now consider various ways in which the temporal profile of origin-destination flows, which we denote as $\mathbf{T}(s)$ can arise. It is these flows that are assignment at each instant $s$ to the paths through the network, so that

$$\sum_{p \in P_{od}} e_p(s) = T_{od}(s) \quad \forall s, \forall od$$

where $P_{od}$ is the set of all paths from origin $o$ to destination $d$.

### 4.2 Fixed exogenous origin-destination flows

The first case that we consider is that the flows $\mathbf{T}$ are specified exogenously. This has the merit that the assignment calculations are performed for origin destination inflows that are fixed, so that the only dimension of choice to be resolved is that of route. However, formulations of this kind have the disadvantage that a complete temporal profile of travel demands is required for the whole network, which will present practical difficulties of data identification and collection. In this case, the costs of travel between each origin-destination pair will usually vary through the study period as there is no mechanism for equilibrating them.

### 4.3 Departure time choice

The dynamic traffic assignment formulation can be extended to include choice of departure time as well as that of route. This has extension the merit that the temporal profile of flows $\mathbf{T}(s)$ is endogenous in this case, which obviates collection of specific data. However, in order for the interval during which travel takes place to be located, this requires time-varying incentives to be associated with travel. Thus by comparison with the case of exogenous flows, the representation of the reason for travel rather than the amount itself is exogenous. In this case, the costs of travel between each origin-destination pair will be equilibrated as part of the choice process that is represented in the model.

### 4.4 Elastic demand

In the case that choice of departure time as well as that of route is considered, the total volume of travel between each origin-destination pair is required either exogenously or endogenously to the model calculations. For each value of the volume of travel between the origin-destination pairs in the network, a certain cost of travel will arise that is equilibrated in some way through the choice processes that are represented. If the total volume is specified exogenously, then part of the solution process will be to determine an appropriate interval during which that amount of travel can take place through the network according to the behavioural assumptions of the model, and the associated cost will be determined by this. On the other hand, the relationship between the equilibrated cost and the volume of travel can be used to represent a performance characteristic of the network. This can be used together with an appropriate demand function to give a formulation that includes supply-demand in which the total volume of travel becomes endogenous.

# 5 Analysis of assignment

## 5.1 Introduction

Having identified satisfactory frameworks for approaches to modelling each of network performance and demand for travel, we now consider the interaction between them that is represented in the form of an assignment principle. Resolution of this interaction between flows loaded onto a network according to a choice principle and the costs that arise as a consequence corresponds to solution of assignment. The issues that we address here in turn are the assignment principle, the scope and interpretation of the choice process, and mathematical formulations required for model representation.

## 5.2 Assignment principles

Wardrop's user equilibrium principle of route choice can be extended to the dynamic form that

> At each instant, the costs incurred by travellers on those routes that
> are used are equal and no more than those on any unused route.

This can be expressed in complementary inequality form after Beckmann (1956) as

$$e_p(s) \left\{ \begin{array}{l} >0 \Rightarrow \tilde{C}_p(s) = \tilde{C}^*_{od}(s) \\ =0 \Rightarrow \tilde{C}_p(s) \geq \tilde{C}^*_{od}(s) \end{array} \right\} \qquad \forall p \in P_{od}, \quad \forall od, \quad \forall s \tag{10}$$

where $\tilde{C}_p(s)$ is the cost incurred in using route $p$ starting at time $s$, and $\tilde{C}^*_{od}(s)$ is the minimum cost of travel from $o$ to $d$ starting at time $s$.

Proceeding after Heydecker and Addison (1996), we consider the rate of change with time of the cost of travel on routes that are in use at a certain time. Differentiating the first case of (10) with respect to time $s$ gives

$$e_p(s) > 0 \quad \Rightarrow \quad \frac{d\tilde{C}_p}{ds} = k_{od}(s) \qquad \forall p \in P_{od}, \quad \forall od, \quad \forall s \tag{11}$$

where $k_{od}(s) = \dfrac{d\tilde{C}^*_{od}}{ds}$ is the common rate of change of costs for all routes in use between origin $o$ and destination $d$ at time $s$.

We now suppose that the cost of travel on each route $p$ consists of the sum of the travel time $\tau_p(s) - s$ and a constant part, denoted as $\beta_p$, which can be used to represent distance and any special features. Without loss of generality, we can express the constant part in units of equivalent travel time so that the cost of travel on route $p$ starting at time $s$ is $\tilde{C}_p(s) = [\tau_p(s) - s] + \beta_p$. Differentiating this with respect to the start time $s$, using the flow propagation relationship (6) to eliminate $\dot{\tau}_p(s)$ and rearranging gives the necessary condition for equilibrium

$$e_p(s) = \frac{g_p[\tau_p(s)]}{\sum_{q \in P_{od}} g_q[\tau_q(s)]} \; T_{od}(s) \qquad \forall p \in P_{od}, \quad \forall od, \quad \forall s \tag{12}$$

where the case of zero inflow can now be included because the corresponding outflow will also be zero.

Although the flows on the right-hand side of (12) occur at time $\tau_p(s)$, which is later that the time $s$ of the path inflow that appears on the left-hand side, in a causally determinate traffic model, the value of the right-hand side is determined by inflows before time $s$. However, we note that in the case of outflow function of the form proposed by Merchant and Nemhauser, the right-hand side is determined by inflows strictly after time $s$, with the result that the equilibrium assignment at time $s$ is determined by later inflows rather than by earlier ones. A consequence of this is that with link traffic models of that kind, calculation of equilibrium assignments must be undertaken in reverse order of time.

As an alternative to the deterministic equilibrium principle (10), Dial (1971) adopted an approach that assigns some traffic to each reasonable path between each origin-destination pair, with decreasing amounts as costs incurred increase. Dial's rule for the assignment of traffic to each path $p$ in the set of reasonable ones $P_{od}$ can be extended to the dynamic case in the form

$$e_p(s) = \frac{\exp\left(-\theta C_p(s)\right)}{\sum_{q \in P_{od}} \exp\left(-\theta C_q(s)\right)} T_{od}(s) \qquad \forall\, p \in P_{od}\,, \forall\, od\,,\ \forall\, s \tag{13}$$

where $\theta$ is a positive parameter that controls dispersion between routes. This formulation has the merit that the assignments are continuous functions of the costs. When exit time functions of the form (7) are used, the costs of travel are continuous in the state variable $x$. In turn, according to (1) the state variable is continuous in time, so that the assignments are continuous in time. This has continuity has practical consequences for the convenient calculation of dynamic assignments according to (13).

### 5.3 Scope of the choice process

As discussed in section 4.3, the dimensions of choice can be extended from that of route alone to include departure time. Thus travellers can select the combination of departure time and route at that time in order most conveniently to meet their travel requirements. In order for this extended formulation to be complete, some time varying incentive is required to localise travel in the time domain. Initially, we suppose that the is a monotonically decreasing cost $h_o(s)$ associated with time $s$ of departure from origin $o$, and a monotonically increasing cost $f_d(t)$ associated with time $t$ of arrival at the destination $d$. Thus the cost of using route $p$ to travel from origin $o$ to destination $d$ departing at time $s$ is $C_p(s) = h_o(s) + \tau_p(s) - s + \beta_p + f_d\left[\tau_p(s)\right]$.

This formulation associates some benefit with time spent at these locations, contrasting with the cost that is associated with time spent travelling. The equilibrium principle in this case is that the total cost associated with travel, including the origin and the destination costs, is minimal and hence constant whenever travel takes place. An immediate consequence of this is that, by contrast with the case in which the departure profiles are exogenous, in this case there is a single cost associated with travel for each volume of travel between an origin-destination pair. This cost can then be used as part of a more extensive transport model that includes further dimensions of choice such as various kinds of elastic demand, modal split and distribution of trips. The effect of including departure time choice together with route choice in a dynamic traffic assignment model is that it then provides a choice-based network performance measure that relates unambiguously the costs of travel to the volumes that are assigned.

The dynamic equilibrium condition (12) of route choice still applies to this extended formulation because equilibrium route choice remains part of it. However, in this case, the rate of change of arrival time, $\dot{\tau}_p(s)$ for each route can be identified (Heydecker and Addison, 2005) using derivatives of the origin and destination-specific cost functions. Thus the equilibrium assignment in this case satisfies

$$e_p(s) = \left[\frac{1 - h_o'(s)}{1 + f_d'\left[\tau_p(s)\right]}\right] g_p\left[\tau_p(s)\right] \quad \forall p \in P_{od},\ \forall od\,,\ \forall s\,. \tag{14}$$

Similarly, Dial's stochastic choice principle can be extended to include the dimension of departure time choice, but in this case using the expected minimum cost at each departure time as a representative cost of travelling at that time. Thus the choice between departure times can be represented as being made according to a continuous logit model using the cost criterion:

$$\breve{C}_{od}(s) = h_o(s) + \frac{1}{\theta} \log_e \sum_{p \in P_{od}} \exp\left[-\theta\left\{C_p(s) + f_d\left[\tau_p(s)\right]\right\}\right]. \tag{15}$$

When departure time choice is introduced into dynamic assignment models, an interesting possibility arises to interface the resulting travel model with analysis of activities at the origin an destination. In this case, the functions $h(.)$ and $f(.)$ can be interpreted as representing the flow of utility associated with time spent (respectively) at the origin and the destination of the journey. In this case, departure form an origin to travel to a destination is induced when the benefit of reaching the destination outweighs that of remaining at the origin: the cost of travel is bourn to achieve this increase in benefit. This approach has been developed by Polak and Heydecker (2006).

### 5.4 Mathematical formulations

In simple cases, the necessary conditions (12) and (14) are sufficient to calculate dynamic equilibrium assignments (respectively, with departure time choice) in closed form: the equilibrium route inflows are expressed in terms of quantities that can be determined from earlier inflows. However, in general networks, more versatile procedures are required. Several approaches have been developed for this, which we note here.

A crucial point in developing numerical methods for solution of dynamic assignments is in the transition from the continuous time formulation of the equilibrium conditions (10) to a discrete time formulation for solution. This is because a the continuous time formulation applies at an instant, whilst a numerical solution method will typically assign a calculated flow $e_p(s)$ to a path $p$ throughout a time increment $[s, s+\Delta s)$. In order of the effect of the flows assigned to influence the costs that are used in determination of the equilibrium criterion, these costs should then be those incurred by a traveller departing at time $s+\Delta s$. Thus supposing the network to be in equilibrium at time $s$, we seek assignments $\mathbf{e}(s)$ to the paths through a network that will apply throughout the incremental time interval $[s, s+\Delta s)$ that achieve equilibrium at time $s+\Delta s$. If costs at time $s$ are used instead, then the resulting assignments will be all or nothing, using only the path that had least cost for each origin-destination pair at time $s$.

This approach can be applied to a mathematical programming formulation based upon Beckmann's (1956) objective function (Han and Heydecker, 2006). In this case, the objective $Z(s)$ that is minimised is calculated at each incremental time interval using the flow $\mathbf{e}(s)$ that is assigned throughout that increment together with the costs $\mathbf{c}(s+\Delta s)$ at the end of it. Although this approach leads to tractable analysis, it has not yet been applied to effectively to substantial networks.

By contrast, the variational inequality formulation developed initially by Smith (1979) and Dafermos (1980) provides a practical approach to calculation of dynamic traffic assignments within the present framework. This was introduced for dynamic traffic assignment by Friesz, Luque, Tobin and Wie (1993) and has been adopted widely since then. It is convenient in this formulation to adopt a path-based expression of the variational inequality, though this requires evaluation of path cost along trajectories according to (9). Thus we seek assignments $\mathbf{e}(s)$ for the incremental time interval $[s, s+\Delta s)$ leading to path costs $\mathbf{C}(s+\Delta s)$ at the end of it that satisfy the variational inequality

$$\left[\mathbf{v}-\mathbf{e}(s)\right]^T \cdot \mathbf{C}(s+\Delta s) \geq 0 \quad \forall\, \mathbf{v} \in D(s)$$

where $D(s)$ is the set of demand-feasible assignments at time $s$. The alternative approach of a link-based formulation has the advantage of more convenient cost calculation for each individual link, but instead requires calculation of link flows according to the flow propagation relationship (6) applied to path flows departing at some earlier time.

The calculation of dynamic stochastic assignments according to Dial's rule (13) is more convenient than that of deterministic assignments (See, for example, Lim and Heydecker, 2005). This is because of the continuity of the assignments in time that was noted in section 5.2. A practical procedure in this formulation is to calculate assignments $\mathbf{e}(s)$ for use during the incremental time interval $[s, s+\Delta s)$ according to the costs $\mathbf{c}(s)$ at the start of the interval. Although this gives rise to some error because the costs are outdated, continuity of the assignment guarantees that the error will tend to zero as the time increment $\Delta s$ does.

Finally, we consider the case of dynamic system optimal assignment, in which the objective is to minimise the total cost associated with travel: $\sum_{od}\sum_{p\in P_{od}}\int_{s} C_p(s)\,e_p(s)\,ds$ . This objective was adopted by Merchant and Nemhauser, and by several other authors since then. If an elastic demand formulation is adopted, then the objective of maximising travellers' surplus is appropriate, though this requires use of some composite measure of travel cost in order to identify the level of demand. When the costs of travel depend on the state $\mathbf{x}(s)$, dynamic system optimal assignment can be formulated as an optimal control problem (see, for example, Friesz, Luque, Tobin and Wie, 1989), though the state-dependent time lags give rise to technical issues. In this case, the formulation can be reduced to an equivalent dynamic user-optimal assignment formulation in which additional components of cost are introduced for cost externalities caused to each of other traffic travelling at the same time and future traffic (see, for example, Chow, 2006).

# 6 Conclusions

The broad view of dynamic traffic assignment that has been presented in this paper has identified several requirements on the components of the models that are used in order for the model to have satisfactory behaviour in all respects. Whilst the formulation of dynamic traffic assignment in continuous time appears to differ from its static counterpart only in that flows and costs are qualified by time, implementation of this in a satisfactory model requires that several distinct dynamic characteristics be treated adequately. In summary these are:

**Development of traffic state**: the accumulation of traffic in parts of the network during periods of overload and its dissipation afterwards should be modelled correctly, observing conservation of traffic.

**Appropriate traffic models**: several requirements arise for the models that are used to represent the dynamics of traffic on links of the network. These include the first-in first-out (FIFO) discipline, and causality, which together restrict the choice of link traffic models that are suitable for this work to relatively few.

**Flow propagation**: the relationship between variation in flows along a vehicle trajectory and variation in travel time that is consequent upon FIFO discipline together with conservation should be respected in the structure of the model.

**Costs of travel**: the costs associated with following a path through the network should be calculated along vehicle trajectories.

**Predictive framework**: in deterministic equilibrium assignments, flows should be calculated incrementally in time. The costs that are associated with flows during these time increments should be ones that are influenced by them, and hence those arising at the late end of the time increment.

Provided that each of these aspects of model formulation is addressed, the way is now open for the development of practical implementations of analytical dynamic traffic assignment that are well founded and can be solved with reasonable computational effort, giving results that stand scrutiny. The models developed in this way should serve reliably in the role of the assignment component of the transport modelling process, and should interface appropriately with other modelling components including the traditional ones of choice of mode, destination and frequency as well as novel ones of activity-based modelling.

# Acknowledgements

The authors are grateful to Richard Allsop for his encouragement and continued interest in the work presented here, and to Yongtaek Lim, Sangjin Han, Jinsu Mun and Andy Chow for their many stimulating discussions on these topics. The work presented here was supported by the EPSRC.

## References


Astaria, A (1996) A continuous time link model for dynamic network loading based on travel time functions. J-B Lesort, ed. *Transportation and Traffic Theory*. Pergamon, Oxford, 79-102.

Beckmann, M, McGuire, C and Winsten, CB (1956) Studies in the Economics of Transportation. New Haven: Yale University Press.

Carey, M (1992) Nonconvexity of the dynamic traffic assignment problem. *Transportation Research*, **26B**, 127–33.

Carey, M, Ge, YE and McCartney, M (2003) A whole-link travel-time model with desirable properties. *Transportation Science*, **37**(1), 89-96.

Chow, AHF (2006) Analysis of dynamic system optimum and externalities with departure time choice. *Proceedings* of the First International Symposium on Dynamic Traffic Assignment, University of Leeds.

Dafermos, SC (1980) Traffic equilibrium and variational inequalities. *Transportation Science*, **14**(1), 42–54.

Daganzo, CF (1994) The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research*, **28B**(4), 269-87.

Daganzo, CF (1995) Properties of link travel time functions under dynamic loads. *Transportation Research* **29B**(2), 95-98.

Dial, RB (1971) A probabilistic multipath traffic assignment model which obviates path enumeration. *Transportation Research*, **5**(2), 83-111.

Friesz, TL, Luque, J, Tobin, RL and Wie, BY (1989) Dynamic network traffic assignment considered as a continuous time optimal control problem. *Operations Research*, **37**(6), 179-91.

Friesz, TL, Bernstein, D, Smith, TE, Tobin, RL and Wie, BY (1993) A variational inequality formulation of the dynamic network user equilibrium problem. *Operations Research*, **41**, 179-91.

Han, S-J and Heydecker, BG (2006) Consistent objective and solutions of dynamic user equilibrium models. *Transportation Research*, **40B**(1), 16-34.

Heydecker, BG and Addison, JD (1996) An exact expression of dynamic traffic equilibrium. J-B Lesort, ed. *Transportation and Traffic Theory*. Pergamon, Oxford, 359-83.

Heydecker, BG and Addison, JD (1998) Analysis of traffic models for dynamic equilibrium traffic assignment. MGH Bell, ed. *Transportation Networks: Recent methodological Advance*. Oxford: Pergamon, 35-49.

Heydecker, BG and Addison, JD (2005) Analysis of dynamic traffic equilibrium with departure time choice. *Transportation Science*, **39**(1), 39-57.

Hurdle, VF (1986) Technical note on a paper by Andre de Palma, Moshe Ben-Akiva, Claude Lefevre, and Nicolaos Litinas entitled "stochastic equilibrium model of peak period traffic congestion," *Transportation Science* **20**(4), 287-9.

Lighthill, MJ and Whitham, GB (1955) On kinematic waves: II. A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society* **229A**, 317-45.

Lim, Y and Heydecker, BG (2005) Dynamic departure time and stochastic user equilibrium problem. *Transportation Research*, **39B**(2), 97-118.

Merchant, DK and Nemhauser, GL (1978) A model and an algorithm for the dynamic traffic assignment problem. *Transportation Science* **12**(3), 183-99.

Mun, J-s (2001) A divided linear travel time model for dynamic traffic assignment. *Proceedings* of the 9th World Conference on Transport Research, Seoul, **D1-05**, 4189.

Nie, X and Zhang, HM (2005) A comparative study of some macroscopic link models used in dynamic traffic assignment. *Networks and Spatial Economics*, **5**(1), 89-115.

Papageougiou, M (1990) Dynamic modelling, assignment and route guidance in traffic networks. *Transportation Research*, **24B**(6), 471-95.

Polak, JW and Heydecker, BG (2006) The dynamic assignment of tours in congested networks with pricing. *Proceedings* of the First International Symposium on Dynamic Traffic Assignment, University of Leeds.

Patriksson, M (1994) The traffic assignment problem: models and methods. Utrecht: VSP.

Smith, MJ (1979) Existence, uniqueness and stability of traffic equilibria. *Transportation Research*, **13B**, 295-304.

Wardrop, JG (1952) Some theoretical aspects of road traffic research. *Proceedings* of the Institute of Civil Engineers, **Part II**, 325-78.

Wie BW, Tobin, RL, Friesz, TL and Bernstein, D (1995) A discrete time, nested cost operator approach to the dynamic network user equilibrium problem. *Transportation Science*, **29**(1), 79-92.

# ANALYSIS AND COMPARISON OF MACROSCOPIC APPROACHES TO DYNAMIC TRAFFIC ASSIGNMENT

*G.E. Cantarella: University of Salerno, Italy (EU)* g.cantarella@unisa.it

## Abstract
The assumption of a within-day stationary transportation system leads to well established static assignment models based on a linear synchronic network model. These models can quite easily be extended to deal with transportation systems with discrete service through diachronic linear network models. On the other hand, generalization to within-day dynamics is not straightforward at all for transportation systems with continuous service, since in this case the demand model is slightly affected but the network model is highly non-linear, possibly involving differential equations. This paper will presents a general analysis of macroscopic approaches to within-day supply models (for transportation systems with continuous service), leading to within-day Dynamic Traffic Assignment when combined with a demand model.

## 1 Introduction
A Traffic Assignment model may be broken down into:

- A SUPPLY MODEL, which expresses how route flows affects route costs, it is made up by:
  - a link TRAFFIC FLOW MODEL to simulate congestion, that is how the number of users (flow, density) on a link affects link travel times and (possibly different) transportation costs;
  - a network model of existing connections, usually through a graph (see also beginning of section 3), expressing route and link FLOW CONSISTENCY, and link and route TIME and cost CONSISTENCY.

- a DEMAND MODEL, which expresses how route transportation costs affects route flows, by modelling route choice behaviour through utility functions and choice functions and including consistency between OD matrix demand flows and route flows.

Assumptions about demand-supply interaction leads to equilibrium models, which may be considered an instance of day-to-day dynamic (DaDy) process models, which only affects demand models.

The assumption of within-day stationary transportation system leads to well established static assignment models based on a linear synchronic network model, and largely used to support strategic planning,. These models can quite easily be extended to deal with within-day non-stationary transportation systems with discrete service through diachronic linear network models. In this case the supply model can sequentially be solved from route flows through link flows, link times and costs, to route times and costs.

On the other hand, generalization to within-day dynamics is not straightforward at all for transportation systems with continuous service, since in this case the network model is highly non-linear, the demand model being only slightly affected. Moreover the number of users (density, flow, ...) on a link depends on the travel time from the origin to the link, whilst due to congestion the number of users affects travel times (speed-density coupling). Hence consistency between route and link flows and link traffic flow models must be simultaneous solved leading to DYNAMIC NETWORK LOADING (DNL). The resulting link travel times allows to compute route travel times according to consistency equations, then transportation costs, which may include other attributes such monetary costs, penalty for early/late arrival (departure) wrt a desired time, etc.. Finally combining such a supply model with demand model yields the within-day DYNAMIC TRAFFIC ASSIGNMENT (DTA). In literature it is common to reserve the expression within-day dynamics (WiDy) to transportation systems with continuous service only, and use the same classification for traffic flow, DNL and DTA models, as in this paper.

After a review of Traffic Flow Theory (section 2), this paper presents an analysis of modelling approaches to DNL (section 3) focusing on macroscopic models (section 4). Section 6 reports considerations about their use for DTA models, whilst section 6 some general comments as well as research perspectives. Only few references are reported in this paper (more references in Cascetta, 2001; a recent review in Simonelli, 2004).

## 2 Definition and Notations

This section briefly reviewed main definitions and notations of (deterministic) traffic flow theory, regarding vehicles (or other kind of users) travelling along a linear facility or waiting to be served at a bottleneck. First, let us observe a *traffic stream*, i.e. cars moving along a highway segment, and consider which variables we can measure. To simplify notation no explicit reference is made to the segment. Let:

$t$ be the time at which the stream is observed;
$x$ be the abscissa of any point along the segment, increasing along traffic direction;
$v_i(x, t)$ be the speed of vehicle $i$ at time $t$ while traversing point $x$.

For traffic observed at point $x$ during the time interval $[t, t+\Delta t]$, some variables can be defined:

$m(x)$ the number of vehicles traversing point $x$;
$\underline{f}(x) = m(x) / \Delta t$ the flow of vehicles crossing point $x$,(vehicles per unit of time);
$\overline{v}_T(x) = \Sigma_{i=1,\ldots,m} v_i(x) / m(x)$ the time mean speed, among all vehicles crossing point $x$.

Similarly, at time $t$ between points $x$ and $x+\Delta x$, the following variables can be defined:

$n(t)$ the number of vehicles at time $t$;
$\underline{k}(t) = n(t) / \Delta x$ the density between points $x$ and $x+\Delta x$ at time $t$, (vehicles per unit of length);
$\overline{v}_S(t) = \Sigma_{i=1,\ldots,n} v_i(t) / n(t)$ the space mean speed at time $t$

and also

$u = f(x)$ the *inflow*, i.e. the entering flow during time interval $[t, t+\Delta t]$;
$w = f(x+\Delta x)$ the *outflow*, i.e. the exiting flow during time interval $[t, t+\Delta t]$.

During time interval [t, t+Δt] between points x and x+Δx, a *general flow conservation equation* holds:

$$\Delta n(x, x+\Delta x, t, t+\Delta t) + \Delta m(x, x+\Delta x, t, t+\Delta t) = 0 \qquad (1a)$$

More generally, the right term may be different from zero, and equal to the number of entering minus exiting vehicles (if any) during time interval $[t, t+\Delta t]$, due to entry/exit points (e.g. on/off ramps), between points $x$ and $x+\Delta x$. Dividing the general flow conservation equation $\Delta t$, yields:

$$\Delta n / \Delta t + \Delta f = 0 \qquad \text{or} \qquad \Delta n / \Delta t + w - u = 0 \qquad (1b)$$

Then, dividing by Δx yields:

$$\Delta k / \Delta t + \Delta f / \Delta x = 0 \qquad (1c)$$

A traffic stream is called *stationary* during a time interval $[t, t+\Delta t]$ between points $x$ and $x+\Delta x$ if flow is (on average) independent of point $x$, and density is independent of time $x$:

$f(x) = f = u = w$
$k(t) = k$

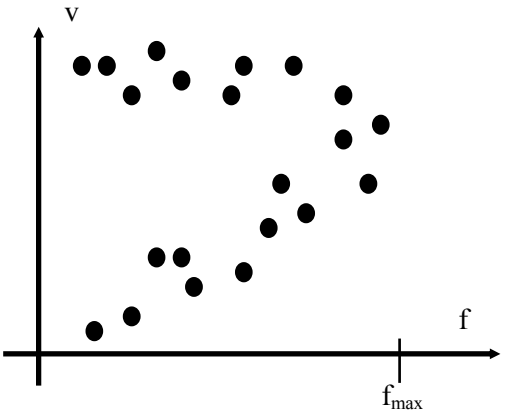
In this case, the time mean speed is independent of location and space mean speed is independent of time:

$\overline{v}_T(x) = \overline{v}_T$
$\overline{v}_S(t) = \overline{v}_S = v$

It is easily noted that under stationary conditions flow conservation equation (1) is rather useless, still it can be easily proved that density and space mean speed must satisfy the *stationary flow conservation equation*:

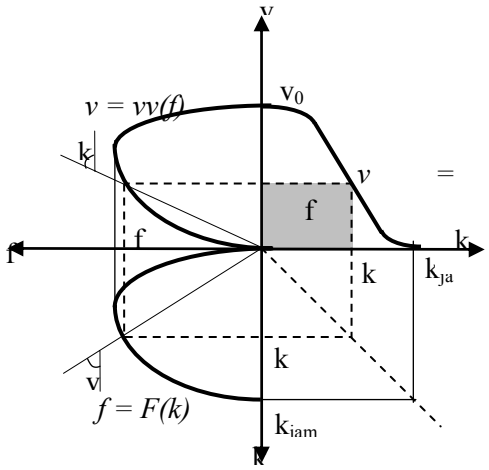
$$f = k\, v \qquad (2)$$

Multiple vehicles using the same facility may interact with each other and the effect of their interaction will increase with the number of vehicles. This effect, called *congestion*, occurs in most transportation systems, generally worsening the overall performances of the facility, such as the mean speed or the travel time, since a vehicle may not be able to move at the desired speed. Congested systems with continuous service can be modelled through (aggregate) deterministic models. In fact, under stationary conditions, aggregate relationships may be observed between any pair of variables: flow, density and speed. Generally, observed values are rather dispersed, an example of a speed-flow observation is given on the right side, usually the maximum value of flow, $f_{max}$, is called *capacity*, Q.




The relationship between speed and flow may be represented through (deterministic) models, as well as those between speed and density and between flow and density, consistency being assured by equation (2):

$$v = V(k) \quad (3)$$
$$v = vv(f) \quad (4)$$
$$f = F(k) \quad (5)$$




These three functions together are also known as the *fundamental diagram* of traffic flow, their general aspects being illustrated on the right side. Analysis of the diagram, as well as, the description of the several models proposed in literature are out of the scope of this paper. Let $L$ be the length of the segment, travel time $tt$ turns out a function of flow:

$$tt = tt(f) = L / vv(f) \quad (6)$$

Now, let us observe a *traffic queue*, i.e. cars waiting to be served at a bottleneck with limited capacity. As in the above approach for traffic streams, some variables can be defined:

$t$ be the time at which the queue is observed;
$t_s$ service time;
$Q = 1/t_s$ capacity;
$m_{IN}$ the number of vehicles joining the queue during time interval $[t, t+\Delta t]$;
$m_{OUT}$ the number of vehicles leaving the queue during time interval $[t, t+\Delta t]$;
$u = m_{IN} / \Delta t$ be the *inflow*, i.e. the flow entering the queue during time interval $[t, t+\Delta t]$;
$w = m_{IN} / \Delta t \leq Q$ be the *outflow*, i.e. the flow exiting the queue during time interval $[t, t+\Delta t]$;
$n(t) \geq 0$ be the queue length, the number of queuing vehicles at time $t$.

During time interval $[t, t+\Delta t]$ a *general flow conservation equation* holds (cfr. 1b):

$$\Delta n / \Delta t + (w - u) = 0$$

Considering both under-saturation, $u < Q$, and over-saturation, $u \geq Q$, conditions a general formula may be derived to describe the evolution over time of the queue length between time $t$ and $t+\Delta t$:

$$n(t+\Delta t) = \mathrm{MAX}\{0, n(t) + (u - Q)\,\Delta t\}$$

From eqn (6), waiting time $tw$ may easily be defined as a function of arrival flow (cfr. 6):

$$tw = tw(u) = n / Q$$

# 3 Approaches to within-day dynamics modelling

For modelling a transportation system, a route can be considered as a sequence of links, each one a part of travel carried out with (approximately) constant features (e.g. a highway segment). Then, connection pattern is modelled through a graph, such that a link is modelled by an arc and a route through a path between two nodes (representing the origin and the destination). It is common practice in literature, and accepted in this paper, to use link and arc as well as route and path as synonymous (being no longer necessary to distinguish the modelled object and the modelling tool to avoid misunderstandings when models are described).

From results of traffic flow theory about observed variables, several modelling approaches can be defined, according to assumptions about variables adopted for arc $a$, not explicitly introduced for simplicity's sake.

- FLOW REPRESENTATION over space $x$ and time $t$
  - disaggregate, direct representation by tracing each vehicle $i$, $x_i(t)$, such that $n(t) \cong n^{OBS}(t)$;
  - aggregate, mono-dimensional fluid approximation, through functions $f(x,t)$, $k(x,t)$, $u(t)$, $w(t)$, such that:

$$\int_{t1}^{t2} f(x,t)\,\mathrm{d}t \cong f^{OBS}(x;\, t_1,\, t_2) \qquad \int_{x1}^{x2} k(x,t)\,\mathrm{d}x \cong k^{OBS}(t;\, x_1,\, x_2)$$

$$\int_{t1}^{t2} u(t)\,\mathrm{d}t \cong u^{OBS}(\, t_1,\, t_2) \qquad \int_{t1}^{t2} w(x,t)\,\mathrm{d}t \cong w^{OBS}(x;\, t_1,\, t_2)$$

    Explicit conditions may need to assure mono-dimensional fluid assumption: a particle may never reach (or overtake) a particle who has entered the link before, otherwise the fluid is no longer mono-dimensional. This condition is often referred to in literature as FIFO rule.

- SPEED REPRESENTATION over space $x$ and time $t$
  - disaggregate, by modelling the speed of each vehicles $i$, $v_i(t)$, such that $v(t) \cong v^{OBS}(t)$;
  - aggregate, through functions $v(x,t)$ (or travel time $tt$) such that

$$\int_{t1}^{t2} v(x,t)\,\mathrm{d}t \cong v_s{}^{OBS}(x;\, t_1,\, t_2)$$

In *macroscopic models* flow is represented through aggregate variables, so far single vehicles are not explicitly traced, and aggregate speed-flow (or other LoS attributes) relations are used, derived from stationary models. In *mesoscopic models*: flow is represented through disaggregate variables, by explicitly tracing each single (or group of) vehicles but aggregate speed-flow relations are used, derived from stationary models. In *microscopic models* flow is represented through disaggregate variables, by explicitly tracing single vehicles and disaggregate speed modelling is adopted based on explicit modelling of driver behaviour of speed adjustment (through well established models of car following, lane changing, overtaking, gap-acceptance, etc.). As already said, analysis of mesoscopic and microscopic models are out of the scope of this paper. Macroscopic approaches, which are the object of this paper, will be analysed below.

| | | Speed representation | |
|---|---|---|---|
| | | Aggregate | Disaggregate |
| Flow Representation | *Continuous* | MACROSCOPIC | - |
| | *Discrete* | MESOSCOPIC | MICROSCOPIC |

# 4 Macroscopic Dynamic Network Loading models

Macroscopic DNL models can be classified according to assumptions about space and time representation, resulting models will be described in the next sub-sections:

- SPACE REPRESENTATION: continuous vs. discrete
- TIME REPRESENTATION: continuous vs. discrete

It should be noted that the case of continuous space and discrete time never occurs for macroscopic models, in a broad sense mesoscopic and microscopic models might be considered belonging to this class.

*3.1 Discrete space and continuous time models*

Discrete space and continuous time macroscopic arc models (MACRO-DC) are specified by a differential equation over time. With reference to arc *a*, let

$L_a$ be the length of arc *a*,
$w_a(t)$ be the outflow from arc *a* at time *t*,
$u_a(t)$ be the inflow into arc *a* at time *t*,
$n_a(t)$ be the number of vehicles within arc *a* at time *t*,
$k_a(t) = n_a(t)/L_a$ be the density within arc *a* at time *t*,
$v_a(t)$ be the speed of a vehicle entering arc *a* at time *t*,
$tt_a(t) = v_a(t)/L_a$ be the travel time of a vehicle entering arc *a* at time *t*.

Assuming that no exit or entry occurs within the arc, this equation, resembling equation (1c), is given below for each arc *a*:

$$\partial n_a(t) / \partial t + w_a(t) - u_a(t) = 0 \qquad (7)$$

then a conservation equation between inflow and outflow is added, taking into account travel time:

$$w_a(t + tt_a(t)) \cdot (1 + \partial tt_a(t) / \partial t) = u_a(t) \qquad (8)$$

together with some boundary conditions. A speed model should also be included specified through a stationary speed-density function (eqn 3) expressed by travel time against number of users:

$$tt_a = L_a / v_a \qquad \text{or} \quad tt_a(t) = L_a / V_a(n_a(t)/L_a) \qquad (9)$$

For one arc, equations (7-10) specify the arc traffic flow model. It worth noting that mono-dimensional fluid assumption is not granted by any travel time function, thus some preliminary condition, usually addressed as FIFO rule, should be imposed about travel time function (Astarita, 1996), such as:

$$(1 + \partial tt_a(t) / \partial t) > 0 \qquad \text{or} \quad 1 > \partial tt_a(t) / \partial t \; tt_a$$

This condition, which also assures non-negative flows, is granted only by quite simple travel time functions. In a broad sense speed-density coupling is not granted since inflow immediately propagates to outflow.

MACRO-DC DNL may be specified by adding to any of the above presented equations (7-9) a flow conservation equation for each node, provided that the so called FIFO rule is assured by the adopted travel time function (see for instance Friesz et al, 1993; Wie et al, 1990; Wu et al, 1998; Xu et al, 1999). A comprehensive formulation and in-depth analysis is reported in chapter 7 in Cascetta (2001). It has been recently addressed whether such models may effectively deal with limited storage capacity of downstream arcs (Wu et al, 1998; Xu et al, 1999).

Solution for real scale applications requires discretisation over time of equation (7), leading back to equation (1b), and equation (8). The effect of immediate propagation of inflow to outflow can be reduced by dividing the arc into smaller segments. This way discrete space and discrete time models are obtained, as described in

the next section 3.3. It should be noted that mono-dimensional fluid assumption as well as speed-density coupling are no longer assure under discretisation.

Recently Rubio-Ardanaz et al (2003) proposed a solution of MACRO-DC by considering a polynomial approximation of involved functions over time; the greater the number of used points the better the approximation is. The degree of best polynomial is still an open issue, even though some indications are provided.

*3.2 Continuous space and continuous time models*

Continuous space and continuous time macroscopic arc models (MACRO-CC) are specified by a differential equation over space and time among flow, $f_a(x, t)$, density, $k_a(x, t)$, and speed, $v_a(x, t)$, functions and a conservation equation. Assuming that no exit or entry occurs at point $x$, these equations, resembling equations (1c) and (2), are given below for each arc $a$:

$$\partial k_a(x, t) / \partial t + \partial f_a(x, t) / \partial x = 0 \qquad (10)$$

$$f_a(x, t) = k_a(x, t) \cdot v_a(x,t) \qquad (11)$$

together with some boundary conditions. A speed model should also be included.

In first order models (MACRO-CC1) a stationary speed-density function is adopted (eqn 3) for each arc $a$:

$$v_a = V_a(k_a) \qquad \text{or} \quad v_a(x,t) = V_a(k_a(x, t)) \qquad (12a)$$

For one arc, equations (10-12a) specify the well-known LWR model (Lightill and Whitman, 1955) whose solution leads to the kinematic wave theory (Richards, 1956). (A review in Daganzo, 1995b.)

In second order models (MACRO-CC2) the speed model is specified through an equation about acceleration, d$v$/d$t$, like the well-known model proposed by Payne (1971):

$$\partial v_a(x, t)/\partial t + v_a(x, t) \cdot \partial v_a(x, t)/\partial x = (V_a(k_a(x,t)) - v_a(x,t))/\tau_{REC} - (\alpha_{ANT}/\tau_{REC})\, \partial k_a(x, t)/\partial x \qquad (12b)$$

where $\tau_{REC}$ and $\alpha_{ANT}$ are parameters to be calibrated. These models have been criticized and revised over the years (see for instance Ross, 1988 and Papageorgiou et al, 1989), declared deceased (Daganzo, 1995c) and resurrected (Aw and Rascle, 2000; Rascle 2002), since their solutions may present some counter-intuitive features. Anyhow they seem better suited for simple motorway networks, rather than complex urban ones.

MACRO-CC DNL may be specified by adding to any of the above presented models a flow conservation equation for each node. Some problems may arise to address this issue, especially when trying to deal with limited storage capacity of downstream arcs. It worth noting that mono-dimensional fluid assumption as well as speed-density coupling are implicitly granted by the differential model (10-12).

Solution for real scale applications requires discretisation over time and space of equation (10) (and 12b) if the case) leading back to equation (1c). This way discrete space and discrete time models are obtained, as described in the next section 3.3. It should be noted that mono-dimensional fluid assumption as well as speed-density coupling are no longer assure under discretisation.

Newell (1993a, 1993b) proposed a solution approach by considering only arc inflow and outflow profiles, thus leading to some sort of MACRO-DC models, described in section 3.1.

*3.3 Discrete space and discrete time models*

Discrete space and discrete time macroscopic arc models (MACRO-DD) are specified by a finite difference equation over space and time. With reference to arc $a$, not explicit introduced for simplicity's sake, let

$v_{max}$ be the maximum (or free-flow) speed,
$\Delta t$ be the time discretisation step,
$\Delta x$ be the space discretisation step, with $\Delta x > v_{max}\, \Delta t$ to avoid that during a time interval flow may travel more than space interval (this condition is greatly useful to implement solution algorithms),

$w_{jk} = u_{j+1,k}$ be the outflow from space interval $j$ during the $k$-th time interval,
$u_{jk} = w_{j-1,k}$ be the inflow into space interval $j$ during the $k$-th time interval,
$n_{jk}$ be the number of vehicles within space interval $j$ at the end of the $k$-th time interval,
$k_{jk} = n_{jk} / \Delta x$ be the (average) density within space interval $j$ at the end of the $k$-th time interval,
$v_{jk}$ be the (space average) speed of vehicles within space interval $j$ during the $k$-th time interval.

Assuming that no exit or entry occurs in a space interval, these equations, to be compared with equations (1b) and (2) or (7) and (8), are given below:

$$(n_{jk} - n_{jk-1}) / \Delta t + w_{jk} - w_{j-1,k} = 0 \quad (13)$$
$$w_{jk} = (n_{jk} / \Delta x) \cdot v_{jk} \quad (14)$$

together with some boundary conditions about $w_{0,k}$ and $n_{j0}$. A speed model should also be included.

In first order models (MACRO-DD1) a stationary speed-density function is adopted (eqn 3) for each arc:

$$v_{jk} = V(n_{jk} / \Delta x) \quad (15)$$

or similar specifications.

MACRO-CC DNL may be specified by adding a flow conservation equation for each node to equations (13-15) for each arc, thus leading to the basis of the well-known cell model (Daganzo, 1994a,b 1995a). This model has been further detailed also including limited storage capacity of downstream arcs.

In second order models (MACRO-DD2) the speed model is specified through the discretisation over time and space of an equation about acceleration. For instance this way METANET model can be obtained (Papageorgiou and Korsialos, 1998).

Some inconsistency may arise among number of users on arc, speed and outflow against intuitive analysis. In particular, mono-dimensional fluid assumption is not generally granted, but the shorter the space discretisation step, the better is the approximation. Moreover, speed-density coupling is only approximately granted, but the shorter the time discretisation step, the better is the approximation.

Apart any other consideration solution for real scale applications can quite easily carried out.

# 5 From DNL to DTA

All the previously described macroscopic DNL models provide arc flows and travel times, given boundary conditions, which may be based on path flows. Then they may be combined with a duly defined model simulating travel time consistency, thus allowing to compute path travel times. These values may be used to define transportation costs, which may include other attributes such monetary costs, penalty for early/late arrival (departure) wrt a desired time. These attributes can be used as input variables for the demand model to obtained path flows consistent with given O-D demand flows and assumed choice behaviour, leading to a full DTA model. This whole procedure has been outlined by Cascetta (2001), but still never be applied to real size cases. Most reported proposals are based on implicit path choice analysis by considering some rather simple approaches, such diversion fractions at nodes, or instantaneous shortest paths revised while-trip, etc.. Thus it still seems an open issue whether and how an explicit demand model may be combined with macroscopic supply models. Therefore macroscopic DTA may not yet be considered a closed issue.

# 6 Conclusions

Even though at a first glance all the macroscopic models may seem equivalent from the solution point-of-view, apart from the polynomial approach for MACRO-DC, is still an open issue which relationships hold among different kinds of models. A more efficient solution approach may be devised by combining space and time discretisation with the polynomial approach.

Anyhow it should be stressed that all the macroscopic models show an radical inconsistency as they are based on stationary speed-density relations. On the other hand these models, when duly specified, are a

generalisation of static models, since under constant demand time profile provide a flow pattern which tend to static one as time approaches infinite.

Other issues worth of further research work seem the extension to multi-user DNL and to some sort of spillback modelling (if feasible), and at for urban networks the effect traffic lights on modelling formulation.

All the above issues should be addressed through theoretical investigations as well as applications to toy and real networks, including a comparison with mesoscopic and microscopic models, and an analysis of the cases when their application can be most effective.

## References

Astarita V. (1996) "A continuous time link model for dynamic network loading based on travel time functions" *Proceeding of the 13th International Symposium on the Theory of Traffic Flow*, Lyon, pp. 87–102.
Aw A. and M. Rascle, (2000) Resurrection of "second order" models of traffic flow?, *SIAM J. Appl. Math.* 60 (3), 916-938.
Cascetta, E., (2001), Transportation systems engineering: theory and methods, Kluwer Academic Publishers, Dordrecht, The Netherlands.
Daganzo C.F., (1994a), The cell transmission model, part I: a simple dynamic representation of highway traffic, UCB-ITS-93-7 300
Daganzo C.F., (1994b), The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory, Transportation Research 28B
Daganzo C.F. (1995a) The cell transmission model 2: network traffic simulation. *Transp. Res.*, vol. 29B, pp. 79-93.
Daganzo C.F., (1995b), A finite difference approximation of the kinematic wave model of traffic flow, Transportation Research 29B
Daganzo C.F. (1995c) Requiem for second-order fluid approximations of traffic flow. Transportation Research 29B(4), pp. 277-286.
Friesz T. L., D. Bernstein, T.E. Smith, R. L. Tobin, and B.W. Wie (1993). A variational inequality formulation of the dynamic network users equilibrium problem. *Operations Research* 41:179-191.
Lighthill M andG. Whitham (1955) On Kinematic waves II. Theory of traffic flow on long crowded roads. Proc. Royal Society,London, Series A, 229, pp.317-345.
Newell G.F., (1993a), A simplified theory of kinematic waves in highway traffic, part I: general theory, Transportation Research 27B
Newell G.F., (1993b), A simplified theory of kinematic waves in highway traffic, part II: queueing at freeway bottlenecks, Transportation Research 27B
Papageorgiou M. and A. Korsialos (1998), Short term traffic forecasting with METANET. Daccord Short Term Forecasting Workshop TU Delft.
Papageorgiou M., J.-M. Bloseville. and H. Hadj-Salem (1989). Macroscopic modelling of traffic flow on the Boulevard Peripherique in Paris. *Transpn. Res,* 23B, p 29-47.
Papola, Bellei, Gentile, (2003), A within-day dynamic traffic assignment model for urban networks.
Payne H.J. (1971), Models of Freeway Tra_c and Control, *Math. Models Publ. Sys., Simulation Council Proc*. 28, Vol. 1, 1971, pp. 51-61.
Rascle, M. (2002) An improved macroscopic model of traffic flow: Derivation and links with the Lighthill-Whitham model. In *Mathematical and Computer Modelling*, 35, p 581-590.
Richards P.I., (1956). Shock waves on the highway. *Oper. Res*. 4, p 42-51.
Ross P. (1988). Traffic Dynamics. *Transpn. Res B.* Vol 22B, p 421-435.
Rubio-Ardanaz J.M., J.H. Wu, M. Florian (2003). Two improved numerical algorithms for the continuous dynamic network loading problem. *Transportation Research Part B* 37 (2003) 171–190.
Simonelli F. (2004). I modelli di offerta. In *I sistemi stradali di trasporto nella società dell'informazione* Gennaro Nicola Bifulco ed., Aracne Editrice, Italia
Wie B. W., T. L. Friesz, and T. L. Tobin (1990). Dynamic user optimal traffic assignment on congested multi-destination networks. *Transportation Research* 24B: 431-442.
Wu J.H., Y. Chen, M. Florian (1998), The continuous dynamic network loading problem: a matematical formulation and solution method, *Transportation Research 32B*.
Xu Y. W., J. H. Wu, M. Florian, P. Marcotte, and D.L. Zhu (1999) New advances in the continuous dynamic network loading problem. *Transportation Science* 33, pp 341-353.

# DYNAMIC QUEUING AND SPILLBACK IN AN ANALYTICAL MULTICLASS DYNAMIC TRAFFIC ASSIGNMENT MODEL*

*Michiel C.J. Bliemer, Delft University of Technology, The Netherlands, m.c.j.bliemer@tudelft.nl*

## Abstract

In this paper a new analytical multiclass dynamic network loading (DNL) model as part of a simulation-based dynamic traffic assignment (DTA) model is proposed. In contrast to many other proposed DNL models, this model will explicitly deal with queuing and spillback without having to rely on link travel time functions as input. As will be illustrated in the paper, using link travel times is likely to under- or overestimate the true travel times in a dynamic model if queues are considered. The proposed DNL model consists of a link model and a node model. The link model computes queue inflows and potential outflows, while the node model determines the actual outflows depending on the node structure. In the end, the link travel times are computed backwards in time. The model has been implemented in the INDY DTA software and an application shows that the approach is viable in real-life networks.

## 1 Introduction

In order to analyze transportation networks for planning purposes, traffic assignment models have shown to be a useful tool. For this reason these models have been applied for many years now. Although static models are still widely used, the theory and practice of dynamic models have evolved significantly over the last 10 years. This resulted in a shift of focus from static traffic assignment to dynamic traffic assignment (DTA) in both research and (commercial) applications.

DTA models typically describe route choice behaviour of travellers on a transportation network and the way in which traffic dynamically propagates through the network. A nice overview of DTA approaches is given in Peeta and Ziliaskopoulos (2001). Two main approaches can be distinguished, namely (i) a pure analytical approach, and (ii) a simulation-based approach. In the pure analytical approach, the DTA problem (typically formulated as a variational inequality problem) is directly solved by using well-known optimization techniques. Examples are models proposed by Ran and Boyce (1996), Chen and Hsueh (1998), and Bliemer and Bovy (2003). These models are usually limited to small hypothesized networks, as they use solution procedures that do not take advantage of the specific characteristics of the transportation problems. On the other hand, simulation-based models are specifically designed for transportation problems and can handle larger and more realistic networks. These simulation-based DTA models are nowadays widely available and can define the problem either on a microscopic level (e.g., PARAMICS, AIMSUN2 with a micro-simulator propagating the network flows), a mesoscopic level (e.g., DYNASMART, INTEGRATION), or on a macroscopic level (e.g., INDY, MARPLE with a dynamic network loading procedure performing the flow propagation).

While microsimulation-based models are best for (small area) urban transport networks and intersections, macrosimulation-based models are most suitable for (large area) motorway networks. As microsimulation models have to store data for all vehicles that are simultaneously on the network, computer memory becomes the limiting factor. The number of variables in macroscopic models is independent of the number of vehicles, therefore these models are more easily scalable and are faster in computation time. This way, dynamic traffic

* The author would like to thank Erik Versteegt, Edwin van Veldhoven, Muriël Poelman, and Ana Barros from The Netherlands Organisation for Applied Scientific Research (TNO) for their contributions and the implementation of the model into the INDY DTA software. INDY can be downloaded and is free to use on networks with maximum 25 zones, see www.tno.nl/indy, and uses the Omnitrans graphical user-interface, which can be downloaded from www.omnitrans.nl.

assignment is even possible for large area networks with millions of vehicles (see e.g. Bliemer et al., 2005, where a country-wide network of the Netherlands is considered). On the other hand, microscopic simulation models can more easily deal with different vehicle types (in terms of different speed characteristics, e.g. cars and trucks), while many macroscopic models are limited to a single vehicle type. Furthermore, queuing and spillback are typically difficult to implement into a dynamic network loading procedure for macroscopic models.

In this paper we will propose an analytical dynamic network loading (DNL) procedure in which queuing and spillback are taken into account in a multiclass setting with different vehicle types. Attempts of others have some strong restrictions, as will be pointed out in Section 2, while the model proposed in this paper does not rely on these (unrealistic) restrictions. This will lead to a completely time-responsive dynamic queuing model without any assumptions on stationary inflows, queue lengths, and/or outflow capacities. The key to the dynamic queuing model is that we do not rely on travel time functions that look forward in time, but only use queuing and exit functions that look backward in time. The formulation correctly deals with time-varying link attributes, such as inflow and outflow capacities and maximum speeds, such that a wide range of dynamic traffic management (DTM) measures can be incorporated.

The outline of the paper is as follows. First, in Section 2 a brief description of other models proposed in the literature to include queuing in analytical dynamic network loading will be given, in which the underlying assumptions for validity will be discussed. In Section 3 the proposed DTA model framework will be introduced, in which the DNL model consisting of two main components: (i) a link model, and (ii) a node model. The link and node model will be described in Sections 4 and 5, respectively. Section 6 describes outcomes of an application on a real-life network using a new version of the INDY model in which the proposed dynamic queuing model is implemented. Conclusions are drawn in the final section.

## 2 Approaches of queuing in dynamic network loading models

Many macroscopic simulation-based DTA models use DNL models that propagate the traffic flow on the links using link travel time functions (see e.g. Astarita, 1996; Wu, 1998; Chabini, 2001; Bliemer et al., 2004), which relate the travel time to the number of vehicles on the link at the time of link entrance. Although these models will generally be able to give good estimates for average travel times and flows, dealing with specific phenomena such as spillback and DTM measures in which changes in capacities play an important role constitute weak points. This is mainly due to the fact that the (forecasted) travel times are assumed fixed while traversing the link, while in reality the travel time is not known before a vehicle exits the link, as there may be dynamically changing queues and dynamically changing outflow restrictions. Also, they often predict delays inside the bottleneck instead of upstream the bottleneck.

In the literature a few possible approaches for dealing with queues and spillback have been proposed. Often, these models consider a horizontal queue (as opposed to the unrealistic vertical queues) on a link that is artificially split up into a free-moving part followed by a queuing part. There are typically two difficulties:
(a) the queue length may change while traversing the link, and
(b) the outflow capacity may change while traversing the link.
By means of trajectories illustrated in Figure 1 we will explain the differences in the approaches and how they deal with the above mentioned difficulties.

Consider a link with a free-moving and a queuing part. In the figure is depicted how the queue grows and shrinks over time. Furthermore, consider a vehicle entering the link at time instant $t$. We are interested to know what time instant this vehicle will exit the link.

Ran and Boyce (1996) proposed a model in which the queue length at link entrance (time $t$) is considered, and then compute the link exit time computing the time in the free-moving part and in the queuing part, assuming that neither the queue length, nor the outflow capacity will change (which we will call an

*instantaneous queue* approach). This leads to exit time $t_I$. However, the queue length may change (due to vehicles entering the queue from the free-moving part and vehicles leaving the queue flowing out of the link). This problem was solved by He (1997) by realizing that all vehicles that are in the free-moving part at time *t* will have entered the queue when our considered vehicle enters the queue. By assuming a fixed outflow capacity, it is possible to determine the exact queue length, which leads to an exit time of $t_{II}$ (so-called *variable queue* approach). Note that their approach is only valid if no overtaking finds place in the free-moving part, hence it will not hold in case of multiple vehicle types where first-in-first-out (FIFO) does not hold between all vehicles. Roels and Perakis (2004) do not require FIFO to hold in the free-moving part, as they look backwards from the tail of the queue. However, they also compute the travel time assuming that the outflow capacity does not change. In this paper we propose an approach that also does not require FIFO to hold in the free-moving part, but can take changing outflow capacities into account (so-called *dynamic queue* approach). For example, if the outflow capacity decreases at $t'$ (e.g., due to spillback, or DTM measures, or even just by changing composition of vehicles in the queue with respect to directions, etc), then the actual exit time will be $t_{III}$. All other described approaches will therefore yield an underestimation of the link travel time. Similarly, it can be shown (assuming a decreasing queue length and an increasing outflow capacity) that the other approaches will give an overestimation of the link travel time. Clearly, the true travel time is not known until the time of exiting the link, hence using link travel time functions may yield incorrect travel times that are not consistent with capacity constraints.

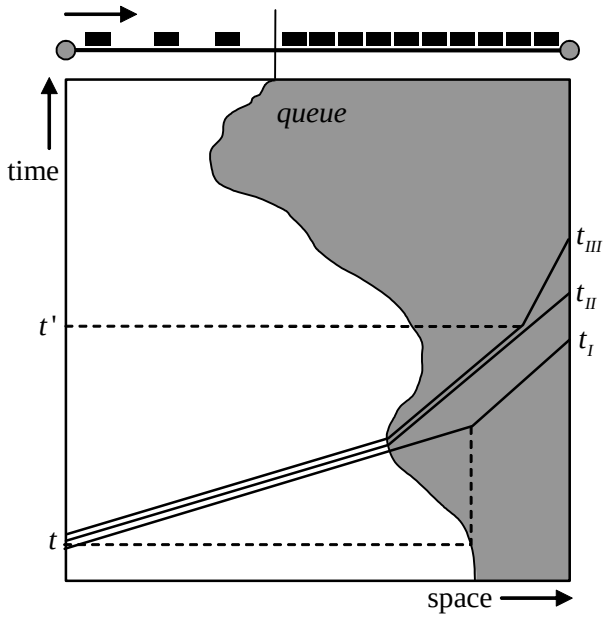



***Figure 1:*** *Different queuing approaches*

Besides approaches splitting links into two parts, other approaches exist, such as the cell transmission model (CTM, Daganzo, 1994), and the recently proposed link transmission model (LTM, Yperman et al., 2006). As these models rely on FIFO, multiple vehicle types cannot be taken into account, although they have the advantage that shockwaves are included.

# 3 Model framework

Consider a given transport network $G=(N,A)$ consisting of nodes *N* and directed links *A* having certain attributes, and a given dynamic vehicle type specific travel demand $D_m^{rs}(k)$ for each origin-destination (OD) pair $(r,s)$, each vehicle type *m*, and each departure time *k*. The DTA model framework, depicted in Figure 2, is a typical route-based framework consisting of three main parts: (1) a route set generation model, (2) a route choice model, and (3) a dynamic network loading model. The first two models are the same as in the INDY model described in Bliemer et al. (2003) and will only be briefly described here. The main focus in this paper will be on the dynamic network loading model.

The *route generation model* aims to determine a set of routes $P_m^{rs}$ for each origin-destination (OD) pair and each vehicle type. Two approaches are used, namely a Monte Carlo approach (adapted from Catalano and Van der Zijpp, 2001) and a static traffic assignment approach. In the Monte Carlo approach the link travel times are assumed stochastic, in which a new fastest path is determined using the current draws for the link travel times and then added to the route set. The static traffic assignment approach uses the well-known deterministic static assignment of the OD matrix (the time period with the highest travel demand) and returns all used routes (and also initial path proportions for the dynamic assignment). If certain links are not accessible for certain vehicle types, then the routes using these links are removed from the corresponding vehicle type specific route set.

The *route choice model* aims to determine a stochastic dynamic multiclass user-equilibrium based on the actual route travel times (or generalized costs including e.g. tolls). In each iteration, new route flow proportions over route set $P_m^{rs}$ are computed using a relative multinomial logit or path-size logit model. This yields new route flow rates $f_m^{rs}(k)$ that are passed on to the dynamic network loading model. In order to speed up convergence, the method of successive averages (MSA) is adopted on the route flow level.

The *dynamic network loading (DNL) model* simulates the route flows $f_m^{rs}(k)$ along the links in the network. This model is at the heart of the DTA model and is also the most computationally intensive part. A completely new DNL model has been developed and added to INDY, which has dynamic queuing possibilities and does not have the drawbacks from most DNL models based on link travel time functions (see the previous section). This DNL model consists of a link model and a node model. Both will be discussed in detail in the next sections.

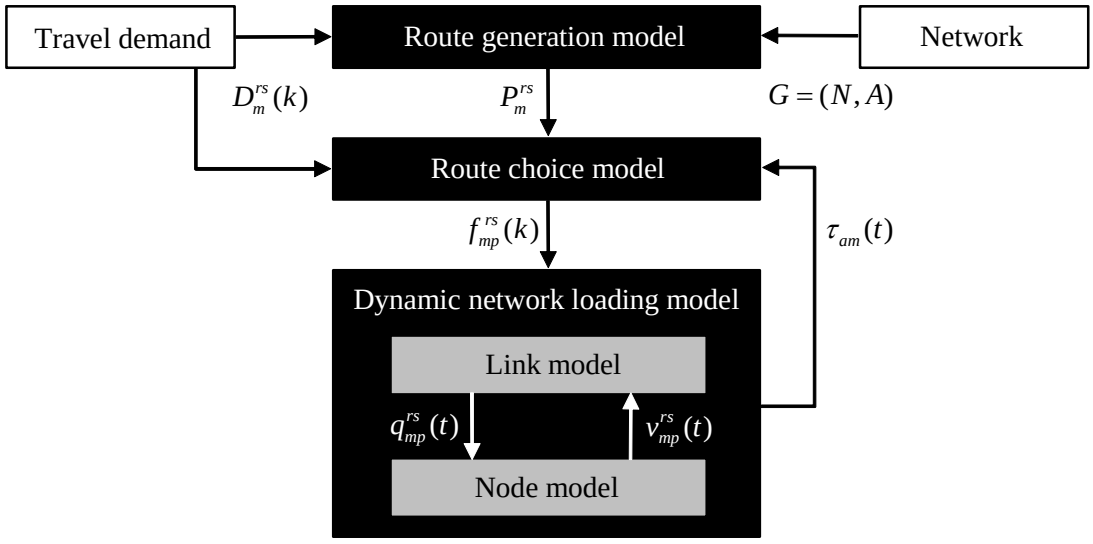



***Figure 2:*** *DTA model framework*

# 4 Link model

The link model is part of the analytical DNL model and describes the propagation of the flow through each link, taking into account different speeds for different vehicle types and a dynamic horizontal queue. The main outcomes are the link inflows, queue inflows, queue lengths, and link travel times.

Consider a link $a \in A$ in network $G$ for which the following attributes are given: link length $L_a$ [km], maximum speeds per vehicle type $\vartheta_{am}^{\max}$ [km/h], and a queue density $J_a$ [pcu[1]/km]. Also, an unrestricted inflow capacity $C_a$ [pcu/h] is given for each link, however this attribute will only be used by the node model,

[1] Passenger car unit (pcu), which in our case it is a measure of road space occupancy of a certain vehicle type compared to passenger cars.

see Section 5. As mentioned in Section 2, we will (artificially) split each link into a free-moving part and a queuing part, conform Figure 3. The lengths of the free-moving part and the queuing part, denoted by $L_a^f(t)$ and $L_a^q(t)$ respectively, are variables in the model. In case there is no queue, $L_a^f(t) = L_a$, while in case the queue covers the entire link (and hence spillback to previous links may occur), $L_a^q(t) = L_a$. Splitting the link into these two parts is also important from a multiple vehicle type point of view. In the queue all vehicle types are assumed to travel at the same speed, which seems, while in the free-moving part vehicle types may travel at different speeds and may overtake each other. Hence, first-in-first-out (FIFO) need not be satisfied among different vehicle types (but typically is assumed to hold within each vehicle class), which has important consequences for the model. The model proposed here explicitly deals with this situation.

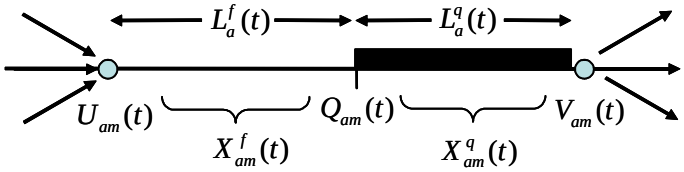



***Figure 3:*** *Link variables*

The variables $U_{am}(t)$, $Q_{am}(t)$, and $V_{am}(t)$ denote the cumulative link inflow, cumulative queue inflow, and cumulative outflow of link *a* at time *t*, respectively. In case there is no queue, $Q_{am}(t) = V_{am}(t)$. The number of vehicles of each vehicle type in the free-moving part, $X_{zm}^f(t)$, and the number of vehicles in the queuing part, $X_{zm}^q(t)$, then are

$$X_{am}^f(t) = U_{am}(t) - Q_{am}(t), \quad \text{and} \tag{1}$$

$$X_{am}^q(t) = Q_{am}(t) - V_{am}(t). \tag{2}$$

The link inflow rate $u_{amp}^{rs}(t)$ is determined by the corresponding outflow rate $v_{a'mp}^{rs}(t)$ of the previous link $a'$ on route *p*, or (in case link *a* is the first link on route *p*) given by the route flow rate $f_{mp}^{rs}(t)$. This leads to the following flow conservation constraint:

$$u_{amp}^{rs}(t) = \begin{cases} v_{a'mp}^{rs}(t), & \text{if } a' \text{ is the previous link on route } p, \\ f_{mp}^{rs}(t), & \text{if } a \text{ is the first link on route } p. \end{cases} \tag{3}$$

Then the vehicle type *m* specific cumulative inflow $U_{am}(t)$ into (and cumulative outflow $V_{am}(t)$ out of) link *a* till time *t* can be computed as

$$U_{am}(t) = \sum_{(r,s)} \sum_{p \in P_m^{rs}} U_{amp}^{rs}(t), \quad \text{with} \quad U_{amp}^{rs}(t) = \int_{\omega=0}^{t} u_{amp}^{rs}(\omega) d\omega, \tag{4}$$

$$V_{am}(t) = \sum_{(r,s)} \sum_{p \in P_m^{rs}} V_{amp}^{rs}(t), \quad \text{with} \quad V_{amp}^{rs}(t) = \int_{\omega=0}^{t} v_{amp}^{rs}(\omega) d\omega. \tag{5}$$

It is important to note that the link outflow rates $v_{amp}^{rs}(t)$ are determined by the node model (see next section) taking capacity constraints into account, and are not determined by a flow propagation constraint as usual in a link model using link travel time functions. As mentioned in Section 2, using flow propagation based on link travel time functions may violate capacity constraints. In our model, only the outflow out of the free-moving part (which is the same as the inflow into the queuing part) is determined by a flow propagation constraint, which does not corrupt any existing capacity constraints at the end of the link. Instead of determining a travel time (for the free-moving part) at the time of link entrance, only a speed is computed per vehicle type. The reason for that is the following. The travel time of the free-moving part cannot be determined at the time of link entrance, as the length of the free-moving part $L_a^f(t)$ may change before reaching the queue (due to new vehicles entering the queue and due to vehicles leaving the queue depending on changing outflow capacities). Therefore, the flow propagation constraint for the free-moving part uses past speeds $\vartheta_{am}(\cdot)$ and determines which vehicles will enter the tail of the queue at time *t*. By looking

backward in time at these speeds (and assuming that they remain constant for each vehicle type while traversing the link), the correct queue lengths can be determined at all times, even in the case of different vehicle types that can overtake. The speeds $\vartheta_{am}(t)$ can be determined as a (decreasing) function of the number of vehicles of all types on the free-flowing part. However, this could potentially violate FIFO within each vehicle type. Choosing $\vartheta_{am}(t) = \vartheta_{am}^{\max}$, which is not unreasonable for the free-moving part, overcomes this problem.

Assume a given link-specific queue density[2] $J_a$ (in passenger car units, pcu, per km). The total number of pcu's in the queue is defined by $\sum_m \rho_m X_{am}^q(t)$, where $\rho_m$ denotes the vehicle type specific pcu-value. The length of the free-moving part, $L_a^f(t)$, can then be determined as

$$L_a^f(t) = L_a - L_a^q(t), \quad \text{where } L_a^q(t) = \frac{\sum_m \rho_m X_{am}^q(t)}{J_a}. \tag{6}$$

Note that if $L_a^q(t) = L_a$, spillback will occur to previous link(s) by restricting the inflow into link *a*, see the node model in Section 5. A vehicle of type *m* entering link *a* at time $\omega$ with speed $\vartheta(\omega)$ will reach the tail of the queue at time *t* if $\left(\omega + L_a^f(t)\right)/\vartheta_{am}(\omega) \le t$. Hence, the cumulative queue inflow is given by

$$Q_{amp}^{rs}(t) = \int_{\omega \in \Omega(t)} u_{amp}^{rs}(\omega) d\omega, \quad \text{with } \Omega(t) = \left\{ \omega \,|\, \omega + \frac{L_a^f(t)}{\vartheta_{am}(\omega)} \le t \right\}. \tag{7}$$

Then, the queue inflow rates $q_{amp}^{rs}(t)$ and the total cumulative queue inflow $Q_{am}(t)$ can be computed as

$$q_{amp}^{rs}(t) = \frac{dQ_{amp}^{rs}(t)}{dt}, \qquad \text{and} \tag{8}$$

$$Q_{am}(t) = \sum_{(r,s)} \sum_{p \in P_m^{rs}} Q_{amp}^{rs}(t). \tag{9}$$

Note that no travel times need to be computed for the flow propagation. However, the link travel times are usually an important output of the model, hence we will compute them afterwards. The computation of the link travel times is performed backwards in time for each link, as it is not until a vehicle leaves the link that the actual travel time for that vehicle is known. Consider a certain vehicle that exits link *a* at time $t_1$. Then due to FIFO in the queuing part (keeping the order of the vehicles in the queue) the time of entering the tail of the queue is $t_2$, where $Q_{am}(t_2) = V_{am}(t_1)$. If FIFO also holds for the free-moving part,[3] then that vehicle entered the link at time $t_3$, where $U_{am}(t_3) = Q_{am}(t_2) = V_{am}(t_1)$. This results in a link travel time of $t_1 - t_3$. In general, the link travel time $\tau_{am}(t)$ for type *m* vehicles entering link *a* at time *t* is given by

$$\tau_{am}(t) = V_{am}^{-1}\left(U_{am}(t)\right) - t. \tag{10}$$

## 5 Node model

As a second part of the analytical DNL model, the node model relates the inflows into and outflows out of each node. In doing this, it takes into account the inflow capacities into the outgoing links, and the potential outflow rates coming from the incoming links. The capacity constraints of links are therefore completely managed by the node model, not by the link model. The outcomes of the node model are the OD-dependent, path-dependent and vehicle type dependent dynamic outflow rates $v_{amp}^{rs}(t)$.

---

[2] Considering this value as a constant may seem a rather restricting assumption, as the queue density may depend on how fast the queue moves (i.e., the outflow rate). However, the link definition with two parts is artificial and as such the queue density may be used as a parameter for calibrating the link travel times, which is the main output of the model. Although the exact moment of spillback may not be completely accurate, it is our belief that this deviation is within acceptable bounds by assuming a constant queue density.

[3] FIFO will hold by definition if the speeds are constant for the free-moving part, as explained before. In case the speed is a general function of the vehicles on the free-moving part, FIFO may not hold. In that case, the time of link entrance cannot be computed directly, but can still be determined based on the history of the speeds.

Consider a node $n \in N$ in network *G*. Each node *n* has a set $B^{\text{in}}(n)$ of incoming links and a set $B^{\text{out}}(n)$ of outgoing links, see Figure 4. Each incoming link $a \in B^{\text{in}}(n)$ has some potential outflow rates $\bar{v}_{amp}^{rs}(t)$, consisting of vehicles arriving at the end of the link (queued or not queued) at time *t*. However, the actual outflow rates $v_{amp}^{rs}(t)$ may be smaller than these potential outflow rates due to inflow capacity restrictions. Each outgoing link $b \in B^{\text{out}}(n)$ has a certain inflow capacity $C_b^{\text{in}}(t)$ which can change over time due to queue spillbacks or DTM measures. The potential outflow rates in the direction of link *b* have to share this limiting capacity.

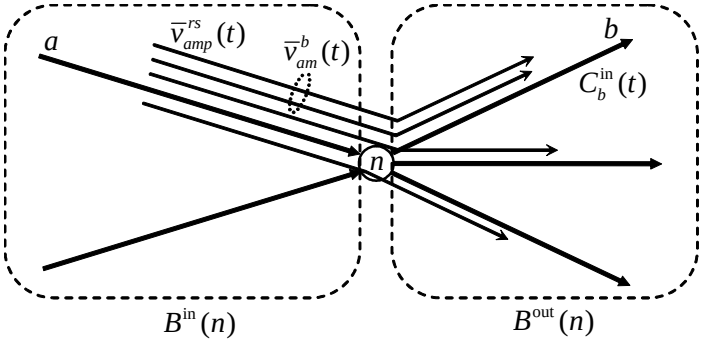



***Figure 4:*** *Node variables*

Let us first define these inflow capacities and potential outflow rates. The inflow capacity $C_b^{\text{in}}(t)$ depends on whether the queue on link *b* is spilling back or not (or if a DTM measure changes the inflow capacity). If there is no spillback (i.e., $L_b^q(t) < L_b$), then the inflow capacity is equal to the unrestricted capacity $C_b$. On the other hand, if there is spillback, then the inflow capacity is set to the current outflow out of that link. Mathematically,

$$C_b^{\text{in}}(t) = \begin{cases} C_b, & \text{if } L_b^q(t) < L_b, \\ \sum_{(r,s)} \sum_m \sum_{p \in P_m^{rs}} \rho_m v_{bmp}^{rs}(t), & \text{otherwise.} \end{cases} \tag{11}$$

In order to compute the potential outflow rates $\bar{v}_{amp}^{rs}(t)$, we have to determine the flow rates at the head of the queue.[4] In case there is no queue, the potential outflow rates are simply equal to $q_{amp}^{rs}(t)$. However, if there is a queue, it will be assumed that all lanes will be used by vehicles in the queue and that the total potential outflow is equal to the capacity of the link,[5] $C_a$. The potential outflow rates are then functions of earlier queue inflow rates $q_{amp}^{rs}(t_a^*(t))$, where $t_a^*(t)$ is the time instant in which the vehicles now at the head of the queue entered the tail of the queue. Since FIFO holds in the queuing part, this time instant can easily be determined by $t_a^*(t) = Q_a^{-1}(V_a(t))$ using the cumulative queue inflow in pcu, $Q_a(t) = \sum_m \rho_m Q_{am}(t)$, and cumulative outflow in pcu, $V_a(t) = \sum_m \rho_m V_{am}(t)$. Of importance is, that the outflow rate proportions are conserved between each OD-pair, each path, and each vehicle type (again, due to FIFO). Therefore, the proportions between the queue inflow rates $q_{amp}^{rs}(t_a^*(t))$ should transfer to proportions of the potential outflow rates sharing the link capacity $C_a$. Hence, the potential outflow rates are determined by

$$\bar{v}_{amp}^{rs}(t) = \begin{cases} q_{amp}^{rs}(t), & \text{if } L_b^q(t) = 0, \\ \dfrac{q_{amp}^{rs}(t_a^*(t))}{\sum_{(r',s')} \sum_{m'} \sum_{p' \in P_m^{rs}} \rho_{m'} q_{am'p'}^{r's'}(t_a^*(t))} C_a, & \text{otherwise.} \end{cases} \tag{12}$$

[4] Without loss of generality, the end of the link is referred to as the head of the queue, even if the queue length is zero.

[5] The assumption made here is that the capacity is reached when there is a queue, while in real life this may not hold. Again we stress here that, just like the queuing density, the capacity is a parameter for calibrating the link travel times (and queue lengths) and may not have a completely similar interpretation in traffic flow theory.

Since multiple paths can use the same outgoing link, the *directional* potential outflow rates $\overline{v}_{am}^{b}(t)$ describing the (vehicle type specific) outflow from link *a* to link *b* will be used (see also Figure 4), defined by

$$\overline{v}_{am}^{b}(t) = \sum_{(r,s)} \sum_{p \in \left\{ P_m^{rs} | b \in p \right\}} \overline{v}_{amp}^{rs}(t). \tag{13}$$

Knowing the boundaries on the inflows and outflows, the actual outflow rates $v_{amp}^{rs}(t)$ should be determined. The determination of these actual outflow rates is not trivial for a general node with multiple incoming and outgoing links and with multiple vehicle types. In the cell transmission model of Daganzo (1995), these outflow rates are computed for a simple merge and a simple diverge. These computations are essentially based on a linear programming (LP) problem in which the throughput of the node is maximized subject to capacity constraints and proportion conservation constraints. Adopting this idea, and extending it to a general node with multiple in- and outgoing links and multiple vehicle types, leads to the following LP formulation:

$$\max_{v_{am}^{b}(t)} \sum_{a \in B^{\text{in}}(n)} \sum_{b \in B^{\text{out}}(n)} \sum_{m} v_{am}^{b}(t) \tag{14}$$

s.t.

$$\sum_{a \in B^{\text{in}}(n)} \sum_{m} \rho_m v_{am}^{b}(t) \le C_b^{\text{in}}(t), \quad \forall b \in B^{\text{out}}(n), \tag{15}$$

$$v_{am}^{b}(t) \le \overline{v}_{am}^{b}(t), \quad \forall a \in B^{\text{in}}(n), \forall b \in B^{\text{out}}(n), \forall m, \tag{16}$$

$$\frac{v_{am}^{b}(t)}{v_{a'm}^{b}(t)} = \frac{\overline{v}_{am}^{b}(t)}{\overline{v}_{a'm}^{b}(t)}, \quad \forall a, a' \in B^{\text{in}}(n), \forall b \in B^{\text{out}}(n), \forall m, \tag{17}$$

$$\frac{v_{am}^{b}(t)}{v_{am}^{b'}(t)} = \frac{\overline{v}_{am}^{b}(t)}{\overline{v}_{am}^{b'}(t)}, \quad \forall a \in B^{\text{in}}(n), \forall b, b' \in B^{\text{out}}(n), \forall m, \tag{18}$$

$$\frac{v_{am}^{b}(t)}{v_{am'}^{b}(t)} = \frac{\overline{v}_{am}^{b}(t)}{\overline{v}_{am'}^{b}(t)}, \quad \forall a \in B^{\text{in}}(n), \forall b \in B^{\text{out}}(n), \forall m \ne m'. \tag{19}$$

The objective function is the total outflow through node *n* at time *t*, Eqns. (15) and (16) describe the outflow and inflow constraints, respectively, and Eqns. (17)–(19) ensure flow proportion conservation. These proportion conservation constraints typically hold for freeway traffic if we assume that flow in a capacity constraint direction also constrains flow in other directions. This is because we assume a single queue on each link. If multiple queues for different directions are required (such as different lanes for different directions at intersections), then multiple links should be created. Constraints can be adapted or added in order to simulate different priorities of different links. It can be shown that solving LP problem (14)–(19) yields the following analytical expression for the solution (see Bliemer, 2005):

$$v_{am}^{b}(t) = \min_{b' \in B^{\text{out}}(n) \setminus \{b' | \overline{v}_{am}^{b'}(t) \ge 0\}} \left\{ \overline{v}_{am}^{b}(t), \frac{\overline{v}_{am}^{b}(t)}{\sum_{a'} \sum_{m} \rho_m \overline{v}_{a'm}^{b'}(t)} C_b^{\text{in}}(t) \right\}. \tag{20}$$

Clearly from this expression, the actual outflow is either the potential outflow if the capacity constraints are not binding, or the actual outflow is a proportion of the available capacity. Knowing the directional outflow rates, the OD-pair and path dependent actual outflow rates $v_{amp}^{rs}(t)$ can be determined by

$$v_{amp}^{rs}(t) = \frac{v_{am}^{b}(t)}{\overline{v}_{am}^{b}(t)} \overline{v}_{amp}^{rs}(t). \tag{21}$$

As an example to illustrate the capacity distribution in the node model, consider a node with two incoming and two outgoing links (See Figure 5). Furthermore, assume that two vehicle types are present. The directional potential outflow rates $\overline{v}_{am}^{b}(t)$ and the inflow capacities $C_b^{\text{in}}(t)$ are given, where the potential outflow rates have already been converted to pcu's for convenience. What can be observed is that link 3 is the bottleneck with a capacity of 2,000 pcu/h, while in total 2,500 pcu/h would like to enter this link. Link 3

will be used at capacity, constraining the outflows out of links 1 and 2 in the direction of link 3. Because the flows from these links are constrained, the flows to link 4 will also be constrained (although 4,500 pcu/h could potentially flow into link 4, only 3,600 pcu/h actually enters link 4).

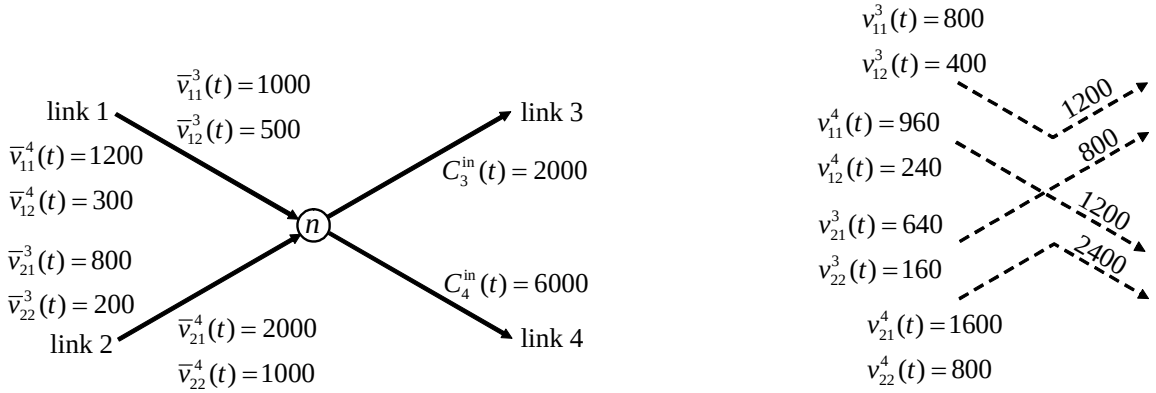



***Figure 5:*** *Application of the node model on a two-link in, two-link out node*

## 6 Application

The proposed model has been implemented (although currently only for one vehicle type) in the INDY DTA software and has been successfully tested on small test networks and correctly builds up queues upstream bottlenecks. The model presented in this paper is defined in continuous time, hence a discretization scheme was needed for implementation. This discretization scheme is not trivial, but it is beyond the scope of this paper to describe the discretization and the algorithm. Details can be found in Bliemer (2005). Here we will just briefly mention an application on a reasonably large real-life network.

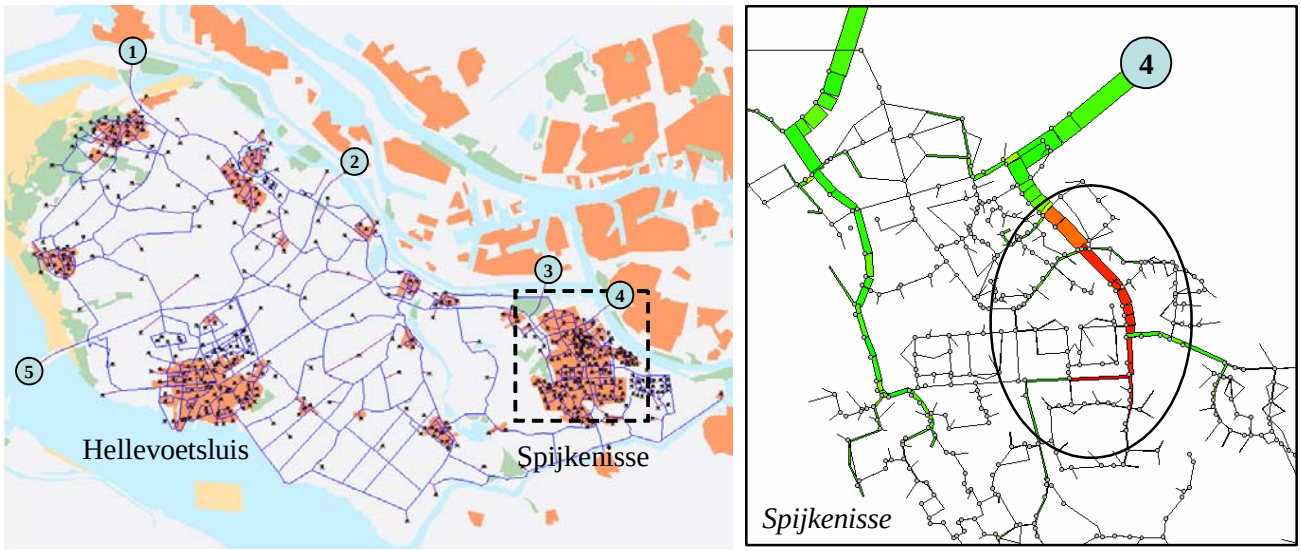



***Figure 6:*** *Voorne-Putten evacuation network*

The application involves an evacuation study in the Voorne-Putten area near Rotterdam in The Netherlands, see Figure 6. This area is surrounded by water and has only five exit points (as indicated) to get to the main land. Hence, in case of a flooding, the approximately 150,000 inhabitants of a few smaller cities have to evacuate using one of these five exit points. The aim of the study was to design evacuation plans for all inhabitants (consisting of the appointed exit points, routes, and departure times) that minimize the total evacuation time. The network consists of approximately 1,500 nodes and 6,000 directed links. In total there are 468 zones, of which 5 destinations and 463 origins. Results showed that evacuation of the city of Spijkenisse was the bottleneck. Long queues in the direction of exit point 4 existed in most scenarios (as indicated in the figure on the right-hand side), spilling back in multiple directions. The dynamic queuing

model in the DNL model seems to give plausible and realistic outcomes. More on the evacuation study can be found in Van Genugten (2005).

# 7 Conclusions

The newly proposed analytical multiclass dynamic network loading model with dynamic queuing has been described, consisting of a link model and a node model. The link model avoids the use of link travel time functions to maintain consistency between the travel times and dynamically changing queues. A general node model has been proposed to enable the distribution of the capacity to the corresponding directions. The model has been implemented in the INDY DTA software and has been successfully run on a fairly large real-life network in the application presented. Special attention needs to be paid to gridlocks, as capacity constrained models typically may suffer from this, and our proposed model is no exception. It should be noted that the proposed model can handle dynamic changes of network attributes (by means of events in INDY), such as capacities, hence traffic lights and dynamic traffic management measures can be simulated.

# References

Astarita, V. (1996) A Continuous Time Link Model for Dynamic Network Loading Based on Travel Time Function. In: J.-B. Lesort (ed.) *Transportation and Traffic Flow Theory*, Pergamon, pp. 79-102.

Bliemer, M.C.J. (2005) INDY 2.0 Model Specifications. Delft University of Technology working report.

Bliemer, M.C.J., and Bovy. P.H.L. (2003) Quasi-Variational Inequality Formulation of the Multiclass Dynamic Traffic Assignment Problem. *Transportation Research B*, 37, pp. 501-519.

Bliemer, M.C.J., Versteegt, H.H., and Castenmiller R.J. (2004) INDY: A New Analytical Multiclass Dynamic Traffic Assignment Model. *Proceedings of the TRISTAN V conference*, Guadeloupe.

Catalano, S.F. and Van der Zijpp, N.J. (2001) A Forecasting Model for Inland Navigation Based on Route Enumeration. *Proceedings of the AET European Transport Conference*, Cambridge, UK.

Chabini, I. (2001) Analytical Dynamic Network Loading Problem: Formulation, Solution Algorithms, and Computer Implementations. Transportation Research Record, 1771, TRB, National Research Council, Washington, D.C., pp. 191-200.

Chen, H.-K, and Hsueh, C.-F. (1998) A Model and an Algorithm for the Dynamic User-Optimal Route Choice Problem. *Transportation Research B.* 32(3), pp. 219-234.

Daganzo, C.F. (1994) The Cell Transmission Model: A Dynamic Representation of Highway Traffic Consistent with Hydrodynamic Theory. *Transportation Research B*, 28(4), pp. 269-287.

Daganzo, C.F. (1995) The Cell Transmission Model, Part II: Network Traffic. *Transportation Research B*, 29(2), pp. 79-93.

He, Y. (1997) A flow-based approach to the dynamic traffic assignment problem: Formulations, algorithms and computer implementations. MSc. Thesis, Massachusetts Institute of Technology, Cambridge MA, USA.

Peeta, S., and Ziliaskopoulos, A.K. (2001) Foundations of Dynamic Traffic Assignment: The Past, the Present and the Future. *Networks and Spatial Economics*, 1(2), pp. 233-265.

Ran, B., and Boyce, D.E. (1996) *Modeling Dynamic Transportation Networks: An Intelligent Transportation System Oriented Approach*. Second edition, Springer-Verlag, Berlin.

Roels, G. and Perakis, G. (2004) An Analytical Model for Traffic Delays. *Proceedings of the TRISTAN V Conference*, Guadeloupe, France.

Van Genugten, W.L.M. (2005) Evacuation modeling: A study to evacuation strategies in large scale evacuations. MSc Thesis, Delft University of Technology, The Netherlands.

Wu, J.H., Chen, Y., Florian, M. (1998) The Continuous Dynamic Network Loading Problem: A Mathematical Formulation and Solution Method. *Transportation Research B*, 32(3), pp. 173-187.

Yperman, I., Logghe, S., Tampère, C., and B. Immers (2006) The Multi-Commodity Link Transmission Model for Dynamic Network Loading. Presented at the 85th Annual Meeting of the Transportation Research Board, Washington DC, USA.

# MULTI-COMMODITY DYNAMIC NETWORK LOADING WITH KINEMATIC WAVES AND INTERSECTION DELAYS

I. Yperman, Traffic and Infrastructure, Katholieke Universiteit Leuven, Belgium
C.M.J. Tampère, Traffic and Infrastructure, Katholieke Universiteit Leuven, Belgium

## 0. Abstract

This paper presents a method for including intersection delays in a flow-based traffic model. The original Multi-Commodity Link Transmission Model (MC LTM), a Dynamic Network Loading (DNL) model that is consistent with kinematic wave theory, is extended with delayed urban intersection models. Travel time as a function of traffic load becomes a monotonically increasing function, which is a desirable property in view of equilibrium calculation in DTA.

## 1. Introduction

Modeling traffic flow propagation is one of the fundamental issues in dynamic traffic assignment (DTA) problems. The problem is known as the dynamic network loading (DNL) problem and consists of finding time-dependent route travel times given the time-dependent route flow rates. DNL models for simulation-based, iterative equilibrium DTA problems, are often classified in microscopic, mesoscopic and macroscopic simulation models.

Microscopic simulation models, such as AIMSUN2 (Barcelo, 2002) describe traffic flow on the level of individual vehicles. Due to the stochastic nature of microscopic simulation models, one simulation run only represents one solution in a whole spectrum of possible solutions. Since route choices are made based on average travel times, one network-loading step should be composed of several simulation runs. However, this procedure is generally too time-consuming in a DTA framework, since one micro-simulation run for a medium-sized network already requires high computational efforts.

Dynamic network loading models that are used in mesoscopic DTA models, such as CONTRAM (Taylor, 2003), DYNASMART (Mahmassani et al., 2001) and DYNAMIT (Ben-Akiva et al., 2002) are less cumbersome computationally, but they are also less precise in the representation of traffic dynamics. CONTRAM uses a travel-time based traffic model that describes a vertical queuing process. The speed-flow functions yield incorrect densities and the physical extent of queues is ignored. DYNASMART and DYNAMIT use flow-based traffic models that propagate individual vehicles on links according to a modified Greenshield type speed-density relationship. Traffic description in these models suffers from drawbacks of realism.

Macroscopic traffic flow theory is for example used in INDY (Bliemer et al., 2004). Ziliaskopoulos et al. (2004) and Szeto and Lo (2006) use the Cell Transmission Model (CTM, Daganzo, 1994) for Dynamic Network Loading. Yperman et al. (2006) have recently presented a DNL that is consistent with Newell's simplified theory of kinematic waves (Newell, 1993): the Link Transmission Model (LTM). Kuwahara and Akamatsu (2001) proposed a similar approach to derive travel time functions, but their travel times do not correspond to actually experienced travel times. LTM shares the underlying theory and favorable properties that make traffic flow representation of the CTM superior to traditional approaches; it provides a first order correct prediction of queue spillbacks and dissipations. In addition, the computational complexity and the numerical discretisation error of the implementation are substantially smaller compared to the CTM. LTM is a multi-commodity model (i.e. traffic flows are disaggregated by route), enabling route choice information to be used within the DNL model.

The original LTM describes traffic conditions on motorway networks with simple intersections. This paper presents an extension of the original LTM with refined intersection

models for urban networks. Urban intersections impose both flow restrictions and intersection delays due to traffic lights and priorities of conflicting flows. After giving a short overview of the original LTM, the paper presents a method for including intersection delays in a flow-based traffic model. A preliminary study on this topic was done by Durlin and Henn (2005). The paper ends with a case study, where the impact of these refinements and the consequences for equilibrium in DTA are illustrated.

## 2. The original LTM

### 2.1. Network representation

Traffic networks consist of homogeneous links $a$, which start at place $x_a^0$ and end at place $x_a^L$. The links can have any length $L_a$ and they are connected to each other via nodes. A route $p$ is a series of links $a$ and nodes $n$ between an origin node $r$ and a destination node $s$. Nodes have no physical length. They act merely as a flow exchange medium. A general motorway network can be represented by a combination of links and some basic nodes. Diverge nodes and merge nodes are respectively used to model diverging lanes/off-ramps and merging lanes/on-ramps.

### 2.2. Simplified theory of kinematic waves

The LTM uses Newell's simplified theory of kinematic waves (Newell, 1993) to propagate traffic on links. This theory is based on the conservation of vehicles concept and it assumes a functional relation between the traffic flow $q$ and density $k$, also known as the fundamental diagram of traffic flow (figure 1). The triangular shaped fundamental diagram is defined by three parameters: a fixed free-flow speed ($v_f$), a maximum flow or capacity ($q_M$) occurring at critical density $k_M$ and a jam density ($k_{jam}$) (see figure 1). Traffic states on the increasing branch of the fundamental diagram ($k < k_M$) hold vehicles travelling with a fixed free-flow speed $v_f$. Traffic states on the decreasing branch ($k > k_M$) are congested. Vehicles travel with a speed $q/k$.

The evolution of traffic conditions on a link is described by a combination of the conservation law and the fundamental diagram. Traffic states travel at a wave speed $dq/dk$. In the simplified theory with a triangular shaped fundamental diagram, there are only two possible values of the wave speed, one positive ($v_f$) for traffic states that hold free-flow traffic and one negative ($w$) for congested traffic states. A forward moving traffic state may intersect with a backward moving state in a shock, separating the free flow and the congested state. The evolution of traffic on the road network is in this model represented by the cumulative number of vehicles $N(x,t)$ that pass the up- and downstream ends $x_a^0$ and $x_a^L$ of each link $a$ by time $t$. Cumulative vehicle numbers $N(x,t)$ are disaggregated by route. Details on this multi-commodity aspect can be found in Yperman et al. (2006). For notational convenience, the theory in this paper is further explained in terms of single commodity flows.

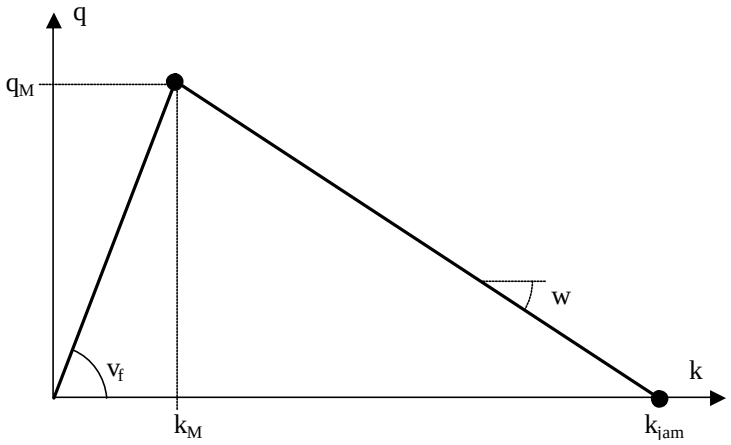



**figure 1: Triangular shaped fundamental diagram**

2.3. Solution algorithm

The LTM algorithm provides a discrete formulation of the continuum kW model of traffic flow. Traffic conditions are updated in successive time steps. For each time interval *Dt*, the method involves three steps:

**Step 1**: "For each link *a*, determine the sending flow at the downstream link end ($x_a^L$) and the receiving flow at the upstream link end ($x_a^0$)".

The sending flow $S_a(t)$ of link *a* at time *t* is defined as the maximum amount of vehicles that could leave the downstream end of this link during *[t, t+Dt]*, if this link end were connected to a traffic reservoir with an infinite capacity.
The receiving flow $R_a(t)$ of link *a* at time *t* is defined as the maximum amount of vehicles that could enter the upstream end of this link during *[t, t+Dt]*, if a traffic reservoir with an infinite traffic demand were connected to this link end.
Sending and Receiving flows are determined based on Newell's simplified theory of kinematic waves. A detailed analysis can be found in Yperman et al., (2006).

$$S_i(t) = \min((N(x_i^0, t+\Delta t - \frac{L_i}{v_{f,i}}) - N(x_i^L, t)), q_{M,i}\Delta t) \quad (1)$$

$$R_j(t) = \min((N(x_j^L, t+\Delta t + \frac{L_j}{w_j}) + k_{jam}L_j - N(x_j^0, t)), q_{M,j}\Delta t) \quad (2)$$

**Step 2: "**For each node n, determine the transition flows $G_{ij}$ (t) from the incoming links $i \in I_n$ to the outgoing links $j \in J_n$, i.e. determine which parts of the sending and receiving flows can actually be sent and received ($I_n$ ($J_n$) is the set of incoming (outgoing) links into node *n*)".

Transition flows $G_{ij}(t)$ are determined by node models. Node models hold some particular priority- and behavioural rules and they always obey to the conservation of vehicles concept. Transition flows for a diverge node with FIFO behavior and for a demand-proportional merge node are as follows:

- diverge node: $G_{ij} = \min_{j' \in J_n}(\frac{R_j S_{ij}}{S_{ij'}}, S_{ij})$ *for all* $j \in J_n$ (3)
- merge node: $G_{ij} = \min(\frac{R_j S_{ij}}{\sum_{i' \in I_n} S_{i'j}}, S_{ij})$ *for all* $i \in I_n$ (4)

**Step 3: "**For all link boundaries $x_a^0$ and $x_a^L$, update the cumulative vehicle numbers *N(x,t)*".

$$N(x_i^L, t+\Delta t) = N(x_i^L, t) + \sum_{j \in J_n} G_{ij} \quad \text{for all } i \in I_n \quad (5)$$

$$N(x_j^0, t+\Delta t) = N(x_j^0, t) + \sum_{i \in I_n} G_{ij} \quad \text{for all } j \in J_n \quad (6)$$

## 3. Urban intersections

The original LTM describes traffic conditions on motorway networks where merge and diverge node models represent basic motorway junctions. To describe traffic conditions on urban networks, we need to include urban intersection models, both for signalized and unsignalized intersections. Signalized and unsignalized intersections impose boundary conditions on the downstream boundaries of their incoming links. Two main effects are to be distinguished on these incoming links *i*: the capacity at the downstream link boundaries ($x_i^L$) is constrained ($cap(x_i^L)$) and vehicles on these links experience intersection delays $D_{int}$. Even if the constrained capacity $cap(x_i^L)$ is on average high enough to handle traffic demand, some vehicles still experience (intersection) delays waiting before red lights or waiting before conflicting priority streams.

To incorporate capacity constraints and intersection delays, there are two possible solution methods: (i) explicitly simulating green and red stages (for signalized intersections) and gaps in conflicting priority streams (for unsignalized intersections), or (ii) considering the average effects of traffic lights and priority streams, without simulating them directly.
The first solution method has three main disadvantages in the context of DTA:

1) Explicit simulations of traffic light stages and gaps in traffic steams impose constraints on the simulation time step. Frequently changing boundary conditions require small simulation time steps. However, small time steps substantially increase computational complexity.
2) Explicit simulations generate small fluctuations in route travel times over time due to stage alternations etc… However, drivers do not take into account these small travel time fluctuations for their route choice. Average travel times are more relevant.
3) Due to their stochastic nature, explicit simulations yield one single solution in a whole spectrum of possible solutions. This single solution depends on an accidental combination of traffic lights offsets, accidental gap distributions in traffic streams etc… Since route choices are made based on average travel times, an average solution would be more relevant.

In an attempt to overcome these disadvantages, we explore the second solution method, where the average effects of traffic lights and priority streams are considered indirectly.

3.1. Capacity constraints and intersection delay formulae.

In the past 5 decades, many formulae for capacity constraints and intersection delays have been proposed, both for signalized and unsignalized intersections.
Signalized intersections impose the following trivial capacity constraints at the downstream link boundaries of their incoming links:

$$cap(x_a^L) = \frac{g}{c} q_{M,a} \tag{7}$$

where $cap(x_a^L) = cap$ = capacity at the downstream end of link $a$ (veh/s)
$g$ = effective green time (s)
$c$ = cycle length (s)
$q_{M,a}$ = capacity of link $a$ (veh/s).

Signalized intersection delay formulae are generally described in terms of a deterministic and stochastic component to reflect both the fluid and random properties of traffic flow. The first, widely used approximate delay formula for signalized intersections was developed by Webster (1958) from a combination of theoretical and numerical simulation approaches:

$$D_{\text{int}} = \frac{c(1-g/c)^2}{2[1-(g/c)(\frac{q_{dem}}{cap})]} + \frac{(\frac{q_{dem}}{cap})^2}{2q_{dem}(1-\frac{q_{dem}}{cap})} \qquad \text{for } q_{dem} < cap \tag{8}$$

where $D_{int}$ = average intersection delay per vehicle (s)
$q_{dem}$ = traffic demand (veh/s).

The first term in equation (8) represents delay when traffic can be considered arriving at a uniform rate (deterministic component), while the second term makes some allowance for the random nature of the arrivals (stochastic component). This is known as the “random delay”, assuming a Poisson arrival process.

For unsignalized intersections, capacity and delay formulae are generally based on gap acceptance theory (Troutbeck and Brilon, 1997). Assuming a constant conflicting priority traffic flow $q_p$ (see figure 2) and an exponential distribution for priority stream headways, the following capacity equation for minor traffic flow was proposed by Drew (1968), Buckley (1962), and by Harders (1968), which these authors however, derived in a different manner:

$$cap(x_a^L) = q_p \frac{e^{-q_p t_c}}{1 - e^{-q_p t_f}} \tag{9}$$

where $cap\ (x_a^L) = cap$ = minor stream capacity (veh/s)
$q_p$ = priority stream flow (veh/s)
$t_c$ = critical gap time (s)
$t_f$ = follow-up times (s).

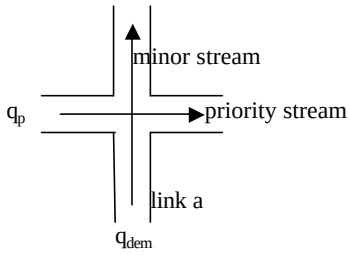



**figure 2 : Simple priority intersection**

A widely used average delay formula has been given by Harders (1968):

$$D_{\text{int}} = \frac{1 - e^{-(q_p t_c + q_{dem} t_f)}}{cap - q_{dem}} + t_f \qquad for\ q_{dem} < cap \tag{10}$$

where $q_{dem}$ = minor stream traffic demand (veh/s).

Formulae (9) and (10) are valid for non-saturated flows ($q_{dem} < cap$) in free-flow traffic states. Adjusted delay formulae are applied for over-saturated flows ($q_{dem} = cap$) in congested traffic states. In that case, arrivals at the intersection are uniform (queue discharge) and only the deterministic component of intersection delay needs to be taken into account. For signalized intersections, intersection delay becomes:

$$D_{\text{int}} = \frac{c(1 - g/c)^2}{2[1 - (g/c)\frac{q_n}{cap}]} \qquad for\ q_{dem} = cap \tag{11}$$

where $q_n$ = flow level in the congested traffic state (veh/s)

For unsignalized intersections, we introduce a reduction factor $a$ to neutralize the influence of random arrivals (validation of factor $a$ belongs to future work):

$$D_{\text{int}} = a\left(\frac{1 - e^{-(q_p t_c + q_n t_f)}}{cap - q_n} + t_f\right) \qquad for\ q_{dem} = cap \tag{12}$$

Intersection delay formulae (8), (10), (11) and (12) are valid for steady states, assuming an infinite time period of stable traffic conditions. In reality, traffic flows are seldom stationary; the period over which demand is sustained, is finite. When traffic demand changes, it takes some time to reach a new equilibrium intersection delay. This topic is discussed in the next section.

3.2. Including intersection delays in a flow-based traffic model

As opposed to travel-time based models such as CONTRAM, it is not self-evident to include intersection delays in a flow based model such as LTM, since travel time is not a basic variable in these models. Travel times are a result in flow based models, they are only determined after vehicles completed their journey. Therefore, intersection delays have to be taken into account indirectly.

We propose to do this by holding up vehicles at the downstream link ends of incoming links, thereby introducing Point-Queues (P-Q's) at these downstream link ends. Vehicles have to pass through these P-Q's before they can exit the link. Times spent in P-Q's correspond to the intersection delays.

P-Q's are determined by their inflow and outflow rate, and by the number of vehicles in the P-Q. Assume there is a P-Q at the end of link $a$ (link $a$ starts at place $x_a^0$ and ends at

place $x_a^L$). The entrance and the exit of this P-Q are respectively indicated by space-coordinates $x_a^L$ and $x_a^{L'}$. P-Q inflow rates $q(x_a^L,t)$ are easily determined: they correspond to link outflow rates in the original LTM (i.e. link outflow rates in case of empty P-Q's). The determination of P-Q outflow rates $q(x_a^{L'},t)$ and the number of vehicles in a P-Q is illustrated here based on the example in figure 3:

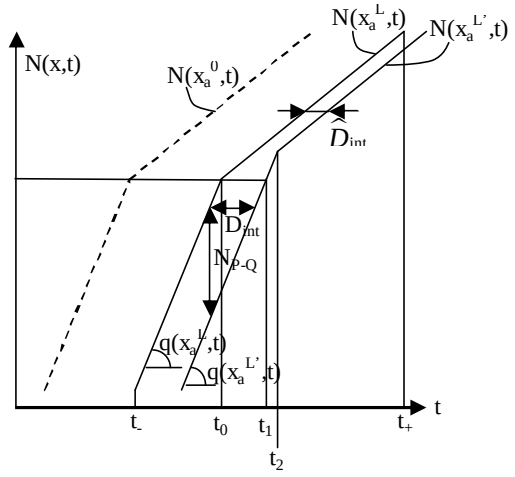



**figure 3: Cumulative vehicle numbers of a P-Q**

- Steady state $[t_-, t_0]$

For steady states, the P-Q outflow rate equals the P-Q inflow rate:

$$q(x_a^L,[t_-,t_0]) = q(x_a^{L'},[t_-,t_0]) \quad (13)$$

This equation is used in the next section; algorithm step 3a; equation (21).

The number of vehicles in the P-Q ($N_{P\text{-}Q}$) is such that each vehicle experiences the presupposed delay $D_{int}$ passing through this P-Q at given flow rate $q(x_a^L,[t_-,t_0])$:

$$N(x_a^L,[t_-,t_0]) - N(x_a^{L'},[t_-,t_0]) = D_{\text{int}} \cdot q(x_a^L,[t_-,t_0]) \quad (14)$$

- Transition state $[t_0, t_2]$

Two steady states are separated by (at least one) transition state. In transition states, the P-Q outflow rate differs from the P-Q inflow rate to adjust the number of vehicles in the P-Q:

$$q(x_a^L,[t_0,t_2]) \neq q(x_a^{L'},[t_0,t_2]) \quad (15)$$

  - Transition state stage 1 $[t_0, t_1]$

Since we assume anisotropic traffic flows where vehicles only react to stimuli ahead of them (traffic conditions behind the vehicle do not influence driver behavior), the P-Q outflow rate in the first stage of the transition state equals the P-Q inflow rate of the original steady state, until the last vehicle of the original steady state has left the P-Q (note that we also assume FIFO-behavior):

$$q(x_a^{L'},[t_0,t_1]) = q(x_a^L,[t_-,t_0]) \quad (16)$$

This equation is used in the next section; algorithm step 3a; equations (22) and (23).

The number of vehicles in the P-Q changes during time interval $[t_0, t_1]$ towards number $N(x_a^L,t_1) - N(x_a^{L'},t_1)$, which corresponds to the original intersection delay $D_{int}$ at new flow rate $q(x_a^L,[t_0,t_1])$.

  - Transition state stage 2 $[t_1, t_2]$

Following equations (8) and (10), smaller (higher) flow rates generally yield smaller (higher) intersection delays. In case of a decreasing (increasing) P-Q inflow rate in transition state stage 1 ($[t_0,t_1]$), the number of vehicles in the P-Q keeps decreasing (increasing) to realize the smaller (higher) intersection delay $\widehat{D}_{\text{int}}$ corresponding to the smaller (higher) flow rate

$q(x_a^L,[t_0,t_1])$. Generally, this new number of vehicles in the P-Q will be reached as soon as possible. However, we want to prevent that a decreasing (increasing) P-Q inflow rate incites an increasing (decreasing) P-Q outflow rate. Therefore, the P-Q outflow rate stays equal during transition state stage 2, both for decreasing and increasing P-Q inflow rates in transition state stage 1:

$$q(x_a^{L'},[t_1,t_2]) = q(x_a^{L'},[t_0,t_1]) \quad (17)$$

This equation is used in the next section; algorithm step 3a; equations (22) and (23).
The evolution of intersection delay during transition states is modeled here as a linear process with constant P-Q outflow rate. It might be interesting to explore the possibility of modeling an exponential process in transition state stage 2, since recent research results (Viti, 2004) indicate an exponential nature of this evolution.

- Steady state *[$t_2$ , $t_+$]*

The transition state continues until the desired number of vehicles in the P-Q corresponding to the new steady state is reached:

$$N(x_a^L,[t_2,t_+]) - N(x_a^{L'},[t_2,t_+]) = \widehat{D}_{\text{int}} \cdot q(x_a^L,[t_2,t_+]) \quad (18)$$

where $\widehat{D}_{\text{int}}$ is the intersection delay in the new steady state.

The example in figure 3 involves non-saturated flows in free-flow traffic states. For over-saturated flows in congested traffic states, the LTM algorithm takes into account that:

- flow restrictions occur at the P-Q entrance and P-Q exit at the same time. The number of vehicles in the P-Q cannot decrease during congestion:
  $$N(x_a^L,t_+) - N(x_a^{L'},t_+) \geq N(x_a^L,t_c) - N(x_a^{L'},t_c) \quad (19)$$
  where $t_c$ is the time on which congestion is initiated ($t_c < t_+$).
  This equation is used in the next section; algorithm step 3a; equation (24).
- delays due to over-saturation itself (i.e. due to flow exceeding capacity), are already taken into account by the original LTM.
- intersection delays occur in addition to the above-mentioned delays.

Other examples (e.g. increasing traffic flow rates, multiple transition states between two steady states, etc…) are treated analogously. The following algorithm is valid in general.

3.3. Solution algorithm including intersection delays

**Step 1**: "For each link *a*, determine the sending flow at the downstream link end (P-Q entry) and the receiving flow at the upstream link end".

Compared to the original LTM algorithm, signalized and unsignalized intersections impose an extra capacity constraint ($cap(x_a^L)$) on the sending flow:

$$S(x_i^L,t) = \min((N(x_i^0,t+\Delta t - \frac{L_i}{v_{f,i}}) - N(x_i^L,t)) \ , \ q_{M,i}\Delta t \ , \ cap(x_i^L)\Delta t) \quad (20)$$

**Step 2**: "For each node *n*, determine the transition flows $G_{ij}(t)$ from the incoming links $i \ \hat{I} \ I_n$ to the outgoing links $j \ \hat{I} \ J_n$, i.e. determine which parts of the sending and receiving flows can actually be sent and received".

This step is the same as in the original LTM algorithm.

**Step 3**: "For each node *n*, determine the transition flows $G_{ij}'(t)$ from the incoming links $i \ \hat{I} \ I_n$ to the outgoing links $j \ \hat{I} \ J_n$, i.e. determine which parts of the sending flows at the P-Q exit (i.e. $S(x_i^{L'},t)$) can actually be sent".

In this step, a distinction is made between non-saturated and over-saturated flows. The argumentation from previous section 3.2. results in the following procedure:

a) For non-saturated flows in free-flow traffic states (i.e. $S_i < G_i$, where $G_i = \sum_{j \in J_n} G_{ij}$):

- determine the sending flow at the P-Q exit:

$$S_a^{'}(t) = \min(cap(x_a^L, t)\Delta t\ ,\ (N(x_a^{L'}, t) - N(x_a^{L'}, t - \Delta t)) \tag{21}$$

if $(N(x_a^L, t) - N(x_a^L, t - \Delta t)) = (N(x_a^{L'}, t) - N(x_a^{L'}, t - \Delta t))$ (steady state)

$$S_a^{'}(t) = \min(cap(x_a^L, t)\Delta t\ ,\ (N(x_a^{L'}, t) - N(x_a^{L'}, t - \Delta t)\ ,\ (N(x_a^L, t + \Delta t) - D_{\text{int}} \frac{N(x_a^L, t + \Delta t) - N(x_a^L, t)}{\Delta t} - N(x_a^{L'}, t)) \tag{22}$$

if $(N(x_a^L, t) - N(x_a^L, t - \Delta t)) < (N(x_a^{L'}, t) - N(x_a^{L'}, t - \Delta t))$ (transition state, decreasing P-Q inflow)

$$S_a^{'}(t) = \min(cap(x_a^L, t)\Delta t\ ,\max((N(x_a^{L'}, t) - N(x_a^{L'}, t - \Delta t)\ ,\ (N(x_a^L, t + \Delta t) - D_{\text{int}} \frac{N(x_a^L, t + \Delta t) - N(x_a^L, t)}{\Delta t} - N(x_a^{L'}, t))) \tag{23}$$

if $(N(x_a^L, t) - N(x_a^L, t - \Delta t)) > (N(x_a^{L'}, t) - N(x_a^{L'}, t - \Delta t))$ (transition state, increasing P-Q inflow)

where $cap(x_a^L, t)$ is determined by equation (7) (for signalized intersections) or equation (9) (for unsignalized intersections), and where intersection delay $D_{int}$ is determined by equation (8) (signalized intersections) or equation (10) (unsignalized intersections), (where $q_{dem} = S_i$).

- determine the transition flows $G_{ij}^{'}(t)$ following the procedure in step 2

b) For over-saturated flows in congested traffic states (i.e. $S_i \geq G_i$):

$$G_i(t) = \max(0\ ,\ N(x_a^L, t + \Delta t) - [N(x_a^L, t_c) - N(x_a^{L'}, t_c)] - D_{\text{int}} \frac{N(x_a^L, t + \Delta t) - N(x_a^L, t)}{\Delta t} - N(x_a^{L'}, t)) \tag{24}$$

where $t_c$ is the time on which congestion is initiated, and $D_{int}$ is determined by equation (11) (for signalized intersections) or equation (12) (for unsignalized intersections), (where $q_n = G_i$).

**Step 4**: "For all link boundaries $x_a^0$, $x_a^L$ and $x_a^{L'}$, update cumulative vehicle numbers $N(x,t)$".

$$N(x_i^L, t + \Delta t) = N(x_i^L, t) + \sum_{j \in J_n} G_{ij} \tag{25}$$

$$N(x_i^{L'}, t + \Delta t) = N(x_i^{L'}, t) + \sum_{j \in J_n} G_{ij}^{'} \tag{26}$$

$$N(x_j^0, t + \Delta t) = N(x_j^0, t) + \sum_{i \in I_n} G_{ij}^{'} \tag{27}$$

(Note that cumulative vehicle numbers are actually disaggregated by route, cf. section 2.2).

## 4. Case study

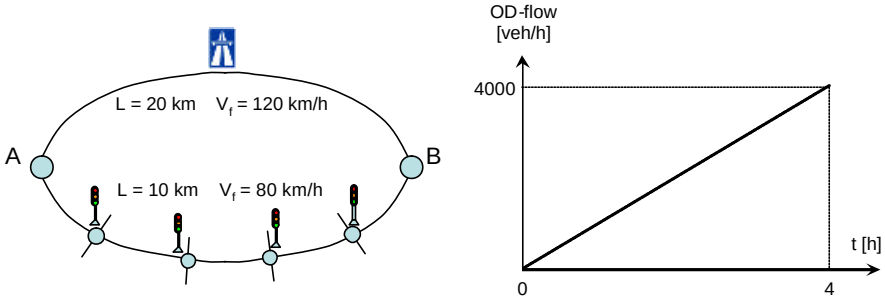



**figure 4: Road network and demand pattern of the case study**

The purpose of this case study is to illustrate how user equilibrium assignment depends on whether or not we account for the delays at intersections. Consider the network of figure 4. Traffic from A to B has two routes at its disposal: one over the ring road (length 20 km; free speed 120 km/h) and one through the city (length 10 km; free speed 80 km/h). We assume that the capacity of any link on both routes is equal to 4000 veh/h.

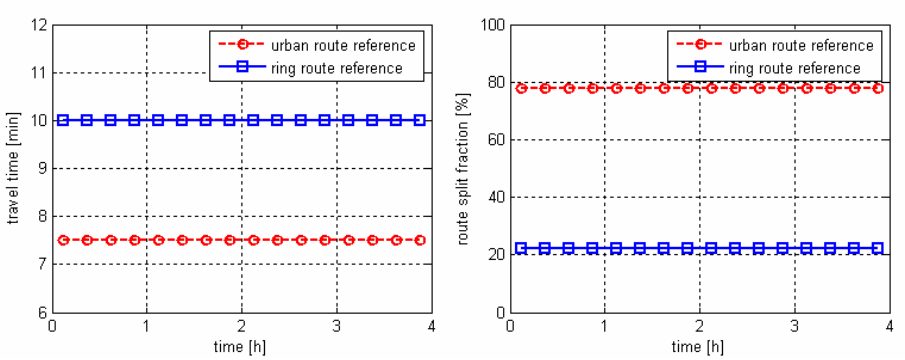



**figure 5: Travel times and route choice in SUE in the reference case (without intersection delay)**

Fig 5 shows the stochastic user equilibrium (SUE) assignment (multinomial logit with cost linear in travel time: $C = -mT$) result for a range of demand levels (demand linearly increasing in time from 0 to 4000 veh/h during 4 hours, leading to no congestion anywhere in the network). In this case delays at intersections on the urban route are neglected. Since traffic is free flowing on both routes, travel times are independent of the demand level and equal to 10 (ring) and 7.5 minutes (urban). With the logit parameter μ = 30, 78% of traffic chooses the urban route, irrespective of the demand.

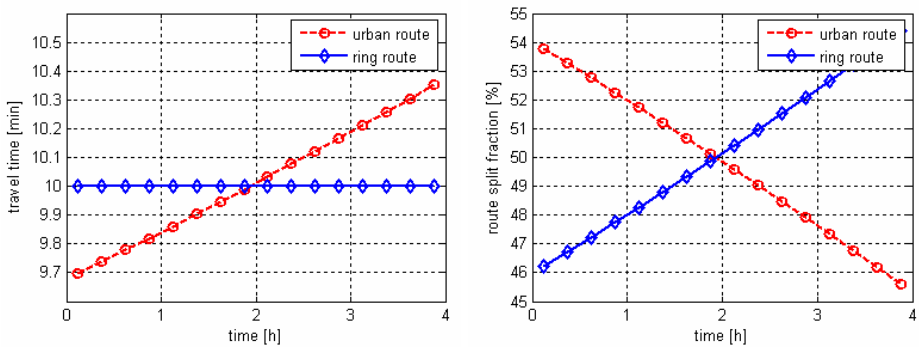



**figure 6: Travel times and route choice in SUE with intersection delay**

Fig 6 shows the SUE assignment for the same demand pattern, however, urban traffic now encounters delay at 4 (non-coordinated) major intersections. Traffic lights control traffic with cycle time 180 s of which 40% green time for the route considered. We have assumed that upstream of the intersections extra lanes are available for through traffic, so that saturation flow during green is 10000 veh/h. Whereas this might be unrealistic, we ensure in this way that the intersections do not act as bottlenecks (average capacity over the node equal to link capacity). The intersections solely impose delays during red, but – with demand levels considered here – all waiting vehicles are served in the consecutive green phase.

Whereas travel time on the ring remains unaffected, the urban route now has become considerably slower. Moreover, free flow travel times are no longer invariant for traffic demand. Minimal travel time is now 9.7 min, increasing to 10.4 min at the highest demand. As a consequence, the ring road is a much more attractive option: 46 to 55% of traffic chooses the ring instead of 22% in the reference case.

We conclude from this case study that accounting for intersection delays in DTA problems has two major effects: (i) travel times over secondary roads increase, hence making them less attractive alternatives as compared to motorways; (ii) travel times over secondary roads become monotonically increasing with demand (even if demand remains below capacity of any link or node that is traveled) instead of being invariant (up to capacity) in the original Link Transmission formulation. This is a more desirable property in view of equilibrium calculation (e.g. uniqueness).

## 5. Conclusions and future research

This paper presents a method for including intersection delays in a flow-based traffic model. The method is used as an extension to the Multi-Commodity Link Transmission Model (MC LTM), a Dynamic Network Loading (DNL) with the following properties:

- the propagation of traffic flows is consistent with kinematic wave theory
- vehicles are disaggregated by route (multi-commodity model)
- high computational efficiency

The extended LTM accounts for capacity constraints and delays imposed by conflicting flows at intersections. As a result, travel times over secondary roads in a network increase, hence rendering them less attractive as compared to motorway alternatives. Moreover, travel time becomes a monotonically increasing function of traffic load instead of being invariant for loads up to capacity. This is a more desirable property in view of equilibrium calculation in DTA (convergence, uniqueness).

The next steps in our research involve applying a similar approach to motorway links: here also a monotonically increasing dependency of travel times versus traffic demand can be obtained by decreasing free flow speed as demand levels approach capacity. Another open issue is to ensure consistency between priority flows and minor flow capacity, which might be mutually dependent on each other if a fraction of the minor flow turns to join the priority flow. Finally, the impact of first order correct congestion dynamics on the existence and uniqueness of equilibriums in networks needs to be examined, as well as the impact on iteration schemes to calculate equilibrium. A comparison with alternative DNL formulations based on travel time functions or simpler queuing models is desired.

## 6. References


Barcelo J. (2002) *Microscopic traffic simulation: A tool for the analysis and assessment of its systems*. In Proceedings of the Half Year Meeting of the Highway Capacity Committee. Lake Tahoe.

Ben-Akiva, M. et al. (2002), Development of a deployable real-time dynamic traffic assignment system. DynaMIT and DynaMIT-P: models and algorithms. Report MIT Intelligent Transportation Systems & Volpe National Transportation Systems Center

Bliemer, M.C.J., H.H. Versteegt and R.J. Castenmiller (2004) INDY: A New Analytical Multiclass Dynamic Traffic Assignment Model. Proceedings of the TRISTAN V conference, Guadeloupe

Buckley D.J. (1962) *Road Traffic Headway Distribution*. In Proceedings 1st ARRB Conf., Vol. 1(1), 153-186.

Daganzo C.F. (1994) The cell-transmission model. A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research* vol. 28B, 269-288.

Drew D.R. (1968) *Traffic Flow Theory and Control*. McGraw-Hill Book Company, New York.

Durlin T. and V. Henn (2005) *A delayed flow intersection model for dynamic traffic assignment*. Advanced OR and AI Methods in Transportation, Proceedings of 10th EWGT Meeting and 16th Mini-EURO Conference, Poznan, Poland, 441-449, Publishing House of Poznan University of Technology.

Harders J. (1968) *Die Leistungsfähigkeit Nicht Signalgeregelter Städtischer Verkehrsknoten (The Capacity of Unsignalized Urban Intersections).* Schriftenreihe Strassenbau und Strassenverkehrstechnik, Vol. 76.

Kuwahara M. and Akamatsu T. (2001) Dynamic user optimal assignment with physical queues for a many-to-many OD pattern. *Transportation Research* 35B, 461-479.

Mahmassani H.S., Abdelghany A.F., Huynh N., Zhou X., Chiu Y-C. and Abdelghany K.F. (2001) *DYNASMART-P User's Guide*. Technical Report STO67-85-PIII, Centre for transportation Research, University of Texas at Austin.

Newell G.F. (1993) A simplified theory of kinematic waves in highway traffic, Part I: General theory, Part II: Queuing at freeway bottlenecks, Part III: Multi-destination flows. *Transportation Research*, vol. 27B, 281-313.

Szeto W.Y. and H.K. Lo (2006) Dynamic Traffic Assignment: properties and extensions, *Transportmetrica*, Vol. 2, No. 1, 31-52

Taylor N.B. (2003) The CONTRAM dynamic traffic assignment model. *Networks and Spatial Economics*, vol. 3, 297–322.

Troutbeck R. and W. Brilon (1997) Unsignalized Intersection Theroy. In *Traffic flow theory: A state of the art report - revised monograph on traffic flow theory*, N. H. Gartner, C. Messer and A. K. Rathi, Eds., Oak Ridge, Tennessee, Oak Ridge National Laboratory.

Viti F. and H.J. Van Zuylen (2004) Modeling Queues At Signalized Intersections. *Transportation Research Record* No. 1883, 68-77.

Webster F.V. (1958) *Traffic Signal Settings*. Road Research Laboratory Technical Paper No. 39, HMSO, London.

Yperman I., S. Logghe, C.M.J. Tampère and L.H. Immers (2006) *The Multi-Commodity Link Transmission Model for Dynamic Network Loading*. In Proceedings 85th Annual Meeting of the TRB, Washington, DC.

Ziliaskopoulos, A.K., Waller, S.T., Li, Y. and Byram, M. (2004) Large-scale dynamic traffic assignment: implementation issues and computational analysis. Journal of Transportation Engineering, ASCE, 130, 585-593.

# A Dynamic Network Loading Model Based On The Variations Solution Procedure

*Blumberg Michal, Department of Industrial Engineering and Management, Ben-Gurion University , Israel, michalb@bgu.ac.il*
*Bar-Gera Hillel, Department of Industrial Engineering and Management, Ben-Gurion University , Israel, bargera@bgu.ac.il*

**Abstract**

This paper explores an analytic macroscopic dynamic network loading model with continuous flows. At the single link level the model is based on the kinematic waves theory with a triangular flow-density relationship. As shown by Daganzo [2005ab], the single link model is equivalent to a variations problem that can be solved on a discrete intra-link time-space mesh by a standard least-cost-path algorithm. Our work extends the single link traffic flow model to road networks by adding merge/diverge rules, and, more importantly, by considering route flows explicitly. Consistent integration between the traffic flow model and the route model is achieved by an iterative procedure, using a novel concept of *anticipated arrival order* (AAO). The AAO describes the number of vehicles that pass a diverge road-node and choose a certain downstream link out of a given number of vehicles that arrive at this road-node. The AAO is not known a-priory, but it is derived from route flows in view of their trajectories, and allows to determine the behavior at diverge road-nodes in the traffic flow model. Computational results for two examples demonstrate the viability of the proposed approach.

## 1 Introduction and problem statement

The dynamic network loading problem (DNLP) is the problem of finding time-dependent link cumulative volumes and flows when time-varying traffic demands for routes are known. DNLP is a sub problem of the dynamic traffic assignment problem (DTAP) which is the problem of finding both the time-varying route choices and the time-dependent link cumulative volumes and flows. Analytic DNLP models strive for mathematical formulations, typically with continuous flows, that maintain various desired properties, such as conservation of flow; First In First Out (FIFO); and above all - realistic traffic behavior.

At the network level, Friesz et al. [1993] presented a general rigorous formulation for the user-equilibrium dynamic traffic assignment with proper consideration of conservation of flow and FIFO. This formulation was further refined and explored by Ran and Boyce [1996], Wu et al. [1998], and Xu et al. [1999]. These models view flow and time as continuous, while space is discrete, as they focus on traffic flows at links' entrances and exits, and ignore the traffic patterns within the link. In principle, these models can be extended to handle any traffic flow model in which the interactions between links occur only at nodes [Carey 2001]. The nature of these formulations has lead researchers to focus on models where either the exit flow or the exit time is a function of the total amount of traffic on the link, despite the well known problems of this simplifying assumption [Daganzo, 1995].

At the link level, Lighthill and Whitham [1955] and Richards [1956] proposed the well respected kinematic wave model, commonly referred to as the LWR model. The LWR model has been studied

by many researchers, including Newell [1993] and Daganzo [2005ab]. The latter proposed to view the model as a continuous variations problem, that can be solved at any desired level of accuracy using sufficiently dense discrete intra-link time-space mesh, also referred to as "solution network." The actual computation required in the solution procedure proposed by Daganzo [2005b] is in fact a computation of least-cost-paths from a single source on an a-cyclic network, which therefore can be performed fairly efficiently. The LWR approach has also been studied in several models that consider route flows and route choice [Kuwahara and Akamatsu, 2001; Lo and Szeto, 2002; Bar-Gera, 2005].

To formulate DNLP mathematically, we consider a road network that is represented by a graph, consisting of a set of nodes $N$ and a set of links $A$. We will refer to the network nodes as 'road-nodes', to distinct between road network nodes and 'mesh-nodes' that are used in discussions of Daganzo's solution procedure. A network route is defined as a set of road-nodes, where every consecutive pair of nodes is a link in $A$. Link $a$ is defined as a pair of nodes $[n_i, n_j] \in N$ where the first road-node is the link tail, $a_{tail}$, and the second node is the link head, $a_{head}$. The link length is denoted by $L_a$. $R = \{r_1, r_2, ..r_k\}$ is the set of routes. Traffic demand for route $r$ is described by the cumulative entrance volume as a function of time $D_r(t)$. A road-node can be: 1) a two-links merge node, where $a_{1head} = a_{2head} = a_{tail}$; 2) a two-links diverge node, where $a_{head} = a_{1tail} = a_{2tail}$; or 3) a trivial node, where $a_{1head} = a_{2tail}$. The traffic flow model is described mainly by the cumulative volume $v_a(x,t)$; $0 \leq x \leq L_a$, which is defined as the total amount of flow (number of vehicles) that passed a location at distance $x$ along link $a$ by time $t$. When the link cumulative volume map is differentiable, its derivatives can be used to find the flow, $f_a(x,t) = \frac{\partial v_a}{\partial t}\big|_{xt}$, and the density, $k_a(x,t) = \frac{\partial v_a}{\partial x}\big|_{xt}$. The purpose of DNLP is to find the cumulative volumes $v$ when route demands $D$ are given.

## 2 Model description

In this research we explore an analytic macroscopic dynamic network loading model with continuous flows. The model is composed of two hierarchal sub models. The inner level model is the traffic flow model; Its objective is to determine link cumulative volumes. Section 2.1 explains how cumulative volume solutions are found for every link independently. Section 2.2 explains the way these are composed to one network cumulative volume solution. The outer level model is the route model, presented in section 2.3. An iterative procedure, described in section 2.4, is used to find consistent solutions for both the traffic flow model and the route model.

The main innovation of the proposed approach is the integration between the traffic flow model and the route model through the road-node *Anticipated Arrival Order* (AAO). In a discrete traffic model, where vehicles are discrete entities, arrival order is defined for each individual vehicle. For example, if 5 vehicles are expected to arrive to a diverge road-node between links $a_1$ and $a_2$, their order can be described by stating, for example, that the 1st, 2nd and the 4th vehicles are directed to link $a_1$, while the 3rd and the 5th vehicles are directed to link $a_2$. The same information can be described by a function $Y_a$, describing the number of vehicles that pass a diverge road-node and choosing link $a$ out of a given number of vehicles that arrive at this diverge road-node. In the above example $Y_{a_1}(1) = 1$; $Y_{a_1}(2) = Y_{a_1}(3) = 2$; and $Y_{a_1}(4) = Y_{a_1}(5) = 3$; while $Y_{a_2}(1) = Y_{a_2}(2) = 0$; $Y_{a_2}(3) = Y_{a_2}(4) = 1$; and $Y_{a_2}(5) = 2$. In our model, since traffic is continuous, we assume that $Y_a$ can be any continuous piecewise linear monotonically non-decreasing function. Precise definition of the

function $Y$ and its computation are given in (17) §2.3. Note that $Y_a$ can be non-increasing, hence its inverse $Y_a^{-1}(x)$ can be multi-valued. To resolve ambiguity, we define $Y_a^{-1}(x) = max\{u : Y_a(u) \leq x\}$.

There is a cyclic relationship between the traffic flow model and the route model since the AAO is essential for determining link cumulative volumes at the inner level, while link cumulative volumes dictate route trajectories, which are required for the computation of AAO at the outer level. This is why an iterative procedure is needed to achieve a consistent solution.

## 2.1 Single link traffic flow model

The single link level model is based on the kinematic waves theory of Lighthill and Whitham [1955] and Richards [1956] (LWR) with the enhancement proposed by Newell [1993], and on a new variational formulation presented by Daganzo [2005a]. The main notion in LWR is the existence of a functional relationship between the flow, $f$, and the density, $k$

$$f(x,t) = F(k,x,t) \tag{1}$$

where the function $F$ is known a-priory from road geometry, while $f$ is computed by the model in view of actual demand. This functional relationship, together with the conservation of flow equation:

$$\partial k/\partial t + \partial f/\partial x = 0 \tag{2}$$

leads to solutions that are based on "characteristic waves", which are lines in time-space, along which density is constant. The pace of the wave, $\omega$ is

$$\omega(k) = [\partial F/\partial k]^{-1} \tag{3}$$

while $(1/\omega)$ is usually called the 'wave velocity'. In the homogeneous case, where $F(k,x,t) = F(k)$, the pace of the wave, the flow and the speed are also constant along a characteristic wave, which is therefore a straight line. The cumulative volume along such a characteristic wave changes at a constant rate of

$$dv = -kdx + fdt = (-k + f\omega)dx \tag{4}$$

When F is concave in k, for every pace value $s$, there is a unique density k(s), such that $[\partial F/\partial k]_- \geq s \geq [\partial F/\partial k]_+$. Daganzo [2005a] shows that the cumulative volume difference between any pair of discrete points $(x_1,t_1),(x_2,t_2)$, $t_1 < t_2$ with a slope $s = \frac{t_2-t_1}{x_2-x_1}$ is bounded from above by

$$v(x_2,t_2) - v(x_1,t_1) \leq \triangle(x_1,t_1;x_2,t_2) = (-k(s) + F(k(s)) \cdot s)(x_2 - x_1) \tag{5}$$

As a direct result, the maximum change in $v$ along a forward piecewise linear time-space path $p = [x_1,t_1;x_2,t_2;...,x_n,t_n]$; $t_1 < t_2 < \ldots < t_n$, denoted by $\Delta(p)$, is:

$$v(x_n,t_n) - v(x_1,t_1) \leq \Delta(p) = \sum_{i=1}^{i=n-1} \triangle(x_i,t_i;x_{i+1},t_{i+1}) \tag{6}$$

Daganzo [2005a] extended the formulation to general continuous time-space paths, and showed that the cumulative volume at any point $(x,t)$ is determined by

$$v(x,t) = min\{B_p + \Delta(p) : \forall p \in P(x,t)\}, \tag{7}$$

where $P(x,t)$ is the set of all forward time-space paths from a boundary $S$, where the cumulative volume is known, to $(x,t)$. $B_p$ is the cumulative value at the beginning of the path.

This model can be solved to any desired accuracy by consideration of a dense discrete intra-link time-space mesh, also referred to as "solution network" [Daganzo, 2005b]. Mesh-nodes are uniformly distributed in time-space according to a model-global time interval, $dt$, and a link-specific spatial interval, $dx_a$, such that $L_a = c_a \cdot dx_a$ for some integer $c_a \in \mathbb{N}$. A simplified flow-density function refers to the case where all uncongested traffic (forwards characteristic waves) move at free flow speed $\omega_a^0$; all backwards characteristic waves have speed $-\omega_a^1$; and the density on each link never exceeds the jam density $k_a^{jam}$, at which speed and flow are zero. In this simplified case, if

$$dx_a/\omega_a^0 = c_0 \cdot dt, \quad dx_a/\omega_a^1 = c_1 dt; \quad \forall a \in A, \quad c_0, c_1 \in \mathbb{N} \tag{8}$$

then a 'lopsided mesh' can be applied, in which mesh arcs are of one of the following forms: $(x_1,t_1),(x1+dx_a,t_1+dx_a/\omega_a^0)$; $(x_1,t_1),(x1+dx_a,t_1-dx_a/\omega_a^1)$; $(x_1,t_1),(x_1,t_1+dt)$. Horizontal arcs in this mesh represent entrance/exit capacity constraints. Note that in order to maintain (8), it may be necessary to use approximate values for $\omega_a^0$, $\omega_a^1$, so that $c_0$ and $c_1$ will indeed be integers.

We can consider every mesh arc as a time-space path with its $\Delta$ value from (5) as the arc cost. Applying any standard algorithm for finding least-cost-paths on a-cyclic networks using the arc costs produces the link level solution, where the least cost to reach a mesh-node is the cumulative volume value. One exception is that the value on the route entrances must be determined according to the demand. If the least cost to a route-entrance mesh-node is lower than the cumulative demand, the model is considered unsolvable.

## 2.2 Network traffic flow model

Cumulative volumes on the entire network are computed in chronological order, assuming road-nodes AAO are known. First, each link is being considered independently, with no consideration of the road network structure constraints (behavior at merge/diverge road-nodes). Suppose $v_a(x,t-1)$ was computed. Temporary volumes, $v_a^{'}(x,t)$, are computed according to (5) and (7). If $0 < x < L_a$ is an internal mesh-node, $v_a(x,t) = v_a^{'}(x,t)$. At a trivial road-node where $a_{1head} = a_{2tail}$, $v_{a_1}(L_{a_1},t) = v_{a_2}(0,t) = min\{v'_{a_1}(L_{a_1},t), v'_{a_2}(0,t)\}$. It remains to describe the behavior at merge and diverge road-nodes.

At merge road-node $n = a_{1head} = a_{2head} = a_{tail}$, an exit of an upstream link is congested when the following occurs simultaneously: the entrance of the downstream link is congested and the link is not using less than its pre-known relative share compared to the other upstream link, which is denoted by $\phi_{a_1/a_2} = 1/\phi_{a_2/a_1}$. Using the concept of 'send' and 'receive' [Daganzo, 1994], where, $Se_a(t) = v_a^{'}(L_a,t) - v_a(L_a,t-1)$ and $Re_a(t) = v_a^{'}(0,t) - v_a(0,t-1)$, the cumulative volumes by

time $t$ are:

$$v_a(0,t) = v_a(0,t-1) + \min\{Se_{a_1}(t) + Se_{a_2}(t), \quad Re_a(t)\} \tag{9}$$

$$v_{a_1}(L_{a_1},t) = v_{a_1}(L_{a_1},t-1) + \min\left\{Se_{a_1}(t), \max\left(\frac{Re_a(t)}{1+\phi_{a_1/a_2}}, Re_a(t) - Se_{a_2}(t)\right)\right\} \tag{10}$$

$$v_{a_2}(L_{a_2},t) = v_{a_2}(L_{a_2},t-1) + \min\left\{Se_{a_2}(t), \max\left(\frac{Re_a(t)}{1+\phi_{a_2/a_1}}, Re_a(t) - Se_{a_1}(t)\right)\right\} \tag{11}$$

At diverge road-node $n = a_{head} = a_{1tail} = a_{2tail}$ the exit of an upstream link is congested whenever the entrance to one of its downstream links is congested. Cumulative volumes are determined using the AAO:

$$v_a(L_a,t) = \min\{v_a^{'}(L_a,t), Y_{a_1}^{-1}(v_{a_1}^{'}(0,t)), Y_{a_2}^{-1}(v_{a_2}^{'}(0,t))\} \tag{12}$$

$$v_{a_i}(0,t) = Y_{a_i}(v_a(L_a,t)); \quad i = 1,2 \tag{13}$$

At the end of this procedure we get a cumulative volumes solution $\bar{V}$ for the entire network that is consistent with the AAO, that is $\bar{V} = V(\bar{Y})$. This solution maintains conservation of flow.

### 2.3 Route flows and trajectories

The route model uses cumulative volumes from the traffic flow model to compute an anticipated arrival order (AAO), in consideration of route trajectories. Route trajectories are described by the arrival time function $\bar{t}(r,n,\tau)$, which defines the arrival time to node $n$ for departure time $\tau$ along route $r$. The trajectories are computed explicitely for a finite set of departure times, $\Theta = \{\tau_1, \tau_2, ..., \tau_m\}; \tau_i = i \cdot \triangle t, \triangle t = c \cdot dt, c \in \mathbb{N}$, and determined by linear interpolation for any other departure time $\tau : \tau_i < \tau < \tau_{i+1}$. Explicit trajectory computation is based on the obvious fact that $\bar{t}(r, a_{2tail}, \tau) = \bar{t}(r, a_{1head}, \tau); \quad \forall a_1, a_2 \in r, a_{2tail} = a_{1head}$, and the computation of link exit time for a given entrance time, which is determined by

$$\bar{t}(r, a_{head}, \tau) = max(min(t : v_a(L_a,t) = v_a(0, \bar{t}(r, a_{tail}, \tau)), \quad \bar{t}(r, a_{tail}, \tau) + L_a\omega_a^0) \tag{14}$$

The second expression of (14) refers to the case where vehicles are moving along the route at free flow speed. The inverse function of $\bar{t}$ is the departure time function $\bar{\tau}(r,n,t) = min\{\tau : \bar{t}(r,n,\tau) \geq t\}$. According to the route-trajectory model, the anticipated cumulative volumes at link $a$ entrance/exit by time $t$ should be:

$$u_a(0,t) = \sum_{r:a\in r} D_r(\bar{\tau}(r, a_{tail}, t)) \tag{15}$$

$$u_a(L_a,t) = \sum_{r:a\in r} D_r(\bar{\tau}(r, a_{head}, t)) \tag{16}$$

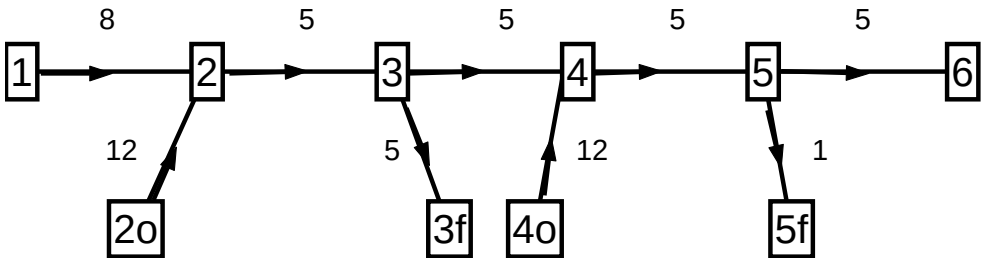



Figure 1: Example network for numerical testing

If the traffic flow model and the route model are perfectly consistent, then $u_a(0,t)$ should be equal to $v_a(0,t)$ and $u_a(L_a,t)$ should be equal to $v_a(L_a,t)$. In any case, we wish to capture the arrival order described by $u$, and use it as input for the traffic flow model. For that purpose we now formally define the anticipated arrival order (AAO), $Y_{a_1}$, for every diverge link $a_1$ such that $a_{1tail} = a_{head}$, as a piecewise linear function such that

$$Y_{a_1}\left(u_a(L_a,t)\right) = u_{a_1}(0,t) \quad \forall t = c \cdot dt; c \in \mathbb{N} \tag{17}$$

The entire process of obtaining the set of AAO's $\bar{Y}$ from the entire set of cumulative volumes $\bar{V}$ is depicted by the function $\bar{Y} = \Pi(\bar{V})$.

### 2.4 The iterative process

As mentioned before, the cyclic relationship between the traffic flow model and the route model requires integration through an iterative process. We start with $\bar{V}^0 = 0$. As a result, the initial AAO, $\bar{Y}^0 = \Pi(\bar{V}^0)$, are based on free flow trajectories. In any subsequent iteration $m$ we let $\bar{V}^m = V(\bar{Y}^{m-1})$, and $\bar{Y}^m = \Pi(\bar{V}^m)$. The process continues until convergance, which we defind as

$$max(|\bar{V}^{m+1} - \bar{V}^m|) \leq \epsilon \tag{18}$$

When the process converges we measure the solution's consistency by

$$DM = \max_a \max_t |u_a(0,t)) - v_a(0,t)| \tag{19}$$

## 3 Numerical results

This section presents numerical results for a network that contains 9 links with 8 potential routes, shown schematically in Fig. 1, together with link lengths in km. The forward wave velocity is 50 km/h and the backward wave velocity is 20 km/h for all links. The jam density is 360 v/km for all links. The exit capacity of link [5-5f] is 500 v/h and the exit capacity of link [5-6] is 5000

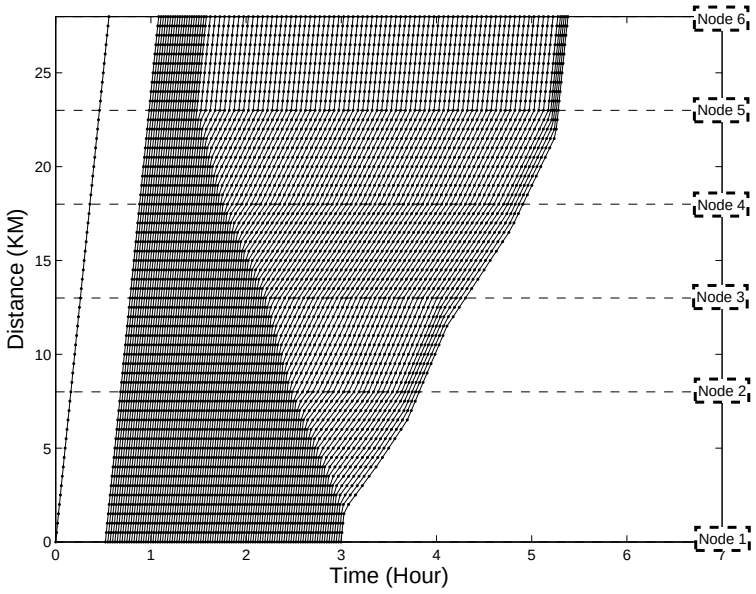



Figure 2: Trajectories along route [1-2-3-4-5-6]

v/h. Given origin-destination (OD) flows for $0.5 \leq \tau \leq 3$ are 4000 v/h from 1 to 6, 1000 v/h from 4o to 5f, 1000 v/h from 2o to 3f and zero otherwise. The network includes two merge and two diverge road-nodes. The pre-known ratio limits of the merge links are $\phi_{[2o,2]/[1,2]} = 0.35$ and $\phi_{[4o,4]/[3,4]} = 0.25$. Mesh-nodes were chosen so the time interval is quite small, $dt = 0.0025h$, and the spatial interval $dx_a = 0.5$km $\forall a \in A$. Route trajectories departure time interval was set to 0.01. The results described here are based on a solution obtained after running the model for 5 iterations.

Fig. 2 presents trajectories along the main route a=[1-2-3-4-5-6]. The time interval between the plotted trajectories departure time was chosen so that each consecutive trajectories present entrance of 100 vehicles to the route. Congestion in this model is created due to a bottleneck located at the exit of link [5-5f]. At time $\tau = 0.5$, vehicles that leave node 1 and travel along route a are moving at free flow speed towards node 6 with a travel time that is equal to 0.56h. Similarly, vehicles that leave node 2o and travel along route b=[2o-3-4-4f] are moving at free flow speed towards node 4f, thus their travel time is 0.44h. At time 0.8625h the first vehicles that travel along route c=[4o-4-5-5f] reach the exit of link [5-5f]. As a result of the relatively high demand for route c (1000 v/h compare to an exit capacity of 500 v/h only) congestion is created at the exit of link [5-5f]. The congestion propagates deeper into the upstream links. At 1.49h congestion reaches link [4-5] and as a result the travel time for vehicles that travel along route a increases. Fig. 3 presents routes travel time as a function of the departure time. At 1.76h, when the congestion reached the entrance to link [4-5], the relative ratio limit on merge road-node 4 determines the exit flow from links [4o-4] and [3-4]. Fig. 4 demonstrates how the exit flow ratio between link [4o-4] and [3-4] is equal to the pre-known ratio limit $\phi_{[4o,4]/[3,4]} = 0.25$, as long as the entrance to link [4-5] is congested. At 2.21h the congestion reaches link [2-3] and route b travel time increases as well. At 2.48h the entrance to link [2-3] is congested so road-node 2 relative ratio limit controls the flow of vehicles from links [2o-2] and [1-2] as can be seen in Fig. 4. After 3 hours, when the congestion already reached link

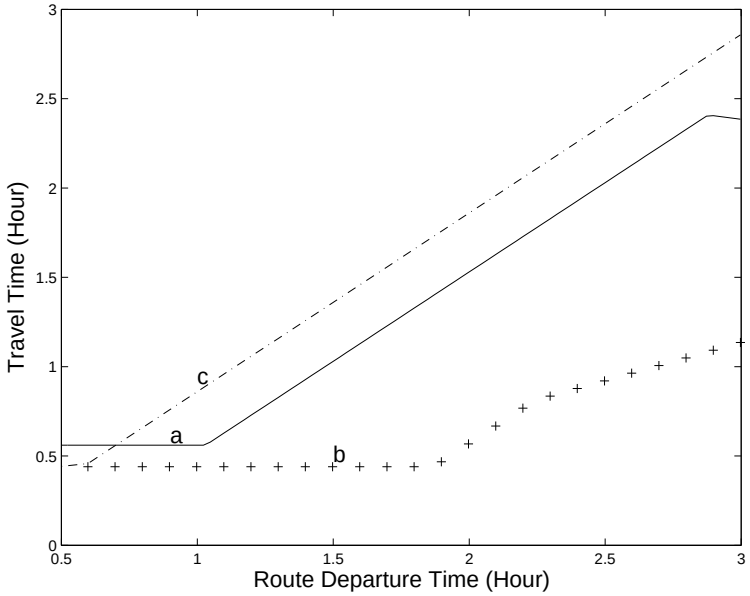



Figure 3: Route travel time as a function of departure time

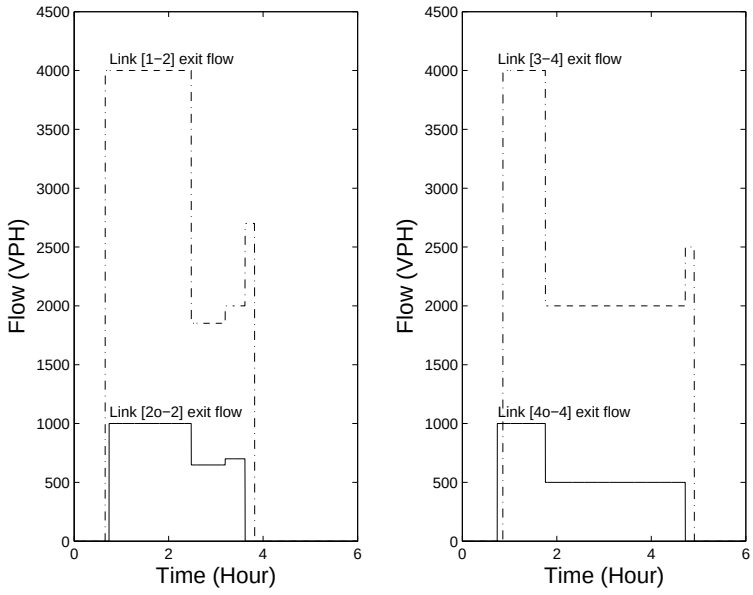



Figure 4: Exit flows on the merges upstream links

[1-2], the demand for all routes drops to zero. This is the point where the congestion starts to dissipate until at time 5.86h the last vehicle gets to its destination.

Consistency measures were computed at four critical points in the network: the entrances to the diverge downstream links [3-3f], [3-4], [5-5f] and [5-6]. As described in the previous section, the consistency measure evaluates the difference between the cumulative volume as determined by the traffic model solution to the value that was determined by the route level solution. The maximum difference was found at the entrance to link [5-6] between 5.19h and 5.2725h and is equal to 12.8577 vehicles. At the end of simulation the difference is equal to zero, meaning, all vehicles reached their destination as dictated from the demand. The convergence measure after 5 iterations was 0.1591.

To examine the way changes in the network characteristics affect the results, $\phi_{[4o,4]/[3,4]}$ was changed from 0.25 to zero. This is equivalent to adding a very simplistic ramp metering that is based on a detector placed just downstream of the merge road-node. Whenever the entrance to the merge link [4-5] is congested, vehicles traveling from link [4o-4] are not allowed to enter link [4-5]. This modification has caused significant changes to the network flow model. The complex behavior that resulted from this modification will be described in a future paper. Despite the complexity of the resulting situation, after only 5 iterations the model achieved a maximum consistency measure of 22.328 vehicles at the entrance to link [5-6], which is still quite acceptable. The results allow to evaluate the effect of adding ramp metering of this type, by considering the total system travel time that decreased from 13376h in the original example to 7584.3h in the modified example.

## 4 Conclusions and future research

In this paper we presented a continuous analytic dynamic network loading model that is based on the established theory of kinematic waves for the description of within-link traffic behavior. The solution at the single-link level relies on Daganzo's reformulation of the model as a continuous variations problem, and uses simple least-cost-path computations on dense discrete time-space mesh to solve the model. Our model relies on trajectories to ensure proper flow along routes. The main innovation of this paper is the integration beween the traffic flow model and the route model using the concept of anticipated arrival order (AAO). We show that AAO capture the most essential information from the route model, and passes it to the traffic flow model in a natural manner, with minimal undesirable disturbances to the traffic flow model. Numerical results on a non-trivial network show that the model produces consistent solutions whithin a relatively small number of iterations.

In order to verify the model performance, it is clearly needed to examine more examples, including examples of greater scale and complexity. Additional possible venues for future research include the evaluation of computational efficiency, and the relaxation of model assumptions. In particular, the examples discussed here assume triangular flow-density relationship, while the variations formulation and the lopsided mesh solution method allow other piecewise linear relationships, which could improve the model ability to replicate realistic traffic behavior.

## Acknowledgements

Financial support from the United States-Israel Binational Science Foundation through grant number 2002145 is greatly appreciated.

## References

Bar-Gera, H. (2005) Continuous and discrete trajectory models for dynamic traffic assignment. *Networks and Spatial Economics.* **5** (1), 41–70.

Carey, M. (2001) Dynamic traffic assignment with more flexible modelling within links. *Networks and Spatial Economics.* **1**, 349–375.

Daganzo, C.F. (1994) The Cell Transmission Model: A Dynamic Representation of Highway Traffic Consistent with the Hydrodynamic Theory. *Transportation Research.* **28B** (4) 269–287.

Daganzo, C.F. (1995) Properties of link travel time functions under dynamic loads. *Transportation Research.* **29B**, 95–98.

Daganzo, C.F. (2005a) A variational formulation of kinematic waves: Basic theory and complex boundary conditions. *Transportation Research.* **39B** (2), 187–196.

Daganzo, C.F. (2005b) A variational formulation of kinematic waves: Solution methods. *Transportation Research.* **39B** (10), 934–950.

Friesz, L.T., D. Bernstein, T.E. Smith, R.L. Tobin, B.W. Wie. (1993) A Variational Inequality Formulation of the Dynamic Network User Equilibrium Problem. *Operations Research*, **Vol. 41**, 179–191.

Kuwahara, M. and T. Akamatsu. (2001) Dynamic user optimal assignment with physical queues for many-to-many OD pattern, *Transportation Research*, **35B**, 461–479.

Lighthill, J.M., and J.B. Whitham. (1955) On Kinematic Waves: II, A Theory of Traffic Flow in Long Crowded Roads. *Proceedings of the Royal Society.* **229A** (1178), 316–345.

Lo, H.K. and W.Y. Szeto. (2002) A cell-based variational inequality formulation of the dynamic user optimal assignment problem. *Transportation Research.* **36B**, 421–443.

Newell, G.F. (1993) A simplified theory of kinematic waves in highway traffic, part I: general theory. *Transportation Research.* **27B** (4), 281–287.

Ran, B., and D. Boyce. (1996) *Modeling Dynamic Transportation Networks*, Second Edition, Springer-Verlag, Berlin.

Richards, P.I. (1956) Shock waves on the highways. *Operations Research.* **4**, 42–51.

Wu, J.H., Y. Chen, M. Florian. (1998) The Continuous Dynamic Network Loading Problem: A Mathematical Formulation and Solution Method. *Transportation Research.* **32B**, 173–187.

Xu, Y.W., J.H. Wu, M. Florian, P. Marcote, D.L. Zhu. (1999) Advances in the Continuous Dynamic Network Loading Problem. *Transportation Science.* **33**, 341–353.

# APPLICATION OF A NEW DYNAMIC TRAFFIC ASSIGNMENT MODEL FOR THE ASSESSMENT OF MOVING BOTTLENECKS

*Ido Juran:* *Technion,Israel ido.juran@pgl.co.il*
*Joseph N. Prashker:* *Technion,Israel trryosy@tx.technion.ac.il*
*Shlomo Bekhor:* *Technion,Israel sbekhor@tx.technion.ac.il*
*Ilan Ishai:* *Technion,Israel iishai@tx.technion.ac.il*

## Abstract
Moving bottlenecks in highway traffic are defined as a situation where a slow moving vehicle or convey disrupt the continuous flow of traffic. The effect of these moving bottlenecks on traffic delays is an important factor in the evaluation of network performance, especially in places that experience high level of heavy vehicles such as industrial areas, harbours, military bases and highway work zones, which cannot be properly assessed by existing transportation planning tools.

This paper describes a new dynamic traffic assignment model that can evaluate the effects of moving bottlenecks on network performance both in terms of travel times and travelling paths. The model assumes that the characteristics of the moving bottleneck such as travelling path, physical dimensions and desired speed are predefined and therefore is suitable for planned conveys. The model is based on a mesoscopic simulation with unique features that allow for the assessment of special dynamic characteristics of moving bottlenecks and the induced moving queue while ensuring all the constraints of traffic flow dynamics such as the first-in-first-out and causality principles. The paper presents numerical examples that provide insights regarding the impact of a moving bottleneck on a traffic network.

## 1 Introduction
A moving bottleneck (MBT) in highway traffic is a well-known phenomena where a slow moving vehicle or convey obstructs the traffic stream and causes delays. Many MBTs, such as military convoys and slow-moving heavy vehicles near work zones, are predictable in terms of their route, physical dimension and desired speed. The ability to evaluate the impact of MBTs on the performance of a traffic network and obtain optimal paths in their presence is important for short-term planning and traffic management.

Several studies (Gazis and Herman (1990), Newell (1998), Muñoz and Daganzo (2002), Daganzo and Laval (2005)) discussed the issue of MBTs. These studies mostly addressed the impact of MBTs on traffic flow within a single link. To the best of our knowledge, there have not been any attempts to evaluate, qualitatively or quantitatively, the impact of MBTs on the performance of transportation networks.

This paper presents a new model that facilitates the analysis of the impact of MBTs on a transportation network in terms of travel times and travelling paths. The model incorporates a Dynamic Traffic Assignment (DTA) procedure that employs a mesoscopic traffic simulation as a dynamic network loader. The model can be used to find a set of optimal travelling paths that minimize traffic delays in the presence of a MBT, and thus can be used as a tool to generate and evaluate guidance strategies.

## 2 Literature Review

### 2.1 Dynamic Traffic Assignment
Dynamic Traffic Assignment models capture the dynamic evolution of traffic flow in networks with time varying demand, under an assumed route choice behaviour. There is an extensive literature on the Dynamic User Equilibrium (DUE) principle, which reflects the objective of drivers to minimize their travelling costs. The Dynamic System Optimum (DSO) problem, which aims to optimize traffic distribution, is also of interest, specifically in situations where traffic may be controlled. A variety of DTA models have been proposed in the last two decades and an extensive review can be found at Ziliaskopoulos and Peeta (2002).

To be reliable, a DTA model should represent the traffic dynamics in a realistic fashion. This requires that four principles should be maintained:

• The FIFO principle that states that on average, traffic should exit a network link in the same order it has entered it.
• The causality principle that states that the link travel time for traffic entering at a given time should depend on the traffic entering at that time and earlier, but not on traffic entering later. Astarita (1996) and Carey et al (2003) emphasized the importance of the causality property in realistic dynamic traffic networks.
• Preventing artificially unreasonable traffic delays at nodes. A situation that might happen in DSO, as discussed by Ghali and Smith (1995) and Carey and Subrahmanian (2000).
• Proper queuing modelling which required to be physical rather than vertical in order to allow the assessment of spillbacks. Daganzo (1998) and Kuwahara & Akamatsu (2001) showed that the two approaches might yield different results since the location of the back of the queue is an important factor in determining shortest paths at each instant of time.

Existing mesoscopic simulation DTA models, such as DYNASMART (Jayakrishnan et al, 1994), DYNAMIT (Ben-Akiva et al, 1997) and CONTRAM (Taylor, 2003) maintain most of the above principles. However, they have two drawbacks in terms of traffic flow. The first is the use of deterministic queue clearance rates to determine queuing delays at intersections while the second is that they fail to preserve that causality principle due to the application of whole-link delay functions. This implies that all the vehicles located simultaneously on the same link travel at the same speed regardless of their position and that speed and density variations along the link are not assessed and therefore queues are associated only with the intersection capacity at the end of the link.

In the scope of the objectives of this paper, these assumptions are even more critical since they do not allow for the integration of a moving bottleneck, which by definition moves at a lower speed than the average traffic and may create a moving queue along the link.

2.2 Moving Bottleneck Theory

Gazis and Herman (1992) were the first to introduce a theoretical analysis of a MBT on a two-lane roadway. They defined three distinct regions around a MBT: a free flow region upstream of the moving queue, a moving queue region and an escape region downstream of the moving queue. Based on the LWR model (Lighthill and Whitham (1955), Richards (1956)), and assuming that the traffic in the moving queue is equally spread in both lanes, they derived the following equation for the traffic flow that passes a MBT:

(1) $q_r = (v_d - v_b) * k_d$

Where $v_b$ is the MBT speed and, $v_d$ and $k_d$ are, respectively, the speed and density within the escape region. The average speed of vehicles in the moving queue (with density $k_1$) was then formulated as:

(2) $v_{mq} = v_b + (v_d - v_b) * \frac{k_d}{2k_1} = v_b + \frac{q_r}{2k_1}$

Newell (1998) further developed the theory of MBTs. Relying on kinematic wave theory, he showed that by using a moving coordinates system, the problem can be simplified and solved as a problem of a static bottleneck on a scaled-down version of the freeway's flow-density curve.

Muñoz and Daganzo (2002) conducted experiments to model the effects of a MBT on traffic using occupancy detectors located along the US Interstate Freeway I-880 in California. The observations supported Newell (1998) and Gazis and Herman (1992) assumption that traffic around a MBT behaves as predicted by the LWR solution to the kinematic wave traffic theory. They extended the research of the previous authors and showed that the maximal flow downstream of a MBT increases monotonically with the bottleneck speed and it is considerably higher than in a stationary bottleneck. To model these findings they offered a simple linear relationship between the downstream capacity flow and the MBT speed:

(3) $q_d = q_r(0) + [Q_d - q_r(0)] * \left( \frac{v_b}{v_f} \right)$

Where $q_r(0)$ denotes the capacity flow of passing lane at a static bottleneck and $Q_d$ denotes the capacity flow of the passing lane when the freeway is experiencing its maximum sustainable flow and $v_f$ is the

link's free flow speed. By analyzing the experimental data, the authors offered the following relationship between the rate in which vehicles pass the MBT, the bottleneck speed and the downstream capacity flow:

(4) $q_r = q_d * (1 - \frac{v_b}{v_f})$

Daganzo and Laval (2005) proposed to model MBTs with a series of pre-defined static bottlenecks with equal capacity. However this method can capture only the effects that the MBT has on the traffic flow but not vice versa. That is, if the MBT has to slow down due to traffic congestion, then the defined time-space series is no longer valid.

In summary, there are relatively few studies that examine the behaviour of traffic flow around a MBT in highways. There has not been an attempt to assess their impact on the performance of a traffic network.

# 3 Model Description

3.1 General Definitions and Overall Structure

The outline of the problem solved by the proposed model modifies the general definition of the DTA problem and can be stated as follows – Given a traffic network consisting of nodes and links, a time dependent OD matrix for the analysis period and one or more MBTs with a pre-specified path, desired speed and physical properties, determine the set of paths that will carry traffic, based on a specified route choice principle. The model can solve both the DSO and the DUE problems but it is described in this paper using the DSO problem since it was created as a tool for optimal route guidance. The model can also assess situations where the MBT travels slower than its desire speed due to traffic congestion.

The proposed model is essentially a mesoscopic simulation-based DTA that incorporates an enhanced network loading procedure. The model receives as an input a traffic network and time varying trip matrices. Each link in the network is characterised by length, free flow speed, jam density and number of lanes, while each node is characterized by intersection control type and specifications. The time varying trip matrices specify the number of trips that wish to travel from each origin to each destination at each (demand) time interval. A typical length of these intervals ranges between 10-15 minutes depending on the available data.

The algorithm has a bi-level structure and it is solved by an iterated averaging algorithm. At each iteration, the upper problem (UP) solves a least marginal travel time path problem with time dependent link travel times and sets a time-space expended temporal sub network that includes only links that are part of these paths. The sub network is stored as a set of binary variables where each variable is determined as follow:

(5) $\psi_{r,d}^{i,t} = \begin{cases} 1 & if\ a\ path\ originating\ from\ r\ at\ time\ d\ crosses\ node\ i\ at\ time\ t \\ 0 & otherwise \end{cases}$

The lower level (LP) solves a dynamic system optimum traffic assignment problem on the defined sub network. LP consists of a least marginal travel time path algorithm and a mesoscopic simulation. At each iteration, the least marginal time paths are determined and stored as a set of splitting rate variables $\Phi$, which determine the turning proportions of traffic at each network node. Each variable $\phi_{rd}^{ij}$ is defined as the fraction of traffic originating node *r* at departure time *d* that leave node *i* to an immediate downstream node *j*. Once the set of splitting rates is determined, the mesoscopic simulation assigns the corresponding OD demand and computes new travel times and marginal times that are used to determine a new set of paths. The Method of Successive Averages (MSA, Powell and Sheffi (1982)) is then used as the averaging procedure and the splitting rates are the averaged variables. LP converges when the splitting rate values of two successive iterations are sufficiently close. Then, the travel times and marginal travel times calculated in last simulation are sent back to UP for next iteration. The entire program converges when the difference between two consecutive sub-networks of two successive solutions of UP is within the convergence criteria.

The proposed algorithm resembles previous works by Janson (1992), Janson & Robles (1995) and Jayakrishnan et al. (1995, 1999), which offered it as a solution to the difficulty of applying averaging algorithms due to the temporal nature of the links-paths incidence matrix in dynamic networks, which might

cause discontinuity of flows. The proposed algorithm performs a full network loading simulation, which assures the continuity of flows but applies the bi-level structure to enhance the MSA convergence. Following the heuristic method offered by Cascetta and Pastorino (2001) of restarting the algorithm at specific points, the bi-level structure restarts the algorithm and determines new directions search based on the optimal solution of the assignment problem at the previous iteration. A detail description of the model formulation and properties can be found in Juran (2005).

3.2 Network Loading Procedure

The network loading procedure is based on a mesoscopic simulation that simulates the movement of packets of traffic using speed-density macroscopic relationships. A packet $p_{r,d,a}^{i,t}$ is defined as a packet that departed origin node *r* at time *d*, and arrived at the tail node *i* of link *a* at time *t*. When a packet is initiated, its load is equal to the total demand from its origin at its departure time to all destinations. The packet follows its pre-defined path and at each node it is divided proportionally to smaller packets according to the splitting rate variables at that node. Each of the smaller packets is headed for different sub set of destinations. The simulation terminates when all packets from all origins and departure times have arrived their destinations. Splitting rates may be adjusted within the course of the simulation to reflect changes in the composition of destinations due to limited available capacity of downstream links.

To allow the integration of a MBT within the simulation and the preservation of the causality principle of traffic flow, the simulation does not use link-based delay functions but rather assigns each packet a unique speed on each link, based on the packet load and the links' characteristics. The basic assumption calculating this speed is that the leading vehicle of the packet is not influenced by vehicles behind it and it may travel at a free flow speed, until it encounters the packet in front. Hence the distance that the leading vehicle travels in one interval is the product of the link's free flow speed $(Sf_a)$ and the length of the time interval:

(6) $\lambda = f * \Delta t * Sf_a$

Since the length of the demand time intervals is relatively long, $\lambda$ is adjusted by factor $f$ that reduces it to the distance that can be traveled in a smaller period of time, usually one minute. The other vehicles in the packet follow the leading vehicle and will be spread uniformly throughout this distance on the link's lanes $nl_a$. Hence, the density $k_{r,d,a}^{i,t}$ and speed $s_{r,d,a}^{i,t}$ of a packet with load $H_{r,d,a}^{i,t}$ is:

(7) $k_{r,d,a}^{i,t} = H_{r,d,a}^{i,t} / (\lambda * nl_a)$

(8) $S_{r,d,a}^{i,t} = Sf_a * \left[ 1 - \left( \frac{k_{r,d,a}^{i,t}}{K_{ja}} \right)^{\beta} \right]^{\alpha}$

Where $K_{ja}$ is the link's jam density and α and β are calibrating parameters. It is important to emphasize that although the packet's speed is calculated, as it was the only packet on the link, its actual progression is subjected to packets traveling ahead in order to ensure FIFO.

The current model employs physical queues by, as common in other models, decomposing each link in the network to two parts, a moving segment and a physical queuing segment located at the end of the link. However since the MBT may create a second (moving) queue, links in which it passes are divided to 4 segments: the bottleneck and the queue it induces (both of them are denoted by the QMBa segment in Figure 1 where the bottleneck is denoted by MB), two moving segments (one behind the moving bottleneck (denoted by M1a) and the second in front of it (M2a) and a queuing segment (Qa) at the end of the link.

Using this approach, the following process is assumed: When a packet enters link a at point A, it moves according to its calculated desire speed throughout segment M1a, constrained by the packet in front. Thus its dynamic motion in this segment is expressed by:

(9) $X_{r,d,a}^{i,t,\tau+1} = \min \left\{ X_{r,d,a}^{i,t,\tau} + S_{r,d,a}^{i,t} * \Delta tsim(\tau) / (K_{ja} * nl_a) \, , \, X_{r',d',a}^{i,t',\tau+1} - L_{r',d',a}^{i,t'} \right\}$

(10) $L_{r,d,a}^{i,t} = H_{r,d,a}^{i,t} / (K_{ja} * nl_a)$

Where $X_{r,d,a}^{i,t,\tau}$ is packet's $p_{r,d,a}^{i,t}$ location at time $\tau$ and $L_{r,d,a}^{i,t}$ is its length. $\Delta tsim(\tau)$ is the length of the simulation time interval at time $\tau$ and packet $p_{r',d',a}^{i,t'}$ is assumed to travel directly in front of $p_{r,d,a}^{i,t}$.

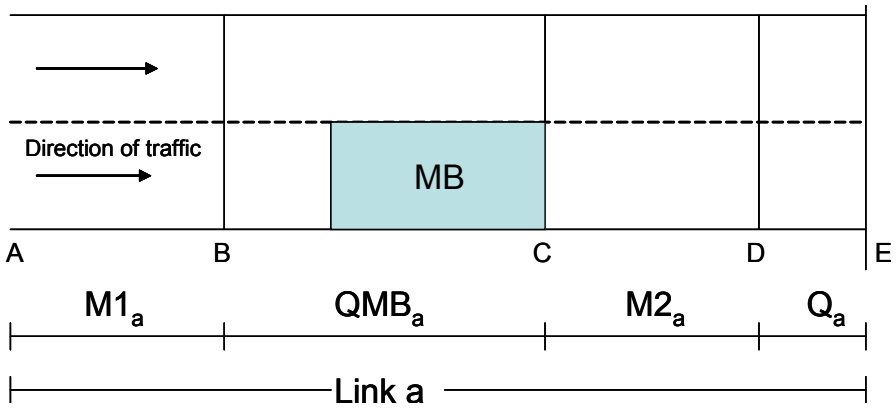



**Figure 1. Four segments link representation**

The simulation time interval (STI) is a second type of time interval that is dedicated only to the simulation. It is required since the demand time interval is too long and cannot capture traffic dynamics. The STI length ranges between 0.5 and 10 seconds and it is continuously recalculated within the simulation.

When the packet reaches point B it joins the moving queue behind the MBT. At this stage, the vehicle adjusts its speed to the speed of the moving queue. The MBT travels in its pre-specified speed unless it encounters a slower packet and is required to adjust its speed. When the vehicle passes the moving queue segment and reaches the front of the bottleneck (point C), it enters an escape region M2 where it first travels at a speed slightly higher than its desire speed in order to compensate on the time it had lost and then returns to its desire speed. The vehicle continues to travel until it reaches (and joins) the queue at the end of the link (point D), if the latter exists at that time. At the queue, the packet moves a different distance at each STI according to the number of vehicles that can exit the link at each time interval ($c_a^\tau$, exit capacity), again constrained by the packet in front:

(11) $$X_{r,d,a}^{i,t,\tau+1} = \min\left\{X_{r,d,a}^{i,t,\tau} + c_a^\tau * \Delta tsim(\tau),\ X_{r',d',a}^{i,t',\tau+1} - L_{r',d',a}^{i,t'}\right\}$$

When the packet reaches the link's head node (point E), the link's travel time is calculated. At this point, the packet either completely exits the link, or when its load is greater than the exit capacity, it is divided to two parts. The first part, with load equal to the exit capacity, exits the link while the second part remains on the link and forms a physical queue. The part of the packet that exits the link is then divided among downstream links according to its splitting rates defined at this intersection.

The temporal exit capacities are recalculated endogenously for each link in each STI for each arriving packet as the minimum of two values: (a) the available input capacity of downstream links (b) the maximum vehicles that can pass through the head node intersection in one time interval. The latter is calculated based on the intersection control type. Four types are defined: signalized intersections, un-signalized intersections, metered ramp and ramps without meter. The exit capacities based on the available input capacity of downstream links are calculated in a link-to-link fashion based on the splitting rate variables in order to reflect the true demand of the packet. For example, the packet may not need to utilize in full the available input capacity of a downstream link. therefore the actual exit capacity, which determines the progression of vehicles in the queue, might be less than the sum of available input capacities of downstream links.

Let $C(l_b, a_e)$ represent the link-to-link transferring capacity from links entering node *i*, ($l_b \in P(i)$), to links exiting node i ($a_e \in O(i)$) at time $\tau$ and assume that $\bar{p}_{r,d,a_e}^{i,t}$ is the last (most backward) packet on link $a_e$. Then the link-link and total exit capacities (the some of the link-to-link transferring capacity) are:

$$(12)\ c^{\tau}_{(l_b,a_e)} = \frac{|\phi^{i,j_e}_{r,d}|}{\sum_{l_b \in P(i)} \phi^{i,j_e}_{r,d}} * K_{ja_e} * nl_{a_e} * (\overline{X}^{i,t,\tau}_{r,d,a_e} - \overline{L}^{i,t}_{r,d,a_e}) \quad \forall a_e = (i, j_e) \in O(i), l_b \in P(i), r \in R, d \in D$$

$$(13)\ c^{\tau}_{(l_b)} = \sum_{a_e \in O(i)} \left[ \frac{\left[\phi^{i,j_e}_{r,d}\right]}{\sum_{l \in P(i)} \phi^{i,j_e}_{r,d}} * k_{ja_e} * nl_{a_e} * (\overline{X}^{i,t,\tau}_{r,d,a_e} - \overline{L}^{i,t}_{r,d,a_e}) \right]$$

Figure 1 illustrates the base state where the bottleneck is located between two moving segments and all four segments exist. However the dynamic motion of the MBT creates other states. For example, if it has just, partially, entered the link, then segment M1 does not exist and the MBT constrains the exit capacities of upstream links. Vehicles entering the link join the moving queue at the side of the bottleneck. A different state, for example, occurs when the MBT joins the queue at the end of the link and segment M2 does not exist. An analysis of sample networks revealed that the progression of a MBT on a link might lead to 14 different feasible states of the link depending on the location of the MBT, the size of the induced moving queue and the size of the queue at the end of the link. All these states have to be addressed within the simulation procedure given the fact that they may be encountered and cannot be ignored.

Considering the above approach, the following three issues are assessed within the simulation:

- The length of each of the four segments in each simulation time interval
- The methods in which vehicles move in each segment
- The marginal travel time within the moving queue

**The length of each of the four segments is continuously updated in each STI in the following manner:**
First, the length of segment Q is determined based on the number of vehicles waiting in the queue at time $\tau$ ($VQ^{\tau}$), the jam density and the number of lanes:

$$(14)\ Q^{\tau} = \frac{VQ^{\tau}}{K_j * nl}$$

Next, the location of the MBT is determined (point C in Figure 1). If the MBT is part of the physical queue and section M2 does not exist, then its location at each STI depends on the link's temporal exit capacities:

$$(15)\ M^{\tau+1}_a = M^{\tau}_a + \Delta t_{sim}(\tau) * c^{\tau}_a / (k_{ja} * nl_a)$$

Alternatively, if the MBT is not within the physical queue, then its location at time $\tau$ is determined by:

$$(16)\ M^{\tau+1}_a = M^{\tau}_a + \Delta t_{sim}(\tau) * \min(v_b, v_d)$$

Where $v_d$ is the traffic speed downstream of the bottleneck.

Once the location of the MBT has been determined, M2 length is calculated as the distance between the upstream end of segment Q (point D) and the location of the MBT (point C). The length of M1 depends on the location of the upstream end of the moving queue and it is discussed later.

## The packets' speed in each segment is determined as follows:

In segment M1 it is assumed that packets travel in their desire speed (while maintaining FIFO). In segment QMB, they travel according to the moving-queue speed. In segment M2 it is assumed that the packets discharged from the moving queue first travel in an escaping speed and then proceed with their desire speed while in queue Q vehicles are progressed according to the temporal exit capacities.

Hence, four variables have to be determined: (a) The speed of the moving queue; (b) The rate in which vehicles pass the MBT; (c) The escaping speed of vehicles discharged from the moving queue and (d) The location of the upstream end of the moving queue.

To determine the speed of the moving queue, Gazis and Herman (1992) formulation (equation 2) was modified to reflect Muñoz and Daganzo (2002) correction that the capacity downstream of the bottleneck is a function of the MBT speed (equation 4). Hence the speed of the moving queue is determined by:

(17) $v_{mq} = v_b + q_d * \left(1 - \frac{v_b}{v_f}\right) \Big/ 2k_1$

The rate in which vehicles pass the MBT is calculated based on Muñoz and Daganzo (2002) theory constrained by the available feasible space, that is:

(18) $q_r^\tau = \min\left\{ q_d * (1 - \frac{v_b}{v_f}) * \Delta t_{sim}(\tau) \,,\, \left( X_{r,d,a}^{i,t,\tau} - L_{r,d,a}^{i,t} - M_a^\tau \right) * k_{ja} * nl_a \right\}$

Packet $p_{r,d,a}^{i,t}$ is assumed to travel just in front of the MBT. Since STI length is very short, it may be take required several intervals for a packet to pass the MBT. Therefore, in each interval, the fraction of the packet is calculated until it has completely passed.

The escaping speed in segment M2 is also calculated based on Muñoz and Daganzo (2002) suggestion that the capacity of downstream flow varies linearly with the MBT speed. Once the downstream capacity flow is determined, the escaping speed $v_d$ is evaluated based on the assumption that a simple Greenshilds (1934) traffic flow relationships can be applied to the road, though other macroscopic relationship can be applied.

Lastly, the location of the upstream end of the moving queue is calculated in each STI based on the MBT location and the number of vehicles in the packets located within the moving queue. Let $\eta_a^\tau$ denote the available space adjacent to the MBT on link a at time $\tau$ and $P_a^\tau$ denote the set of all packets located on link a at time $\tau$, then the length of the moving queue behind the MBT at time $\tau$ is calculated as:

(19) $\xi_a^\tau = \dfrac{\sum_{p_{r,d,a}^{i,t} \in P_a^\tau} H_{r,d,a}^{i,t} * \delta_{r,d,a}^{i,t,\tau} - \eta_a^\tau}{k_1 * nl_a}$

(20) $\eta_a^\tau = \left\{ \min(\lambda_b, M_a^\tau, L_a - \gamma_a^\tau) \right\} * k_1 * (nl_a - \varpi_b)$

(21) $\delta_{r,d,a}^{i,t,\tau} = \begin{cases} 1 & \textit{if packet } p_{r,d,a}^{i,t} \textit{ is in the moving queue at time } \tau \\ 0 & \textit{otherwise} \end{cases}$

Where $\gamma_a^\tau$ is the location of the rear end of the MBT on link a at time $\tau$ and $\varpi_b$ is its width. Once the length of the moving queue is determined, the location of its upstream end at time $\tau$ is given by:

(22) $o_a^\tau = \gamma_a^\tau - \xi_a^\tau$

## The marginal travel time within the moving queue is evaluated as follows:

Both link travel times and marginal travel times are measured within the course of the simulation. As each packet possess different traffic characteristics, these values are evaluated at the packet level rather than at the link level (which is an average of packets' characteristics). Travel times are measured by recording the entrance and exit time of the packet from the link and calculating the difference between them. The link's marginal travel times is evaluated as the difference between the travel time that the packet experienced and the travel time experienced by a virtual packet which travel in parallel with the actual packet and carries a specified additional number of vehicles, That is, for each packet that enters a link at a given time, a virtual packet is created and the difference in travel time between them is used to evaluate the link marginal time. The travel time difference is the sum of the difference in travel times in the moving segments, the delay in the moving queue and the delay in the queue at the end of the link.

The additional delay caused by each vehicle in the queue at the end of the link is based on deterministic queuing models with a fixed discharge rate. At the moving queue, the model assumes that the additional delay caused by each vehicle is a decreasing function of the bottleneck speed with two boundary points: (1) the bottleneck speed is zero where it acts as a stationary bottleneck that induces ω seconds delay for each additional vehicle, and (2) the bottleneck speed is equal to the approaching packet's speed where it does not

induce any delay. Using these two boundary points, the following function is applied to calculate the additional delay of each vehicle within the moving queue:

$$(23)\ \omega_m = \frac{\omega}{S_{r,d,a}^{i,t\ 2}} * V_b^2 - \frac{2\omega}{S_{r,d,a}^{i,t}} * V_b + \omega$$

# 4 Experimental Study

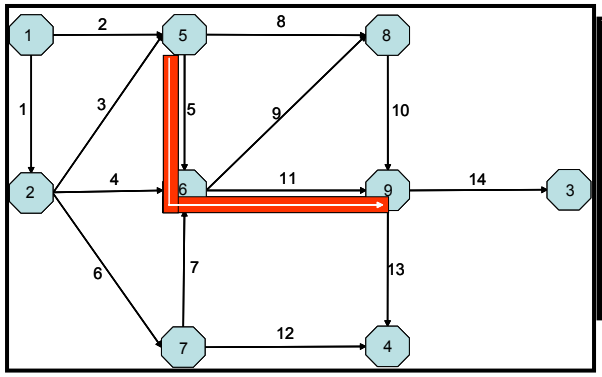
The model was applied in two experimental studies. The first study, which did not include a MBT, was performed to verify the validity of the developed DTA model and included experiments of both DUE and DSO, using several degrees of traffic demand. The experiments showed that the DSO traffic flow pattern is most effective when the traffic network experiences moderate to high congestion levels. When the network observes low or extremely high levels of congestion, the DSO pattern does not produce significantly better results than DUE in terms of total network travel times. These findings are consistent with the previous studies of Mahmassani and Peeta (1993) and Wie et al (1995).

The second experimental study was carried out in order to derive insights regarding the impact that a MBT has on a traffic distribution among the available paths in the network and on travel times.

## 4.1 The Effect of a MBT on Path Utilization

To observe the effects that a MBT has on path utilization, experiments were conducted on the sample network illustrated in Figure 2. It consists of 14 links, 9 nodes, 2 origins (nodes 1 & 2) and 2 destinations (3 & 4). The trip matrix represents an hourly demand of 6400 trips divided evenly among the 4 OD pairs. This matrix was multiplied by several loading factors that generated several levels of traffic demand. The matrix was applied to the network in 12 time intervals of 5 minutes each, in a gradual profile. The MBT departs node 5 at interval 3 and travels to node 9 via node 6. Tests were performed with several MBT speeds and several degrees of traffic demand. All tests were based on the DSO assignment principle.

An analysis of the results revealed that all traffic originating node 1 (from all departure intervals) for both destinations passed through node 9 using one of 3 paths: path #11 (with links 2-8-10 and free flow time of 50), path #12 (links 2-5-11 and free flow time of 40) and path #13 (links 2-5-9-10 and free flow time of 56). Figure 3 compares the fraction of traffic that uses each path in networks with and without a MBT (base network) as a function of network load. In the base network, as expected, at low traffic (LF=0.8) trips are divided between paths 11 and 12 with more trips assigned to path 12 that has a lower free flow travel time. No traffic is assigned to path 13. At higher demand levels, the use of paths 11 and 13 increases at the expense of path 12 due to increased interactions with traffic originated node 2. In the MBT network the results are different. Since the MBT travels through link 11, traffic is significantly diverted from path 12 to the alternative paths, mostly path 11, specifically in low demand levels. Therefore path 11 usage is significantly higher in the MBT network (45%) than in the Base network (30%), in low traffic (LF=0.8). As the average network congestion increases, the distribution of trips between the three available paths becomes similar in the two networks and at high demand levels (LF=1.5, 1.8 and 2.0) the utilization of paths is almost identical for the two networks. Hence the effects of the MBT on the paths utilized by traffic departing origin 1 is significant at low to moderate congestion but hardly noticed in higher levels. A similar analysis of traffic originating node 2 yielded the same conclusion, that is, the MBT affects the optimal distribution of trips in the network as it diverts traffic from links in its path especially in low to moderate congestion.




| ARC | FN | TN | KJAM | FFTT | LEN | FFSPEED | LANES |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 100 | 8 | 12 | 90 | 2 |
| 2 | 1 | 5 | 100 | 12 | 18 | 90 | 2 |
| 3 | 2 | 5 | 100 | 16 | 24 | 90 | 2 |
| 4 | 2 | 6 | 100 | 14 | 21 | 90 | 2 |
| 5 | 5 | 6 | 100 | 10 | 15 | 90 | 2 |
| 6 | 2 | 7 | 100 | 20 | 30 | 90 | 2 |
| 7 | 7 | 6 | 100 | 12 | 18 | 90 | 2 |
| 8 | 5 | 8 | 100 | 20 | 30 | 90 | 2 |
| 9 | 6 | 8 | 100 | 16 | 24 | 90 | 2 |
| 10 | 8 | 9 | 100 | 18 | 27 | 90 | 2 |
| 11 | 6 | 9 | 100 | 18 | 27 | 90 | 2 |
| 12 | 7 | 4 | 100 | 16 | 24 | 90 | 2 |
| 13 | 9 | 4 | 100 | 14 | 21 | 90 | 2 |
| | 9 | 3 | 100 | 24 | 36 | 90 | 2 |

**Figure 2. Sample network**

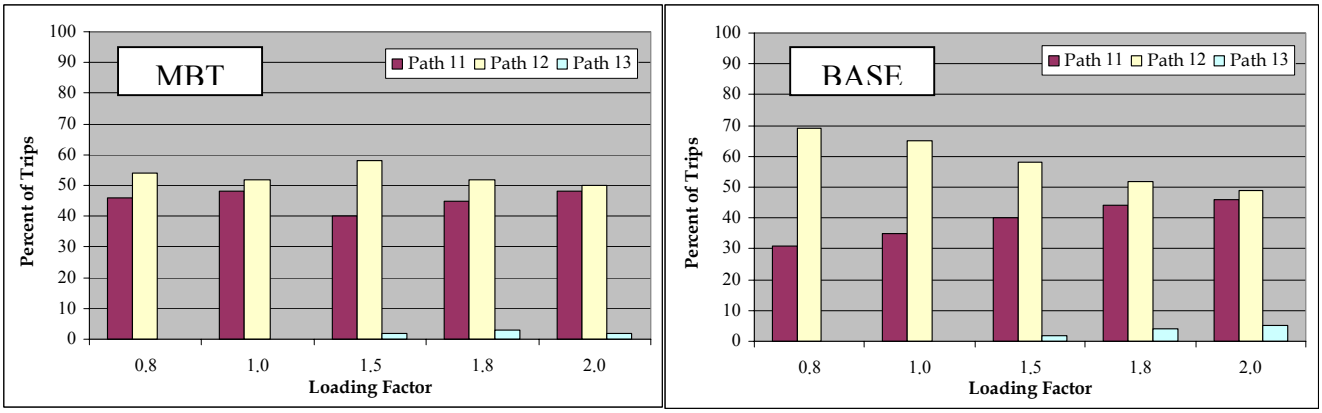



**Figure 3. Trip distribution on paths from origin 1**

4.2 The Effects of a MBT on Network Travel Times

To observe the effects of a MBT on network travel times, experiments have been conducted on the Tel Aviv metropolitan network using data obtained from the DOT. The network has 182 links including freeways, regional highways, arterials and suburban Freeways. The network includes 22 traffic zones that create 462 OD pairs. The OD matrix, that has a total of 173,397 trips, reflects demand of a peak hour in the afternoon. The network, traffic zones and their respective productions and attractions trips are presented in Figure 4.

The MBT travels south on road no. 4 from point A to point B. It is 0.5 km long and 1-lane wide. Tests were performed with two MBT speeds – 40 and 50 kph. As in the sample network experiment, the network was tested under different levels of loading factors leading to several degrees of traffic congestion ranging from very light to very heavy traffic. Each OD matrix was loaded to the network in 10 intervals of 6 minutes each, in a gradual profile. The total network travel time and average speed are reported in table 1.

The results indicate two things: The first shows, in accordance with the result obtained in the sample network tests, that the disruption of the MBT to the average traffic has an inverse relationship with its speed. That is, as the speed of the MBT increases, the total travel time spent in the network decreases. This result seems logical for two reasons: (a) as the bottleneck travels faster, its speed becomes closer to the average traffic speed and therefore it creates less interruption to traffic and (b) as the MBT's speed is higher it arrives its destination faster and therefore affects the traffic in fewer time intervals.

The second conclusion suggests that the disruption of the MBT is a function of overall network congestion. When the network experiences very light traffic (LF=0.1), the MBT has practically no effect on network travel times since there are very few vehicles that might be affected. As the level of demand increases, the disruption to traffic grows since the MBT affects more vehicles. This continues until the network is loaded with a certain demand level in which the additional network delay imposed by the MBT is the highest (LF=0.6). Beyond this point, the opposite phenomenon occurs as the MBT effects decline with higher levels of demand. This happens since as of this point, the average network speed decreases more rapidly and becomes closer to the bottleneck's velocity. Hence, beyond the maximal point, as average congestion grows the MBT causes less delay to traffic. Figure 5 illustrates this result. It depicts the relative additional network travel time induced by the MBT with respect to a network with out a MBT. The figure clearly indicates that the MBT that travels in a speed of 50 kph causes less delay than the one that travels in 40 kph and that their disruption to traffic is a function of the demand level.

| LF | Total Network Travel Time (hr) | | | Average Network Speed (kph) | | |
|---|---|---|---|---|---|---|
| | Base | 40 | 50 | Base | 40 | 50 |
| 0.1 | 3,142 | 3,144 | 3,143 | 79.8 | 79.7 | 79.8 |
| 0.3 | 10,684 | 10,854 | 10,763 | 74.0 | 71.6 | 71.7 |
| 0.5 | 13,894 | 14,650 | 14,579 | 73.1 | 71.6 | 71.6 |
| 0.6 | 15,067 | 16,344 | 15,819 | 72.1 | 69.7 | 71.3 |
| 0.8 | 19,308 | 20,095 | 19,878 | 66.8 | 66.1 | 66.4 |
| 1.0 | 23,054 | 23,822 | 23,666 | 61.7 | 61.5 | 60.9 |
| 1.1 | 25,414 | 26,021 | 25,907 | 59.8 | 59.5 | 59.7 |
| 1.3 | 32,058 | 32,639 | 32,552 | 53.5 | 53.8 | 53.3 |

**Table 1. Tel Aviv network, total network travel time and average speed**

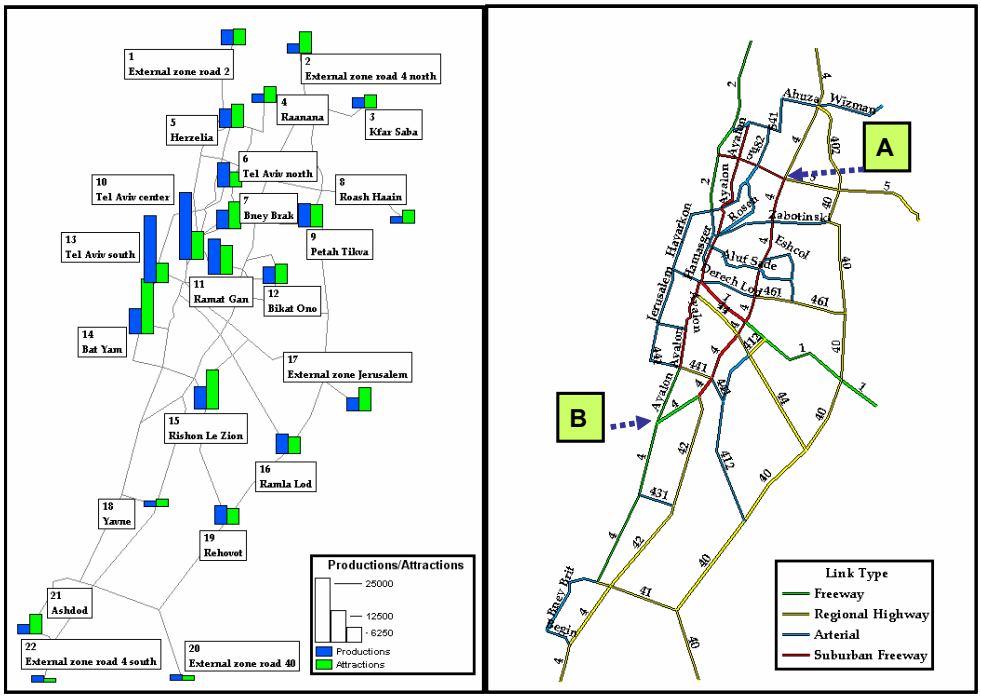



**Figure 4. Tel Aviv metropolitan network and demand data**

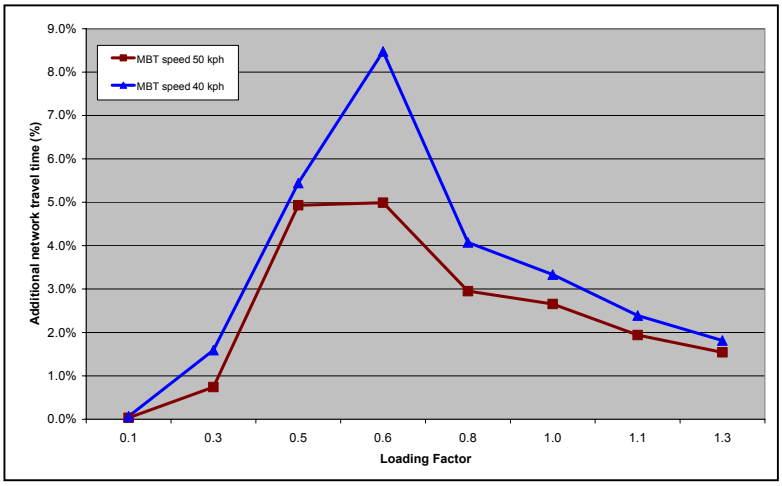



**Figure 5. Relative additional network travel time induced by the MBT**

# 5 Summary and Conclusions

This paper presented a dynamic traffic assignment model that can assess the phenomena of moving bottlenecks and moving queues and evaluates their impact on network performance and network congestion. The model incorporates an enhanced mesoscopic simulation-based network loading procedure that is able to assess the complex dynamic interactions between the MBT and the traffic in an efficient and original methodology while maintaining all dynamic constrains of traffic flow including the causality principle.

The paper presents results of an experimental study conducted with several levels of demand and MBT speeds. The study found that a MBT affects both the optimal distribution of traffic and travel times in the network. The study also found that the MBT disruption to traffic is a function of the average network congestion and it may be possible to define a certain level of demand where the bottleneck induces a maximum of additional network travel times. The MBT effect has also an inverse relationship with its desire speed where MBT that travel at higher speeds cause less delay, except for very light and very heavy traffic conditions where the bottleneck speed has very little impact

The experimental study appears to indicate that the model developed in this research yields a reasonable forecasting of the impact of a MBT on the traffic flow and network performance. However it should be noted that it is based on the existing theory and practical research of traffic mechanism in the vicinity of a MBT, which is limited. Further experimental studies are needed to develop a complete theory of MBT and moving queue phenomena based on observations. This theory may then be incorporated in the developed model.

## References


1. Astarita V. (1996) A Continuous time link model for dynamic network loading models based on travel time function. *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*. Elsevier Science Publishers, pp. 79-102
2. Ben-Akiva M. et al. (1997) Simulation Laboratory for Evaluating Dynamic Traffic Management Systems. *ASCE Journal of Transportation Engineering*. Vol. 123, 283-289
3. Carey M. and Subrahmanian E. (2000) An Approach to Modeling Time Varying Flows on Congested Networks. *Transportation Research.* Vol. 34B, pp. 157-183
4. Carey M., Ge Y.E. and McCartney M. (2003) A Whole-Link Travel Time Model with Desirable Properties. *Transportation Science.* Vol. 37, No.1, pp. 83-96
5. Cascetta, E., and M. Postorino (2001) Fixed-Point Models for Estimating or Updating Origin/Destination Matrices from Traffic Counts. *Transportation Science.* Vol. 35, pp. 134-147
6. Daganzo C. F. (1998) Queue Spillovers in Transportation Networks with a Route Choice. *Transportation Science*. Vol. 32, No. 1, pp. 3-11
7. Daganzo C.F. and Laval J.A. (2005) On the Numerical Treatment of Moving Bottlenecks. *Transportation Research*. Vol. 39B, pp. 31–46
8. Gazis, D.C. and Herman, R. (1992) The Moving and 'Phantom' Bottlenecks. *Transportation Science*. Vol. 26, pp. 223-229
9. Ghali M.O. and Smith M.J. (1995). A Model for the Dynamic System Optimal Traffic Assignment Problem. *Transportation Research.* Vol. 29B, pp. 155-170.
10. Janson B.N. (1992). Dynamic Traffic Assignment with Schedule Delay. *Paper Presented at 72nd Annual Meeting*. Transportation Research Board, Washington DC.
11. Janson B. N. and Robles J. (1995) Quasi-Continuous Dynamic Traffic Assignment Model. *Transportation Research Record*. 1493, pp. 199-206, Transportation Research Board, Washington DC.
12. Jayakrishnan R., Mahmassani H.S. and Hu T., (1994) An Evaluation Tool for Advanced Traffic Information and Management Systems in Urban Network. *Transportation Research*. Vol. 2C, pp. 129-147
13. Jayakrishnan R., Tsai W. K. and Chen A., (1995) A Dynamic Traffic Assignment Model with Traffic Flow Relationships. *Transportation Research*. Vol. 3C, pp. 51-72
14. Jayakrishnan R., Chen A. and Tsai W. K., (1999) Freeway and Arterial Traffic Flow Simulation Analytically Embedded in Dynamic Assignment. *Transportation Research Record.* Vol. 1678, pp. 242-250, Transportation Research Board, Washington DC.
15. Juran I. (2005) *Improvement of Dynamic Work- Zone Operations Using ITS*. Ph.D. thesis, Israel Institute of Technology.
16. Kuwahara M. and Akamatsu T., (2001) Dynamic User Optimal Assignment with Physical Queues for Many-to-Many OD Pattern. *Transportation Research.* Vol. 35B, pp. 461-479
17. Lighthill M.J. and Whitham J.B., (1955) On Kinematic Waves. I Flow Movement in Long Rivers. II A Theory of Traffic Flow on Long Crowded Road. *Proceedings Of Royal Society*, A229, pp. 281-345
18. Mahmassani, H.S. and Peeta, S. (1993) Network Performance under System Optimal and User Equilibrium Dynamic Assignments: Implications for ATIS. *Transportation Research Record.* Vol. 1408, pp. 83-93, Transportation Research Board, WashingtonDC
19. Muñoz J.C. and Daganzo C.F. (2002) Moving Bottlenecks: A Theory Grounded on Experimental Observations. *Transportation and Traffic Theory in the 21st Century*. Elsevier Science Publishers, pp. 441-461
20. Newell G.F. (1998) A Moving Bottleneck. *Transportation Research.* Vol. 32B, pp. 531-537
21. Powell W.B., and Sheffi, Y. (1982) The Convergence of Equilibrium Algorithms with Predetermined Step Sizes. *Transportation Science.* Vol. 16, pp. 45-55
22. Richards P.I., (1956) Shockwaves on the Highway. *Operations Research.* Vol. 4, pp. 42-51
23. Taylor N.B. (2003). The CONRAM Dynamic Traffic Assignment Model. *Networks and Spatial Economics.* Vol. 3, pp. 297-322
24. Wie, B.W. et al. (1995) A Comparison of System Optimum and User Equilibrium Dynamic Traffic Assignments with Schedule Delays. *Transportation Research*. Vol. C3, pp. 389-411.
25. Ziliaskopoulos A.K., and S. Peeta. Foundations of Dynamic Traffic Assignment: The Past, the Present and the Future. *Presented at 82nd Annual Meeting*, Transportation Research Board, Washington, D.C., 2002

# A GRAPH-BASED FORMULATION FOR THE SINGLE DESTINATION DYNAMIC TRAFFIC ASSIGNMENT PROBLEM

*Georgios Kalafatas* : *Purdue University, Indiana, USA,* gkalafat@purdue.edu
*Srinivas Peeta* : *Purdue University, Indiana, USA,* peeta@purdue.edu

## Abstract
The cell transmission model (CTM) has been used in the past to develop a linear formulation for the single destination dynamic traffic assignment problem, and show the existence of a minimum cost flow sub-structure. Here, the fundamental equations of the CTM are revisited to develop a complete formalized graph structure, a graph theoretic version of the CTM. An analytical discussion is provided for the modelling of backward propagating traffic waves. The graph structure is proved to be acyclic. Further, it is illustrated that every pair of arc-disjoint paths are also node-disjoint. The trade-offs between modelling accuracy and computational benefits are highlighted and computational experience is presented.

## 1 Introduction
The cell transmission model (Daganzo, 1994, 1995) has been widely used for a variety of planning applications. The single destination system optimal dynamic traffic assignment (DTA) formulation (Ziliaskopoulos, 2000), based on the cell transmission model (CTM), enabled a linear formulation of dynamic flows in traffic networks through a linear approximation of the fundamental traffic flow-density relation. Related work (Li et al, 2003) identified a graph sub-structure by excluding the constraint that addresses the propagation of the backward traffic waves which models the heavily congested traffic conditions. Also, following Daganzo (1994), the CTM-based DTA literature has always used this constraint in its empirically-calibrated format. In our paper, we propose a formal graph structure with hard constraints for traffic flow propagation, including for the backward traffic waves. The graph structure is proved to be acyclic. Further, it is illustrated that every pair of arc-disjoint paths are also node-disjoint. These special topological properties permit the implementation of polynomial algorithms of the acyclic minimum cost flow problem for the system optimal (SO) and user equilibrium (UE) dynamic traffic assignment problems.

The proposed graph structure is achieved through the following steps. Initially, all the constraints are integrated and formalized to a capacitated graph structure with arc costs independent of the flow. Then, the hard constraint for the propagation of backward traffic waves is obtained by explicitly noting that the total number of vehicles in a cell is less than the maximum number of vehicles that can exist in the cell at maximum occupancy. The system optimal (Ziliaskopoulos, 2000) and the user equilibrium (Ukkusuri, 2002) objectives are simply instances of this graph structure with transformed arc costs. If needed, the original CTM empirical constraint for backward traffic wave propagation can be incorporated through a Lagrangian relaxation, while still obtaining an acyclic minimum cost flow structure.

The identified graph structure is shown to be acyclic, and that every pair of arc-disjoint paths is also node-disjoint. The acyclic property is due to the time-expanded nature of the graph; a flow in a time-expanded network cannot loop in time. It allows the exploitation of the acyclic shortest path algorithm (Ahuja et al, 1993). An arc-disjoint path is always node-disjoint because every node of the graph has either the in-degree or the out-degree equal to 1; there can be no second arc-disjoint path crossing any node of the graph. The coexistence of the minimum cost flow structure and the acyclic structure allows for the efficient implementation of algorithms with shortest path sub-structures for the minimum cost flow problem (such as successive shortest paths, primal-dual, out-of-kilter, capacity scaling, repeated capacity scaling and the enhanced capacity scaling (Ahuja et al, 1993)).

The proposed graph-based formulation (GBF) reduces the complexity of the single destination dynamic traffic assignment problem from general linear programming to acyclic minimum cost flow. Hence, its main contribution to the DTA literature is that it directly bridges DTA and graph theory, providing computationally efficient mechanisms to solve a wide range of DTA problems.

## 2 The single destination system optimal DTA formulatiom (SDSODTA)

This section summarizes Ziliaskopoulos (2000) single destination system optimal dynamic traffic assignment formulation. It is a linear model based on the CTM. It provides the starting point for the comprehensive transition to the graph-based formulation, and is hence analytically described here using notation consistent with our paper.

The network is represented by the set of cells $i \in C$, and the set of cell connectors $j \in E$. A cell belongs to one of three cell types: the subset of origin cells $C_R \subset C$ (source cells), the subset of destination cells $C_S \subset C$ (sink cells), or the subset of intermediate cells $C_G \subset C$. The set of the successor cells of cell $i \in C$ is $\Gamma(i)$ and the set of the its predecessor cells is $\Gamma^{-1}(i)$. The maximum occupancy of a cell $i \in C$ in time interval $t \in T$ is $N_i^t$, and the maximum inflow or outflow is $Q_i^t$. For cell $i \in C$ in time interval $t \in T$, the free-flow speed is $v_i^t$, the traffic wave's backward propagation speed is $w_i^t$, and the $w_i^t / v_i^t$ ratio is $\delta_i^t$. The constant discretization time interval is $\tau$. The demand (inflow) at a source cell $i \in C_R$ in time interval $t \in T$ is $d_i^t$.

The variables of the model are the number of vehicles $x_i^t$ in cell $i \in C$ in time interval $t \in T$, and the number of vehicles $y_j^t$ routed by cell connector j in time interval $t \in T$. These variables are non-negative real numbers. The linear programming formulation for the SDSODTA is expressed as follows:

Minimize: $$\sum_{t \in T} \sum_{i \in C \setminus C_S} \tau \cdot x_i^t \tag{1}$$

Subject to:

$$x_i^t = x_i^{t-1} - \sum_{j \in \Gamma(i)} y_j^{t-1} + \sum_{j \in \Gamma^{-1}(i)} y_j^{t-1} \qquad \forall i \in C \setminus C_R, \forall t \in T \tag{2}$$

$$x_i^t = x_i^{t-1} - \sum_{j \in \Gamma(i)} y_j^{t-1} + d_i^t \qquad \forall i \in C_R, \forall t \in T \tag{3}$$

$$\sum_{j \in \Gamma(i)} y_j^t \le x_i^t \qquad \forall i \in C, \forall t \in T \tag{4}$$

$$\sum_{j \in \Gamma(i)} y_j^t \le Q_i^t \qquad \forall i \in C, \forall t \in T \tag{5}$$

$$\sum_{j \in \Gamma^{-1}(i)} y_j^t \le Q_i^t \qquad \forall i \in C, \forall t \in T \tag{6}$$

$$\sum_{j \in \Gamma^{-1}(i)} y_j^t \le \delta_i^t \left( N_i^t - x_i^t \right) \qquad \forall i \in C, \forall t \in T \tag{7}$$

$$x_i^t \ge 0 \qquad \forall i \in C, \forall t \in T \tag{8}$$

$$y_j^t \ge 0 \qquad \forall j \in E, \forall t \in T \tag{9}$$

The objective (Equation 1) of the formulation is to minimize the total time spent in the network. It consists of the total time spent by travelers in all cells other than the destination cells. The parameter $\tau$ is included for highlighting the physical meaning of the objective function.

Equation (2) represents the mass conservation constraint for all cells other than source cells. The number of vehicles $x_i^t$ in a non-source cell $i \in C \setminus C_R$ in time interval $t \in T$ equals the number of vehicles $x_i^{t-1}$ in it the previous time interval plus the inflow from incoming cell connectors $j \in \Gamma^{-1}(i)$, minus the outflow on the outgoing cell connectors $j \in \Gamma(i)$. Equation (3) represents mass conservation at source cells $i \in C_R$ and

introduces demand $d_i^t$ in time interval $t \in T$. Constraints (4) to (7) are capacity constraints for cells and cell connectors. Constraint (4) models defines free-flow region. Constraints (5) and (6) constrain the inflow and outflow, respectively, in the capacitated region. Constraint (7) addresses the over-congested region, and captures the backward traffic wave effects through the empirical parameter $\delta$. Constraints (8) and (9) are the non-negativity constraints.

This formulation ((1)-(9)) is an exact equivalent of Ziliaskopoulos' model. However, its notational simplicity allows a straightforward elucidation of the conceptual generalization to the graph-based formulation, as discussed hereafter.

## 2. The graph-based formulation for the single destination dynamic traffic assignment problem

This section defines, describes and analytically introduces the graph-based formulation for the single destination dynamic traffic assignment problem (GBSDDTA or briefly GBF). The concept that enables the development of the graph-based formulation is introduced hereafter in three simple steps: (i) new variables are introduced, (ii) the mass balance constraints are reconstructed and the complete well-formalized set of constraints is summarized, and (iii) the objective functions for the SO and UE cases are built for the single destination dynamic traffic assignment problem (SDDTA). A complete comparison between the proposed GBF and the SDSODTA demands an analytical discussion on the modeling of the constraint (7) responsible for backward propagating traffic waves, which is given in the following section. Finally, the formulation for the generalized SDDTA problem is summarized, with its exact graph structure illustrated in Figure 1.

The total number of vehicles that advance into cell $i \in C$ in time interval $t \in T$ is defined to be ${y_{IN\,i}}^t$, which is equal to the sum of the flows $y_j^{t-1}$ of the incoming cell connectors $j \in \Gamma(i)$ (Equation 10). Accordingly, the total number of vehicles in cell $i \in C$ in time interval $t \in T$ that advance to the next cells is defined to be ${y_{OUT\,i}}^t$, which is by definition equal to the sum of the flows $y_j^t$ of the outgoing cell connectors $j \in \Gamma(i)$ (Equation 11). The number of vehicles in cell $i \in C$ in time interval $t \in T$ that do not advance to the next cells is denoted by $z_i^t$, and is equal to the difference between the total number of vehicles $x_i^t$ minus the number of vehicles ${y_{OUT\,i}}^t$ that advance to the next cells for each cell $i \in C$ and time interval $t \in T$ (Equation 12). The corresponding definitional equations are:

$${y_{IN\,i}}^t = \sum_{j \in \Gamma^{-1}(i)} y_j^{t-1} \qquad \forall i \in C, \forall t \in T \qquad (10)$$

$${y_{OUT\,i}}^t = \sum_{j \in \Gamma(i)} y_j^t \qquad \forall i \in C, \forall t \in T \qquad (11)$$

$$z_i^t = x_i^t - {y_{OUT\,i}}^t \qquad \forall i \in C, \forall t \in T \qquad (12)$$

The number of vehicles $z_i^t$ in cell $i \in C$ in time interval $t \in T$ can also be viewed, from a mathematical programming standpoint, as the strictly non-negative slack variable added to the left hand side of constraint (4) for formulating the equivalent equality. This makes the definitional equation (12) mathematically equivalent to constraint (4). Constraints (2) and (3), considering definitional equations (10) and (11), can now be rewritten as follows:

$$x_i^t = x_i^{t-1} - {y_{OUT\,i}}^{t-1} + {y_{IN\,i}}^t \qquad \forall i \in C \setminus C_R, \forall t \in T \qquad (13)$$

$$x_i^t = x_i^{t-1} - {y_{OUT\,i}}^{t-1} + d_i^t \qquad \forall i \in C_R, \forall t \in T \qquad (14)$$

Constraints (13) and (14) will now be unified in order to represent the mass balance equation for all types of cells. The inbound traffic demand $d_i^t$ in cells other than the origin cells $i \in C \setminus C_R$ in time interval $t \in T$ is by definition:

$$d_i^t = 0 \qquad \forall i \in C \setminus C_R \qquad (15)$$

The total inflow ${y_{IN\,i}}^t$ in origin cells $i \in C_R$ from other cells in time interval $t \in T$ is by definition:

$$y_{IN\,i}^{\ t} = 0 \qquad \forall i \in C_R \qquad (16)$$

Accordingly, the total outflow $y_{OUT\,i}^{\ t}$ in destination cells $i \in C_S$ to other cells in time interval $t \in T$ is by definition:

$$y_{OUT\,i}^{\ t} = 0 \qquad \forall i \in C_R \qquad (17)$$

By adding $d_i^t$ to the right hand side of equation (13) and $y_{IN\,i}^{\ t}$ to the right hand side of equation (14), we simply get the same equation defined for two independent and complementary subsets $C_R \subset C$ , $C \setminus C_R \subset C$ of the set of all cells $C$ . Considering equations (13) to (17) we have the mass balance equation (18) at the cell level, which is equivalent to the original constraints (2) and (3), and therefore it replaces them.

$$x_i^t = x_i^{t-1} - y_{OUT\,i}^{\ t-1} + y_{IN\,i}^{\ t} + d_i^{t-1} \qquad \forall i \in C, \forall t \in T \qquad (18)$$

Finally, we apply constraint (12) for cell $i \in C$ in time interval $(t-1) \in T$ , and we substitute $y_{OUT\,i}^{\ t-1}$ in the mass balance equation (18), the following holds:

$$x_i^t = x_i^{t-1} - \left(x_i^{t-1} - z_i^{t-1}\right) + y_{IN\,i}^{\ t} + d_i^t \Leftrightarrow$$

$$x_i^t = z_i^{t-1} + y_{IN\,i}^{\ t} + d_i^t \qquad \forall i \in C, \forall t \in T \qquad (19)$$

The physical interpretation of constraint (19) is that the number of vehicles $x_i^t$ in cell $i \in C$ in time interval $t \in T$ is equal to the sum of the vehicles $z_i^{t-1}$ that existed in the same cell $i \in C$ in the previous time interval $(t-1) \in T$ and did not advance to the next cell(s), the number of vehicles $y_{IN\,i}^{\ t}$ that advance into cell $i \in C$ from the predecessor cells and the inbound traffic demand $d_i^t$ in cell $i \in C$ in time interval $t \in T$ . It is noted here that in order to retain the feasibility of the model, $d_i^t$ has to be less or equal to $N_i^t$ . This is true for origin cells, since an infinite capacity is assumed for them. For intermediate cells (ordinary, merging, diverging) $d_i^0 \le N_i^0$ values can be used to represent initial traffic conditions, without loss of generality. Constraints (12) and (19) are equivalent to constraint (18), and can replace it as now representing the mass balance equations. Concluding, we have proved that the original mass balance constraints (2) and (3) and the free-flow speed constraint (4) are equivalent to constraints (12) and (19) and therefore will be replaced by them.

In order to enhance the generalization capabilities of the proposed formulation, the outbound traffic demand of cell $i \in C$ in time interval $t \in T$ is defined to be $b_i^t$ and it is added to the right hand side of equation (12). The following holds for all non-destination cells:

$$b_i^t = 0 \qquad \forall i \in C \setminus C_S \qquad (20)$$

The SDSODTA based on the CTM, as originally formulated by Ziliaskopoulos (2000), did not include the parameter of outbound traffic demand $b_i^t$ . Traffic flow was propagated to the destination cells only because destination cells were not contributing to the system optimal objective. When Lo (1999) used the same formulation for traffic signal coordination, the objective used was the minimization of the total delay. It can be observed that in a network with loops, the objective could reach a non-inferior optimal value by just propagating flow at free flow speed in the loops – without any delay occurring – and still never allowing the vehicles to reach the destination cells. The proposed introduction of the $b_i^t$ parameter explicitly constrains vehicles to reach their destination, and further contributes to the formalization of the problem in the classical structure of the minimum cost flow problem. A feasibility restriction of the minimum cost flow problem is that the total inbound traffic demand $d_i^t$ and the total outbound traffic demand $b_i^t$ have to be: (i) equal, and (ii) in time intervals when there exists a feasible set of routes. The first restriction is accounted directly by the definition of the problem, by considering the definitional equation (21):

$$\sum_{i \in C_S} \sum_{t \in T} b_i^t = \sum_{i \in C_R} \sum_{t \in T} d_i^t \qquad (21)$$

Since the inbound traffic $d_i^t$ is usually fixed, the second restriction can be handled by assigning all outbound traffic $b_i^t$ in the last time interval $|T|$ . Finally, in scenarios where all destinations are equivalent, as in the

evacuation problem, a single super-destination cell can added and connected to all destination cells or directly replace the destination cells and retain their cell connectors.

The transformations made are adequate to provide the exact set of constraints ((22) to (31)) of the graph-based formulation. They are summarized here:

$$y_{IN\,i}^{\ t} = \sum_{j \in \Gamma^{-1}(i)} y_j^{t-1} \qquad \forall i \in C, \forall t \in T \qquad (22)$$

$$x_i^t = z_i^{t-1} + y_{IN\,i}^{\ t} + d_i^t \qquad \forall i \in C, \forall t \in T \qquad (23)$$

$$x_i^t = z_i^t + y_{OUT\,i}^{\ t} + b_i^t \qquad \forall i \in C, \forall t \in T \qquad (24)$$

$$y_{OUT\,i}^{\ t} = \sum_{j \in \Gamma(i)} y_j^t \qquad \forall i \in C, \forall t \in T \qquad (25)$$

$$y_{IN\,i}^{\ t} \le Q_i^t \qquad \forall i \in C, \forall t \in T \qquad (26)$$

$$x_i^t \le N_i^t \qquad \forall i \in C, \forall t \in T \qquad (27)$$

$$z_i^t \le N_i^t \qquad \forall j \in E, \forall t \in T \qquad (28)$$

$$y_{OUT\,i}^{\ t} \le Q_i^t \qquad \forall i \in C, \forall t \in T \qquad (29)$$

$$y_j \le Q_j^t \qquad \forall i \in C, \forall t \in T \qquad (30)$$

$$x_i^t, y_j^t, z_i^t \ge 0 \qquad \forall i \in C, \forall j \in E, \forall t \in T \qquad (31)$$

Constraints (22) to (25) are the mass balance constraints, constraints (26) to (30) are the arc capacity constraints, and constraints (31) are the non-negativity constraints. It is interesting to note that in the exact formalized graph structure of Figure 1, each mass balance constraint corresponds to a node of the graph representation. Capacity constraint (27), although definitional, will be explicitly discussed in the next section.

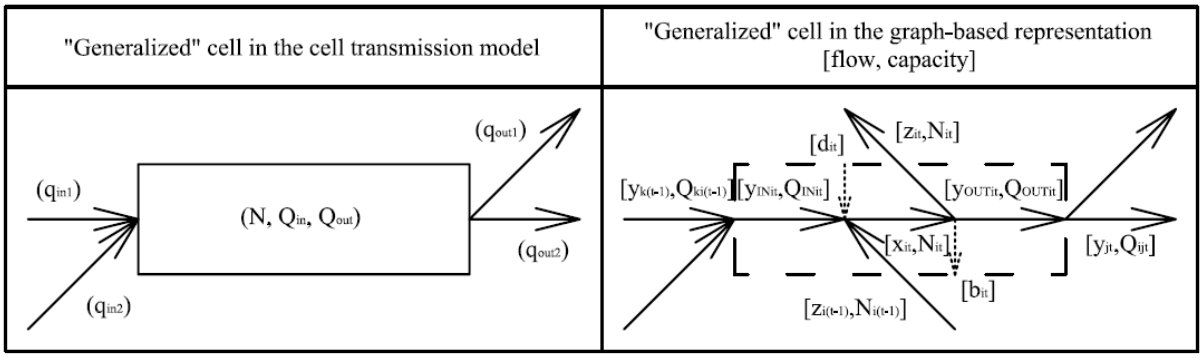



Figure 1. “Generalized” cell representations.

The generalized objective function in the GBF is presented in equation (32). The weights of the flows in the objective function are selected appropriately for modeling the SO or UE objectives.

$$\sum_{t \in T} \left( \sum_{i \in C} \left( c_{1i}^{\ t} \cdot x_i^t + c_{2i}^{\ t} \cdot y_{IN\,i}^{\ t} + c_{3i}^{\ t} \cdot y_{OUT\,i}^{\ t} + c_{4i}^{\ t} \cdot x_i^t \right) + \sum_{j \in E} c_{5\,j}^{\ t} \cdot y_j^t \right) \qquad (32)$$

The SO objective has been formulated for the CTM in Ziliaskopoulos (2000). The SO objective (1) is to minimize the total time spent in the network. It is noticed that from all the flows that appear in the GBF, only the flows representing the number of vehicles $x_i^t$ in cell $i \in C \setminus C_s$ in time interval $t \in T$ are accounted for with weight $c_{1i}^{\ t} = \tau$. All other weights are zero. These are the exact flows that “count” the total vehicle-hours spent in the network.

The single destination user-equilibrium dynamic traffic assignment problem (SDUEDTA) has been formulated for the cell transmission model by Ukkusuri (2002). The objective is to minimize the weight sum

of exit cell connector flows, where the weights are strictly increasing with time and it is lexicographically better to route each unit of flow earlier to the sink cells than later. The SDUEDTA problem's objective is:

$$\sum_{t\in T}\sum_{j\in\Gamma^{-1}(i),i\in C_S} M^t y_j^t \tag{33}$$

The constraints for routing the flow in the network are the same as in the single destination system optimal dynamic traffic assignment problem. The equivalent objective function in the GBF is:

$$\sum_{t\in T}\sum_{i\in C_S} M^t y_{INi}^{\ t} \tag{34}$$

The formulation retains the minimum cost flow problem structure as in the SDSODTA, differing only in the arc weights. It is noticed that from all the flows that appear in the GBF, only the flows representing the number of vehicles $y_{INi}^{\ t}$ exiting the network to destination cell $i\in C_S$ in time interval $t\in T$ are accounted for with a weight $c_{2i}^t = M^t$ lexicographically increasing with time.

## 3 Backward propagating traffic waves

In this section, the replacement of the original constraint (7) for handling backward propagating traffic waves with constraint (29) is justified. We specifically focus on constraint (29) as it has never appeared in the existing literature. This is because it has been considered redundant with the use of constraint (7). The hard constraint (29) can be extended as follows:

$$x_i^{t+1} \le N_i^{t+1} \overset{(23)}{\Leftrightarrow} \left(z_i^t + y_{INi}^{\ t+1} + d_i^{t+1}\right) \le N_i^{t+1} \overset{(24)}{\Leftrightarrow} \left(x_i^t - y_{OUTi}^{\ t} - b_i^t\right) + y_{INi}^{\ t+1} + d_i^{t+1} \le N_i^{t+1} \Leftrightarrow$$

$$y_{INi}^{\ t+1} \le N_i^{t+1} - x_i^t + \left(y_{OUTi}^{\ t} + b_i^t - d_i^{t+1}\right) \qquad \forall i\in C, \forall t\in T \tag{35}$$

The hard constraint (35), which is equivalent to constraint (29), differs from the well-established constraint (7). Certain remarks justify the transformation:

(i) Constraint (35) is the direct outcome of the hard-physical constraint (29); it simply has to hold.

(ii) Parameter $N_i^{t+1}$ in (35) refers to a later time interval than in (7). In the general case that $N_i$ remains the same across all time intervals, there is no difference between constraints (35) and (7). It is considered that (31) is a more exact representation. It allows to physically bound flows in case of sudden reductions in the cell's maximum occupancy (accidents, contra-flow operations). It is noted that the original CTM is a discrete model, outcome of a "discretization" process of the continuous equations of traffic flow theory. The use of adjacent discrete intervals in both time and space for all modeling elements needs to be discussed in each case, and accordingly selected after comparing the case-specific tradeoffs. Moreover, the CTM was directed mostly towards simulation applications, where simplicity precludes "looking" one time step ahead.

(iii) Variable $y_{OUTi}^{\ t}$ does not exist in (7), because "the effects of the outflow should only be noticed upstream after some time" (Daganzo, 1994). Constraint (35) alone does not allow for the existence of the time lag in the CTM. However, if relatively large time intervals are used, the effect of the outflow can be too significant to be avoided.

(iv) Parameters $b_i^t$ , $d_i^{t+1}$ are equal to zero for $i\in C_G$ , $t\in T$ . When $i\in C_R\cup C_S$ , $t\in T$ , at least one of them is zero and the constraint simply holds because origin and destination cells have infinite maximum occupancy $N_i^t$. This further implies that this constraint, by definition, is always satisfied, without even becoming active, thus it is redundant for $i\in C_R\cup C_S$ ; no backward propagating traffic waves can be created from source and destination cells with infinite maximum occupancy.

(v) Parameter $\delta_i^t$ , which is responsible for enhancing the modeling realism of the backward traffic wave speed, does not exist in (35), so backward propagating traffic waves with speed equal to the free-flow speed are allowed. Nevertheless, let us examine a typical traffic pattern of uniform initial traffic density $Q_{\mathrm{MAX}i}^{\ t}$

(where maximum flow $Q_{\mathrm{MAX}i}^{t}$ is achieved) along a corridor of maximum density $N_{\mathrm{MAX}i}^{t}$ with a full-stop enforced at its end. The GBF will generate a backward propagating traffic wave with speed equal to $Q_{\mathrm{MAX}i}^{t} \Big/ \left( N_{\mathrm{MAX}i}^{t} - Q_{\mathrm{MAX}i}^{t} \right)$, which for freeway cells is approximately equal to 1/5 – close to a typical $\delta_i^t$ value ranging from 1/4 to 1/6.

(vi) The additional use of a Lagrangian relaxation for constraint (7) retains the minimum cost flow structure of the formulation, while it allows: (i) the time lag in the backward propagation of waves, and (ii) backward speed less than the free-flow speed. It is noted that constraint (7) is from an empirical calibration (Daganzo, 1994). This implies that it is not a hard constraint unlike (35), and therefore it may be acceptable for it to be "slightly" violated.

## 4 Graph properties

It has been proved that the GBF has a minimum cost flow (MCF) structure. The MCF structure allows for well-known algorithms to be used. Such algorithms include pseudo-polynomial (successive shortest paths, the primal-dual, out-of-kilter), weakly polynomial (capacity scaling, cost scaling, double scaling), strongly polynomial algorithms (minimum mean cycle-canceling, repeated capacity scaling, enhanced capacity scaling), and the traditional network simplex method (Ahuja et al., 1993). Furthermore, it is proved in the following section that the GBF's network structure is: (i) acyclic, and (ii) an edge-disjoint path is also a node-disjoint path. These properties allow for increased computational efficiencies in the application of the MCF algorithms.

**Proposition 1**: The GBF's network is acyclic. Proof: It is proven by contradiction. It is assumed that there is a loop in the network structure. The loop will allow a flow to pass twice from an arc representing the number of vehicles in a cell $i \in C$ in a time interval $t \in T$. The flow entering that arc arrives from previous time intervals and exits in future time intervals, which are exclusively connected by the single direction that time flows; a flow can never go back in time in order to enter the same cell $i \in C$ in a time interval $t \in T$ if it has been there before. This contradicts the initial assumption and completes the proof. The physical interpretation is that there is no path allowing a flow to return back in time. This property allows the implementation of the reaching shortest path algorithm for acyclic networks (Ahuja et al., 1993), which has a running time complexity of O(m). This property, in accordance with the special graph structure of the GBF can further classify the structure as a time-expanded network. A potential application is in the field of fast real-time traffic-responsive routing of emergency vehicles.

**Proposition 2**: An edge-disjoint path in the GBF's network is always node-disjoint. Proof: It is proven by contradiction. It is assumed that there is a node in the GBF's network structure such that it belongs to at least two edge-disjoint paths. Then the inflow degree of this node is at least two and the outflow degree is at least two, since there are at least two different links starting from and ending to this node. As seen in Figure 1, all nodes have at least either the inflow or outflow degrees strictly equal to one. This contradicts the initial assumption and completes the proof. This property allows standard solution methodologies for edge-disjoint paths to be implemented for node-disjoint paths also. Furthermore, it is useful in the identification of spatio-temporal disjoint paths for secure and redundant routing of emergency response vehicles.

## 5 Computational experience

A test network (Figure 2) is built and the linear SDSODTA formulation based on the CTM and the graph-based formulation are compared. The SDSODTA based on the CTM is solved with the Simplex method and the SDSODTA based on the GBF is solved with the Network Simplex method. The Network Simplex method does not have the best worst-case complexity for minimum cost flow problems. However, it is selected as a benchmark as it is well established, widely available, and generally performs well on average. The computing environment consists of a Sun Ultra Enterprise server E6500 with 26 400-MHz UltraSparc II processors under the multi-user Solaris 7 operating environment with 23 GB of RAM, 131 GB of swap space, and 8 MB of cache. The GAMS modeling language and CPLEX's solver were used.

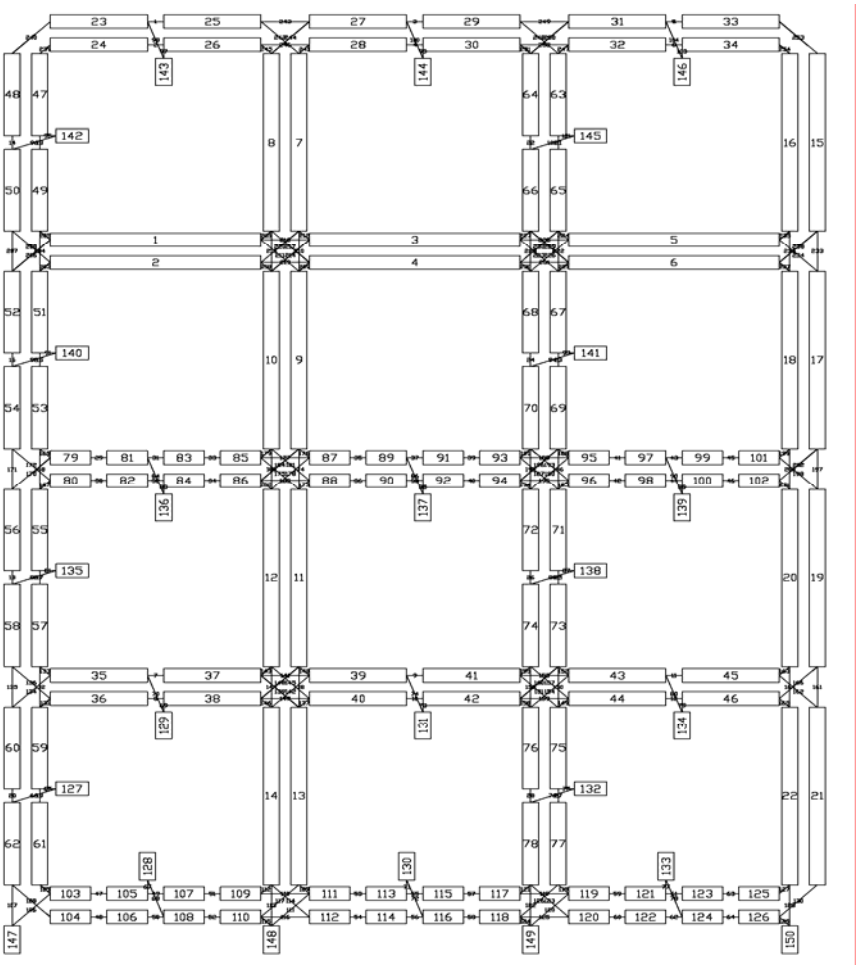


Figure 2. Test network.

The test network consists of a 3x4 grid network that replicates a dense urban environment with highways (long cells), arterials (medium length cells) and side streets (short cells), as described in Table 1. The single destination formulation is capable of modeling the evacuation problem (Daganzo, 1994). Therefore, the experimental setup can be seen as an evacuation routing problem. From an evacuation standpoint, the bottom of the network represents the boundary of the evacuation zone from which vehicles move to the safety zone. Evacuees are assigned uniformly to 20 sources cells and routed to the safety zone. In order to evaluate the formulations' performance for different problem sizes, the problem is solved for 500 to 5000 evacuees in increments of 500. The effect of population size on the problem size is in terms of time-expansion; although the same network topology is used, the number of time intervals needed to solve the problem is increases proportionally to the population size.

Table 1. Cell characteristics of the test network.

| Cell Type | Highway | Arterial | Side Street | Source | Destination |
|---|---|---|---|---|---|
| Cell IDs | 1-22 | 23-78 | 79-126 | 127-146 | 147-150 |
| Free flow speed (miles/h) | 70 | 35 | 20 | - | - |
| Time interval (sec) | 10 | 10 | 10 | 10 | 10 |
| Cell length (feet) | 1000 | 500 | 250 | - | - |
| Number of lanes | 3 | 2 | 1 | 3 | 3 |
| Maximum flow in evacuation operations (veh/hr/lane) | 1440 | 1260 | 1080 | 1440 | 1440 |
| Maximum cell flow (veh/time step) | 12 | 7 | 3 | 12 | 12 |
| Maximum number of vehicles per cell (veh/cell) | 108 | 36 | 9 | infinite | infinite |

The computational results are illustrated in Figure 3. Both the linear and the graph-based formulations reach the same SO objective level for each population size. This means that the GBF produces non inferior results for the SDSODTA problem, although it is expected to overestimate the speed of the backward propagating traffic density wave. The reason is that a cell's maximum flow is still achieved at light and intermediate traffic density levels other than the high traffic densities at which backward propagating waves occur. These are non-inferior solutions for the SDSODTA, where traffic is not assigned to high traffic densities. In other words, there are non-inferior solutions for the SDSODTA where constraint (7) remains inactive at system optimal levels. For the largest problem instance of a population size equal to 5000, the percentage reduction in computational times is 78.90%. It is interesting to note that for the smallest instance of the problem, there is an increase in the computational time. This is explained by the fact that the GBF uses more variables than the linear formulation, despite the better worst-case complexity.

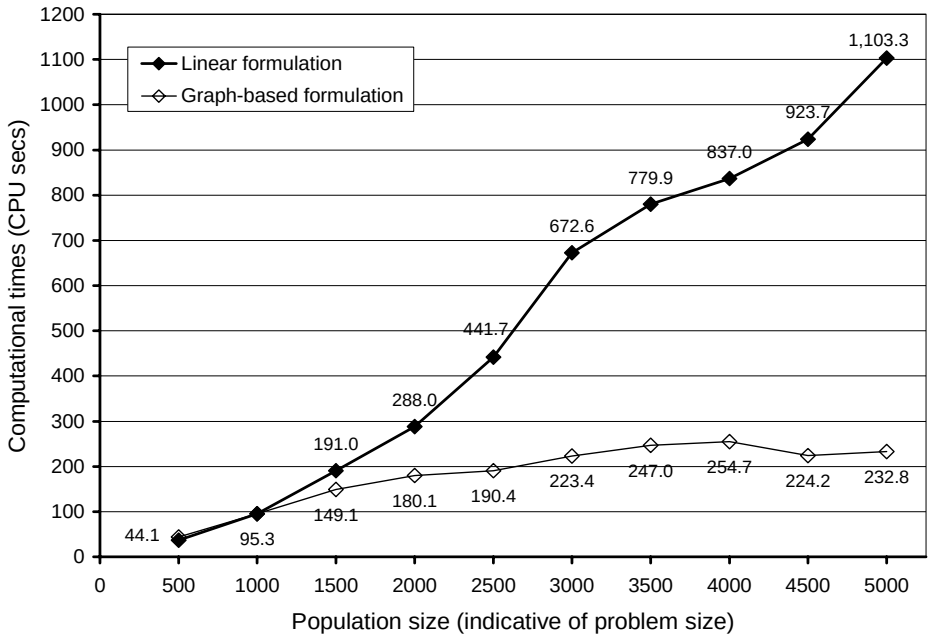



Figure 3. Computational time as a function of the problem size for the linear and the graph-based formulation of the SDSODTA.

## 6 Conclusions and extensions

This study introduced a graph-based formulation for the SDDTA. It is a hard-constrained, acyclic minimum cost flow dynamic traffic assignment formulation based on the CTM which produces non-inferior solutions for the single destination system optimal dynamic traffic assignment problem at significantly reduced computational times. Therefore, the importance of the graph-based formulation is that it has a solid computational base directly related to well-known structures and algorithms of the acyclic minimum cost flow problem derived from the operations research domain.

The graph-based computational base can be applied to formalize the structure and improve the performance of several planning schemes, as they have been already modeled in a CTM format. The traffic signal coordination problem has been modeled in a CTM base (Lo, 1999). Under a graph-based perspective, the traffic signal coordination problem is equivalent to the network design problem. We simply need to solve for the capacity of cell outflow arcs. Contra-flow operations for the network re-design problem have been modeled using a CTM basis (Tuydes and Ziliaskopoulos, 2003; Kalafatas, 2005). It is also equivalent to the network design problem. We simply solve for the capacity of cell inflow, cell outflow and cell occupancy arcs. Hence, the ability to formalize these problems to the network design problem allows for more efficient solution methodologies from the operations research domain.

In on-going research, the multiple destination dynamic traffic assignment problem is formulated in the GBF context, explicitly accounting for the first-in-first-out property. Also, a better approximation of the

fundamental flow-density relation is being developed to improve the compliance of the GBF to well-known empirical observations. Finally, the graph structure of the GBF provides the bridge to graph theory and operations research domains for the direct adoption of graph properties and solution methodologies.

## References

Ahuja, R. K., Magnanti T. L. and Orlin J. B. (1993) *Network Flows: Theory Algorithms and Applications*. New Jersey: Prentice Hall.

Daganzo C. F. (1994) The Cell Transmission Model: A Simple Dynamic Representation of Highway Traffic Consistent with the Hydrodynamic Theory. *Transportation Research B*, 28(4), 269-287.

Daganzo C. F. (1995) The Cell Transmission Model, Part II: Network Traffic. *Transportation Research B*, 29(2), 79-93.

Li Y., Waller S.T. and Ziliaskopoulos A. (2003) A Decomposition Scheme for System Optimal Dynamic Traffic Assignment Models. *Networks and Spatial Economics*, 3(4), 441-456.

Lo H. K. (1999) A Novel Traffic Signal Control Formulation. *Transportation Research A*, 33, 433-448.

Ukkusuri, S. (2002) *Linear Programs for the User Optimal Dynamic Traffic Assignment Problem*, M.S. Thesis, University of Illinois at Urbana-Champaign, IL.

Ziliaskopoulos A. K. (2000) A Linear Programming Model for the Single Destination System Optimum Dynamic Traffic Assignment Problem. *Transportation Science*, 34(1), 37-49.

# THE ENHANCED LAGGED CELL-TRANSMISSION MODEL

*W.Y.Szeto: Department of Civil, Structural and Environmental Engineering, Trinity College Dublin, Ireland (szetow@tcd.ie)*

## Abstract
The Lagged Cell Transmission Model (L-CTM) is the enhanced version of the cell transmission model. This model assumes that the flow-density relation is non-concave, which allows capturing rather dense traffic in queues coasting towards the end of the queues. However, this paper shows that this model can yield negative densities and densities higher than the theoretical jam density. To cope with this, this paper improves L-CTM by introducing one more term in each sending and receiving functions of the model. The resulting model is proved to have to yield nonnegative densities not greater than the jam density, and includes CTM and L-CTM as the special cases. Numerical studies are also set up to illustrate the accuracy of the proposed model.

## 1 Introduction
Dynamic traffic assignment (DTA) models have many applications, ranging from offline transport planning, design and policy evaluation, to online intelligent transportation system (ITS) applications such as real-time traffic control, time-varying tolling, and traveller information services. While offline applications only require DTA models giving an accurate prediction or evaluation, online applications require the models obtaining a reasonable accurate solution within a very short period of time to manage the sudden change in traffic conditions and provide real-time traffic information to drivers. DTA models for real-time applications have to strike a balance between solution speed and modelling accuracy. These two issues relate to the approaches of developing one of the important components of DTA, the dynamic network loading model.

Many approaches have been developed (e.g., Adamo et al., 1999; Addison and Heydecker, 1995; Balijepalli and Watling, 2005; Ben-Akiva et al., 1998; Bliemer and Bovy, 2003; Cantarella et al., 1999; Carey, 1987; Friesz et al., 1993; Gentile et al., 2005; Janson, 1998; Jayakrishnan et al., 1995; Jin and Zhang, 2005; Lam and Huang, 1995; Peeta and Mahmassani, 1995; Perakis and Roels, 2004; Rubio-Ardanaz et al., 2001; Ran and Boyce, 1996; Smith, 1993; Tong and Wong, 2000; Wu et al., 1998). These approaches have been reviewed and compared recently in Szeto and Lo (2005). In particular, one of the approaches that receive most attention is based on the hydrodynamic theory of Lighthill and Whitham (1955) and Richards (1956), probably because this approach can capture realistic traffic dynamics, such as shock wave, queue formulation, queue dissipation, and queue spillback, and can describe the actual traffic situation with a desirable level of accuracy. The hydrodynamic theory views traffic as fluid. The movement of traffic is then described by a partial differential equation, an empirical flow density relation, and a definitional relationship of flow, density, and speed. These three equations form the Lighthill-Whitham-Richards (LWR) model.

The model can be solved by the method of characteristics, but the solution procedure is very tedious even for a single link. To solve the model efficiently, current efforts focus on deriving a discrete version of the LWR model and computing solutions using finite difference schemes. In particular, two methods receive wide attention: Newell's (1993) solution scheme and Daganzo's (1994) cell transmission model (CTM). Newell's solution scheme is able to solve the LWR model for a simple link based on a triangular flow density relation. His scheme is easily extended to deal with trapezoidal flow-density relations and general networks (Kuwahara and Akamatsu, 2001; Yperman et al., 2006), but cannot handle general nonconcave flow-density relations.

Daganzo's CTM is developed based on a trapezoidal flow-density relation, but this scheme also works when the flow-density relation is a triangular one. The scheme can also be applied to general networks (Daganzo, 1995a,) and extended to consider general flow-density relations (Daganzo, 1995b), priority vehicles and special lanes (Daganzo, 1997; Daganzo et al., 1997). In this aspect, CTM is better than the Newell's approach. However, CTM requires dividing a link into many homogenous segments called cells, and all operations are calculated at the cell level rather than at the node level as in Newell's method. This greatly increases the computational burden.

To reduce the computation burden while maintaining a certain level of accuracy to suit the needs of online ITS applications, various modifications to CTM has been proposed (e.g., Boel and Mihaylova, 2006; Ishak et al., 2006; Muñoz et al., 2003; Ziliaskopoulos and Lee, 1997). These modifications allow variable cell lengths, and hence require few cells, and reduce the computation burden. However, their developments are based on triangular or trapezoidal flow-density relations, which cannot model the coasting effect of rather dense traffic in queues (rather dense traffic in queue appears to be coasting towards the end of the queue). To capture this effect, non-concave relations must be adopted in the LWR model (Daganzo, 1999).

Daganzo (1999) improves CTM by allowing variable cell lengths. This relaxation leads to an enhanced version of CTM, called the Lagged Cell Transmission Model (L-CTM), which is suitable for online applications. This model has many desirable properties. First, like the CTM version proposed in Daganzo (1995b), this model is developed based on a general flow-density relation, and hence able to capture the coasting effect when a non-concave flow-density relation is adopted. Second, this model can converge to the LWR model when the lattice spacing tends to zero. Third, L-CTM is proved to be second order accurate when the flow density relation is triangular or trapezoidal and the optimal lag is used. Fourth, L-CTM is shown to be more accurate than CTM in term of flow prediction. However, L-CTM can yield negative densities and densities higher than the maximum density, the jam density, as will be shown in Section 2 using two counter examples. Negative densities are unrealistic as this implies that the number of vehicles is negative. Densities to be higher than the jam density are not realistic too since this allows highways or roads to hold vehicles more than the storage capacity.

To get rid of this, this paper enhances L-CTM, by introducing two terms in it, one in the sending function and one in the receiving function. The resulting model, called the Enhanced Lagged Cell Transmission Model (EL-CTM), is proved to give densities between zero and the jam density inclusively, and can retain the most of desirable features of L-CTM, including allowing variable cell lengths and applicable to non-concave flow density relations. This model also includes CTM and L-CTM as the special cases, and is shown to give more accurate solutions than L-CTM.

The rest of the paper is organized as follows: Section 2 briefly reviews L-CTM for the sake of completeness. Section 3 describes the two counter examples. Section 4 formulates and discusses the proposed EL-CTM. Section 5 is the numerical studies. Finally, section 6 is the concluding remarks.

## 2 The lagged cell transmission model

Consider a highway, which is partitioned into cells. These cells can have different lengths. The flow-density relation $T(k)$ of this highway is represented by two monotone functions, the sending function $S(k)$ and the receiving function $R(k_{jam}-k)$. $S(k)$ is non-decreasing with respect to the density $k$, which has a value of zero when $k$ equals zero (i.e., $S(0)=0$), and has a maximum value of $Q_{\max}$ when the density $k$ is greater than or equal to the optimal density $k_o$ (i.e., $S(k)=Q_{\max}, k \geq k_o$). $R(k_{jam}-k)$ is non-decreasing with respect to $k_{jam}-k$, whose value is zero when $k$ equals the jam density $k_{jam}$ (i.e., $R(0)=0$) and is equal to $Q_{\max}$ when $k$ is less than or equal to the optimal density $k_o$ (i.e., $R(k_{jam}-k)=Q_{\max}, k \leq k_o$). The minimum of these two non-decreasing functions, $Min(S(k), R(k_{jam}-k))$, forms the unimodal $T(k)$ curve with the maximum flow $Q_{\max}$, and gives $S(k)$ when $k \leq k_o$ and $R(k_{jam}-k)$ otherwise. This unimodal curve can be non-concave depending on the choices of $S(k)$ and $R(k_{jam}-k)$.

Consider also a time-space plane, where the t-axis is for time and the x-axis is for the highway considered. A rectangular lattice with spacings $\varepsilon$ and $d_j$ is then overlaid on this (t, x)-plane, where $d_j$ is the length of cell $j$ and $\varepsilon$ the length of each discrete time interval. The x-coordinate of the lattice point $x_j$ represents the centre of cell $j$, and the t-coordinate $t_i$ the time when the average cell density $K(t_i, x_j)$ for cell $j$ is evaluated by L-CTM. This evaluation is based on flow conservation and the flow-density relation adopted. Flow conservation is expressed as:

$$K\left(t+\varepsilon,x_j\right)=K\left(t,x_j\right)-\frac{\varepsilon}{d_j}\left[Q\left(t+\varepsilon/2,x_j+d_j/2\right)-Q\left(t+\varepsilon/2,x_j-d_j/2\right)\right], \quad (1)$$

where $Q\left(t+\varepsilon/2,x_j+d_j/2\right)$ is the average flow that advances from cell $j$ to cell $j+1$, and $Q\left(t+\varepsilon/2,x_j-d_j/2\right)$ the average flow that advances from cell $j-1$ to cell $j$.

The flow advanced is the minimum of the local demand from cell $j$, $S\left(K\left(t-f_j\varepsilon,x_j\right)\right)$ and the local supply from the downstream cell $j+1$, $R\left(k_{jam}-K\left(t-l_{j+1}\varepsilon,x_{j+1}\right)\right)$:

$$Q\left(t+\varepsilon/2,x_j+d_j/2\right)=Min\left(S\left(K\left(t-f_j\varepsilon,x_j\right)\right),R\left(k_{jam}-K\left(t-l_{j+1}\varepsilon,x_{j+1}\right)\right)\right), \quad (2)$$

where $S(.)$ and $R(.)$ follow earlier definitions.

$f_j$ and $l_j$ in $(2)$ are respectively the two non-negative lags for cell $j$, which must satisfy:

$$0\le f_j\le\frac{1}{2}\left[\frac{d_j}{\varepsilon\left|S_{k,\max}\right|}-1\right] \text{ and }, \quad (3)$$

$$0\le l_j\le\frac{1}{2}\left[\frac{d_j}{\varepsilon\left|R_{k,\max}\right|}-1\right] \quad (4)$$

for stability reasons, where $\left|S_{k,\max}\right|$ and $\left|R_{k,\max}\right|$ are the maximum of the absolute forward and backward wave speeds, respectively; that is, they correspond to the maximum of the absolute values of the slopes of the sending and receiving functions. The lag $f_j$ is used to forestall the deterioration of accuracy where large cells are used whereas the lag $l_j$ is used to account for the difference of the speeds of waves emitted from the sources to the lattice points where the average densities are calculated.

To minimize the second order error, $f_j$ and $l_j$ are set such that

$$f_j=\frac{1}{2}\left[\frac{d_j}{\varepsilon\left|S_{k,\max}\right|}-1\right], \text{ and} \quad (5)$$

$$l_j=\frac{1}{2}\left[\frac{d_j}{\varepsilon\left|R_{k,\max}\right|}-1\right], \quad (6)$$

hold, but this can result in non-integer $f_j$ and $l_j$. In this case, an interpolation of densities is required. In the extreme case, when $f_j$ and $l_j$ are set to zero for all cells, L-CTM becomes CTM if all cell lengths are the same.

## 3 Two counter examples

L-CTM (1)-(4) has many desirable properties, including being able to capture the coasting effect of rather dense traffic in queues, allowing variable cell lengths which suits for online applications, providing second-order accurate density estimation, and converging to the LWR model when the lattice spacing tends to zero. However, L-CTM has two undesirable properties: it can yield negative densities as shown in counter example 1 below, and it can give densities to be greater than the jam density, as illustrated in counter example 2.

*Counter example 1: L-CTM predicts negative densities*

Consider a homogenous highway with the maximum flow of 30 veh/min and the jam density of 180 veh/mile. The flow-density relation is triangular (The forward and backward wave speeds are 1 mile/min and -0.2 mile/min). The inflow to the highway is zero veh/min. The initial density profile follows $k\left(0,x_j\right)=j-1$. The ultimate objective of this counter example is to illustrate that L-CTM can yield negative densities.

Since L-CTM requires valid densities in addition to the initial densities when lags are positive, the exact results are calculated by CTM first as CTM can give the exact result predicted by the LWR theory in this case. These results can also be used to illustrate how well L-CTM predicts the densities. The highway is discretized into homogenous 11 cells, and the results are shown in Table 1. The first row is for time. The first column shows the cell number. The last column shows the average densities of the last 5 cells, which are used for comparison purposes later. The shaded numbers are the input densities. The arrows show the wave and traffic propagation directions. The slope of each arrow is 1, which exactly matches the forward wave speed.

Table 1 Densities predicted by the LWR theory and CTM in example 1

| cell \ time | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | average density of cells 7-11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 8 |
| 1 | 0 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 7 |
| 2 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 6 |
| 3 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 5 |
| 4 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 4 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 5 | 3 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 2 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 1.2 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0.6 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.2 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 2 Densities predicted by L-CTM in example 1

| cell \ time | 1 | 2 | 3 | 4 | 5 | 6 large cell |
|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 2 | 3 | 4 | 5 | 8 |
| 1 | 0 | 0 | 1 | 2 | 3 | 4 | 7 |
| 2 | 0 | 0 | 0 | 1 | 2 | **3** | 6 |
| 3 | 0 | 0 | 0 | 0 | 1 | 2 | 5 |
| 4 | 0 | 0 | 0 | 0 | 0 | 1 | 4 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | **3** |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 1.2 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0.6 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0.2 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | **-0.04** |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | **-0.16** |
| 12 | 0 | 0 | 0 | 0 | 0 | 0 | **-0.20** |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | **-0.19** |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | **-0.16** |

L-CTM is then used to predict the densities over time. The highway is discretized into 7 cells. The cell lengths of the first 6 cells are the same and the cell length of the last cell is five times longer. Based on this setting, we calculate the lag according to (5). We can obtain the lag $f_7$ to be equal to 2, (and the rest are zero) Hence we need two more sets of input data in addition to the boundary data of the densities of cell 1 as shown in the shaded region in Table 2. The input densities for cells 1-6 can be computed using the LWR theory and the last one is obtained by the last row of the table 1. We can then predict the densities over time using L-CTM. As you can see, the underlined numbers in boldface are the same since the lag is 2. Moreover, because we draw the table in the correct scale, we can see that the slopes of the arrows in table 2 are exactly the same as those in table 1. This matching means that the traffic and wave propagations agree with the LWR theory. In addition, the densities of cell 7 match with those in the last row of table 1 quite well at the beginning, but do not at the end. More importantly, L-CTM generates negative densities, as shown by the bolded numbers! It happens as the flow (the number of vehicles per time interval) leaving cell $j$ at time $t$ is predicted based on the density at $t - f_j\varepsilon$, $K\left(t - f_j\varepsilon, x_j\right)$, not the current density $K\left(t, x_j\right)$ but the flow present on that cell is a function of $K\left(t, x_j\right)$. The flow leaving can be greater that the flow present on that cell. A negative density arises when the difference is greater than the flow entering that cell and the current flow on that cell is close to zero.

*Counter example 2: L-CTM predicts densities higher than the jam density*

This example considers the same highway as in example 1. This highway is discretized into 11 homogenous cells again but the initial density profile is $K(0,x_j)=\begin{cases}160+2j,\ j=1,..,10\\ 180,\ j=11\end{cases}$. Based on this setting, we calculate the lag $l$ according to (6), which equals 2. Since the lag $l$ is two, we again need two more set of data, which can be computed using the LWR theory. The densities for cell 1 and cell 11 are also computed using the LWR theory, which serve as the correct boundary conditions. L-CTM is used to compute the densities and they are shown in Table 3. Like Table 2, the first row is the time index; the first column is the cell number; the traffic propagates in the direction of increasing cell number; the shaded numbers are the input values, and the arrow is the wave propagation direction. The slope is equal to -0.2 which matches with the backward wave speed. The underlined numbers in bolded face also match exactly each other. This shows that L-CTM can in general give a very accurate result compared with the LWR result. However, the numbers in italic do not match exactly, contrasting to the LWR results that they must be the same. This is due to the diffusion effect introduced by the numerical approximation near the shock region. More importantly, L-CTM can give the densities higher that the jam density! Again, this is because of the lag introduced. The allowable number of vehicles entering the cell at time $t$ depends on the densities $l_j\varepsilon$ time unit earlier, which can be greater than the actual available space at time $t$. When the number of vehicles leaving the cell is smaller than the difference between predicted and actual available spaces, and the number of vehicles in that cell is very close to the maximum allowable number, the densities evaluated can be greater than the jam density.

Table 3 Densities obtained by L-CTM in example 2

| cell \ time | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 162 | 164 | 166 | 168 | 170 | **172** | 174 | 176 | 178 | *180* | 180 |
| 1 | 162.4 | 164.4 | 166.4 | 168.4 | 170.4 | 172.4 | 174.4 | 176.4 | 178.4 | 180 | 180 |
| 2 | 162.8 | 164.8 | 166.8 | 168.8 | 170.8 | 172.8 | 174.8 | 176.8 | 178.8 | 180 | 180 |
| 3 | 163.2 | 165.2 | 167.2 | 169.2 | 171.2 | 173.2 | 175.2 | 177.2 | 179.2 | 180 | 180 |
| 4 | 163.6 | 165.6 | 167.6 | 169.6 | 171.6 | 173.6 | 175.6 | 177.6 | 179.52 | 180 | 180 |
| 5 | 164 | 166 | 168 | 170 | **172** | 174 | 176 | 178 | *179.76* | 180 | 180 |
| 6 | 164.4 | 166.4 | 168.4 | 170.4 | 172.4 | 174.4 | 176.4 | 178.4 | 179.92 | 180 | 180 |
| 7 | 164.8 | 166.8 | 168.8 | 170.8 | 172.8 | 174.8 | 176.8 | 178.78 | **180.02** | 180 | 180 |
| 8 | 165.2 | 167.2 | 169.2 | 171.2 | 173.2 | 175.2 | 177.2 | 179.14 | **180.06** | 180 | 180 |
| 9 | 165.6 | 167.6 | 169.6 | 171.6 | 173.6 | 175.6 | 177.6 | 179.44 | **180.08** | 180 | 180 |
| 10 | 166 | 168 | 170 | **172** | 174 | 176 | 178.00 | *179.69* | **180.08** | 180 | 180 |
| 11 | 166.4 | 168.4 | 170.4 | 172.4 | 174.4 | 176.4 | 178.38 | 179.87 | **180.06** | 180 | 180 |
| 12 | 166.8 | 168.8 | 170.8 | 172.8 | 174.8 | 176.8 | 178.75 | 180.00 | **180.05** | 180 | 180 |
| 13 | 167.2 | 169.2 | 171.2 | 173.2 | 175.2 | 177.20 | 179.09 | **180.08** | **180.03** | 180 | 180 |
| 14 | 167.6 | 169.6 | 171.6 | 173.6 | 175.6 | 177.60 | 179.39 | **180.12** | **180.02** | 180 | 180 |

## 4 The enhanced lagged cell transmission model

To get rid of these undesirable properties, one may adopt truncation or round off when the densities are out the range of $[0,k_j]$. Nevertheless, this is not a good idea, as flow conservation will not be satisfied, and this will cause some flows to disappear. In this paper, we propose to introduce two terms in (2), one in the sending function and one in the receiving function:

$$Q(t+\varepsilon/2,x_j+d_j/2)=Min\left(S'\left(K(t,x_j)\right),R'\left(k_{jam}-K(t,x_{j+1})\right)\right), \tag{7}$$

where

$$S'\left(K(t,x_j)\right)=Min\left(S\left(K(t-f_j\varepsilon,x_j)\right),\frac{d_j}{\varepsilon}K(t,x_j)\right), \text{ and} \tag{8}$$

$$R'\left(k_{jam}-K(t,x_{j+1})\right)=Min\left(R\left(k_{jam}-K(t-l_{j+1}\varepsilon,x_{j+1})\right),\frac{d_{j+1}}{\varepsilon}\left[k_{jam}-K(t,x_{j+1})\right]\right). \tag{9}$$

$S'(k)$ is the modified sending function which is formed by the sending function used in L-CTM, $S(k)$ and the newly added term, $\frac{d_j}{\varepsilon}K(t,x_j)$. This additional term accounts for the current available number of vehicles in cell $j$ at time $t$, $d_jK(t,x_j)$. With this term, the maximum outflow of cell $j$ cannot be greater than the available number of vehicles. Similarly, $R'(k)$ is the modified receiving function, formed by $R(k)$

and the additional term $\frac{d_{j+1}}{\varepsilon}\left[k_{jam}-K\left(t,x_{j+1}\right)\right]$ to account for the current available space in the downstream cell $j+1$, $d_{j+1}\left[k_{jam}-K\left(t,x_{j+1}\right)\right]$ and to ensure the inflow to cell $j+1$ must be less than or equal to the available space. By introducing the two terms, we have a new model called EL-CTM formed by (1), (3), (4), (7)-(9), which can guarantee that the densities must be within $\left[0,k_{jam}\right]$ as shown below.

*Proposition 1: The density $K\left(t,x_j\right), \forall t>T$ obtained by EL-CTM must be within the range $\left[0,k_{jam}\right]$ if $K\left(t,x_j\right), \forall t\in\left[0,T\right]$ is in this range.*

Proof:

Given that $K\left(t,x_j\right), \forall t\in\left[0,T\right], \forall x$ is in the range $\left[0,k_{jam}\right]$, we know that $Q\left(K\left(t,x_j\right)\right), \forall t\in\left[0,T\right], \forall x$ must be nonnegative according to (8) and (9). Based on this fact, we consider two cases when $t=T$ : i) the uncongested case ( $K\left(t,x_j\right)\le k_0$ ) and ii) the congested case ( $K\left(t,x_j\right)>k_0$ ).

For the uncongested case, we show that $K\left(t,x_j\right)$ must be non-negative in this region. From (1), we have:

$$\begin{aligned} K\left(t+\varepsilon,x_j\right) &= K\left(t,x_j\right)-\frac{\varepsilon}{d_j}\left[Q\left(t+\varepsilon/2,x_j+d_j/2\right)-Q\left(t+\varepsilon/2,x_j-d_j/2\right)\right] \\ &\ge K\left(t,x_j\right)-\frac{\varepsilon}{d_j}\left[Q\left(t+\varepsilon/2,x_j+d_j/2\right)\right] \\ &\ge K\left(t,x_j\right)-\frac{\varepsilon}{d_j}S'\left(K\left(t,x_j\right)\right). \end{aligned} \tag{10}$$

The first inequality arises because $Q\left(t+\varepsilon/2,x_j-d_j/2\right)$ must be non-negative, and the second inequality arises because $K\left(t,x_j\right)\le k_0$. When $S\left(K\left(t-f_j\varepsilon,x_j\right)\right)<\frac{d_j}{\varepsilon}K\left(t,x_j\right)$, or

$$K\left(t,x_j\right)-\frac{\varepsilon}{d_j}\left[S\left(K\left(t-f_j\varepsilon,x_j\right)\right)\right]>0, \tag{11}$$

$S'\left(K\left(t,x_j\right)\right)=S\left(K\left(t-f_j\varepsilon,x_j\right)\right)$ according to (8). (10) then becomes $K\left(t+\varepsilon,x_j\right)\ge K\left(t,x_j\right)-\frac{\varepsilon}{d_j}\left[S\left(K\left(t-f_j\varepsilon,x_j\right)\right)\right]$, which is greater than zero because of (11). On the other hand, when $S\left(K\left(t-f_j\varepsilon,x_j\right)\right)\ge\frac{d_j}{\varepsilon}K\left(t,x_j\right)$, $S'\left(K\left(t,x_j\right)\right)=\frac{d_j}{\varepsilon}K\left(t,x_j\right)$. (10) becomes $K\left(t+\varepsilon,x_j\right)\ge K\left(t,x_j\right)-\frac{\varepsilon}{d_j}\left[\frac{d_j}{\varepsilon}K\left(t,x_j\right)\right]=0$. Thus, $K\left(t+\varepsilon,x_j\right)$ must be nonnegative when $t=T$.

For the congested case, we have:

$$\begin{aligned} K\left(t+\varepsilon,x_j\right) &= K\left(t,x_j\right)-\frac{\varepsilon}{d_j}\left[Q\left(t+\varepsilon/2,x_j+d_j/2\right)-Q\left(t+\varepsilon/2,x_j-d_j/2\right)\right] \\ &\le K\left(t,x_j\right)+\frac{\varepsilon}{d_j}\left[Q\left(t+\varepsilon/2,x_j-d_j/2\right)\right] \\ &\le K\left(t,x_j\right)+\frac{\varepsilon}{d_j}R'\left(k_{jam}-K\left(t,x_j\right)\right), \end{aligned} \tag{12}$$

since $Q\left(t+\varepsilon/2, x_j + d_j/2\right)$ is nonnegative and $K\left(t,x_j\right) > k_0$ . According to (9), $R'\left(k_{jam} - K\left(t,x_j\right)\right) = R\left(k_{jam} - K\left(t - l_j\varepsilon, x_j\right)\right)$ when $R\left(k_{jam} - K\left(t - l_j\varepsilon, x_j\right)\right) < \frac{d_j}{\varepsilon}\left[k_{jam} - K\left(t,x_j\right)\right]$ or when

$$K\left(t,x_j\right) + \frac{\varepsilon}{d_j}\left[R\left(k_{jam} - K\left(t - l_j\varepsilon, x_j\right)\right)\right] < k_{jam}. \tag{13}$$

Hence, (12) becomes $K\left(t+\varepsilon,x_j\right) \le K\left(t,x_j\right) + \frac{\varepsilon}{d_j}\left[R\left(k_{jam} - K\left(t - l_j\varepsilon, x_j\right)\right)\right]$, which is less than $k_{jam}$ due to (13). On the other hand, $R'\left(k_{jam} - K\left(t,x_j\right)\right)$ is equal to $\frac{d_j}{\varepsilon}\left[k_{jam} - K\left(t,x_j\right)\right]$ when $\frac{d_j}{\varepsilon}\left[k_{jam} - K\left(t,x_j\right)\right]$ is less than $R\left(k_{jam} - K\left(t - l_j\varepsilon, x_j\right)\right)$. Hence, (12) becomes:

$$K\left(t+\varepsilon,x_j\right) \le K\left(t,x_j\right) + \frac{\varepsilon}{d_j}\left[\frac{d_j}{\varepsilon}\left[k_{jam} - K\left(t,x_{j+1}\right)\right]\right] = k_{jam}.$$

Hence, $K\left(t+\varepsilon,x_j\right)$ must be less than or equal to the jam density at $t = T$ .

Combining the results of cases (i) and (ii), $K\left(t+\varepsilon,x_j\right)$ must be within in the range of $\left[0,k_j\right]$ at $t = T$ . Using this fact, we know $Q\left(K\left(T+\varepsilon,x_j\right)\right)$ must be nonnegative. We repeat the above analysis at each $t = T+\varepsilon, T+2\varepsilon, \ldots$ starting from $t = T+\varepsilon$ sequentially. We can then find that $0 \le K\left(t,x\right) \le k_{kam}$ when $t > T$ if this holds when $\forall t \in \left[0,T\right]$. This completes the proof. □

EL-CTM becomes CTM if both lags equal zero and all cell lengths are the same, because the two additional terms are always redundant. This is pointed out by the following proposition.

*Proposition 2: EL-CTM becomes CTM if both lags are equal to zero and all cells have the same length.*
Proof:
When $f_j = l_j = 0$ , $d_j = d$ , (8) and (9) become:

$$S'\left(K\left(t,x_j\right)\right) = Min\left[S\left(K\left(t,x_j\right)\right), \frac{d}{\varepsilon}K\left(t,x_j\right)\right], \text{ and} \tag{14}$$

$$R'\left(k_{jam} - K\left(t,x_{j+1}\right)\right) = Min\left[R\left(k_{jam} - K\left(t,x_{j+1}\right)\right), \frac{d}{\varepsilon}\left(k_{jam} - K\left(t,x_{j+1}\right)\right)\right]. \tag{15}$$

The first term in the square brackets in (14) must be less than the second term because

$$\frac{d}{\varepsilon} \ge \left|S_{k,\max}\right| > \frac{\int_0^{K\left(t,x_j\right)} S_k\left(z\right)dz}{K\left(t,x_j\right)} = \frac{S\left(K\left(t,x_j\right)\right)}{K\left(t,x_j\right)}. \tag{16}$$

The first inequality is due to (3). The second inequality means that the maximum absolute forward wave speed must be greater than the average forward wave speed. The last equality is obtained by simplifying the numerator. Rearranging the terms in (16) gives $S\left(K\left(t,x_j\right)\right) < \frac{d}{\varepsilon}K\left(t,x_j\right)$. This implies:

$$S'\left(K\left(t,x_j\right)\right) = S\left(K\left(t,x_j\right)\right), \tag{17}$$

and the second term in the brackets in (14) is redundant.

Similarly, the first term in the square brackets in (15) must be less than the second term because

$$\frac{d}{\varepsilon} \ge \left|R_{k,\max}\right| > \frac{\int_0^{k_{jam} - K\left(t,x_{j+1}\right)} R_k\left(z\right)dz}{k_{jam} - K\left(t,x_{j+1}\right)} = \frac{R\left(k_{jam} - K\left(t,x_{j+1}\right)\right)}{k_{jam} - K\left(t,x_{j+1}\right)}. \tag{18}$$

Rearranging the terms gives $R\left(k_{jam} - K\left(t,x_{j+1}\right)\right) < \frac{d}{\varepsilon}\left(k_{jam} - K\left(t,x_{j+1}\right)\right)$, which implies:

$$R'\left(k_{jam} - K\left(t, x_{j+1}\right)\right) = R\left(k_{jam} - K\left(t, x_{j+1}\right)\right). \quad (19)$$

Substituting (17) and (19) into (7) gives $Q\left(t+\varepsilon/2, x_j + d_j/2\right) = Min\left(S\left(K\left(t, x_j\right)\right), R\left(k_{jam} - K\left(t, x_{j+1}\right)\right)\right)$, which is the rule used in CTM. This completes the proof. □

As both EL-CTM and L-CTM are more general than CTM and L-CTM has the undesirable properties in estimating density near the jam density region or zero density region, one may question whether CTM will generate negative densities or densities larger than the jam density. The answer is no, as CTM does not consider the lags. The proof can be found in Daganzo (1995b).

One of the features of EL-CTM is that when the adjustment terms are redundant, so that EL-CTM becomes L-CTM and retain all the desirable properties of L-CTM. When they are not, they can help improving the accuracy of density prediction, as shown in the two examples in the next section.

## 5 Numerical studies

*Example 1. EL-CTM predicts the exact solution*

The setting of Example 1 is the same as that of counter example 1. The highway is discretized into 7 cells in which the first 6 cells have the same length but the last one has the cell length five times longer. The result is obtained by EL-CTM and completely satisfies the LWR theory as shown in Table 1.

*Example 2. EL-CTM predicts a better result than L-CTM*

Similarly, the setting of example 2 follows that of counter example 2, but this example computes the absolute errors produced by EL-CTM and L-CTM, which are defined as the density estimated by the model minus the correct density $K\left(t, x_j\right) = \min\left[160 + 2\left(j + t/5\right), 180\right]$. The absolute errors are shown in Tables 4 and 5. The positive (negative) values mean that the estimated densities are lower (higher) than the jam density. According to these tables, one can see that EL-CTM produces a better result. First, EL-CTM always estimates densities less than or equal to the jam density, as shown by the fact that the errors are always non-negative. Second, its error region (the region where the numbers are non-zero) is much smaller. Third, its maximum error in each cell is smaller. Fourth, the magnitudes of errors (absolute values) are always equal to or less than the corresponding values estimated by L-CTM.

## 6. Concluding remarks

This paper shows that L-CTM can give densities outside the feasible range, and proposes a model to get rid of this. The proposed model can never yield densities outside the feasible range and includes some desirable features of L-CTM, including allowing variable cell lengths and capturing the coasting effect of traffic in queues by using a non-concave flow density relation. The proposed model can include CTM and L-CTM as special cases, and can produce more accurate results than L-CTM.

## References

Adamo, V., Astarita, V., Florian, M., Mahut, M. and Wu, J.H. 1999. Modeling the Spillback of Congestion in Link Based Dyanmic Network Loading Models: A Simulation Model with Application. In A. Ceder (Ed.), Transportation and Traffic Theory, Pergamon-Elservier, New York, 555-573.

Addison, J.D. and Heydecker, B.G. 1995. Traffic Models for Dynamic Assignment, In N.H. Gartner and G. Improta (Ed.), Urban Traffic Networks: Dynamic Flow Modeling and Control. London: Spring, 213-231.

Balijepalli, N.C. and Watling, D.P., 2005. Doubly Dynamic Equilibrium Distribution Approximation Model for Dynamic Traffic Assignment. In Proceedings of 16th International Symposium on Transportation and Traffic Theory, Maryland, USA.

Ben-Akiva M., Koutsopoulos H.N., Mishalani R. 1998. DynaMIT: A Simulation-Based System for Traffic Prediction. Paper presented at the DACCORD Short Term Forecasting Workshop, Delft. (See also its.mit.edu).

Bliemer M.C.J. and P.H.L. Bovy. 2003. Quasi-Variational Inequality Formulation of the Multiclass Dynamic Traffic Assignment Problem. Transportation Research Part B 37 501–519

Table 4 the exact result minus the corresponding L-CTM result

| time\cell | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 3 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.08** | 0.00 | 0.00 |
| 5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.24** | 0.00 | 0.00 |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.08** | 0.00 | 0.00 |
| 7 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.02** | -0.02 | 0.00 | 0.00 |
| 8 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.06** | -0.06 | 0.00 | 0.00 |
| 9 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.16** | -0.08 | 0.00 | 0.00 |
| 10 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.31** | -0.08 | 0.00 | 0.00 |
| 11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.02** | **0.13** | -0.06 | 0.00 | 0.00 |
| 12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.05** | 0.00 | -0.05 | 0.00 | 0.00 |
| 13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.11** | -0.08 | -0.03 | 0.00 | 0.00 |
| 14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.21** | -0.12 | -0.02 | 0.00 | 0.00 |
| 15 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.01** | **0.36** | -0.13 | -0.01 | 0.00 | 0.00 |
| 16 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.04** | **0.17** | -0.12 | 0.00 | 0.00 | 0.00 |
| 17 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.08** | **0.02** | -0.10 | 0.00 | 0.00 | 0.00 |
| 18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.15** | -0.08 | -0.07 | 0.00 | 0.00 | 0.00 |
| 19 | 0.00 | 0.00 | 0.00 | 0.00 | **0.01** | **0.25** | -0.13 | -0.05 | 0.00 | 0.00 | 0.00 |
| 20 | 0.00 | 0.00 | 0.00 | 0.00 | **0.03** | **0.40** | -0.16 | -0.03 | 0.00 | 0.00 | 0.00 |
| 21 | 0.00 | 0.00 | 0.00 | 0.00 | **0.05** | **0.20** | -0.16 | -0.02 | 0.00 | 0.00 | 0.00 |
| 22 | 0.00 | 0.00 | 0.00 | 0.00 | **0.10** | **0.04** | -0.14 | -0.01 | 0.00 | 0.00 | 0.00 |
| 23 | 0.00 | 0.00 | 0.00 | **0.01** | **0.18** | -0.07 | -0.12 | 0.00 | 0.00 | 0.00 | 0.00 |
| 24 | 0.00 | 0.00 | 0.00 | **0.02** | **0.29** | -0.14 | -0.09 | **0.01** | 0.00 | 0.00 | 0.00 |
| 25 | 0.00 | 0.00 | 0.00 | **0.04** | **0.43** | -0.18 | -0.06 | **0.01** | 0.00 | 0.00 | 0.00 |

Table 5 the exact result minus the corresponding EL-CTM result

| time\cell | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 3 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.08** | 0.00 | 0.00 |
| 5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.24** | 0.00 | 0.00 |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.08** | 0.00 | 0.00 |
| 7 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 8 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 9 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.08** | 0.00 | 0.00 | 0.00 |
| 10 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.24** | 0.00 | 0.00 | 0.00 |
| 11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.08** | 0.00 | 0.00 | 0.00 |
| 12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.08** | 0.00 | 0.00 | 0.00 | 0.00 |
| 15 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.24** | 0.00 | 0.00 | 0.00 | 0.00 |
| 16 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.08** | 0.00 | 0.00 | 0.00 | 0.00 |
| 17 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 19 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.08** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.24** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 21 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.08** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 22 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 24 | 0.00 | 0.00 | 0.00 | 0.00 | **0.08** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 25 | 0.00 | 0.00 | 0.00 | 0.00 | **0.24** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Boel, R. and L. Mihaylova .2006. A Compositional Stochastic Model for Real Time Freeway Traffic Simulation. Transportation Research Part B 40 319–334

Cantarella, G.E., Cascetta E., Adamo V., and Astarita V., 1999 A Doubly Dynamic Traffic Assignment Model. Proceedings of 14th International Symposium on Transportation and Traffic Theory, 373-396.

Carey M. 1987. Optimal Time Varying Flows on Congested Networks. Operations Research 35(1), 58-69.

Daganzo C.F. 1994. The Cell-transmission Model. A Dynamic Representation of Highway Traffic Consistent with The Hydrodynamic Theory. Transportation Research 28B, 269-288.

Daganzo C.F. 1995a. The Cell Transmission Model, Part II: Network Traffic. Transportation Research 29B, 79-94.

Daganzo, C.F. 1995b. A Finite Difference Approximation for the Kinematic Wave Model. Trans. Res. 29B(4) 261-276.

Daganzo, C.F. 1997. A Continuum Theory of Traffic Dynamics for Freeways with Special Lanes. Trans Res. 31B (2), 83-102.

Daganzo, C.F. 1999. The Lagged Cell Transmission Model. A. Ceder (Ed.), Transportation and Traffic Theory, Pergamon-Elservier, New York, 81-103

Daganzo, C.F., Lin, W.H. and del Castillo, J.M. 1997. A Simple Physical Principle for the Simulation of Freeways with Special Lanes and Priority Vehicles. Trans. Res. 31B (2), 105-125.

Friesz, T.L., Bernstein, D.H., Smith, T.E., Tobin, R.L., and Wie, B.W. 1993. A Variational Inequality Formulation of the Dynamic Network User Equilibrium Problem. Operations Research 41, 179-191.

Gentile G, Meschini L, Papola N. 2005. Macroscopic Arc Performance Models with Capacity Constraints for Within-Day Dynamic Traffic Assignment. Transportation Research Part B 39 (2005) 319–338

Ishak S, Alecsandru C., Seedah D. 2006. Improvement and Evaluation of the Cell-Transmission Model for Operational Analysis of Traffic Networks: A Freeway Case Study. Proceedings of 85th TRB annual meeting

Janson B.N. 1998. Convergent Algorithm for Dynamic Traffic Assignment. Transportation Research Record 1328, 69-80.

Jayakrishnan, R., Tsai, W.K., and Chen, A. 1995. A Dynamic Traffic Assignment Model with Traffic Flow Relationship. Transportation Research 3C, 51-82.

Jin W.L. and Zhang H.M. 2004. A Multicommodity Kinematic Wave Simulation Model of Network Traffic Flow. Transportation Research Record 1883:59-67, 2004.

Kuwahara M. and Akamatsu T. 2001. Dynamic User Optimal Assignment with Physical Queues for a Many-to-many OD Pattern. Transportation Research 35B, 461-479.

Lam, W.H.K., and Huang, H.J. 1995. Dynamic User Optimal Traffic Assignment Model for Many to One Travel Demand. Transportation Research 29B, 243-259.

Lighthill M.J. and Whitham G.B. 1955. On kinematic waves (II): A Theory of Traffic Flow on Long Crowded Roads. Proc. Roy. Soc., A. 229, 281-345.

Liu, Y. Lai X.R. and Chang, G.L. 2005. A Two-Level Integrated Optimization Model for Planning of Emergency Evacuation: A Case Study of Ocean City under Hurricane Evacuation. Proceedings of 84th TRB meeting.

Muñoz, Laura, X.T. Sun, R Horowitz, L. Alvarez. 2003. Traffic Density Estimation with the Cell Transmission Model. Proceedings of the American Control Conference 3750-3756

Newell G.F. 1993. A Simplified Theory of Kinematic Waves in Highway Traffic, Parts I - III. Transpn Research 27B, 281-313.

Peeta S and Mahmassani. H. S. 1995. Multiple User Classes Real-Time Traffic Assignment for Online Operations: A Rolling Horizon Solution Framework. Transpn. Res.-C, Vol. 3. No. 2. pp. 83-98, 1995

Perakis G. and Roels G. 2004. An Analytical Model for Traffic Delays and the Dynamic User Equilibrium Problem. Accepted for publication in Operations Research, see also http://web.mit.edu/~georgiap/www/Perakis_Roels.pdf (Oct 2005).

Rubio-Ardanaz, J.M., Wu, J.H., and Florian, M. 2001. A Numerical Analytical Model for the Continuous Dynamic Network Equilibrium Problem with Limited Capacity and Spill Back. 2001 IEEE Intelligent Transportation Systems Conference Proceedings, 263-267.

Ran B. and Boyce D.E. 1996. A link-based Variational Inequality Formulation of Ideal Dynamic User Optimal Route Choice Problem. Transportation Research 4C(1), 1-12.

Richards P.I. 1956. Shockwaves on the Highway. Operations Research 4, 42-51.

Smith, M.J. 1993. A New Dynamic Traffic Model and the Existence and Calculation of Dynamic User Equilibria on Congested Capacity-Constrained Road Networks. Transportation Research 27B, 49-63.

Szeto W.Y. and Lo, H.K. 2005. Dynamic Traffic Assignment: Review and Future Research Directions. Journal of Transportation Systems Engineering and Information Technology, 5(5), 85-100.

Tong, C.O. and Wong, S.C. 2000. A Predictive Dynamic Traffic Assignment Model in Congested Capacity-Constrained Road Networks. Transportation Research 34B, 625-644.

Wu, J.H., Chen, Y., and Florian, M. 1998. The Continuous Dynamic Network Loading Problem: A Mathematical Formulation and Solution Method. Transportation Research 32B, 173-187.

Yperman I., Logghe S. and Immers L.H. Tampere C. 2006. The Multi-Commodity Link Transmission Model for Dynamic Network Loading. Proceedings of 85th Annual TRB meeting.

Ziliaskopoulos, A. K. and Lee, S. 1997. A Cell Transmission Based Assignment-Simulation Model for Integrated Freeway/Surface Street Systems. Transportation Research Record 1701, 12-23.

# ROUTE GENERATION AND DYNAMIC TRAFFIC ASSIGNMENT FOR LARGE NETWORKS

*Michiel Bliemer: Delft University of Technology, The Netherlands, m.c.j.bliemer@tudelft.nl*
*Henk Taale: Delft University of Technology, The Netherlands, h.taale@tudelft.nl*

## Abstract
In this paper two existing dynamic traffic assignment models (INDY and MARPLE) are described and tested for a large road network (the Dutch main road network). The models consist of three modules: route set generation, route choice and network loading. The focus of the paper is on the route generation process, which generates an a priori route set. It is shown that both models are able to quickly generate routes for large networks with normal computer resources. The variance allowed in the link travel times is an important variable within this process as well as the allowed route overlap. For the DTA process itself, it can be concluded that it is also feasible to use a dynamic model on large networks, even using relatively small time steps. Both computation time for a single iteration and internal memory usage are within reasonable limits for both models. MARPLE is faster, but uses more internal memory. Typically, only a limited number of iterations are needed for convergence to a stochastic dynamic user-equilibrium due to the use of an a priori route set.

## 1 Introduction
In The Netherlands transport and traffic policy heavily relies on traffic management. Building new roads is either too expensive or too difficult due to spatial and environmental conditions. Road pricing will not be feasible the coming years, so traffic management is the key direction in which solutions for the increasing congestion problems have to be found. Traditionally, traffic management is local: locally there is a problem and it is solved with a local traffic management measure, mostly without considering the effects on the rest of transportation system or other side effects. Also, in most cases, motorways and urban roads are operated and maintained by different road authorities. In practise, these authorities are only responsible for their own part of the network and do not have the incentive to cooperate. In The Netherlands this problem has been recognised and a method for regional cooperation has been developed (Rijkswaterstaat, 2003), together with a tool to support the steps described in this method (Taale *et al.*, 2004). Part of this tool is a dynamic traffic assignment model, called MARPLE (Model for Assignment and Regional Policy Evaluation). MARPLE is the result of research at the Delft University of Technology on the combined traffic assignment and control problem. It is a route-based macroscopic dynamic traffic assignment (DTA) model, which uses travel time functions to propagate traffic through the network. The assignment can either be deterministic or stochastic. Earlier, research on multi-class models led to the multi-class DTA model INDY (Bliemer *et al.*, 2003; Bliemer, 2005). INDY is also a route-based macroscopic DTA model and is similar to MARPLE, but uses a different dynamic network loading procedure and is completely multi-class.

Performing DTA calculations is very cumbersome and time consuming, especially for larger networks. In most DTA models, this is due to the large number of (dynamic) shortest path calculations for successive time periods and successive iterations. Apart from the computational burden, the DTA solution approaches using successive shortest path search have a number of theoretical drawbacks.

This paper compares two DTA approaches (INDY and MARPLE) offering computational and theoretical advantages by applying so-called a priori route choice sets established in advance of the route choice modelling and consequent dynamic network loading (DNL). In this approach, the set of attractive feasible paths for all time periods is generated only once at the beginning of the DTA process. The behavioural bases and computational principles of this effective and efficient choice set generation approach are elaborated. The use of a priori route set generation has advantages for the travel choice modelling in DTA, such as the easy treatment of non-linearities present in the generalized cost function and unlimited flexibility in choice model types to be adopted. Complexities such as physical overlap among routes and other dependencies can be handled satisfactorily with a-priori route set generation.

## 2 Dynamic traffic assignment framework

In dynamic traffic assignment (DTA) we are interested in route choice proportions, link loads, and travel times on the network. In this paper we focus on a (stochastic) dynamic user-equilibrium assignment on large networks. Peeta and Ziliaskopoulos (2001) give an overview of different DTA approaches in which they mainly distinguish pure analytical approaches and simulation-based approaches. Pure analytical approaches (e.g., Ran and Boyce, 1996; Bliemer and Bovy, 2003) have a nice mathematical structure and can be solved using standard optimization techniques, however they are typically restricted to small hypothesized networks. On the other hand, simulation-based DTA models are designed for transportation problems and are able to handle real-life networks. A wide range of simulation-based DTA models have been proposed, which range from microscopic models (e.g., PARAMICS, AIMSUN, VISSIM) to macroscopic models (e.g., VISUM, INDY, MARPLE). Microscopic models that base computations on individual vehicles can include a lot of detail in car movements, but also require a significant amount of computer memory and computation time in case many vehicles are present at the same time on the network. This makes them suitable for small area (urban) networks. Macroscopic models rely on aggregate variables on links of the network, such that the memory usage and computation time are independent of the number of vehicles. Hence, they are very suitable for large area (freeway) networks in which less detail is required in vehicle movements.

In this paper we discuss the feasibility of DTA in large networks using two macroscopic simulation-based models, MARPLE (Taale, 2004) and INDY (Bliemer *et al*., 2003; Bliemer, 2005). Both models are route-based and use route sets that are a priori generated, thereby avoiding time-consuming (dynamic) fastest paths computations while running the DTA model. The framework of both models is the same and is depicted in Figure 1.

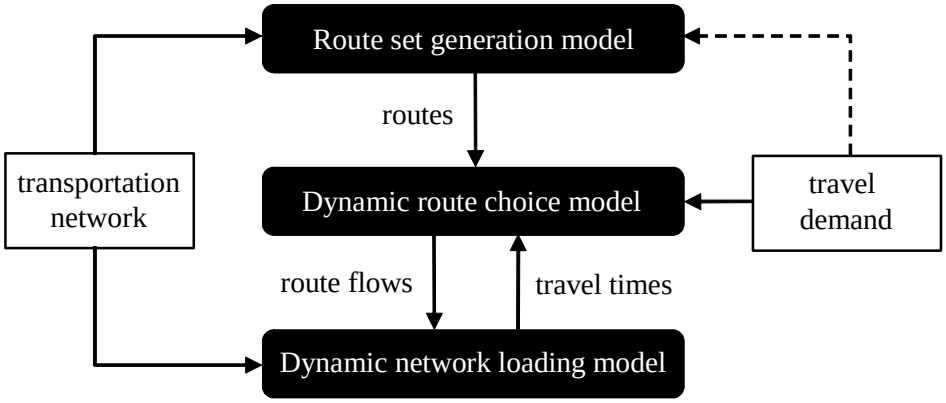



***Figure 1:*** *DTA components of MARPLE and INDY*

MARPLE as well as INDY consider three components: (i) a route set generation model, (ii) a dynamic route choice model, and (iii) a dynamic network loading model. Given the transportation network, routes are generated by the route set generation model. These routes are typically generated without any knowledge of the travel demand, as will be explained in the next section, only the set of origin-destination (OD) pairs are passed to the route generation model. Once the route set is determined, route choice proportions are computed for each OD pair and each departure time interval based on a generalized dynamic route travel costs (which includes route travel times and maybe other attributes). Multiplying these proportions by the (given) dynamic OD demand for each departure time, we obtain route flows. Finally, these route flows are propagated along each route on the network by using a dynamic network loading (DNL) model. The DNL model is the most computationally-expensive model in the framework and performs the traffic simulation using a system of analytical equations.

Each of the DTA components will be discussed in the following sections, with a special focus on the route set generation procedure. As large networks may require special settings of the components in the DTA model, each section will briefly discuss such settings that may enable faster computations.

# 3 Route set generation model

The first component is the route set generation model. Given a transportation network consisting of nodes and (directed) links, and given a set of OD pairs, a set of routes is determined for each OD pair. It is important to note that the route set generation is only performed once. Instead of finding the (dynamic) shortest routes continuously as part of the assignment and simulation process, it is assumed that the road user chooses a route from an a-priori set of routes (which has been argued to be consistent with route choice from a behavioural point of view). The advantage is that the routes can be computed in advance. The disadvantage is that the set of routes is fixed and does not change during the assignment. Hence, it is important to generate a sufficiently large set for each OD pair such that at least all used routes are included. Clearly, it is not known in advance which routes will be used, therefore the generated route set will typically be larger than the set of used routes (although the number of unused routes should preferably be kept to a minimum). Ideally, one should check after the assignment with a dynamic fastest path algorithm if there are any routes on the network that have a lower travel time than the routes in the route set. If so, then these routes should be added to the route set and the DTA model should be run again.

Different route set generation models could be considered (Bliemer, 2001), such as:

(a) *All acyclic routes* – All routes without cycles are enumerated, independent on their length and travel time. This would generate a huge route set with many unrealistic alternatives. For real-life networks, this approach is not feasible;
(b) *k-shortest routes* – In this approach, the fastest route, the 2nd fastest route, till the *k*-th fastest route will be determined. The number of routes for each OD pair will be exactly *k*, which may for certain OD pairs not be sufficient, or could be too large (yielding unrealistic routes);
(c) *Essentially least-cost routes* – All routes with a travel time within a certain threshold from the fastest route are considered. In this case, the number of routes may be different for each OD pair and would not generate unrealistic routes;
(d) *Most probable routes* – Using Monte Carlo simulations, a set of routes is generated in which the link travel times are assumed to be a random variable (e.g., representing congestion). Same as in the previous approach, the number of routes per OD pair is an outcome, not an input.

MARPLE and INDY both adapt a form of route generation approach (d), based on the most probably routes. It avoids the problem of an exploding number of routes, the problem of having too few or too many (unrealistic) routes, and the problem of only assuming free-flow travel times. Moreover, approach (d) is very easy to implement and the computation time is low (even for large networks).

In the route generation, the link travel times $\tau_a$ for each link *a* are assumed be random variables:

$$\tau_a = \tau_a^0\left(1+\left|\varepsilon_a\right|\right), \qquad \text{where } \varepsilon_a \sim N(0,\sigma^2). \tag{1}$$

Given the free-flow travel time $\tau_a^0$, a positive random component $\varepsilon_a$ is added to this free-flow travel time. This random component is assumed to be normally distributed with zero mean and a variance $\sigma^2$. A higher variance is likely to lead to higher link travel times. If one chooses $\sigma = \frac{1}{3}$, then it is unlikely that the link travel times are longer than twice the free-flow travel time (since $\Pr(\left|\varepsilon_a\right| < 3\sigma) = 0.997$). By iteratively drawing different error components for each link and finding the fastest route based on the updated link travel times (1), routes can be generated. The higher the number of iterations and the higher $\sigma$, the more routes will be generated. Initially the variance is set to zero, to ensure that the actual fastest route is always included. Then in INDY, over the number of iterations, the variance can be increased, which will increase the probably of finding a new route (with also an increasing detour). MARPLE ensures that the detour is not too large by keeping the variance constant, resulting in only allowing routes that are within a certain threshold from the fastest route (as proposed in the essentially least cost approach).

Although routes that take an off-ramp and immediately an on-ramp again are perfectly valid, they are usually considered unwanted in the assignment. In order to prevent this type of routes and in order to create a set of

sufficiently different routes, an overlap filter is used to remove routes that have too much overlap with previously generated routes.

*Route set generation algorithm* (for both MARPLE and INDY)

| | |
|---|---|
| Input: | Network with for each link free-flow travel times, and a set of OD pairs with positive travel demand. Parameters: maximum variance $\sigma$, the maximum allowed number of routes per OD pair, $K$, the maximum percentage of route overlap, $\gamma$, and the maximum number of iterations, $N_1$. |
| Output: | A route set $P^{rs}$ for each OD pair $(r,s)$. |
| Step 1: | Set $n := 1$. Set $\sigma^{(n)} := 0$. |
| Step 2: | Compute link travel times $\tau_a^{(n)} = \tau_a^0\left(1+\left\|\varepsilon_a^{(n)}\right\|\right)$, with independent draws $\varepsilon_a^{(n)} \sim N\left(0,\left(\sigma^{(n)}\right)^2\right)$. |
| Step 3: | For each OD pair $(r,s)$, find the currently fastest path $p^{rs,(n)}$ using Dijkstra's algorithm. |
| Step 4: | For each OD pair $(r,s)$, if $p^{rs} \notin P^{rs}$, and $\left\|P^{rs}\right\| < K$, and the overlap in route length is not more than $\gamma$, then set $P^{rs} := P^{rs} \cup p^{rs}$. |
| Step 5: | If $n = N_1$, then stop. Otherwise, set $n := n+1$ and set $\sigma^{(n)} := \sigma$ (in MARPLE) or $\sigma^{(n)} := \sigma/(N_1 - n + 1)$ (in INDY). Continue with Step 2. |

In INDY there is also an alternative route set generation procedure implemented based on static traffic assignment. If this procedure is selected, the OD matrix for the time period with the highest travel demand (with an optional multiplication factor) is selected and a static deterministic user-equilibrium assignment is performed on the network. The used routes for each OD pair will be stored, yielding the route set, and also the route flow proportions are stored. The latter information can be used in the dynamic route choice model as initial route flow proportions, which may speed up the convergence.

For large networks, a priori route set generation has several advantages. First of all, fastest paths only have to be computed only once, and the route set could be used as input in subsequent studies. Secondly, starting with a set of routes instead of with a single route in the first iteration (which is common for most assignment models that find fastest paths during the assignment) enables the distribution of the OD flows over multiple routes already from the first iteration. This may significantly speed up the convergence of the DTA model.

In the route set generation model, some settings can be chosen that reduce the computation time. As previously mentioned, by using an overlap factor we try to rule out routes that may not contribute much in terms of extra capacity, but do add to the computation time. More importantly, we should aim to reduce the number of OD pairs. Obviously, for each OD pair routes have to be generated, hence if we are able to reduce the number of OD pairs, also the number of routes will decrease. Needless to say that no route generation is necessary for OD pairs with zero flow, hence these OD pairs can be removed. However, there are possibly many OD pairs with very little flow (e.g., less than 1 vehicle per hour). By removing also these OD pairs the number of OD pairs can be reduced significantly without removing much travel demand (as will be shown in the case study in Section 6, we were able to remove more than 50% of the OD pairs, which account for less than 1% of the total travel demand). Removing low demand OD pairs should be done with care, as they are likely to be long distance trips (and therefore to many links on the network), hence removing this travel demand may have some impact on the DTA outcomes. Another advantage of removing OD pairs is that the number of used links decreases, which speeds up the dynamic network loading

## 4 Dynamic route choice model

Once for each OD pair $(r,s)$ the set of routes, $P^{rs}$, has been generated, the travellers have to choose between these alternative routes. In this section we will assume that the route choice behaviour is based on a stochastic dynamic user-equilibrium in which no traveller thinks he or she can be better off by unilaterally changing routes. As commonly assumed, travellers will choose the route alternative with the lowest actually experienced (perceived) generalized cost. This generalized cost may consist of route travel times travel, route toll costs, etc. For simplicity, here we assume that the travel cost solely consists of travel times, although this can easily be extended. Let $\theta_{ap}^{rs}(k,t)$ denote the dynamic link-route incidence indicator, which equals one if

vehicles departing at time *k* on route *p* from *r* to *s* enter link *a* at time *t*, or zero otherwise. This indicator depends on the link travel times on the network. Then the actually experienced route cost is

$$c_p^{rs}(k) = \sum_{a \in p} \theta_{ap}^{rs}(k,t) c_a(t), \tag{2}$$

where $c_a(t)$ is the cost of link *a* when entering the link at time *t*. Each route cost may be perceived differently by different travellers, hence the route cost is assumed to be a random variable by adding an unobserved random term to the route costs. Under the assumption that all routes are independent and that the unobserved term is extreme value type I distributed, the route choice proportions $\psi_p^{rs}(k)$ are given by the well-known multinomial logit (MNL) model. If routes are overlapping, then the independence assumption may not hold. Therefore, a commonality factor (Cascetta, 1996) in MARPLE or a path size factor (Hoogendoorn *et al.*, 2005) in INDY is added to each route cost, denoted by $F_p^{rs}$. Then, the route choice proportions for each departure time *k* can be computed as

$$\psi_p^{rs}(k) = \frac{\exp\left[-\mu\left(c_p^{rs}(k) + F_p^{rs}\right)\right]}{\sum_{p'} \exp\left[-\mu\left(c_{p'}^{rs}(k) + F_{p'}^{rs}\right)\right]}. \tag{3}$$

Let the travel demand for OD pair $(r,s)$ departing at time *k* be given by $D^{rs}(k)$. The route flows $f_p^{rs}(k)$ can be determined by

$$f_p^{rs}(k) = \psi_p^{rs}(k) D^{rs}(k). \tag{4}$$

The route choice is an iterative process in which the route flows are determined based on route costs, which may change again due to changes in the route flows. In each iteration *n*, new route flows are computed using Eqns. (3) and (4), which will be averaged with route flows from previous iterations to speed up convergence. The method of successive averages (MSA) is used in INDY with an averaging weight of $\zeta^{(n)} = 1/n$, while in MARPLE the weights are chosen as $\zeta^{(n)} = \alpha_1 \exp(-\alpha_2 n) + \alpha_3 / n$ (with some $\alpha_i \geq 0$) such that the weights for small *n* are larger than those in MSA and smaller for larger *n* (which may improve the rate of convergence). Faster iterative averaging schemes as proposed by Bottom and Chabini (2001) have been investigated, but these accelerated averaging schemes seem to work predominantly if a larger number of iterations is needed. In our case, where we use an a-priori route set, convergence is typically reached within a small number of iterations (5 to 20). Accelerated averaging therefore did not show a significant improvement, but it did require an extra set of route flows to be stored in memory, which may be problematic for large networks.

The convergence can be checked using multiple criteria. The first one is that the absolute route flow differences between iterations are going to zero. Although this criteria is useful to see that at some point the algorithm can be terminated, it does not guarantee that the algorithm has reaches a (stochastic) dynamic user-equilibrium. This is due to the fact that applying MSA or other averaging schemes yield smaller route flow changes in each iteration by default, so is not a good measure for convergence to a user-equilibrium. A better measure is the dynamic relative duality gap, defined by

$$G(k) = \frac{\sum_{(r,s)} \sum_{p \in P^{rs}} \sum_{k} f_p^{rs}(k)\left(c_p^{rs}(k) - \pi^{rs}(k)\right)}{\sum_{(r,s)} \sum_{k} D^{rs}(k) \pi^{rs}(k)}, \quad \text{where} \quad \pi^{rs}(k) \equiv \min_{p \in P^{rs}} \left(c_p^{rs}(k)\right). \tag{5}$$

For each departure time *k*, this relative gap function should decrease. In a stochastic assignment it will not go to zero, but it should stabilize at some value somewhat greater than zero.

*Route choice algorithm* (for both MARPLE and INDY)

| | |
|---|---|
| Input: | Route sets $P^{rs}$, dynamic travel demand $D^{rs}(k)$, route travel cost functions, and the maximum number of iterations, $N_2$. |
| Output: | Dynamic route flows $f_p^{rs}(k)$. |
| Step 1: | Set $n := 1$. Base the initial travel costs $c_a^{(n)}(t)$ on the free-flow travel times. |
| Step 2: | Compute the route travel costs $c_p^{rs,(n)}(k)$ as in Eqn. (2) based on the current link travel costs $c_a^{(n)}(t)$ and current link travel times. |
| Step 3: | Compute new intermediate route flows $\tilde{f}_p^{rs,(n)}(k)$ using Eqns. (3) and (4) with $c_p^{rs,(n)}(k)$. |
| Step 4: | Compute new route flows $f_p^{rs,(n)}(k) = f_p^{rs,(n-1)}(k) + \zeta^{(n)}\left(\tilde{f}_p^{rs,(n)}(k) - f_p^{rs,(n-1)}(k)\right)$. |
| Step 5: | Perform dynamic network loading (see Section 5) to compute new link travel costs $c_a^{(n+1)}(t)$. |
| Step 6: | If $n = N_2$ or if other convergence criteria are satisfied (maximum route flow changes, maximum relative duality gap), then stop. Otherwise, set $n := n+1$ and return to Step 2. |

# 5 Dynamic network loading model

The third component of both DTA models is the dynamic network loading module. The dynamic routes flows $f_p^{rs}(k)$ from the assignment module are loaded onto the network and flows and travel times are calculated. The remainder of this section describes the DNL algorithms for INDY[1] and MARPLE

*Dynamic network loading algorithm* (for INDY)

| | |
|---|---|
| Input: | Dynamic route flows $f_p^{rs}(k)$, link capacities, maximum speeds, speed-density functions. |
| Output: | Dynamic link travel times $\tau_a(t)$ and link costs $c_a(t)$. |
| Step 1: | Consider an empty network. Set $t := 1$. |
| Step 2: | For each link *a*, compute the current cumulative outflow $V_a(t)$ out of the link, using the inflow history of the link. |
| Step 3: | For each link *a*, compute the current cumulative inflow $U_a(t)$ into the link, using the dynamic route flows $f_p^{rs}(k)$ and the outflows of previous links. |
| Step 4: | For each link *a*, compute the current number of vehicles $X_a(t)$ on the link, $X_a(t) = U_a(t) - V_a(t)$. |
| Step 5: | For each link *a*, compute the current link travel time $\tau_a(t)$ based on the number of vehicles $X_a(t)$ on the link, using speed-density relationships. |
| Step 6: | While not all travel demand has been assigned yet and while the network is not empty yet, set $t := t+1$ and return to Step 2. |

*Dynamic network loading algorithm* (for MARPLE)

| | |
|---|---|
| Input: | Dynamic route flows $f_p^{rs}(k)$, link capacities, maximum speeds, flow-travel time functions. |
| Output: | Dynamic link travel times $\tau_a(t)$ and link costs $c_a(t)$. |
| Step 1: | Initialise the network and give every link an initial flow consistent with the demand. |
| Step 2: | For each link *a* and simulation period *k*, compute the free flow travel time and the change in capacity according to the signal control plans (controlled links only). Set $t := 1$. |
| Step 3: | For each link *a*, compute the travel time and delay with the travel time functions (different link types have a different travel time function). |
| Step 4: | For each link *a*, compute the outflow and the remaining space. |
| Step 5: | For each node, compute the inflows and outflows, based on the routes flows $f_p^{rs}(k)$ and splitting rates. |
| Step 6: | For each link *a*, compute the inflow and correct it for the available space on the link. |
| Step 7: | For each link *a*, if necessary adjust the outflow, because of downstream queues. |
| Step 8: | For each link *a*, compute the link flows and queues. |
| Step 9: | While not all travel demand has been assigned yet, set $t := t+1$ and return to Step 3. |

[1] In INDY, two different dynamic network loading procedures are implemented. The one presented here uses link performance functions in order to propagate the flow over the routes. The other procedure uses a dynamic queuing approach which takes also spillback into account, see also Bliemer (2006). Although this second procedure is able to handle capacity constraints more accurately, it needs significantly more computation time. Hence, we restrict ourselves here to the first procedure, which is applicable for large networks.

A time step of 10 seconds needs 10 times less computation time than a time step of 1 second. For most DTA models the length of the time step should be smaller than the shortest travel time that is needed to traverse any link. MARPLE has certain heuristics implemented to circumvent this restriction, such that larger time steps can be chosen and thus to speed up calculations. In INDY the restriction can be dealt with in two different ways. A first obvious way is to automatically increase the lengths of all links such that the link travel times are equal to the time step set by the user. This may lead to somewhat longer trips, slightly overestimating the travel times. Another way is that INDY ignores congestion on the links with free-flow travel times less than the time step. In this case, the travel times will be slightly underestimated.

## 6 Case study on a large network

In this case study the possibilities of MARPLE and INDY on a large network are illustrated. MARPLE is programmed in Matlab, while INDY is implemented in C++. Both models are able to use the Omnitrans graphical user-interface for creating and editing networks and travel demand matrices, for running the models, and for viewing the (dynamic) output variables on the network. Both models were applied to the road network of the Dutch National Model ("Landelijk Model Systeem", LMS), shown in Figure 2. This road network consists of 345 zones (foreign zones are not included), 25,341 one-directional links and 17,963 nodes. The OD matrix contains 24 hours of travel demand on approximately 109,000 non-zero cells and is divided into 24 matrices with one hour demand.

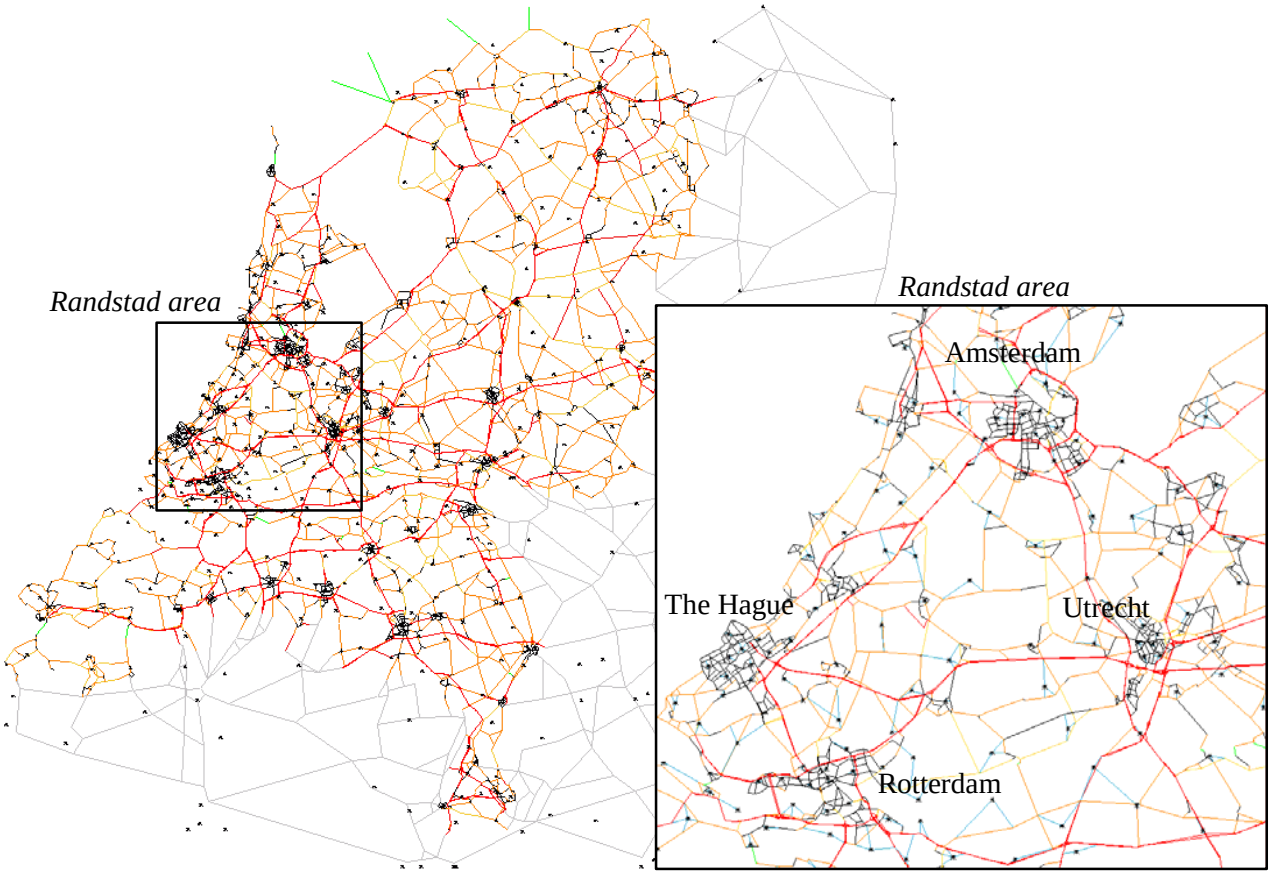



***Figure 2:*** *Road network of the Dutch National Model*

The route generation module was tested by varying the maximum percentage of route overlap, $\gamma$, the maximum number of iterations, $N_1$ and for MARPLE the variance parameter. For the route overlap parameter, percentages of 70%, 80% and 90% were used, in combination with 10 or 25 iterations. For $\sigma = 0.33$ (longest routes is not longer than twice the shortest route), MARPLE generated far less routes than INDY. Therefore, the influence of the variance $\sigma$ was tested by doing the same runs for $N_1 = 10$ with $\sigma = 0.66$ (longest route

is not longer than three times the shortest route). The results, in terms of number of generated paths and calculation times[2], are shown in Table 1.

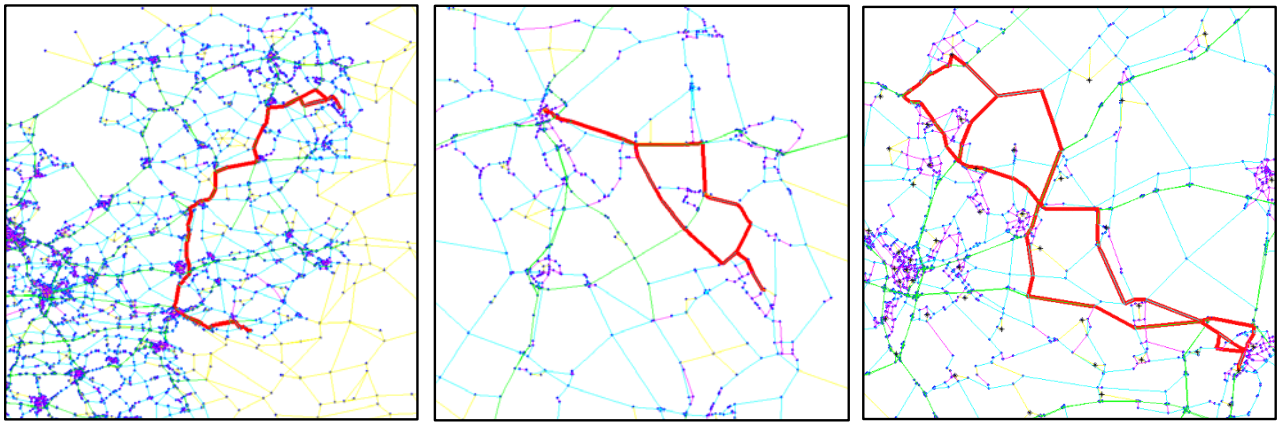
Overall, INDY generates more paths in less time. That INDY generates more paths is due to the increasing variance used in INDY. Choosing $\sigma = 0.66$ gives comparable results for MARPLE and INDY. INDY can also generate a route set using a static assignment. Some results from performing a static assignment (using 25 iterations) are also presented in Table 1. Different loading factors ($L$) of the travel demand were used, where a higher loading factor leads to a more congested network and therefore yields a higher number of used routes.

Some results of the route generation algorithm are shown in Figure 3. This figure illustrates that the number of generated routes for an OD pair is not fixed, but depends on the OD pair and the network. Figure 3(a) shows an OD pair with only two routes that have a large overlap. So, only one route remains in the route set. Figure 3(b) also shows an OD pair with two routes, but these routes are clearly different having a low overlap factor, hence both routes remain in the route set. Finally, the OD pair in Figure 3(c) has many routes.

***Table 1:*** *Results of route generation*

| Monte Carlo | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | $N_1 = 10$ | | | $N_1 = 25$ | |
| | | $\gamma = 0.70$ | $\gamma = 0.80$ | $\gamma = 0.90$ | $\gamma = 0.70$ | $\gamma = 0.80$ | $\gamma = 0.90$ |
| **INDY** | *paths* | 261,324 | 328,372 | 449,134 | 345,278 | 468,984 | 716,733 |
| | *CPU* | ± 4 min. | ± 4 min. | ± 5 min. | ± 12 min. | ± 13 min. | ± 14 min. |
| **MARPLE** | *paths* | 174,507 | 205,744 | 276,810 | 203,462 | 254,292 | 379,240 |
| $\sigma = 0.33$ | *CPU* | ± 18 min. | ± 18 min. | ± 18 min. | ± 43 min. | ± 43 min. | ± 43 min. |
| **MARPLE** | *paths* | 237,901 | 308,447 | 469,642 | | | |
| $\sigma = 0.66$ | *CPU* | ± 21 min. | ± 23 min. | ± 28 min. | | | |

| Static assignment | | | | | |
|---|---|---|---|---|---|
| | | $L = 0.8$ | $L = 1.0$ | $L = 1.2$ | $L = 1.4$ |
| **INDY** | *paths* | 201,292 | 227,541 | 285,257 | 364,179 |
| | *CPU* | ± 5 min. | ± 5 min. | ± 5 min. | ± 6 min. |



(a) (b) (c)

***Figure 3:*** *Example routes generated for different OD pairs*

[2] A Pentium 4 CPU 3.2 Ghz with 2GB of internal RAM is used for all computations.

For the same network an assignment was done with a single iteration to show that it is indeed possible to run a DTA model for such large networks. From the OD matrix about 1% of the demand is removed resulting in about 47,750 OD pairs remaining. Due to this reduction in the OD matrix, also the number of used links in the route set decreases. About 72% of the links and 83% of the nodes are used. The unused links and nodes are omitted from the calculations. The adjusted OD matrix was used in the assignment. The time step in the DNL model is set to 10 seconds for INDY and 30 seconds for MARPLE. The results for the computation time and the internal memory usage are shown in Table 2. For the calculation the length of the time step is taken into account.

***Table 2:*** *Computation time and internal memory usage*

| | INDY | MARPLE |
|---|---|---|
| # OD pairs | 47,628 | 47,965 |
| # routes | 133,169 | 147,042 |
| CPU time route generation | 3 min. | 22 min. |
| CPU time DTA per iteration | 5 hrs. | 1 hr. |
| Internal memory usage | 580 MB | 1223 MB |

From this table it is clear that both models can handle this large network using a reasonable amount of computer resources. Since INDY is implemented in C++ it is more efficient in terms of internal memory usage than the Matlab implementation of MARPLE. MARPLE is faster for the DTA itself. This is probably caused by the differences in the dynamic network loading. MARPLE computes split rates before the dynamic network loading, which speeds up the simulation by making it completely link-based, while INDY remains route-based, explicitly keeping track of the origins and destinations of each link flow. Note that the travel demand period is 24 hrs, while the longest trip in the network takes approximately 4hrs, hence simulating a total time of 28 hrs. Therefore, INDY and MARPLE are simulating traffic through the network 28 and 5.5 times faster than real-time, respectively.

## 7 Discussion and conclusions

In this paper two dynamic traffic assignment models have been tested for a large road network. Both models are the result of PhD research at the Delft University of Technology, but were developed for different purposes and have different characteristics. MARPLE and INDY are both route-based macroscopic DTA models. MARPLE only considers one vehicle type and uses travel time functions, while INDY is completely multiclass and uses speed-density functions.

From the research it can be concluded that the route set generation module of both models is able to determine a route set for large networks, such as the one tested: the network of the Dutch National Model. With some extra simplifications, both models are also able to do a dynamic assignment for this network, using a normal PC with no special additions. This means that using a DTA model for large networks is within reach for all kinds of applications, including planning and traffic management (Taale and Westerman, 2005).

## References

Bliemer, M.C.J. (2001) Analytical Dynamic Traffic Assignment with Interacting User-Classes: Theoretical Advances and Applications Using a Variational Inequality Approach. PhD Thesis, Delft University of Technology, Delft, The Netherlands.

Bliemer, M.C.J. (2005) INDY 2.0 Model Specifications. Delft University of Technology working report.

Bliemer, M.C.J. (2006) Dynamic Queuing and Spillback in an Analytical Dynamic Traffic Assignment Model. *Proceedings of the 1st International Symposium on Dynamic Traffic Assignment*, Leeds, UK.

Bliemer, M.C.J., and Bovy, P.H.L. (2003) Quasi-Variational Inequality Formulation of the Multiclass Dynamic Traffic Assignment Problem. *Transportation Research B*, 37, pp. 501-519.

Bliemer, M.C.J., Versteegt, H.H., and Castenmiller, R.J. (2004) INDY: A New Analytical Multiclass Dynamic Traffic Assignment Model. *Proceedings of the TRISTAN V conference*, Guadeloupe.

Bottom, J. and Chabini, I. (2001) Accelerated Averaging Methods for Fixed Point Problems in Transportation Analysis and Planning. *Proceedings of the TRISTAN IV Conference*, São Miguel, Portugal.

Cascetta, E., Nuzzolo, A., Russo, F. and Vitetta, A. (1996) A Modified Logit Route Choice Model Overcoming Path Overlapping Problems: Specification and Some Calibration Results for Interurban Networks, In: *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, Lyon, France, pp. 697-711.

Hoogendoorn, S., Van Nes, R., and Bovy, P.H.L. (2005) Path Size Modeling in Multimodal Route Choice Analysis. *Transportation Research Record*, 1921, pp. 27-34.

Peeta, S., and Ziliaskopoulos, A.K. (2001) Foundations of Dynamic Traffic Assignment: The Past, the Present and the Future. *Networks and Spatial Economics*, 1(2), pp. 233-265.

Ran, B., and Boyce, D.E. (1996) *Modeling Dynamic Transportation Networks: An Intelligent Transportation System Oriented Approach*. Second edition, Springer-Verlag, Berlin.

Rijkswaterstaat (2003) *Handbook Sustainable Traffic Management*, AVV Transport Research Centre.

Taale, H., Westerman, M., Stoelhorst, H. and Van Amelsfort, D. (2004) Regional and Sustainable Traffic Management in The Netherlands: Methodology and Applications, In: *Proceedings of the European Transport Conference 2004*, October 4-6, 2004, Strasbourg, France, Association for European Transport.

Taale, H. and Westerman, M. (2004) The Application of Sustainable Traffic Management in The Netherlands, In: *Proceedings of the European Transport Conference 2005*, October 3-5, 2005, Strasbourg, France, Association for European Transport.

# TIME AND SPACE DISCRETIZATION IN DYNAMIC TRAFFIC ASSIGNMENT MODELS

*Guido Gentile*, Klaus Nöekel°, Lorenzo Meschini**

*Dipartimento di Idraulica Trasporti e Strade - Università degli Studi di Roma "La Sapienza"
Via Eudossiana 18, 00184 Roma [guido.gentile, lorenzo.meschini]@uniroma1.it
°PTV – Planung Transport Verkehr AG
Stumpfstr.1, D-76131 Karlsruhe Klaus.Noekel@ptv.de

## INTRODUCTION

Dynamic Traffic Assignment is the problem of loading a road network with time varying demand flows from origins to destinations on shortest paths. In order to model this correctly there are several issues that shall be addressed:

1) the presence of vehicle queues due to capacity reductions at intersections (bottlenecks), and the fact that vehicles' speed is a function of the their distance, i.e. of the flow density (fundamental diagram), make the link travel times depend on the link flows (congestion), and since flows are time varying also performances (travel times and costs) are such;
2) queues' spillover on the backward links further reduce their capacity, making the arc performance model non-separable both in time and space;
3) users departing from a given origin at a certain instant will perceive the cost of each link composing the path that they have chosen to reach their destination at the time when they will actually travel on it, which is given by the concatenation of the exit times of the previous links in the sequence;
4) the presence of tolls makes the problem of determining the dynamic shortest paths more complex than that of finding the fastest routes;
5) different perception of link costs by different users requires to introduce random arc cost errors and to model stochastic route choices;
6) high congestion in the peak hour induces users to modify their desired departure time;
7) the combination of congestion with the possibility of choosing the route and the departure time yields an equilibrium problem, where a flow and performance pattern is sought such that no user has an advantage in modifying his travel decisions.

Many of the existing models introduce rough approximations on some of these regards, so that few of them are completely satisfactory in every respect.

The main differences among the approaches to DTA modelling pertain to the fundamental question of discretization, which has deep implications also on the algorithmic side. This is a crucial aspect of the solution to the problem, since it implies a trade-off between accuracy and efficiency. In general, the more the discretization is dense, the more the resulting model is accurate and the more its procedures are time consuming. Indeed, computation efficiency is still a main goal in DTA modelling because, depending on the solution approach and on the dimension of the road network, the simulation can be even slower than the real system. On the other hand, DTA is often involved in real time applications, such as route guidance and

traffic control, and will be more and more utilized in the future for planning purposes where several hours of the day are to be simulated.

We can identify three different levels of discretization:

1) flow discretization – traffic flows are indeed constituted by individual vehicles, which can be modelled as single entities, whose positions can be tracked on the network;

2) space discretization – it is quite natural in traffic assignment to discretize the road network into links, but in order to reproduce traffic conditions in detail some models split each link into segments or "cells";

3) time discretization – most DTA models require a dense time discretization (say 1 second) since they exploit "short time intervals" (the exit time from a link or cell during any time interval falls into the next time interval or later), while few models allow for "long time intervals" (say 10 minutes).

In the last years we have focussed our efforts (Bellei, Gentile and Papola, 2005; Gentile, Meschini and Papola, 2005; Bellei et. al., 2005; Gentile, Meschini and Papola, 2006) to address all the above issues jointly by formalizing within-day DTA in the framework of functional analysis, that is to consider as model variables functions of time instead of vector of numbers. However, when implementing, it is anyway necessary to introduce some sort of time discretization. Our aim is then to reduce discretization as much as possible. Therefore we adopt macroscopic flow models for whole links and long time intervals. On these bases the equilibrium can be formulated as a fixed point problem in terms of link inflow temporal profiles and solved through the MSA or with some other better performing algorithm.

This approach made it possible to achieve a good compromise between accuracy in reproducing the relevant traffic phenomena and efficiency in solving large instances of DTA. The accuracy of the representation is suitable for transport planning and flow or performance pattern estimation, while other methodologies are more appropriate for the detailed analyses needed in traffic signal setting and geometric design of links or intersections. The efficiency of the computation is suitable for simulating with a standard PC the traffic of huge urban and regional networks during several hours. Indeed, the complexity of the model is that of the corresponding static assignment multiplied by the number of (long) time intervals introduced; which makes real time prediction of traffic conditions for large cities at hand.

In this paper we will discuss the role of discretization in DTA, by presenting a classification of the existing approaches from this perspective. Then we will outline our model, with the aim of placing it into the above classification and supporting the soundness of its mathematical formulation, underlying assumptions and algorithmic features.

DTA models have two fundamental components: the *network loading* model, which propagates all user trips on the current route choice pattern yielding an arc flow pattern consistently with the resulting travel time pattern, where queuing and traffic congestion must be taken into account; b) the *route choice* model, which reproduces the behaviour of a driver travelling from an origin to a destination on the road network in terms of path (and departure time) decisions, that are taken on the basis of the performance pattern. These two sub-models are somehow iterated in sequence until an equilibrium is reached.

## THE NETWORK LOADING MODELS

Time discretization can be generally of two types: *short intervals* (a few seconds), *long intervals* (some minutes). Also the space discretization can be of two types: *short arcs* (each road link is divided into segments of 10-100 meters), *long arcs* (each link between two intersections is considered as a whole). In the following we will see which models can adopt the "long features, that are more convenient from a computational point of view.

But the main criteria to distinguish among network loading models is probably flow discretization. The specific purpose of each methodology (from intersection geometry design, to traffic signal setting, until transport planning) legitimates a great variety of approaches on this respect: from *micro-simulation*, where the behaviour of each single vehicle (essentially, longitudinal and lateral acceleration) is connected with the relative speed and position of the adjacent vehicles through car following and lane changing models; to *meso-simulation*, where groups of vehicles (in case, a single vehicle) are handled as dense point packets, while the interaction among vehicles is reproduced though macroscopic flow models; until *macro-simulation*, where the continuous flow paradigm is fully accepted. To be noticed that the concept of meso-simulation is somewhat vague in the literature, and the definition given above is merely our interpretation of this idea.

There are two types of micro-simulation: *time-based*, which leads usually to the assumption that a vehicle can traverse at the most one node during a single clock, thus involving also for simple models short time intervals, especially on urban networks; *event-based*, which does not require to move each vehicle at every clock, but only when a relevant circumstance takes place (such as: a vehicle has reached the head of the link if no queue is present, a new slot becomes available at the tail of a link in spillback, ecc.) and other future events are produced. Although the latter approach may be somewhat more complicated to implement, if a simple model with few types of events under control is considered, it may lead to relevant savings of computing resources.

Micro and meso simulation models are both affected by pseudo-random processes to generate vehicles from centroids, and to assign them to a specific path or to divert them to a specific link at each node, in case the routing is performed according to splitting rates instead of paths. This implies that the expected traffic flow pattern could be actually retrieved only as an average among many simulation runs, which is almost never the case, because each run is nowadays too expensive in terms of computing resources to be replicated, so that the output of the model is indeed the result of a multivariate random variable. However, one could observe that traffic flows on road links are usually made up by hundreds of vehicles, so that the variances of the above variables should be relatively small.

Micro simulation (Barcello and Casas, 2002) usually requires lots of data which are often not readily available, while meso and macro simulation do not require more data than static traffic assignment – besides the modulation of the demand in time – which is a great advantage from the applicative point of view. Moreover, with respect to static traffic assignment DTA models do not require to calibrate the parameters of the volume-delay functions (e.g. alpha and beta of the BPR). The heavy data requirements of micro simulation lead to the opportunity of meso simulation, which still moves single vehicles but according to the

much simpler macroscopic flow models. Two examples of this approach are Mahmassani (2001) and Mahut *et al.* (2004), respectively adopting a time based and an event based mechanism.

Meso and macro simulation models consider typically two distinct flow conditions: hypocritical and hypercritical. Hypocritical flow conditions are usually reproduced according to the *Theory of Kinematic Waves* of Lightill, Whitham and Richards, which requires short arcs, or its first order approximation, called the Simplified TWK (Newell, 1993). Hypercritical flow conditions are usually reproduced according to *deterministic queue theory* (Arnott De Palma and Lindsey, 1990; Ghali and Smith, 1993), where often the concept of storage capacity is introduced to comply with the spillback phenomenon.

Many DTA models (e.g. Astarita, 1996; Ran *et al.*, 1997; Friesz *et al.*, 1993; Tong and Wong, 2000; Papageorgiou, 1990; Carey, Ge and McCartney, 2003) evaluate instead travel times through *volume-delay function* that consider the speed corresponding on the fundamental diagram to the current average density of the link. This approach appears convenient from a computational point of view, since it doesn't require space discretization and leaves a certain freedom in choosing the time discretization. However, it can be easily shown that any model which is separable with respect to time will reproduce satisfactorily neither the formation and dispersion of vehicle queues, nor the interaction of vehicles under hypocritical conditions.

Much better results are obtained applying *space continuous models* (e.g. METANET, Messmer and Papageorgiou, 1990, which derives from a second order approximation to the TKW; the Cell Transmission Model, Daganzo, 1994, 1995a, which is consistent with the first order approximation), that are formulated through differential equations in time and space and solved through finite difference methods. Such models yield accurate results, but their numerical solution relies on a dense space and time discretization, thus requiring considerable computing resources. A suitable compromise between accuracy and efficiency can be obtained by revisiting the STKW in terms of cumulative flows (Daganzo, 1997), which allows for preserving a whole link approach by implicitly taking into account the propagation of flow states along the link. This model, originally developed for urban arcs with a triangular fundamental diagram, has been recently extended to represent extra urban arcs with any concave fundamental diagram (Gentile, Meschini and Papola, 2005) and to address the case of spillback congestion (Gentile, Meschini and Papola, 2006).

Finally we present a rough discussion about complexity. We take as reference the road network of Rome, characterized by the following dimensions for a simulation period of 3 hours during the morning peak:

| | | |
|---|---|---|
| nZ | number of centroids | 500 |
| nA | number of link | 10.000 |
| nT | average number of trips travelling at the same time | 500.000 |
| nC | average number of cells for each link | 10 |
| nIL | number of cells for each link | 10 |
| nIS | number of short time intervals (DTS = 6 sec) | 1800 |

Each network loading costs the following number of operations, depending on the approach:

| | | |
|---|---|---|
| micro simulation | a1 * nT * nIS = | a1 * 900.000.000 |
| meso simulation | a2 * nT * nIS = | a2 * 900.000.000 |
| macro simulation | a3 * nIL * nA * nZ = | a3 * 150.000.000 |

| | | |
|---|---|---|
| CTM without destinations | a4 * nIS * nA * nC = | a4 * 180.000.000 |
| CTM with destinations | a5 * nIS * nA * nZ = | a5 * 9.000.000.000 |

Clearly micro simulation requires a much higher number of operations to move each vehicle with respect to meso simulation, so that a1 >> a2. Macro simulation requires just few operations to propagate the flow forward (network flow propagation should be iterated to compute the dynamic network loading, but this can be avoided leaving the task to the overall equilibrium procedure as illustrated later), so that we can reasonably state a difference of one order of magnitude between meso and macro for a large network. Clearly for smaller networks this difference is higher. Moreover DTL can go up to say 15 min, while DTS can go down to 1 sec, so that the above difference can go up to two order of magnitudes.

## THE ROUTE CHOICE MODELS

Vehicles may divert at nodes: either following a given path; or on the successive link given by the solution of the dynamic shortest path problem; or consistently with some splitting rates, specific of each destination and time varying. In connection to this, the flows resulting from the successive iterations of the network loading are to be averaged. One possibility is to consider path explicitly, and assign the demand to the resulting proportions. A second possibility is to follow an implicit path enumeration approach, that is averaging directly the link flows, and assign the demand all or nothing to shortest paths. A variant of this approach is to average the splitting rates, typically when path choice is stochastic. In any case, the dynamic shortest path problem lies at the heart of within-day dynamic traffic assignment.

We now focus on the route choice model in presence of time-varying arc costs and travel times, but without the possibility of waiting at nodes, under the assumptions of perfect information and rational behaviour of the drivers. This leads to deterministic models where drivers travel only on dynamic shortest paths, i.e., paths with minimum cost. The cost of a path is given by the sum of the costs of its arcs, each one evaluated at the time when the user travelling along the path enters it, while the travel time of a path is given by the concatenation of the exit times of its arcs. As usual in transportation applications, we consider the many-to-one version of the dynamic shortest path problem, since the main concern of the user is to reach his destination.

Most of the literature address the discrete case (Pallottino and Scutellà, 1998), where it is assumed that the time varies in a finite set of values and that the exit time from each arc evaluated at one of these instants belongs to the set. In this case, the dynamic shortest path problem can be studied by introducing the so-called *space-time network*, or *diachronic graph*, where each node of the original network is replicated into a vertex for every instant, and an edge connects the vertex relative to the initial node of each arc at every instant with the vertex of its final node at the corresponding exit time. The space-time network has no cycles, and any chronological ordering of its vertices provides a topological ordering. Exploiting this property, it is possible to solve efficiently the dynamic shortest path problem, by performing reverse chronological visits of the space-time network. For more details about discrete formulations of the dynamic shortest path problem in transportation networks, see De Palma, Hansen and Labbé (1993) and Nachtigall (1995).

The shortest paths may contain cycles, since it may be convenient to wait around on the network with the aim of traveling on certain arcs later at a lesser cost. Moreover, in order to find the dynamic shortest paths arriving to a given destination at a particular instant from each origin it is necessary to handle multiple cost labels for each node, one for every instant. The most efficient algorithms to solve this problem introduce a list of vertices to be examined with a bucket (Dial, 1969) for each instant. A vertex is extracted iteratively from the latest non-empty bucket, so that each vertex of its backward star is updated in terms of shortest path solution and inserted in the proper bucket, if its label can be improved through that edge. All labels are initialized to infinity except the label of the origin vertex, which is initialized to zero and inserted in an empty list.

A shortest path is said to satisfy the *concatenation property* if its two sub-paths from the origin to any intermediate node and from this node to the destination are themselves shortest paths. In general, this does not hold in the dynamic context. However, if the FIFO rule is satisfied for each arc, i.e., no overtaking can occur on road links, then the concatenation property holds true with respect to travel times (Ziliaskopoulos, 1994). Moreover, if the arc costs are proportional to the travel times, this can be extended to costs, thus inducing to study the deterministic model where drivers travel only on dynamic fastest paths, i.e., paths with minimum travel time (Dreyfus, 1969; Cooke and Halsey, 1966; Ziliaskopoulos and Mahmassani, 1993).

Based on the FIFO rule, the fastest paths cannot contain cycles. Moreover, any classical algorithm with only one label for each node can be applied in order to find on the space-time network the fastest *trajectories* arriving to a given destination at a particular instant from each origin (Kaufmann and Smith, 1993), thus solving the dynamic fastest path problem. In this case the bucket list paradigm, applied to the original nodes instead that to the vertices of the space-time network, can be adopted to implement efficiently the label setting approach.

Often is the applications to DTA the dynamic shortest path problem is to be solved for every arriving time. For this case, Chabini (1998) proposed an algorithm that addresses all the temporal instances jointly, and finds the next arc of the trip from any given node at each particular instant in order to reach the destination at a later time with the minimum cost. This is simply achieved by comparing, among each arc exiting that node, the sum between the arc cost at that instant and the minimum cost from the final node of the arc to the destination, evaluated at the corresponding exit time from the arc. The above approach underlies a *temporal-layer* structure of the solution rather than a trajectory structure, since the temporal perspective is local.

The latter algorithm eliminates by its nature the need to handle multiple cost labels and, based on the acyclic property of the time-space network, requires a simple visit of all its vertices in reverse chronological order, while no particular order is to be maintained when processing the vertices of a same temporal layer, which leads to an optimal running time. The only drawback of the above approach is common to all discrete models based on the paradigm of the diachronic graph. This is the need to introduce a great number of time intervals having a short length in order to reproduce the arc travel times with the required accuracy, since these are multiples of such a length (usually 1 second, in urban contexts), which results in time-consuming procedures.

Continuous models, which do not introduce the space-time network but operate directly on variables that are functions of time, have been analyzed by Orda and Rom (1990-1991). However their exact solution approach seams to be unfeasible for large networks. We have explored the possibility of allowing some approximations when operating on the temporal profiles in the algorithm, with the aim of achieving fast heuristics. Specifically, we dealt with temporal profiles that are piecewise linear functions of time defined on a pre-specified set of instants, and keep their form during any operation accepting the necessary distortions. With this aim the trajectory approach has been simply extended, while the temporal-layer approach has been more significantly tailored, to the case where costs and travel times are functions of time having a continuous domain and range. Indeed, to avoid a dense time discretization in the temporal-layer approach, we have renounced to the acyclic property of the space-time network, by letting the exit time from any given arc at the beginning of a particular interval to fall within that same interval. This leads to the possibility of cycles in the solution relative to a same temporal layer and therefore introduces the need to make some approximation. Specifically, we force a label setting strategy and implement it through a bucket list.

In both approaches, the reference instants on which basis the piecewise linear temporal profiles are defined can then identify time intervals of several minutes. This, on one side, allows to capture the features of road traffic dynamic that are relevant for most planning purposes, and on the other side, results in much shorter computational times. The more the time intervals are short, the more the proposed continuous models tend to coincide with the corresponding discrete models. This property has been exploited in the validation of the proposed heuristics both in terms of results and computing times.

Finally we present a rough discussion about complexity, taking as reference the case of Rome:

| | | |
|---|---|---|
| discrete shortest trajectories | nA * (nIS) * nIL * nZ = | 270.000.000.000 |
| discrete fastest trajectories | nA * nIL * nZ = | 150.000.000 |
| discrete shortest temporal layer | nA * nIS * nZ = | 9.000.000.000 |
| continuous fastest trajectories | nA * nIL * nZ = | 150.000.000 |
| continuous shortest temporal layer | nA * nIL * nZ = | 150.000.000 |

## OUTLINE OF THE PROPOSED MODEL

Recently, we proposed a new model for within-day DTA on road networks, where a user equilibrium is expressed as a fixed point problem in terms of arc flow temporal profiles. The model integrates spillback congestion into an existing formulation of DTA based on continuous-time variables and implicit path enumeration, which is capable of representing explicitly the formation and dispersion of vehicle queues on road links, but allows them to overcome the arc length.

By extending to the dynamic case the concept of Network Loading Map (NLM), stated in Cantarella (1997) for the static case, it is no more needed to introduce the CDNL (Xu et. al., 1999) as a sub-problem of the DTA, because the coherence through the arc performance model between the travel times and the flows loaded on the network consistently to given path choices can be attained jointly with the equilibrium.

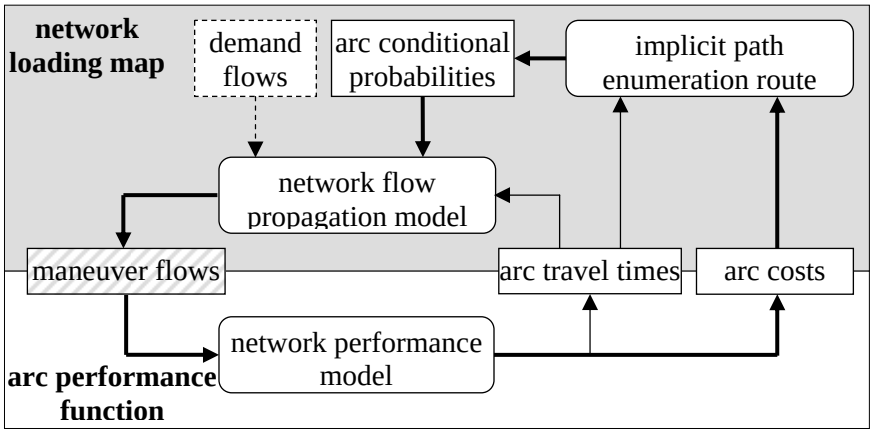



Figure 1. Scheme of the fixed point formulation for the DTA.

The propagation of congestion among adjacent arcs is achieved through the introduction of time-varying exit capacities which limit the inflow on downstream arcs in such a way that their storage capacities are never exceeded. This approach is consistent with the definition provided by Adamo *et al.* (1999), where the spillback phenomenon is characterized as a hypercritical flow state, either propagating backwards from the final section of an arc and reaching its initial section, or originating on the latter, that reduces the capacities of the arcs belonging to its backward star and this way influences their flow states. To determine the temporal profile of these exit capacities requires solving, on the supply side of the DTA, a system of spatially non-separable macroscopic flow models based on the theory of kinematic waves, which describe the dynamic of the spillback phenomenon and yield consistent network performances for given arc flows.

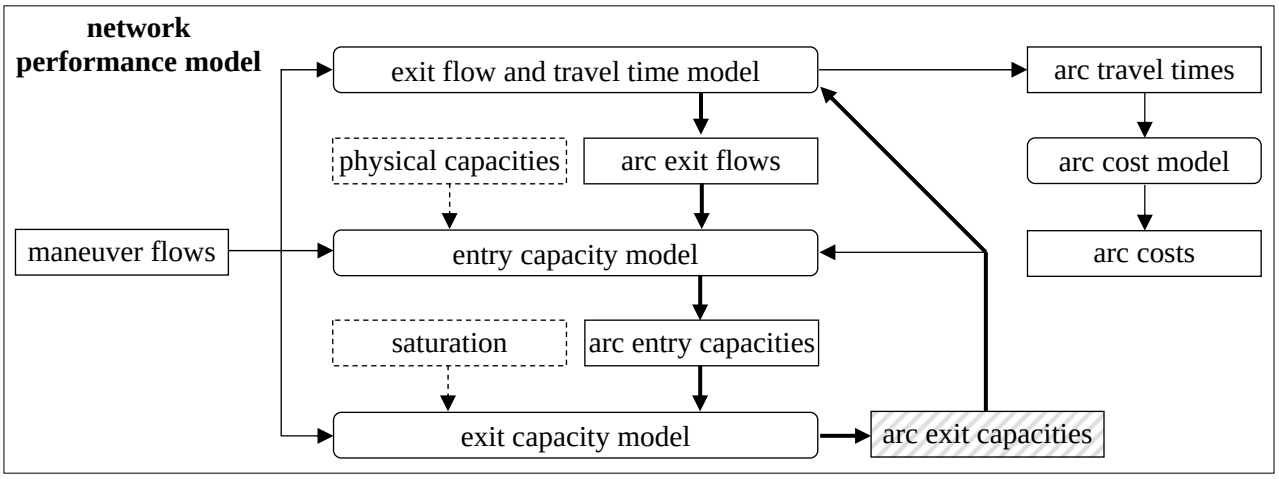



Figure 2. Scheme of the fixed point formulation for the NPM to simulate spillback.

## CONCLUSION

Combining these results, we have been capable of proposing a macroscopic DTA model allowing for long intervals and long arcs. This model, called DUE, has been implemented in an existing commercial software for transport analyses (Visum – by PTV) and will then soon be available to many practitioners.

**REFERENCES**


1. Adamo V., Astarita V., Florian M., Mahut M., Wu J.H. (1999) “Modelling the spill-back of congestion in link based dynamic network loading models: a simulation model with application”, in Proceedings of the 14th International Symposium on Transportation and Traffic Theory, ed. A. Ceder, Elsevier Science, Amsterdam, The Netherlands.

2. Arnott R., De Palma A., Lindsey R. (1990) “Departure time and route choice for the morning commute”, Transportation Research B 24, 209-228.

3. Astarita V. (1996) “A continuous time link model for dynamic network loading based on travel time function”, in Proceedings of the 13th International Symposium on the Theory of Traffic Flow, Lyon, 87-102.

4. Barceló J., Casas J. (2002) “Heuristic dynamic assignment based on microscopic traffic simulation”, in Proceedings of the 9th Meeting of the Euro Working Group on Transportation, Bari, Italy.

5. Bellei G., Gentile G., Papola N. (2005) “A within-day dynamic traffic assignment model for urban road networks”, Transportation Research B 39, 1-29.

6. Bellei G., Gentile G., Meschini L. and Papola N. (2005) “A demand model with departure time choice for within-day dynamic traffic assignment”, European Journal of Operational Research, in press - corrected proof - available online.

7. Carey M., Ge Y. and McCartney M. (2003) “A whole-link travel-time model with desirable properties”, Transportation Science 37, 83-96.

8. Chabini I. (1998) “Discrete dynamic shortest path problems in transportation applications: complexity and algorithms with optimal run time”, Transportation Research Record 1645, 170-175.

9. Cooke L.L. and Halsey E. (1966) “The shortest route through a network with time-dependent intermodal transit times”, Journal of Mathematical Analysis and Applications 14, 492-498.

10. Daganzo C.F. (1994) “The cell transmission model: a dynamic representation of highway traffic consistent with hydrodynamic theory”, Transportation Research B 28, 269-287.

11. Daganzo C.F. (1995a) “The cell transmission model, part II: network traffic”, Transportation Research B 29, 79-93.

12. Daganzo C.F. (1997) “Fundamentals of transportation and traffic operations”, Pergamon, Oxford, UK, chapter 4.

13. De Palma A., Hansen P. And Labbé M. (1993) “Commuters’ paths with penalties for early or late arrival times”, Transportation Science 24, 276-286.

14. Dial R.B. (1969) “Algorithm 360: shortest path forest with topology ordering”, Communications of the A.C.M. 12, 632-633.

15. Dreyfus S.E. (1969) “An appraisal of some shortest-path algorithms”, Operations Research 17, 395-412.

16. Friesz T.L., Bernstein D., Smith T.E., Tobin R.L., Wie B.W. (1993) “A variational inequality formulation of the dynamic network user equilibrium problem”, Operations Research 41, 179-191.

17. Gentile G., Meschini L., Papola N. (2005) “Macroscopic arc performance models with capacity constraints for within-day dynamic traffic assignment”, Transportation Research B 39, 319-338.

18. Gentile G., Meschini L., Papola N. (2006) “Spillback congestion in dynamic traffic assignment: a macroscopic flow model with time-varying bottlenecks”, submitted to Transportation Research B.

19. Ghali M.O., Smith M.J. (1993) “Traffic assignment, traffic control and road pricing”, in Transportation and Traffic Theory, ed. C.F. Daganzo, Elsevier, Amsterdam, The Netherlands, 147-169.

20. Kaufman, D.E., Smith, R.L. (1993) “Fastest Path in time dependent networks for intelligent vehicle-highway system applications”, IVHS Journal 1, 1-11.

21. Orda A. and Rom R. (1990) “Shortest-path and minimum-delay algorithms in network with time-dependent edge length”, Journal of the ACM 37, 607-625.

22. Orda A. and Rom R. (1991) “Minimum weight paths in time-dependent network”, Networks 21, 295-

320.

23. Mahmassani H. (2001) “Dynamic network traffic assignment and simulation methodology for advanced system management applications”, Networks and Spatial Economics 1, 267-292.

24. Mahut M., Florian M., Tremblay N., Campbell M., Patman D., Krnic McDaniel. Z. (2004) “Calibration and application of a simulation-based dynamic traffic assignment model”, presented at the TRB 2004 Annual Meeting, Washington D.C.

25. Messmer A., Papageorgiou M. (1990) “METANET: a macroscopic simulation program for motorway networks”, Traffic Engineering & Control 31, 466-470.

26. Nachtigall K. (1995) “Time-depending shortest path problems with applications to railway networks”, European Journal of Operational Research 83, 154-166.

27. Newell G.F. (1993) “A simplified theory of kinematic waves in highway traffic, part I: general theory; part II: queuing at freeway bottlenecks; part III: multi-destination flows”, Transportation Research B 27, 281-313.

28. Pallottino S., Scutellà M.G. (1998) “Shortest path algorithms in transportation models: classical and innovative aspects”, in Equilibrium and advanced transportation modelling, ed.s P. Marcotte, S. Nguyen, Kluwer Academic Publisher, Dordrecht, The Netherlands, 245-281.

29. Papageorgiou M. (1990) “Dynamic modelling, assignment and route guidance in traffic networks” Transportation Research B 24, 471-495.

30. Ran B., Rouphail N.M., Tarko A., Boyce D.E. (1997) “Toward a class of link travel time functions for dynamic assignment models on signalised networks”, Transportation Research B 31, 277-290.

31. Tong C.O., Wong S.C. (2000) “A predictive dynamic traffic assignment model in congested capacity-constrained road networks”, Transportation Research B 34, 625-644.

32. Xu Y.W., Wu J.H., Florian M., Marcotte P., Zhu L.H. (1999) “Advances in the continuous dynamic network loading problem”, Transportation Science 33, 341-353.

33. Ziliaskopoulos A.K. (1994) “Optimum path algorithms on multidimensional networks: analysis, design, implementation and computational experience”, Ph.D. dissertation, University of Texas at Austin.

34. Ziliaskopoulos A.K. and Mahmassani H.S. (1993) “Time-dependent shortest path algorithm for real-time intelligent vehicle highway system applications”, Transportation Research Record 1408, 94-100.

# A SIMULATION-BASED DYNAMIC TRAFFIC ASSIGNMENT MODEL: DYNAMEQ

Michael Florian: CRT, University of Montreal and INRO Montreal, QC, Canada
Michael Mahut: CRT, University of Montreal and INRO Montreal, QC, Canada
Nicolas Tremblay: CRT, University of Montreal and INRO Montreal, QC, Canada

## ABSTRACT

This paper describes the model, solution algorithm and some selected applications of a simulation-based dynamic traffic assignment model based on equilibrium concepts. The model is formulated as a variational inequality and a solution algorithm is developed for a discretized version of the model. It is most useful, in its current form, for off-line applications, such as the testing of control strategies and scenario analyses in planning studies that must consider temporal flow variations and are not well served by the application of static models. The determination of time-dependent path input flows is modeled as a master problem inspired from a simplicial decomposition approach. The determination of path travel times for a given set of path flows is the network-loading sub-problem, which is solved using the simplified traffic simulation approach of Mahut. The reported applications of this model include references to studies carried out in several cities around the world and some detailed model calibration results for Montreal. The results show that this model is applicable to medium-size networks with very reasonable computing times

***Index Terms*** **– dynamic traffic assignment, method of successive averages, traffic simulation, queuing models**

## 1. INTRODUCTION

This paper provides a summary description and some selected applications of an equilibrium based dynamic traffic assignment model. In order to provide a common terminology to the various models, it is convenient to refer to the main components of any dynamic traffic model: the route-choice mechanism, the determination of the path input flows and the network-loading mechanism. The latter is the method used to represent the evolution of the traffic flow over the links of the network once the route choice and the path input flows have been determined.

Some of the most popular dynamic traffic models today are those based on the representation of the behavior of each driver regarding car following, gap acceptance and lane choice. These are micro-simulation models.
The successful use of micro-simulations is commonly limited to relatively small size networks. Their application has been hindered for medium-to-large networks by the relatively high computation time and the effort required for a proper model calibration. Nevertheless, micro-simulation models are popular and their use is enhanced by traffic animation graphics that capture the attention of non-technical staff. Some of these are micro-simulation models are CORSIM (http://www.fhwa-tsis.com/corsim_page.htm), INTEGRATION (Van Aerde, 1999), AIMSUN2 (Barceló et al, 1994) (http://www.tss-bcn.com), VISSIM (http://www.ptv.de), PARAMICS (http://www.quadstone.com) and DRACULA (http://www.its.leeds.ac.uk/software/dracula/). MITSIM (Yang, 1997) (http://web.mit.edu/its/products.html) is an academic research model that has been used in several studies in Boston, Stockholm and elsewhere.

The aim of handling larger networks with reasonable computational times has led to the development of so-called "mesoscopic" approaches to traffic simulation. The aim is to obtain a traffic representation that still captures the basic temporal congestion phenomena, but models the traffic dynamics with less fidelity. Some of the earliest examples of such an approach are CONTRAM (Leonard et. al., 1989) (www.contram.com) and SATURN (Dow and Van Vliet, 1979) which are commercially available packages that are used in England and elsewhere.
Recently, the development of mesoscopic simulation models for off-line dynamic traffic assignment has become an area of significant research activity, as witnessed by the United States FHWA Dynamic Traffic Assignment Project. (http://www.dynamictrafficassignment.org). The development of DYNASMART (Mahmassani et al., 2001) and DYNAMIT (Ben-Akiva et al., 1998) (http://web.mit.edu/) are two significant developments. Another approach to the network loading algorithm is that based on cellular automata theory (Nagel and Schreckenberg, 1992), which has been implemented in the TRANSIMS software (http://transims.tsasa.lanl.gov), developed by the Los Alamos National Laboratories in the USA. The advance of vehicles is carried out by using local rules for each vehicle that determine the next cell to be occupied and the speed of the vehicle. Ziliaskopoulos and Lee (1997) developed a dynamic traffic assignment model based on the cell transmission model which is a particular solution method for the classical macroscopic traffic flow model. Another macroscopic flow based model is that of Gentile et al (2005).

Other dynamic traffic assignment models that have their roots in macroscopic traffic flow theory (Lighthill and Whitham, 1955) (Richards, 1956) such as METACOR (Diakakis and Papageorgiou, 1996) and METANET (Messmer

et al., 2000a), These models are based on the work of Papageorgiou (1990) and Messmer et al., 2000b respectively. The route choice in METANET is carried out iteratively by splitting proportions at nodes of the network, with the restriction that only two arcs can originate at a given node. The network loading method is based on a second order (p.d.e.) traffic flow model.

Another line of research is that of analytical dynamic traffic assignment models, which have their roots in the mathematical programming and optimal control approaches. Their aim is to introduce temporal dimensions to extensions of static network equilibrium models which are based on link travel time functions. There is a very large body of literature that contains academic contributions made by using this approach. A good recent reference is a special issue of *Networks and Spatial Economics* ( Vol 1, Issue 3/4 , 2001). The paper by Friesz et al (1993) provides a formulation the equilibrium dynamic traffic model which serves as the basis for the algorithm developed here.

The network loading method presented in this paper is based on a traffic simulation model that was designed to produce reasonably accurate results with a **minimum** number of parameters and a **minimum** of computational effort (Mahut, 2000, Astarita et al., 2001). The underlying structure of the network model and the car mover have more in common with microscopic than with mesoscopic approaches, as the model is designed to capture the effects of car following, lane changing and gap acceptance.

The paper is structured as follows. The next section is dedicated to the exposition of the network and demand representation used in the model; then the formal statement of the model and its discretized version are given. The next sections are dedicated to the statement of the algorithm and to the description of the network loading method. Applications of the model are then given and some conclusions end the paper.

## 2. NETWORK REPRESENTATION, TRAFFIC CONTROL AND DEMAND

The network definition required for this DTA model requires somewhat more information than that required for static network equilibrium models, yet somewhat less than is generally required for micro-simulation traffic models.

Since the underlying traffic model moves individual vehicles on discrete lanes, each link must be defined by a number of lanes. Each lane furthermore is defined by an access code that determines which classes of vehicles may use the lane (e.g., taxi, bus, HOV, etc…). A length and speed limit furthermore define each link. At each node (intersection) of the network, a turn is defined for each permitted movement from an incoming link to an outgoing link. Each turn is defined by an access code and a saturation flow rate per lane. Unlike micro-simulation models, the network definition does not require geometrical information such as lane width, turning angles, and the dimensions of intersections.

The model also permits the specification of detailed traffic control information such as (pre-timed) signal timing and ramp metering plans. Traffic control specifications furthermore require the number of lanes associated with each turning movement, and the lanes (on both the upstream and downstream links) that may be used for executing a turn. These data may vary with the signal phase rather than being fixed for each turn.

Vehicle attributes (or parameters) can be broken down into two distinct categories: physical attributes and routing attributes. The physical attributes are the effective length (based on vehicle spacing at jam density), and the driver/vehicle response time. Together, these parameters yield the jam density and negative wave speed associated with each vehicle class.

Routing attributes include the vehicle class identifier, which determines which lanes and turns of the network may be used by the class, and identifies any class-based routing strategy that may be defined. For instance, the class car uses different routing rules than the class bus, which travels along fixed itineraries and has mandatory stops. A demand matrix by class contains the flow in vehicles per hour for each origin-destination pair. The matrices are “time-sliced” in the sense that flow rates may be specified for given time intervals. The demand is considered to be fixed in this version of the model.

## 3. DYNAMIC TRAFFIC ASSIGNMENT – THE MODEL

Two different approaches are commonly used to emulate the path choice behavior of drivers: dynamic assignment *en route* and dynamic *equilibrium* assignment. In this work, the approach taken is to seek an approximate solution to the dynamic equilibrium conditions. Extensions of the model for *en route* assignment
are indicated at the conclusion of the paper.

In the equilibrium assignment problem, only pre-trip path choices are considered. The path choices are modeled as decision variables governed by a user optimal principle where each driver seeks to minimize the used path travel time. All drivers have perfect access to information, which consists of the travel times on all paths (used and unused). The solution algorithm takes the form of an iterative procedure designed to converge to these conditions.

The mathematical statement of the dynamic equilibrium problem is in the space of path flows $h_k(t)$, for all paths *k* belonging to the set $K_i$ for an origin-destination $i \in I$, at time *t*. The time-varying demand rates are denoted $g_i(t)$. The path flow rates in the feasible region $\Omega$ satisfy the conservation of flow and non-negativity constraints for $t \in T_d$, where (*0, $T_d$*) is the period during which the temporal demand is defined. That is

$$\Omega = h(t) : \sum_{k \in K_i} h_k(t) = g_i(t), i \in I; h_k(t) \geq 0 \quad \text{for almost all } t \in T_d. \qquad (1)$$

The definition of user optimal dynamic equilibrium is given by the temporal version of the static (Wardrop) user optimal equilibrium conditions, which are:

$$\begin{aligned} &s_k(t) = u_i(t) \ \textit{if} \ h_k(t) > 0 \\ &s_k(t) \geq u_i(t) \ \textit{otherwise} \end{aligned} \quad \text{for all } k \in K_i, i \in I, \textit{ for almost all } t \in T_d \qquad (2)$$

where: $h_k \in \Omega, u_i(t) = \min_{k \in K_i} \{s_k(t)\}$ for almost all $t \in T_d$ and $s_k(t)$ is the path travel time of path *k*. Friesz et al (1993) showed that these conditions are equivalent to a variational inequality problem, which is to find $h^* \in \Omega$ such that

$$\left(S(h^*), h - h^*\right) \geq 0, \forall h \in \Omega \qquad (3)$$

The solution approach adopted for solving the dynamic network equilibrium model (1)-(3) is based on a temporal discretization into time periods $\tau = 1, 2, ..., \left|\frac{T_d}{\Delta t}\right|$, where $\Delta t$ is the chosen duration of a time interval. This results in the time discrete model:

$$\begin{aligned} &s_k^\tau = u_i^\tau \ \textit{if} \ h_k^\tau > 0 \\ &s_k^\tau \geq u_i^\tau \ \textit{if} \ h_k^\tau = 0 \end{aligned} \quad \text{for all } k \in k_i, \ i \in I, \tau = 1, 2, ..., \left|\frac{T_d}{\Delta t}\right| \qquad (4)$$

where the feasible set of time dependent flows $h_k^\tau$ belong to

$$\Omega^\tau = h_k^\tau : \sum_{k \in K_i} h_k^\tau = g_i^\tau, i \in I, \textit{all } \tau; \qquad (5)$$

$$h_k^\tau \geq 0, k_i, i \in I, \textit{all } \tau,$$

which can be shown to be equivalent to solving the discretized variational inequality.

$$\sum_\tau \sum_{k \in K} s_k^\tau(h^\tau) \left(h_k^\tau - h_k^\tau\right) \geq 0 \qquad (6)$$

where $K = \bigcup_{i \in I} k_i$ where $h^\tau$ is the vector of path flows $(h_k^\tau)$ for all *k* and τ.

The demonstration of existence and uniqueness of a solution to this model depends on the properties of the mapping ***s(h[g]),*** that is the dependence of link and path travel times on the path input flows and the dependence of the path input flows on the demands. Since the properties of this mapping are not easily verified due to the fact that it is the output of a simulation model and not an analytical transformation, no claims are made about the existence or the uniqueness of a solution. The equilibrium principle is used as a **guide** in computing an approximate solution of the time discrete variational inequality.

## 4. THE DTA ALGORITHM

The solution algorithm used here consists of two main components other than the computation of the temporal shortest paths: a method to determine a new set of time-dependent path input flows, given the experienced path travel times at the previous iteration, and a method to determine the actual link flows and travel times that result from a given set of path inflow rates.. The algorithm furthermore requires a set of initial path flows. The general structure of the algorithm is shown schematically in Figure 1.

The path input flows $h_k^\tau, k \in K$ are determined by a variant of the method of successive averages (MSA), which is applied to each O-D pair *I* and time interval $\tau$ .An initial feasible solution is computed by assigning the demand for each time period to a set of successive shortest dynamic paths. Starting at the second iteration, and up to a pre-specified maximum number of iterations, *N*, the time-dependent link travel times after each loading are used to determine a new set of dynamic shortest paths that are added to the current set of paths.
At iteration $n, n \le N$ , the volume assigned as input flow to each path in the set is $g_i^\tau / n$ , $i \in I$, *all* $\tau$ . After that, for iteration $m, m > N$ , only the shortest among *used* paths is identified and the path input flow rates are redistributed over the known paths. The MSA step size may be determined by various rules. In the statement of the algorithm below the step size chosen is *1/n.*

If the flow of a particular path decreases below a small predetermined value then the path is dropped and its remaining flow is distributed to the other used paths. This heuristic approach is akin to the restricted simplicial decomposition algorithm of Lawgphonpanich and Hearn (1987) for the solution of the static network equilibrium model with fixed demand. The algorithm is summarized below:

EQUILIBRIUM DTA ALGORITHM

- *Step 0* *Initialization (l=1):*
  Compute temporal shortest paths based on free-flow travel times;
  load the demands to obtain an initial solution; *l=l+1*
- *Step 1* *Reallocation of input flows to paths:*
  *Step 1.1* If $l \le N$
  Compute a new dynamic shortest path;
  assign to each path *k* the input flow $g_i^\tau / l$
  *Step 1.2* If $l > N$
  Identify the shortest among used paths;
  redistribute the flows as follows:

$$h_k^l(\tau) = \begin{cases} h_k^{l-1}(\tau)\left(\frac{l-1}{l}\right) + \frac{g_i(\tau)}{l}, & \text{if } s_k^l(\tau) = u_k^l(\tau) \\ h_k^{l-1}(\tau)\left(\frac{l-1}{l}\right) & \text{otherwise} \end{cases} \qquad (7)$$

  Drop path *k* for time period $\tau$ if $h_k^l(\tau)$ < a predefined small number.

- *Step 2* *Stopping rule:*
  If $l \le L \; or \; NGap \le \varepsilon \Rightarrow STOP$;

  otherwise return to Step 1

While no formal convergence proof can be given for this algorithm, since the network loading map does not have an analytical form, a measure of gap, inspired from that used in static network equilibrium models may be used for qualifying a given solution. It is the difference between the total travel time experienced and the total travel time that would have been experienced if all vehicles had the travel time (over each interval $\tau$ ) equal to that of the current shortest path.

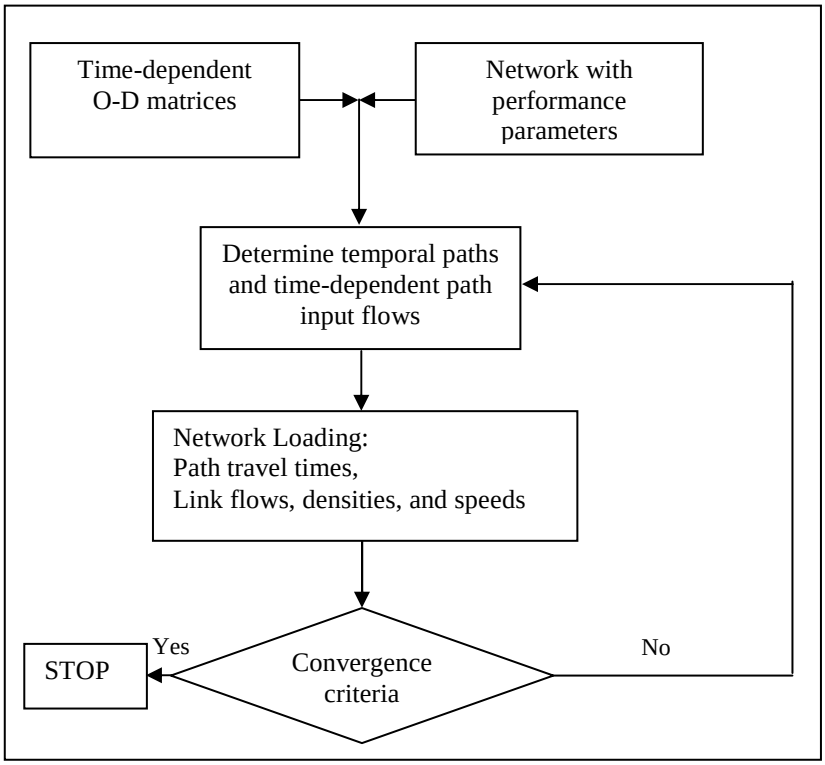



**Figure 1. Structure of the solution algorithm for the DTA model**

A Normalized (or Relative) Gap for each departure time interval $\tau$ may be computed as

$$RGap^{\tau}(n) = \frac{\sum_{i \in I} \sum_{k \in k_i} h_k^{\tau}(n) s_k^{\tau}(n) - \sum_{i \in I} g_i^{\tau} u_i^{\tau}(n)}{\sum_{i \in I} g_i^{\tau} u_i^{\tau}(n)} \quad (8),$$

where $u_i^{\tau}(n)$ are the lengths of the shortest paths at iteration *n*. A normalized gap of zero would indicate a perfect dynamic user equilibrium flow. Clearly this is a fleeting goal to aim for with any dynamic traffic assignment.

It is very important to note that this model, even though its general formulation is very similar to flow based models, is in fact a discrete vehicle model. The network loading procedure, as realized by the event based simulation, moves **individual cars** on the links of the network. This is addressed in the next section.

## 5. THE NETWORK LOADING METHOD

In the simple case of a homogeneous traffic stream (all vehicles with the same physical characteristics) on one-lane link, the simplified car following model underlying the traffic simulator can be solved as follows:

$$T(n,0) = \max\left[ T(n,0) + \frac{L}{V},\ T(n-1,0) + \left(\tau + \frac{\lambda}{V}\right),\ T\left(n - \frac{L}{\lambda}, L\right) + \tau\left(\frac{L}{\lambda}\right) \right]$$

Where $\lambda$ is the effective vehicle length, $\tau$ is the driver/vehicle response time, and ***n*** is the vehicle number by order of arrival to the lane. ***L*** is the length of the link (multiple of $\lambda$), ***V*** is the free-flow speed, ***T(n,0)*** is the time at which vehicle ***n*** enters the link, and ***T(n,L)*** is the time at which vehicle ***n*** exits the link. The above relationship demonstrates how the model is solved in continuous-time using an event-based algorithm. This model yields a steady state triangular flow/density relationship where $Q=1/(\lambda/V+\tau)$, $K=1/\lambda$ and $W=\lambda/\tau$, is the negative wave speed.

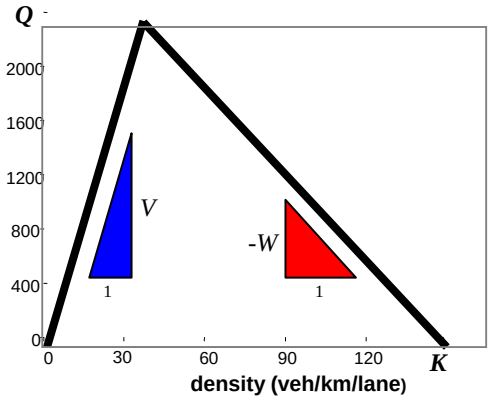



**Figure 2. Flow-Density Relationship**

The simulation, or car mover, that serves for the network loading, is a discrete-event procedure which moves **individual** vehicles through a network defined at the level of individual lanes. The computational effort per link required by this model is strictly proportional to the **number** of vehicles to pass through it, **regardless** of their travel times. Another special property of this model is that traffic dynamics are modeled **without** the (longitudinal) discretization of links into segments or cells. As a result, the procedure only explicitly calculates the time at which each vehicle crosses each node on its path. This leads to a **drastic** reduction in computational effort when compared to microscopic discrete time approaches, where the computational effort is a function of the total travel time experienced by the drivers. This is made possible by the results obtained by Mahut (2000) which demonstrate that the traffic dynamics obtained by the solution of the car following model used do not require the (longitudinal) discretization of links.

The underlying mechanism of congestion in the model is the crossing, merging and diverging of vehicle trajectories. Simply stated, whenever two vehicles pass the same point in space there must be a minimum time separation between them. Conflicts may occur on links, such as weaving movements, or at nodes, due to the interference of vehicles among unprotected movements. Vehicle conflicts are resolved in traffic simulation models by gap-acceptance rules (Barcelo et al., 1994) (Van Aerde, 1999), which are based on one of the two vehicles having priority over the other, and the specification of a time-gap parameter. In continuum traffic models, the approach is to specify the maximum low-priority flow as a function of the prevailing high-priority flow (Leonard et al., 1999). In the model used here, a relatively simple gap-acceptance model has been implemented to determine precedence between vehicle conflicts at nodes, while a FIFO (first-in-first-out) rule is applied on links.

Once conflicts are identified and resolved, the resulting delays are propagated upstream by the car-following model. The *amount* of delay propagated from one vehicle to the next is exactly as would be determined by a standard queuing approach. What is different is *where* and *when* a vehicle in queue experiences each of the delays (or residuals thereof) that are propagated from downstream. This difference is due to the fact that the model employed here rigorously respects the finite speed at which delays propagate in actual traffic, sometimes called the negative wave speed. The positive (forward) wave speed is given by the speed limit. Mahut (2000) provides a detailed description of the model.

The pre-trip path information is complemented by a set of lane choice rules in order to provide the necessary information to identify conflicts between vehicle trajectories. In this way one can capture queue *spil- back* upstream and queue *spill-over* to adjacent lanes. The addition of look-ahead rules based strictly on local information has been shown to significantly improve the reality of the model outputs for some specific though not uncommon network topologies (Barceló, 2000) (Ben Akiva et al., 2000).

A problem associated with traffic control is that of preventing *gridlock*, or *deadlock*, in traffic networks. This situation occurs when a sequence of stopped vehicles forms a cycle in a network, and thus each driver is ultimately waiting for his/her own vehicle to move. This problem was addressed by developing an adaptive deadlock prevention algorithm that identifies when there is a high risk of a deadlocked cycle occurring. The algorithm explores the network starting at any given node using a depth-first search and continues as long as certain conditions are met. The additional computational effort is minimal.

The vehicle time update mechanism is done as follows:

- Each update of a vehicle at (upstream of) a node consists of the following:
  - next link determined by the path
  - arrival and departure lanes on next link:

        - look-ahead rules followed by local rules
    - calculation of time at which vehicle may execute movement

- After all vehicles have been updated at a node, select the next vehicle to move at the node (next event):
    - dead-lock avoidance algorithm:
        - gives priority to vehicles on high density cycles
    - determine precedence among conflicting vehicles
        - gap-acceptance modeling

In the next section some applications of this DTA model are reported.

## 6. APPLICATIONS

This dynamic traffic assignment algorithm was coded in C++ using an object-oriented approach. The code runs under Linux and Win XP/2000. The model was first applied in Stockholm and then in Calgary, Canada. The Calgary application and model calibration was reported in Mahut et al (2004). Other applications have been accomplished in Basel, Switzerland, in California, U.S.A., in Montreal and elsewhere. We give below the relative sizes of some of these applications and the computing times per iteration on an 1.6 Ghz Intel Centrino notebook computer operating under Win XP.

| City | Zones | Nodes | Links | No. of vehicles | Time/iteration |
|---|---|---|---|---|---|
| Stockholm | 114 | 1,191 | 2,080 | 108,000 | 1.5 min |
| Calgary | 76 | 402 | 752 | 12,563 | 6.4 sec |
| Basel | 62 | 376 | 642 | 47,768 | 10.7 sec |
| Montreal | 232 | 966 | 2,563 | 232,667 | 7.5 sec |
| Bakersfield | 64 | 368 | 582 | 47,023 | 51.2 sec |

We present next some details of the Montreal application. The static planning model for the Montreal Region consists of 1,400 zones, 15,000 nodes and 33,500 links. A sub-area identified in pink in Figure 3 indicates the corridor that was identified for the study of an improved urban arterial facility. The demand was obtained by computing a traversal matrix for the sub-area by using the static model and then adjusting it by using link counts. More detail was added where necessary and some 320 traffic signals were coded by the staff of the City of Montreal. The resulting model has 232 zones, 966 nodes, 2,563 links and the total vehicle demand is 232,667. Three classes of vehicles were used: private cars, light trucks and heavy trucks. The data collection, coding and calibration effort took some nine months. The results of the calibration are shown in Figure 4 and a typical convergence result for 10 time intervals for the 6:30 – 9:30 AM peak period are shown in Figure 5. The calibration results are nearly as good as those obtained for the regional static model. Figure 6 shows a snapshot of a dynamic assignment results animation and Figure 7 is a snapshot of a scenario comparison. The facility that is improved is indicated with an arrow.

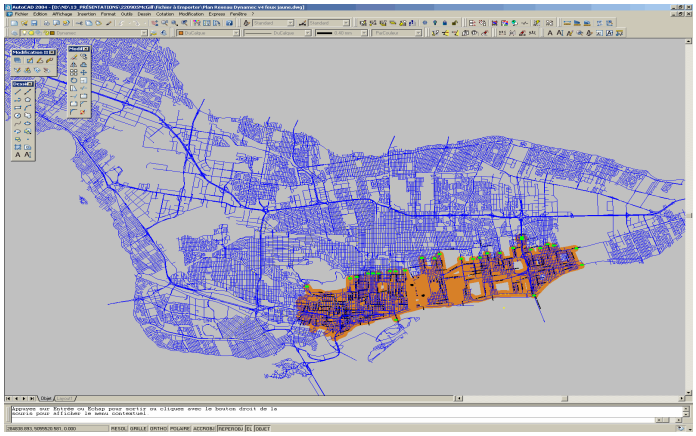


**Figure 3. The Notre Dame corridor sub-area**

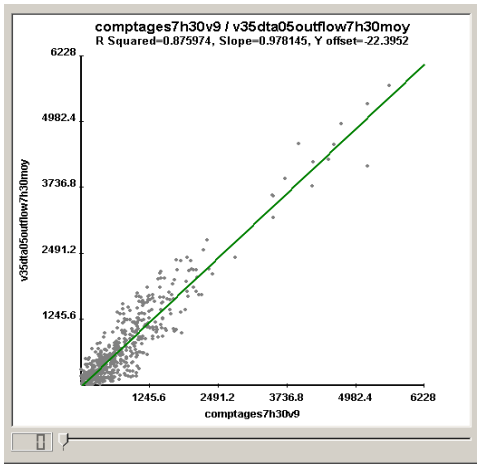



**Figure 4. AM calibration scattergram**

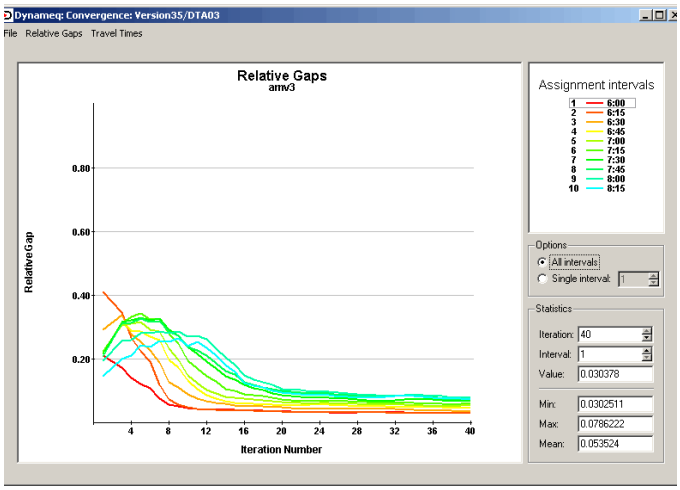



**Figure 5. AM period convergence**

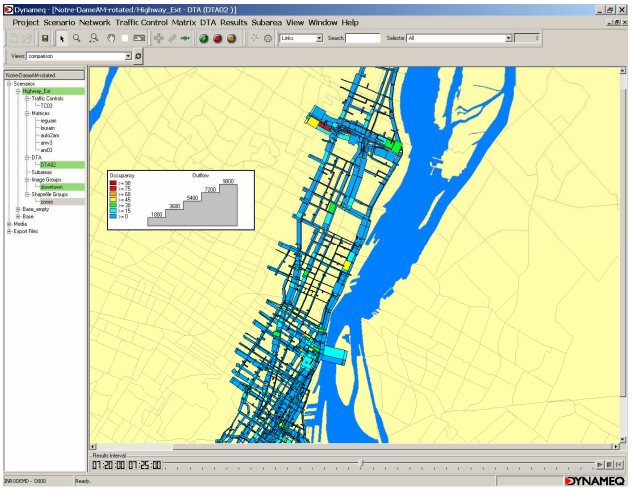



**Figure 6. AM flows colored by occupancy**

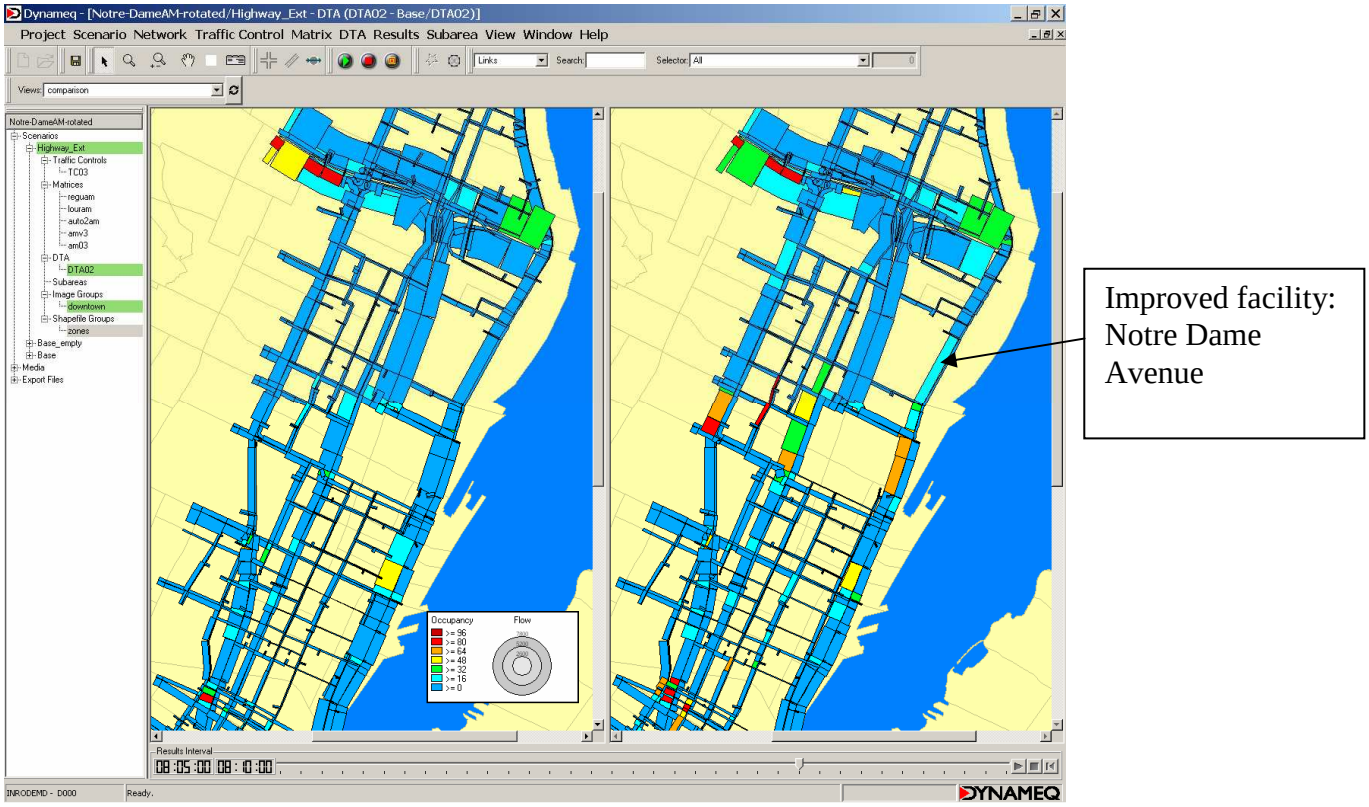



**Figure 7. Scenario comparison : left window after improvement.**

## 7. CONCLUSIONS

A dynamic traffic assignment model, which uses a method of successive averages to determine pre-trip dynamic equilibrium path choices combined with an event-based traffic simulation model, was successfully applied to, validated and calibrated for several medium-sized networks. The results indicate that an acceptable level of convergence can now be obtained for a medium-size network, using a realistic traffic model with a reasonable amount of computing time and memory usage.

The method has excellent potential for use in practice for a variety of planning applications that require the explicit consideration of temporal traffic flows and queues off-line. The model may have potential for further development as an on-line tool, due to the low computation times and memory requirements. Its computational efficiency is at least one

order of magnitude faster than microscopic traffic simulation models. One of the next developments is the adaptation of the method for *en route* assignment and route guidance.

***Acknowledgement* – This work was partially sponsored by an individual operating grant of the Natural Sciences and Engineering Council of Canada (NSERC).**

## 8. REFERENCES

Astarita, V., Er-Rafia, K., Florian, M., Mahut, M., Velan, S. (2001). Comparison of Three Methods for Dynamic Network Loading. *Transportation Research Record*, 1771:179-190.

Barceló, J. (2000). The Role of Traffic Simulation in Advanced Traffic Management Systems. Presented at the Spring meeting of INFORMS, Salt Lake City, USA. May 7-10, 2000.
Barceló, J., Ferrer, J.L., and Grau, R. (1994). AIMSUN2 and the GETRAM Simulation Environment. *Internal Report*, Departamento de Estadistica e Investigacion Operativa. Universitat Politecnica de Catalunya. See also http://www.tss-bcn.com.

Ben-Akiva, M., Koutsopoulis, H., Toledo, T. (2000). MITSIMLab: Recent Developments & Applications. Presented at the Spring meeting of INFORMS, Salt Lake City, USA. May 7-10, 2000.

Ben-Akiva, M., Koutsopoulos, H.N., Mishalani, R. (1998). "DynaMIT: A Simulation-Based System for Traffic Prediction". Paper presented at the DACCORD Short Term Forecasting Workshop, Delft, The Netherlands. See also its.mit.edu.

Diakaki, C. and Papageorgiou, M. (1996). Integrated Modelling and Control of Corridor Traffic Networks usingthe METACOR Modelling Tool. Dynamic Systems and Simulation Laboratory, Technical University of Crete. Internal Report No. 1996-8. Chania, Greece. pp. 41.

Dow,P. and Van Vliet, D., (1979). Capacity Restrained Road Assignment. *Traffic Eng. Control*, 20: 296-305.

Friesz, T., Bernstein, D., Smith, T., Tobin, R., Wie, B. (1993). A variational inequality formulation of the dynamic network user equilibrium problem. *Operations Research*, 41:179-191.

Gentile G., Meschini L., Papola N. (2005) Macroscopic arc performance models with capacity constraints for within-day dynamic traffic assignment, *Transportation Research B* 39, 319-338;

Lawphongpanich, S. and Hearn, D.W. (1984). Simplicial Decomposition of the Asymmetric Traffic Assignment Problem. Transportation Research B, 17: 123-133.

Lighthill, M.J. and Whitham, G.B. (1955). On kinematic waves I: Flood movement in long rivers. II: A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London*, A229:281-345.

Leonard, D.R., Gower, P. and Taylor, N.B. (1989). CONTRAM: Structure of the Model. *Transport and Road Research Laboratory (TRRL) Research Report 178*. Department of Transport, Crowthorne. See also http://www.contram.com/.

Mahmassani, H.S., Abdelghany, A.F., Huynh, N., Zhou, X., Chiu, Y.-C. and Abdelghany, K.F. (2001). DYNASMART-P (version 0.926) User's Guide. Technical Report STO67-85-PIII, Center for Transportation Research, University of Texas at Austin.

Mahut, M. (2000). Discrete flow model for dynamic network loading. Ph.D. Thesis, Département d'informatique et de recherhe opérationelle, Université de Montréal. Published by the Center for Research on Transportation (CRT), University of Montreal.

Mahut, M., Florian, M., Tremblay, N., Campbell, M., Patman, D. and McDaniel., Z. (2004) Calibration And Application Of A Simulation-Based Dynamic Traffic Assignment Model, *Transportation Research Record* 1876: 101-111.

Messmer, A. (2000a). METANET A Simulation Program for Motorway Networks (Documentation). Dynamic Systems and Simulation Laboratory, Technical University of Crete. Chania, Greece.

Messmer, A (2000b). METANET-DTA An Exact Dynamic Traffic Assignment Tool Based on METANET. Dynamic Systems and Simulation Laboratory, Technical University of Crete. Chania, Greece. pp. 37.

Nagel, K. and Schreckenberg, M. (1992). A cellular automaton model for freeway traffic. *Journal de Physique I France*, 2:2221-2229.
Papageourgiou, M. (1990). Dynamic Modelling, Assignment and Route Guidance in Traffic Networks. Transportation Research, 24B(6), 471-95.

Richards, P.I. (1956). Shock waves on the highway. *Operations Research*, 4:42-51.

Van Aerde, M. (1999). INTEGRATION Release 2.20 for Windows: User's Guide. MVA and Assoicates, Kingston, Canada.

Yang, Q. (1997). A Simulation Laboratory for Evaluation of Dynamic Traffic management Systems. Ph.D. Thesis, Massachusetts Institute of Technology. See also http://web.mit.edu/its/products.html.

Ziliaskopulos, A.K. and Lee, S. (1997). A Cell Transmission Based Assignment-Simulation Model for Integrated Freeway/Surface Street Systems. *Transportation Research Record* 1701, 12-23.

# TRAFFIC ASSIGNMENT ON NETWORKS WITH TIME-VARYING FLOWS, WHILE APPROXIMATING CONTINUUM FLOWS ON LINKS

Malachy Carey: Queen's University, Belfast, m.carey@qub.ac.uk

## Abstract

In modelling traffic flows varying over time on road networks it would seem desirable to handle the flows using models based on traffic flow theory. The cell transmission model and a more general finite difference approximation to the LWR model were introduced for that purpose. They both approximate the widely accepted LWR traffic flow model and both have the virtue that they can handle spillbacks, traffic controls, moving queues in a way that is consistent with traffic flow theory. Both were initially formulated as simulation models, with route choice taken as fixed. Here we wish to investigate system marginal costs, economic externalities and optimal congestion tolls hence need a system optimising model with route choice treated as variable. A system optimising assignment model incorporating the CTM was developed some years ago. The present paper develops a system optimising model for traffic networks using a finite difference approximation to the LWR model to handle the flows. This allows nonlinear flow-density functions and we assume multiple origins with a single destination. We also derive system marginal costs, economic externalities and optimal congestion tolls from the model, extend the model to elastic demands, and establish links with previous dynamic traffic assignment (DTA) models.

## 1 Introduction

In this paper we present and analyse a system optimum model for dynamic traffic assignment in which the traffic flows are modelled by approximation to a widely accepted traffic flow model, due to Lighthill and Whitham (1955), Richards (1956) and referred to a the LWR model. Since the latter model is continuous in time and space it is not analytically or computationally tractable for general network modelling and it is more convenient to approximate it by a finite difference approximation, as in Daganzo (1995b). Daganzo (1994) developed a model, referred to as the cell transmission model (CTM), that approximates the LWR model for a single link when the flow-density function is assumed to be triangular or trapezoidal, as in Fig 1. In Daganzo (1995a) he extended the CTM to a network, and in Daganzo (1995b) further extended the analysis to allow general nonlinear flow-density functions, as in Fig 2. He refers to the latter model, with nonlinear flow-density functions, as a finite difference approximation to the kinematic wave model, or LWR model, rather than referring to it as a CTM. For each of these models he showed that as the discretisation of time and space is refined to the continuous limit, the model solves the LWR model.

In the above papers, and in various later papers, the CTM and the finite difference approximation (FDA) to the LWR model are presented as simulation models in which traffic at a junctions and intersections merges and diverges in fixed proportions at each point in time (Daganzo (1995a)), which implies that route choice is fixed. Later there has been substantial interest in embedding the CTM in traffic assignment models for user equilibrium (e.g. Lo (1999), Lo and Szeto (2002), Szeto and Lo (2004)). An important reason for this interest is that in traffic assignment models route choice, and hence the proportions of traffic using the various links, are endogenously determined rather than being prespecified or constrained by fixed merge or diverge proportions.

Ziliaskopoulos (2000) reformulated the CTM as a set of linear constraints and hence developed a linear programming model for the single-destination system optimum DTA problem for a network. The present paper also introduces a system optimizing formulation but assumes a general nonlinear form for the flow-density functions for some or all links, rather than the triangular or trapezoidal form assumed in the CTM. The form of flow-density function assumed in the CTM can be thought of as a special case of the nonlinear form. Also, in this paper we are concerned with deriving system marginal costs, externalities and optimal tolls. Since optimal tolls require a system optimising formulation rather than a user equilibrium formulation, that is what we develop in this paper.

In Section 2 we outline Daganzo's finite difference approximation to the LWR model and re-write the max function as inequalities. In Section 3 we extend this formulation from a single discretised link to a network of such links, and formulate the traffic network assignment problem as a convex nonlinear programme. The merge and diverge proportions at junctions are determined endogenously within this programme, so as to minimize travel costs or maximise traveller net benefits.

This system optimising formulation can lead to the phenomenon of "holding back" of flows, which has been identified by a number of authors as a feature common in models that seek to optimize traffic flows on a network over time. For example, by holding traffic back on some links it may be possible to hold traffic density downstream to a level that allows maximum flow and throughput, that is, at the peak of the flow density function. Conversely, letting traffic flow without any "holding back" could cause traffic density downstream to move onto the congested downward sloping part of the flow-density function, which would restrict traffic flow. This holding back of traffic can be interpreted as a form of traffic flow control (as in Carey (1987) and Ziliaskopoulos (2000)), which could be implemented by variable speed controls, ramp metering, etc. The methods for implementing such flow controls are not necessarily currently available in practice.

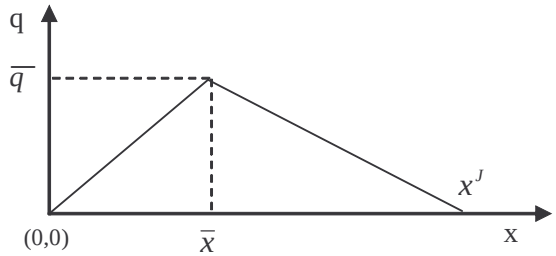



Fig. 1(a). A 'triangular' flow-density or flow-occupancy f'n.

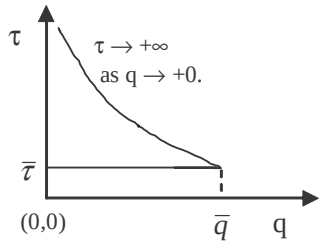



Fig. 1(b). Corresponding link travel-time f'n.

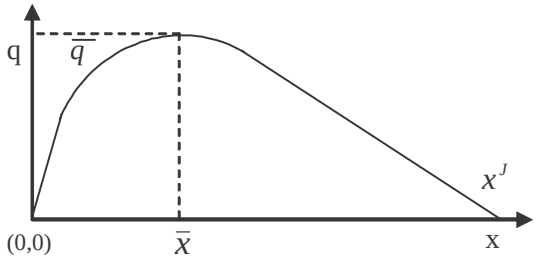



Fig. 2. A nonlinear flow-density or flow-occupancy function.

## 2 A finite difference approximation to the LWR model

The LWR traffic flow model (Lighthill and Whitam (1955) and Richards (1956) assumes that the flow at each point in space and time ($z$,$t$) depends only on the density at that point, and not at any later or earlier points, hence can be set out as a flow-density equation $q(z,t) = Q(k(z,t), z, t)$. Here, as is common, we assume that the link is homogeneous over space and time, which reduces the flow-density function to

$$q(z,t) = Q(k(z,t)) \tag{1}$$

In the LWR model this is combined with a conservation or continuity equation

$$\frac{\partial q(z,t)}{\partial z} = -\frac{\partial k(z,t)}{\partial t} \tag{2}$$

Daganzo (1995b) presented a discretised form of the LWR model (1)-(2) as follows. Divide the time span into time intervals $t$ = 1, …, $T$, each of length $\Delta t$, and divide the link into $j$ = 1, …, $J$, segments or cells such that the

free-flow travel time for the each cell is one time interval. We can assume that the given link is homogeneous, so that all cells will be of the same length $d$. Let $k_{jt}$ denote the cell density, which can be assumed constant along the cell length or can be taken as the mean density in the cell. The flow-density function for a cell can then be written as $q_{jt} = Q_j(k_{jt})$, but it is convenient here to work in terms flow-occupancy $x_{jt} = k_{jt} d$ rather than flow-density $k_{jt}$. Substituting $k_{jt} = x_{jt} / d$ in the flow-density function gives the flow-occupancy function denoted $g_j(x_{jt})$. From the latter construct two functions $g_j^+(x_{jt})$ and $g_j^-(x_{jt})$: $g_j^+(x_{jt})$ is obtained by taking the upward sloping part of $g_j(x_{jt})$ and extending it to the right via a horizontal straight line from its peak, and $g_j^-(x_{jt})$ is obtained by taking the downward sloping part of $g_j(x_{jt})$ and extending it back to the vertical axis via a horizontal straight line from its peak. Then, as in Daganzo (1995b), for consistency with the continuous LWR model the number of vehicles exiting from cell $j$ in time interval $t$, should satisfy

$$v_{jt} = \min\{ g_j^+(x_{jt}),\ g_{j+1}^-(x_{j+1,t}) \} \qquad (3)$$
$$= \min\{(\text{sending capacity of cell } j \text{ at time } t),\ (\text{receiving capacity of cell } j+1 \text{ at time } t)\}$$

Except for the final cell on a link, the outflow $v_{jt}$ from cell $j$ in any time interval equals the inflow $u_{j+1,t}$ to next downstream cell $j$+1 in the same interval, thus

$$v_{jt} = u_{j+1,t} \qquad (4)$$

as illustrated in Fig 3. The number of vehicles in cell $j$ in time interval $t$+1 is the number present in time interval $t$ plus the inflow minus the outflow in interval $t$, thus,

$$x_{j,t+1} = x_{jt} + u_{jt} - v_{jt} \qquad (5)$$

Equations (3)-(5) comprise a finite difference approximation to the LWR model. Equation (3) can be rewritten as $v_{jt} \le g_j^+(x_{jt})$ and $v_{jt} \le g_{j+1}^-(x_{j+1,t})$ if we assume for the moment that the outflow $v_{jt}$ is held at the maximum consistent with these inequalities, so that one or other of the inequalities is a strict equality. Hence (3)-(4) can be rewritten as

$$v_{jt} \le g_j^+(x_{jt}) \text{ and } u_{j+1,t} \le g_j^-(x_{j+1,t}) \qquad (6.1)$$
$$\text{where the flow } v_{jt} = u_{j+1,t} \text{ is assumed held at the maximum consistent with (6.1).} \qquad (6.2)$$

The finite difference approximation to the LWR model now consists of (4)-(6). An advantage of the inequalities in (6.1), for the purposes of a mathematical programming model below, is that they represent convex sets, since $g_j^+(x_{jt})$ and $g_j^-(x_{j+1,t})$ can be assumed to be concave functions. In contrast, (3) is a nonlinear equation hence represents a nonconvex set.

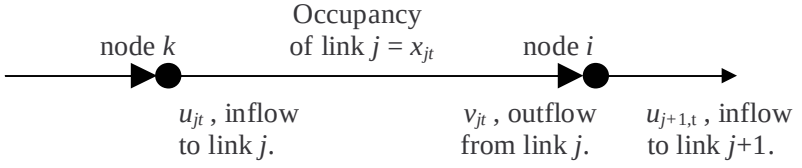



Fig 3. Inflow, outflow and occupancy ($u_{jt}$, $v_{jt}$ and $x_{jt}$) for cell or link $j$.

# 3 A system optimising DTA model, based on a finite difference approximation to the LWR model

Consider a network consisting of a set of nodes $N^O$ connected by a set of directed links $A^O$, with individual nodes and links denoted $i \in N^O$ and $j \in A^O$ respectively. Let each of the original links in the network be divided into cells, introduce an artificial node between each pair of neighbouring cells, and treat each cell as a link between these neighbouring nodes. Denote this expanded set of links as $A$ and the expanded set of nodes as $N$. The set $N$ consists of $N^O$ and the new/ artificial nodes between cells. Let $B(i)$ denote the set of links immediately before node $i$ (pointing into node $i$), and $A(i)$ denote the set of links immediately after node $i$ (pointing out of node $i$).

*Extending the cell conservation equation* (5) *to a network*. To extend (5) to the network, simply rewrite it for all cells $j \in A$ in the network, thus

$$x_{j,t+1} = x_{jt} + u_{jt} - v_{jt} \qquad \forall j \in A. \tag{7}$$

*Extending the node conservation equation* (4) *to a network*. The equations (4) apply to the new nodes introduced between cells on the links of the original network, since for each of those nodes there is only a single in-link and out-link. To ensure conservation at the original nodes $i \in N^O$ of the network, let the sum of the inflows to each node equal the sum of the outflows from the node, thus

$$\sum_{j\in A(i)} u_{jt} = D_{it} + \sum_{j\in B(i)} v_{jt} \qquad i \in N^O$$

where $D_{it}$ is the exogenous travel demand from node $i$ to the destination. For simplicity in the later discussion it is convenient to collect into a single set the conservation equations (8) for nodes $i \in N^O$ and the conservation equations of type (4) that apply to nodes ($i \in N$, $i \notin N^O$), and write all as a single set, thus

$$\sum_{j\in A(i)} u_{jt} = D_{it} + \sum_{j\in B(i)} v_{jt} \qquad i \in N \tag{8}$$

At the new nodes along the original links (nodes $i \in N$, $i \notin N^O$), $D_{it} = 0$ and (8) reduces to (4), i.e. $v_{jt} = u_{j't}$ where $j'$ is the (single) cell immediately after cell $j$. We can assume an artificial link exiting from the destination and $D_{it} = 0$ at the destination node.

*Extending the exit flow equations* (6) *to a network*.

Applying (6) to all cells on all links we have the following. For each cell $j$ the outflow $v_{jt}$ should not exceed the sending capacity $g_j^+(x_{jt})$, thus,

$$v_{jt} \le g_j^+(x_{jt}). \qquad \forall j \in A \tag{9.1}$$

and for each cell $j$ the inflow $u_{jt}$ should not exceed the inflow capacity or receiving capacity $g_j^-(x_{jt})$, thus,

$$u_{jt} \le g_j^-(x_{jt}) \qquad \forall j \in A. \tag{9.2}$$

and

$$\text{either (9.1) or (9.2) is a strict equality} \qquad \text{for each } j \in A \tag{9.3}$$

Condition (9.3) is needed to be consistent with (6.2) and hence (3). However, we wish to construct a mathematical programming model for system optimizing and in a mathematical programme a constraint such as (9.3) converts the programme to a nonconvex nonlinear programme or a combinatorial problem (integer programme). Fortunately, in the mathematical programme that we construct later below, condition (9.3) is

frequently satisfied in any solution of the mathematical programme without having to be imposed as an explicit constraint, and any deviation from satisfying (9.1) can be interpreted as a system optimising flow control.

In the literature on the CTM, and on the finite difference approximation (FDA) to the LWR, it is common to discuss behaviour at merge and diverge nodes separately from ordinary connecting nodes along a link. This is done to handle "splitting" proportions that are prespecified at merge and diverge nodes. However, that is not needed here since the splitting proportions are not prespecified but are endogeneous to the optimisation model set out below. For example, suppose that a single cell $j$ immediately precedes one of the original nodes $i \in \mathrm{N}^{\mathrm{O}}$ and there is more than one link pointing out of this node. In that case, the outflow from cell $j$ flows immediately into a set of cells $j' \in A(i)$ that diverge from the node $i$ at the exit of cell $j$. Equation (12.4) ensures that the outflow from cell $j$ equals the sum of the inflow to the cells $j' \in A(i)$ pointing out from cell $j$, and equation (9.2) ensures that, for each of these cells $j' \in A(i)$, the inflow $u_{j't}$ will not exceed the cell's receiving capacity, that is, it will satisfy $u_{j't} \le g_{j'}^{-}(x_{j't})$ for all $j' \in A(i)$.

*A system optimising DTA model*

We can now set up a system optimising dynamic traffic assignment model, consisting of minimising the network travel costs for all users, subject to the constraints (7)-(9). The time spent by a user in cell $j$ in time interval $t$ is 1 (in whatever units time is measured) hence the time spent by all users $x_{jt}$ in cell $j$ in time interval $t$ is 1 $x_{jt}$, the total time spent by all users on the network is $\sum_{j\in A, t=1}^{t=T} x_{jt}$ and the total cost of this user time is

$$C = \sum_{j\in A, t=1}^{t=T} c_{jt}\, x_{jt} \tag{10}$$

where $c_{jt}$ is the users' cost per unit of time spent in link $j$ in time interval $t$. To obtain the system optimising flows on the network, minimise (10) subject to (7)-(9) and nonnegativity of all the variables. This is set out more formally as follows.

SO: Minimise (10)

subject to, for all time intervals $t = 1, \ldots, T$,

$$(\alpha_{jt}^{+} \ge 0) \qquad v_{jt} \le g_j^{+}(x_{jt}) \qquad \forall j \in A \tag{11.1}$$

$$(\alpha_{jt}^{-} \ge 0) \qquad u_{jt} \le g_j^{-}(x_{jt}) \qquad \forall j \in A \tag{11.2}$$

$$(\beta_{jt}) \qquad x_{j,t+1} = x_{jt} + u_{jt} - v_{jt} \qquad \forall j \in A \tag{11.3}$$

$$(\gamma_{it}) \qquad \sum_{j\in A(i)} u_{jt} = D_{it} + \sum_{j\in B(i)} v_{jt} \qquad \forall i \in N \tag{11.4}$$

$$x_{jt} \ge 0,\ u_{jt} \ge 0,\ v_{jt} \ge 0 \qquad \forall\ j \in A. \tag{11.5}$$

The term in brackets before each equation (i.e., $\alpha_{jt}^{+} \ge 0$, $\alpha_{jt}^{-} \ge 0$, $\beta_{jt}$ and $\gamma_{jt}$) is a dual variables or Lagrange multipliers corresponding to that equation and will be used later.

It is interesting to consider what happens if the equations (11.2) are dropped from the above model SO. If the data is such that, in solutions of the model, the link occupancies $x_{jt}$ are never in the domain of the downward

sloping part of the exit-flow or flow-density functions $g_j(x_{jt})$, for all links and time periods, then equations (11.2) are redundant and can be dropped. Without (11.2), the above model becomes formally the same as the DTA model of Merchant and Nemhauser (1978a, 1978b) except that, in the latter, the equations (11.1) are written as strict equalities whereas here they relaxed to inequalities. The MN model with (11.1) as inequalities was introduced by Carey (1987), where it was noted that it converts the nonconvex optimisation model of MN to a convex model. That model is formally the same as the above model, except that the MN model has usually been applied to networks with each link treated as a whole link. However, the MN model can equally be applied after first discretising whole links into cells and discretising time as in the above model SO.

*Solving programme* SO.

The above convex nonlinear programme SO is linear except for the concave functions in (11.1) and (11.2). It can easily be solved in various ways. It can be solved by using standard linear programming packages if we first piecewise linearise the concave functions in (11.1) and (11.2). Many commercial LP (or mathematical programming) packages include a facility for automatically performing such piecewise linearization. Alternatively, the programme SO can be solved using available commercial packages for solving convex programming problems with nonlinear constraint (e.g., Minos, Conopt, or other solvers available with GAMS). Or special purpose solution algorithms can be devised to take advantage of the special structure of the model. As already noted, the programme SO is similar to the form of DTA model formulated in Merchant and Nemhauser (1978), except for the additional constraints (11.2). Various algorithms were devised to take advantage of the structure of the latter model and these could perhaps be extended to the present model. The model SO also has another special feature which would speed up its solution: most of the links $j$ in SO were formed by discretising the given original links in the network, hence cells that have only single links (cells) pointing in and out of them. That simplifies the structure and greatly increases the sparsity of the matrix of constraint coefficients, which may make standard math programming algorithms competitive with special purpose algorithms.

## 4 Properties of the model: marginal costs, externalities and optimal tolls

To investigate the properties of solutions of Programme SO, we use the Kuhn-Tucker optimality conditions which can be set out as below. These conditions are necessary and sufficient to characterise an optimal solution of SO since the objective function and constraint set of SO are convex. These are convex since the objective function is convex, the constraints (11.3) and (11.4) are linear, and the constraints (11.1) and (11.2) represent convex sets since both are "≤" constraints with a r.h.s. consisting of a convex function, $g_j^+(x_{jt})$ and $g_j^-(x_{jt})$.

The Kuhn-Tucker conditions for Programme SO consist of: For $t = 1, \ldots, T$,

equations (11.1)-(11.5) (12.0)

Complementarity for the pairs of inequalities in (11.1) and in (11.2).

$(v_{jt} \geq 0)$ $\beta_{jt} \leq \gamma_{it} + \alpha_{jt}^+$, $j \in B(i), i \in N$ (12.1)

$(u_{jt} \geq 0)$ $-\beta_{jt} \leq -\gamma_{kt} + \alpha_{jt}^-$, $j \in A(k), k \in N$ (12.2)

$(x_{jt} \geq 0)$ $\alpha_{jt}^+ g_j^+(x_{jt}) + \alpha_{jt}^- g_j^-(x_{jt}) \leq c_{jt} + (\beta_{jt} - \beta_{j,t-1}), \ \forall j \in A$ (12.3)

Complementarity for the pairs of inequalities in (12.1)-(12.3).

.

Complementarity or 'complementary slackness' means that, in a solution of the KT conditions, if either one of a pair of inequalities is a strict inequality then the other one must be a strict equality. To interpret these optimality conditions we first consider the system marginal costs (s.m.c.s). In a mathematical programme the system marginal cost (s.m.c.) of varying a parameter, such as $D_{it}$ in (11.4), is the change in the optimal value of the objective function per unit change in the value of the parameter (such as $D_{it}$), while holding all other parameters fixed. This s.m.c. is given by the value of Lagrange multipliers or dual variables associated with the constraint in an optimal solution of the programme. Thus in Programme SO

$\gamma_{it}$ the optimal value of Lagrange multiplier associated with equation (11.4), is the s.m.c. per unit increase in the exogenous inflow $D_{it}$

**Proposition**: In an optimal solution of programme SO consider the set of time-space paths starting from any node *i* at time *t* and travelling to the destination. The s.m.c. of traversing these time-space path is the same for all time-space paths that are utilised, is less than or equal to the s.m.c. of traversing any time-space paths that are not utilised, and is give by the value of the dual variable $\gamma_{it}$ in the optimal solution.

**Proof:** Since the s.m.c. $\gamma_{jt}$ depends only on *i* and *t*, it is independent of the time-space path, or paths, that the traffic may take from node *i* to the destination. It follows that if the traffic from node *i* at time *t* utilises several time-space paths then the s.m.c. of using each of these paths is the same. Further, the s.m.c. for any unused path must be higher than for used paths, otherwise the total system costs could be reduced by switching marginal traffic to this currently unused path, which would contradict the assumption that the solution is optimal. ■

In the DTA literature, two different definitions of user equilibrium (UE) have been used. In an "ideal" user equilibrium the travel times to the destination are computed along the time-space paths actually followed by users, that is, they are computed from the link travel times that will be experienced by users when they arrive at the link in question. In an "instantaneous" user equilibrium the travel times to the destination, starting from node *i* at time *t*, are computed from the link travel times that hold at the current time *t*. Though these two concepts are usually applied to user equilibrium rather than system optimum, we can adapt them to the latter by replacing "link travel time" with "system marginal cost" in the definitions, to define an "ideal" SO and an "instantaneous" SO. It follows from Proposition 1 that the solution of the model SO satisfies the "ideal" SO definition and the "instantaneous" SO definition, and everything in between.

*Externalities and optimal tolls*

For congested road traffic, an additional or marginal traveller tends to cause an increase in travel times or costs for some or all users and this increase in costs is referred to as the externality caused by the additional user. It is normally assumed that, when deciding whether or when to travel, each road user takes account only of the travel time or cost that he or she would incur (or perceive themselves to incur), and does not take account of any congestion externality. The total cost, or system marginal cost (s.m.c.) caused by an additional (marginal) user of a road link or cell is thus:

(s.m.c of using a link) = (cost incurred by a marginal/ additional user of the link)
+ (congestion externality caused by the marginal/ additional user).

As usual, we define the optimal congestion toll to be exactly equal to the congestion externality. The reason for that is of course that, if this toll is imposed on each user, then the total cost incurred by each user (own cost + toll) becomes equal to the s.m.c. so that the user takes account of the full s.m.c. when making travel decisions. In other words, the user is induced to "internalise" the externality.

### 4.1 S.m.c.s, externalities and optimal tolls for each link

There are various s.m.c.s that we can define associated with using a cell. For example, at time *t* the s.m.c. of travelling to the destination from the node *k* at the beginning of a cell is $\gamma_{kt}$ and from the node *i* at the exit of the

cell is $\gamma_{it}$, hence the s.m.c. of letting a vehicle (a marginal unit of traffic) enter at node *k* instead of node *i* at time *t* is ( $\gamma_{kt} - \gamma_{it}$ ). Similarly, the s.m.c. of letting a vehicle enter at node *k* at time t–1 instead of at time *t* is ( $\gamma_{k,t-1} - \gamma_{it}$ ). The cost of a vehicle being on the link for this extra time period (from t–1 to t) is $c_{jt}$, hence the "externality" associated with letting a vehicle enter at node *k* at time t–1 instead of at time *t* is ( $\gamma_{k,t-1} - \gamma_{it}$ )– $c_{jt}$. However, the above are not the s.m.c.s or externalities with which we are normally concerned in traffic assignment and in obtaining optimal tolls.

The discussion in this section applies whether we are considering one of the original links in the network, or a link that consists of one of the cells into which the original links have been subdivided, or a sequence of such cells. For simplicity we refer to each of these as a link. We noted above that the externality is the s.m.c. of using a link minus the cost of using it incurred, or perceived, by a single user. To derive an expression for the s.m.c. associated with using a given link, we first recall how this is computed in a static assignment model. In a static assignment model we find the s.m.c. of travelling from the beginning of link *j* to the destination and the s.m.c. of travelling from the exit of link *j* to the destination and subtract the latter from the former to give the s.m.c. of traversing link *j*. This procedure can be extended from a static traffic model to a dynamic traffic model, in which travel times and s.m.c.s may vary over time, as follows. Find the s.m.c. of travelling from the node at the beginning of link *j* at time *t* to the final destination, find the time ($e(jt) = t + \tau(t)$ ) at which a vehicle will exit from link *j* if it entered it at time *t*, and find the s.m.c. of travelling from the exit of link to the destination, setting out at time $e(jt)$. Subtracting the latter s.m.c. from the former gives the s.m.c. of traversing link *j*, i.e., the s.m.c. of traversing the link = s.m.c.(k,t) – s.m.c.(i, $t + \tau(t)$ ) where *k* and *i* are the nodes at the entrance and exit respectively of link *j*.

In the case of static traffic assignment model, and some DTA models, we can use the Kuhn-Tucker optimality conditions to derive an analytic expression for the s.m.c. of using a link. Unfortunately, for the present DTA model it seems that it is not possible to do that. The reason is, the Kuhn-Tucker conditions for this model do not yield an analytic expression for the time $e(jt)$ at which traffic exits from the link and, as noted above, the time $e(jt)$ is needed in order to derive the s.m.c. for using the link. To compute $e(jt)$ we first solve Programme SO, use the optimal solution of SO to compute the cumulative inflow curve at the beginning of link *j*, the cumulative outflow curve at the exit of link j, and use these two curves to obtain the exit time $e(jt)$ from link *j*: this well-known procedure is set out in more detail in Appendix 1. The travel time on link *j* is thus obtained from a computational procedure or set of steps, rather than simply evaluating an analytic expression.

Given the exit time $e(jt)$ from link *j*, the s.m.c. of travelling to the destination from node *i* at the exit of link *j* at time $e(jt)$ is $\gamma_{i,e(jt)}$. Recall that the s.m.c. of travelling to the destination from node *k* at the entrance of link *j* at time *t* is $\gamma_{it}$. Hence the s.m.c. of a vehicle using link *j*, entering link *j* at time *t*, is ( $\gamma_{k,t} - \gamma_{i,e(jt)}$ ). Note that the exit time $e(jt)$ may or may not have an integer value, as traversing the link travel time may take a non integer number of time intervals. Hence $\gamma_{i,e(jt)}$ may not be the dual variable associated with equation (11.4) for a particular time interval *t*. To compute $\gamma_{i,e(jt)}$ we can interpolate between the values of $\gamma_{kt'}$ and $\gamma_{k,t'+1}$ where $t' < e(jt) < t' + 1$, using linear interpolation.

The actual travel time (travel cost) incurred on link *j* by an individual user who is about to enter link *j* at time *t* is $e(jt) - t = \tau(t)$. Subtracting this from the s.m.c. of using link *j* gives, for traffic entering link *j* at time *t*, we obtain

externality associated with using link *j*
= (s.m.c. of using link *j*) – (travel time or cost incurred by an individual user on link *j*)
= ( $\gamma_{k,t} - \gamma_{i,e(jt)}$ ) – ($e(jt) - t$).

In view of the discussion just before this section 4.1, the above externality is also the optimal congestion toll for traffic using link *j*, if entering link *j* at time t.

## 5. Extending to cost-elastic travel demands

In the discussion so far we have assumed that the travel demands $D_{it}$ are fixed, that is, they do not depend on any of the other variables in the problem. However, in practice the origin-destination (O-D) travel demands may depend on the travel times or costs incurred or perceived by users, and the latter travel time or cost depend on the traffic flows. Such travel demands are usually referred to as elastic or cost-elastic, can be introduced into the above model, and into Programme SO, by a slight extension of the well-known method used in static traffic assignment models. Recall that, in the latter, elastic travel demands are introduced by adding, to the objective function, the integral of the inverse travel demand functions.

Let the O-D travel demand from node *i* in time interval *t* be $D_{it} = d_{it}(c_{it})$ where $c_{it}$ is the cost incurred/ perceived by a user travelling from node *i* to the destination, and let the inverse of this be $c_{it} = c_{it}(D_{it})$, where $c_{it}(.)$ denotes $d_{it}^{-1}(.)$. (In an Appendix we consider a different form of elastic demand.) The integral of this, inverse function summed over all demand nodes and time intervals is $I = \sum_{t=1}^{T} \sum_{i \in N^C} \int_{D_{it}=0}^{+\infty} c_{it}(D_{it})\, dD_{it}$, which can be interpreted as a measure of benefit to travellers.

To maximise net benefit (i.e. the above travel benefit function minus the travel costs (10)), proceed as follows: add the negative of the benefit function (i.e. $-I$) to the objective function of Programme SO, treat $D_{it}$ as a variable rather than a constant in constraints (11.4) and leave Programme SO otherwise unchanged. Now consider how this affects the Kuhn-Tucker (K-T) optimality conditions for Programme SO, that are set out just after SO above. Since the demand functions $D_{it} = d_{it}(c_{it})$ can be assumed decreasing or non increasing in $c_{it}$, the inverse functions $c_{it} = c_{it}(D_{it})$ are decreasing or non increasing in $D_{it}$, hence the integral *I* is a concave function and the negative of the integral is a convex function. It follows, as before, that the K-T conditions are necessary and sufficient to characterise an optimal solution of Programme SO. The K-T conditions are as before except that there is now an additional set of conditions,

$$(D_{it}) \qquad \gamma_{it} = c_{it}(D_{it}), \qquad \forall\ i \in N \tag{12.4}$$

and inverting the latter gives $D_{it} = d_{it}(\gamma_{it})$. We saw in Proposition 1 that the s.m.c. of traversing any utilised time-space path from node *i* to the destination, setting out at time *t*, is given by the dual variable $\gamma_{it}$. Thus, (12.4) states that the travel demand $D_{it}$ at each origin node increases up to the point where the s.m.c. $\gamma_{it}$ of an additional trip is just equal to the price $\gamma_{it} = c_{it}(D_{it})$ that travellers are willing to incur to sustain that level of demand $D_{it} = d_{it}(\gamma_{it})$.

## 6 Concluding remarks

In the above we set out a system optimising model for a traffic network in which the flows within links are handled by a finite difference approximation to the LWR model. From the model we showed how to obtain system marginal costs (s.m.c.s), economic externalities and optimal congestion tolls for each link and path of the network, and found that these are easily computed. We extended the model to allow cost responsive travel demands, that is, let the travel demands realised at each origin node, at each point in time, depend on the current s.m.c. of travelling from there to the destination. The results obtained for the fixed demand case continue to hold for the cost elastic or responsive demands. We also noted links between the model and one of the oldest models developed for dynamic traffic assignment, the Merchant-Nemhauser model. At present we are running computational experiments applying the model to small networks.

## References

Carey, M. (1987). Optimal time-varying flows on congested networks. *Operations Research* **35(1)**, 56--69.
Daganzo, C.F. (1994). The cell-transmission model: a simple dynamic representation of highway traffic. *Transp Res* **28B(4)**, 269--287.
Daganzo, C.F. (1995a). The cell-transmission model, Part II: Network traffic. *Transp Res **29B(2)***, 79--93.
Daganzo, C.F. (1995). A finite difference approximation of the kinematic wave model of traffic flow. *Transp Res **29B(4)***, 261-276.
Lighthill, M. J. and Whitham G.B. (1955). On Kinematic waves. I: Flow movement in long rivers II: A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society A* **229**, 281-345.
Lo, H.K. (1999) A dynamic traffic assignment formulation that encapsulates the cell transmission model. In A. Ceder (ed.) *Transportation and Traffic Theory*, Pergamon, Oxford, pp. 327-350.
Lo, H.K and W.Y. Szeto (2002). A cell-based variational inequality formulation of the dynamic user optimal assignment problem. *Transportation Research B*(36), 421–443.
Merchant, D. K. and Nemhauser, G. L. (1978a). A model and an algorithm for the dynamic traffic assignment problem. *Transportation Science* **12(3)**, 183-199.
Merchant, D. K. and Nemhauser, G. L. (1978b). Optimality conditions for a dynamic traffic assignment model. *Transportation Science* **12(3)**, 200-207.
Richards, P.I. (1956). Shock waves on the highway. *Operations Research* **4,** 42-51.
Szeto, W.Y. and Lo, H.K. (2004). A cell-based simultaneous route and departure time choice model with elastic demand. Transportation Research Part B, 38, 593-612.
Ziliaskopoulos, A.K. (2000). A linear programming model for the single destination system optimum dynamic traffic assignment problem. *Transportation Science* **34(1)**, 37-49.

## **Appendix 1:** *Obtaining cell, link, path travel times from a solution of Programme* SO

The travel time $\tau_{jt}$ and exit time $e(jt) = t + \tau(t)$ for a cell or link or path, which is used in Section 4.1, is not immediately available from the solution of Programme SO, since the variables in SO are all flow variables. However, $\tau_{jt}$ and $e(jt)$ are easily computed from the solution of SO since, as is well-known, travel times can be computed from cumulative inflow and outflow curves, which are obtained as follows. The same method is used to find the travel time and exit time for a cell, link or path, but we here set it out for a cell. The solution of SO provides values of $u_{jt}$ and $v_{jt}$ for $t = 1, \ldots, T$ (the values of $u_{jt}$ and $v_{jt}$ for $t = 0$ are taken as given) and from these we obtain the cumulative inflows $U_{jt} = \sum_{t'=0}^{t} u_{jt'}$ and outflows $V_{jt} = \sum_{t'=0}^{t} v_{jt'}$ for any. We can assume that $U_{jt}$ and $V_{jt}$ represent discrete points on continuous-time functions $U_j(t)$ and $V_j(t)$. A continuous $U_j(t)$ can be constructed by interpolation (we can assume linear interpolation) between the points $(t, U_{jt})$, $t = 1, \ldots, T$, and a continuous $V_j(t)$ by interpolation between the points $(t, V_{jt})$. We can also assume that traffic exits from a cell $j$ in the same time order that it entered it, that is, first-in-first-out (FIFO). In that case, the cumulative inflow up to time $t$ (i.e. $U_j(t)$) will all have exited by exactly time $t+\tau_j(t)$ where $\tau_j(t)$ is the travel time of the last vehicle that entered at time $t$ (i.e. $U_j(t) = V_j(t+\tau_j(t))$). Hence, to find the cell traversal time, for traffic that enters it at time $t$, take the cumulative inflow $U_j(t)$ up to time $t$ and find the time $e(jt)$ by which this amount $U_j(t)$ has all exited from the cell: the cell exit time is then $e(jt)$–t = $\tau_j(t)$. This can be illustrated graphically by drawing $U_j(t)$ and $V_j(t)$ curves, with time $t$ on the horizontal axis: $\tau_j(t)$ is the horizontal distance between the two curves, starting from the $U_j(.)$ curve at time $t$.

# THE DEVELOPMENT OF A PROBIT-BASED STOCHASTIC DYNAMIC TRAFFIC ASSIGNMENT MODEL

*Andrea Rosa: Napier University, United Kingdom, a.rosa@napier.ac.uk*
*Mike Maher: Napier University, United Kingdom, m.maher@napier.ac.uk*

## Abstract
Building on the authors' previous research on static assignment, the paper formulates a probit-based dynamic traffic assignment model. The departure time choice model allows for correlation between the tripmaker's perception errors associated with similar departure times, whilst the path choice model allows for correlation between the tripmaker's perception errors associated with overlapping paths. The cell transmission model is proposed to model the propagation of traffic flow through the network over the modelled period. Previous work by the authors on numerical approximation methods for evaluating choice proportions in a MVN discrete choice model is employed to produce a number of different approaches to tackling the simultaneous departure time – path choice loading problem. Initial results are described from some, as yet, limited numerical examples.

## 1 Introduction
This paper presents ongoing work for the development of a Dynamic Traffic Assignment (DTA) model representing departure time choice and route choice with a stochastic model set within the random utility framework.

The introduction of stochastic choice models in a DTA model allows the replication of phenomena such as multi-routeing due to different cost perceptions and has been considered in previous studies. In particular, Ran and Boyce (1996) gave formulation and algorithms for a DTA model with stochastic route choice. Han (2003), still working on DTA including route choice only, discussed a logit stochastic network loading procedure that extended the one given by Ran and Boyce (1996) and was solved with a solution algorithm underpinned by the Dynamic Stochastic User Equilibrium formulation of Ran and Boyce.

We aim to model a Dynamic Stochastic User Equilibrium (DSUE) flow pattern where no user can improve their perceived travel cost by unilaterally changing route or departure time. This definition extends on that given for DSUE by Han (2003) which, in turn, was based on the SUE definition given by Daganzo and Sheffi (1977).

An analysis of the desirable properties of route and departure time choice models included in a DTA model, has led us to put forward for both choices the use of the multinomial probit (MNP) model. This, however, means dealing with the efficient solution of the probit choice function, which cannot be written in closed form. In this paper we explore the possibility of solving the MNP models of departure time and route choice by using analytical approximations building on our previous work on static traffic assignment (Maher, 1992, Maher and Hughes, 1997; Rosa and Maher, 2002; Rosa, 2003).

The paper focuses the formulation of the route and departure time choice models and on the difficulties and methods related to calculating the resulting large probit models. The general structure of the model under development is also presented. Section 2 describes how costs are formed in the model, detailing the elements of the systematic part of the costs. The route choice model and the departure time choice model are outlined respectively in section 3 and 4, relating the formulation adopted to the properties desirable in such models. Section 5 introduces briefly the analytical methods for solving the multinomial probit choice model that are considered in this work while, in section 6, their application for dynamic stochastic loading is discussed. Finally, ongoing and further model developments are delineated in section 7.

## 2 The model structure, the travels costs and the traffic model
In this paper the choice of departure time from an origin and of route to the given destination on a road network is modelled within the random utility framework. We assume that there is a fixed number of drivers who will travel during the time modelled and that each of them has a finite set of routes and departure times

to choose from as well as a preferred arrival time at destination. The finite set of departure times results from the division of the period under study into time slices of equal duration, within each of which the rate of traffic flow departing from the origin is obtained dividing the flow to depart in the time slice by its duration (i.e. the flow rate is stationary within each time slice).

According to the definition of DSUE given above, each driver chooses a combination of route $j$ and departure time $s$ so as to minimise his/her perceived overall cost of travelling defined as $C_{sj} = c_{sj} + e_s + \varepsilon_j$. In this expression $c_{sj}$ is a systematic cost due to network conditions encountered by a driver who has chosen to travel along path $j$ leaving at a time $s$; thus "ideal" or "realistic" travel costs are used, i.e. the driver chooses according to the costs which will actually be experienced during the journey rather than those current at the time of entering the network. The random perception errors $e_s$ and $\varepsilon_j$ associated respectively with departure time choice and path choice are assumed independent of each other but capture, through the correlations between $e_s$ and $e_t$ the similarity between departure times $s$ and $t$ in the perception of the travellers and, through the correlation between $\varepsilon_j$ and $\varepsilon_k$ the similarity between paths $j$ and $k$ (that is, the topology of the network). The latter term is therefore analogous to that used in static traffic assignment.

The systematic cost $c_{sj}$ , in more detail, is defined as the sum of a travel cost or travel time, obtained as the sum of the travel times experienced along each of the links that make up a route and of a late/early arrival penalty.

Much DTA literature discusses how to best represent the traffic behaviour along a network arc. Our aim is to look mainly at the choice components of a DTA model; therefore we will employ as a traffic model the Cell Transmission Model (CTM) of Daganzo (Daganzo, 1994, 1995) which, whilst simple, is realistic enough to replicate key features of such real traffic queues and their formation and dissipation, ensures causality (conditions experienced by traffic are not caused by future events) and a FIFO vehicle processing (within the refinement of its time divisions).

The CTM divides the time into time ticks, during which the events of the model take place: flow enters or exits the network and moves from one of the cells representing an arc of the road network to the next. The flow travels through the network in packets, each of which assumed to contain vehicles of equal characteristics (as *e.g.* path to travel along, desired arrival time). Packets are released at the cells representing the origins and are further divided, as necessary, when the whole packet content of flow cannot move from a cell to the next.

Since we have assumed that the flow rate leaving an origin during a time slice is stationary, the traffic flow departing during a time slice is divided into packets of equal content departing at each of the CTM time ticks that make up the time slice.

As the traffic flow is represented in the CTM by packets, our account of costs is kept accordingly: the cost of travelling along a path at a time slice is obtained by recording the time employed by each packet of flow departed during that time slice to reach the destination along that path.

Each packet of flow will also incur in a late/early arrival penalty which, as for other models in the literature is given by a piecewise linear function (see *e.g.* Szeto and Lo, 2004, who refer to Small (1982) for an empirical confirmation of the validity of this assumption). In particular, if $dat_i$ is the desired arrival time for a traffic packet $i$ that reaches its destination at the actual arrival time $aat_i$ and $ib_i$ is the indifference interval for that traffic packet, the late/early arrival penalties $leap_i$ can be written as:

$$leap_i = \begin{cases} \lambda_e \left[ (dat_i - ib_i) - aat_i \right] & \text{if } aat_i < dat_i - ib_i \\ 0 & \text{if } dat_i - ib_i \le aat_i \le dat_i + ib_i \\ \lambda_l \left[ aat_i - (\mathrm{dat}_i - ib_i) \right] & \text{if } dat_i - ib_i < aat_i \end{cases} \quad (1)$$

where $\lambda_e$ and $\lambda_l$ are, respectively, unit costs of early and late arrivals.

The travel time and the late/early arrival penalty associated to each packet of flow reaching the destination are then summed to define the total cost incurred by that packet of flow. Finally, a weighted average of the travel cost experienced by each flow packet that left during a time slice and reached the destination along a

certain path will give the total cost of travelling at that time and along that path considered by the route and departure time choice models. The weights considered are the amounts of flow in each packet.

# 3 The route choice model

The route choice model is defined by the distribution of the random term used to represent the variation in the perception of costs of travelling along a set of paths. Ideally, such a random term should be able to account correctly for the similarity of the alternatives due to the topology of the road network. This challenge has been faced by modellers from the beginning of studies on static stochastic traffic assignment and much of the recent work on choice models for route choice has tried to overcome the limitations of the logit model, widely adopted and convenient to use but unsuitable to capture network structure, since it calculates choice proportions accounting only for absolute cost differences and cannot account for options' similarity (as illustrated by Sheffi, 1985, pp.302-305). One alternative, also investigated from the beginning of static stochastic traffic assignment and able to represent network structure correctly, is to use the probit model, with arc costs assumed to be Normally distributed.

We adopt this latter alternative, thus we assume that the cost of travelling along a network arc can be represented as Normally distributed with the variance of the travel time related to a fixed characteristic of the arc, such as its length, exactly as in probit static traffic assignment (Daganzo and Sheffi, 1977). The Normal distribution of travel cost perceptions can be justified by considering that it represents a mean perception of the costs by a large number of drivers.

As in the static traffic assignment case, considering arc costs as independently Normally distributed results in a multivariate Normal distribution of the path costs, whose covariance matrix is able to capture the network topology. The path choice model is therefore multinomial probit model.

Different ways of accounting for the choice set, the paths between an OD pair, relate the choice models to the methods for solving them when applied to a network. Methods considering paths directly or implicitly using arc data have been investigated in the literature. In this study we assume that the set of paths between an OD pair can be enumerated prior to the DTA calculations and is of limited size, thus not posing problems for its representation in the solution algorithm.

# 4 The departure time choice model

Also the models for departure time choice (DTC) appeared in the literature have included instances of the logit model applied to time divided in slices and also in this case the need to represent realistically the similarity of the alternatives has been long recognised and has lead to develop and employ more complex models. In fact, the significance of accounting for correlation amongst alternatives is remarked by an early work as that of McFadden *et al.* (1977) who pointed out that it is actually more stringent the smaller are the time slices employed.

In general, a satisfactory choice model for DTC should be one in which the correlation between the perception errors of the costs at times $s$ and $t$, $e_s$ and $e_t$, tends to 1 when $|s - t|$ tends to 0, and decreases monotonically as $|s - t|$ increases. Also, the results of a choice model used for DTC should not be unduly sensitive to the discretisation of the time period being modelled but should remain consistent as the discretisation of the time is refined so that the model defined over a discrete set of departure times becomes, in the limit, a model defined over continuous time. We use a continuous time equivalent of a discrete time AR(1) time series model, in which the correlation between $e_s$ and $e_t$ is of the form $\exp(-\lambda|s - t|)$. Given the correlation between the costs of travelling at any two times, it is thus possible to calculate $\lambda$, the rate of decrease of the correlation over time, and obtain the correlation for the costs of the departure time alternatives defined over discrete intervals.

Such an AR(1) process results in costs that are MVN distributed and thus in a multinomial probit choice model. A comparison of the effect of choosing such an AR(1) formulation as opposed to a logit-based one can be found in Maher and Rosa (2006). That work highlights the substantial differences in the choice patterns between the AR(1) formulation used here and a corresponding logit formulation, also pointing out the effect of different systematic cost patterns. Maher and Rosa (2006) also compare choice patterns directly obtained on given time discretisations with those obtained from finer ones, remarking the good aggregation

properties of the multinomial probit model which, however, are contrasted by the practically perfect aggregation experienced with the logit model.

Potentially, the departure time choice framework outlined in this section could be applied by assuming as time slice the duration of the single 'time tick' of the CTM, which we have selected as underlying traffic model. However, in practice such a refinement is limited by the precision and computation effort required by method used to solve the MNP model.

# 5 Analytical solutions of the MNP model

Having the determined multinomial probit models as desirable models for departure time and route choice poses the problem of solving the choice function, which cannot be written in closed form. The solution of the MNP choice function by simulation has been favoured in much of the transportation literature, probably due to the excessive computational effort entailed by the numerical integration methods explored and to discouraging results on the precision of the Clark approximation.

However, being able to solve the choice function numerically would be of advantage to develop an efficient solution algorithm for the DTA problem.

A recent review of analytical methods for solving the MNP model, appeared partly in Rosa and Maher (2002) and more extensively in Rosa (2003), pointed at the work of Genz (1992, 1993) for a more viable numerical integration approach to calculate the MNP choice function and used it to obtain reference results and investigate the precision of several approximation methods. Some of those approximations had not been previously considered in transportation, but had been devised in other fields where MVN integrals are evaluated. Rosa (2003) suggested that the method of Mendell and Elston (1974), also treated by Kamakura (1989), and the method of Tang and Melchers (1987) should be further considered for practical applications since they give a good trade off between the accuracy of the results and the computational effort required.

In part of the work reported in this paper, we consider solving the MNP choice function with the method of Mendell and Elston. Its application relies on turning a MNP choice problem into its equivalent problem in difference with respect to the alternative whose choice probability is calculated. The corresponding MVN integral (which is of one dimension smaller that the original problem) is then evaluated by applying the formulae of the approximation. While, potentially, there is no limit to the size of the choice set that can be dealt with, the approximation may become less precise for larger choice sets.

We also use the approximation of Clark (1961) which is essential for one of the stochastic network loading methods put forward below since it is able to calculate mean and variance of the minimum - approximated as Normal - of the costs of a number of options. This is required if we wish to carry out the stochastic network loading by considering explicitly the marginal choice of, say, departure time and the conditional choice of path and therefore we need to quantify the aggregate cost of all options available in the conditional choice (in the example, all paths available when departing during a time slice). The approximation of Clark is based on approximating as Normal the minimum of two Normal distributions; applying it repeatedly allows us to compare the cost of an option with the approximated minimum of all the others in the choice set and therefore evaluate its choice probability. Similarly, applying the approximation until all options in the choice set are included in the calculations allows us to obtain the approximated distribution of the minimum cost of the choice set. Considering the minimum of a number of Normal variates as Normal is only an approximation which sometimes is so imprecise as to give largely inaccurate results (especially for small actual choice probabilities and for large choice sets).

# 6 The dynamic stochastic network loading problem

If we were interested in the development of a DTA model incorporating only a MNP model of path choice (therefore not allowing for the choice of departure time) the development of a path-based MNP dynamic stochastic network loading method would be a simple extension of that for the static case. The dimension of the choice problem would be given by the number of paths between origin-destination pairs and the approximation methods investigated by Rosa and Maher (2002) could be employed.

However, we set out to develop a DTA model including also departure time choice. The likely large number of time slices to be considered in the model imposes a significant increase on the dimension of the choice set and could result in decreased accuracy of the MNP approximations used. Although the solution of the choice model will be in any case approximate, in the following we consider different ways of structuring the calculations in an attempt to reduce the likely inaccuracies.

The first and simplest method for carrying out the dynamic stochastic loading is to consider all options together, all paths available between an OD pair for all departure times, solve the relevant MNP choice model and release the traffic from the origin accordingly. Each entry of the vector of systematic costs is formed as described in section 2, whilst the covariance matrix is obtained by summing the covariance matrix for the path costs at all time slices and the covariance matrix for the time slice similarities and considering that the two MVN random terms are independent of each other. Thus, if $v_{st}$ is the entry of the covariance matrix of the perception error for departure times $s$ and $t$ and $w_{jk}$ is the entry of the covariance matrix of the perception error for paths $j$ and $k$, $\mathrm{Cov}(C_{sj}, C_{tk}) = v_{st} + w_{jk}$.

This method is simple and has the further advantage that does not impose a hierarchy between path and departure time choice, which – we believe – should be avoided since it seems artificial. However, it requires one calculation of the choice proportions between $SJ$ choices (where $S$ is the number of departure times, and $J$ is the number of paths). Considering the full choice set is not practical for realistic networks where its size would be large especially as a result of the number of time slices and the quality of the MNP approximation would decrease.

An alternative method would be to consider that the time slice-path choice pattern necessary for carrying out a dynamic stochastic network loading could be obtained in two equivalent ways: by considering a marginal choice of path and a conditional choice of departure time or, alternatively, by considering a marginal choice of departure time and a conditional choice of path. Taking into account both ways of conducting the calculations allows us to attain the final result by applying the MNP solution method only to calculate the two sets of conditional choice probabilities, therefore employing it on choice sets much smaller than the whole one.

In more detail, if we call $r_s$ the marginal probability of choosing to depart at time $s$ and $a_{sj}$ denotes the probability of travelling along path $j$ conditional on the choice of departing at time $s$, the probability that a driver leaves at $s$ and travels along $j$ can be written as:

$$p_{sj} = a_{sj} r_s \tag{2}$$

The same probability $p_{sj}$ can be calculated starting from $c_j$, the marginal probability of choosing path $j$, and the conditional probability $b_{js}$ of departing at $s$ conditional on the choice of travelling along path $j$:

$$p_{sj} = b_{js} c_j \tag{3}$$

Having introduced this notation, we can also write that the overall probability of choosing a route is equal to the sum of the probability that it is chosen by the proportion of travellers departing at each time slice:

$$\sum_s a_{sj} r_s = c_j \tag{4}$$

and, similarly, it is possible to write that the total probability of choosing a time slice is equal to the sum of the probability that it is chosen by the proportion of drivers travelling along each path:

$$\sum_j b_{js} c_j = r_s \tag{5}$$

In matrix form the (4) and (5) result, respectively:

$$\mathbf{A}^{\mathrm{T}} \mathbf{r} = \mathbf{c} \tag{6}$$

and

$$\mathbf{B}\,\mathbf{c} = \mathbf{r} \tag{7}$$

where $\mathbf{r}$ the vector of marginal departure time choices, $\mathbf{c}$ is the vector of marginal route choices, $\mathbf{A}$ is the matrix of route choices conditional on the departure time choices and $\mathbf{B}$ is the matrix of departure time choices conditional on the route choices.

Substituting the expression of $\mathbf{c}$ from (6) into (7), it results:

$$\mathbf{B}\,\mathbf{A}^{T}\,\mathbf{r} = \mathbf{r} \tag{8}$$

We can solve for **r** by recognising the linear dependency amongst the *m* equations in (8). Thanks to this dependency we can write the matrix **M** formed as $[BA^T - I]$ with the last row replaced by a row of 1 and the vector **h**, of size m, with all entries equal to 0 except the *m*th which is set to 1, and use them to write:

$$\mathbf{M}\,\mathbf{r} = \mathbf{h} \tag{9}$$

Which readily gives:

$$\mathbf{r} = \mathbf{M}^{-1}\,\mathbf{h} \tag{10}$$

allowing us to use the entries of **A** and **r** as in (2) to obtain the departure time-path choice pattern.
Working similarly, we can also replace **r** given by (7) into (6) and calculate **c**, which then allows us to solve the departure time-path choice pattern using (3).

The advantages of this method are that it reduces the size of the choice set to be dealt with in MNP calculations while continuing to avoid imposing a hierarchy between path and departure time choice. It requires *S* calculations of the proportions in a *J*-choice set (to find the matrix **A**), and *J* calculations of the proportions in an *S*-choice set (to find **B**). Moreover, while it disaggregates the original problem into two problems considering marginal and conditional choices, it avoids calculating the marginal choices directly, thus obviating the need to deal with the aggregate costs of the options considered in the conditional choices. The disadvantage of this method that it relies on the consistency between the two alternative ways of calculating the overall choice proportions set out in (2) and (3) which is not necessarily assured.

A third option to derive the proportion of drivers choosing each alternative path-departure time combination would be to impose a hierarchy of choice and calculate separately marginal and conditional choice probabilities. For instance, we could consider the choice of path conditional on that of time slice. Then, the probability of choosing each of the available paths at a time slice would be calculated as well as the distribution of the minimum of the path costs for that time slice. Using these aggregate costs the choice probability of each time slice could then be obtained and employed to calculate the choice proportions for all departure time-path combinations.

To apply this method we require the minimum of the cost of the conditional choice options. In the MNP case this can be calculated, although approximately, by the method of Clark (1961). Moreover, to complete the calculations of the distribution of the minimum of the costs for each time slice we also need to calculate their covariance, which result from the overall covariance structure assumed and can be obtained, again, with the method of Clark.

For instance, if we assume that the choice of route is conditional on that of path, the cost of travelling a time *s* can then be written as:

$$Y_s = \min_j (C_{sj}) \tag{11}$$

and $p_j^{(s)}$ is assumed to indicate the choice proportion for path *j*, given that departure time *s* has been chosen.
By using Clark we can write the covariance between the distribution of minimum cost of travelling at time *s* and the variate representing the cost of travelling at any departure time-path combination (*t*, *k*):

$$\mathrm{Cov}(Y_s, C_{tk}) = \sum_j p_j^{(s)}\,\mathrm{Cov}(C_{sj}, C_{tk}) = \sum_j p_j^{(s)}\,(v_{st} + w_{jk}) \tag{12}$$

Therefore, if $Y_t = \min_k (C_{tk})$ is the minimum cost of travelling along all paths available at time *t*, that are chosen with proportion $p_k^{(t)}$, the covariance between the minimum costs of departing at time *s* and at time *t*

$$\mathrm{Cov}(Y_s, Y_t) = \sum_k p_k^{(t)}\,\mathrm{Cov}(Y_s, C_{tk}) = \sum_j \sum_k p_j^{(s)}\; p_k^{(t)}\,(v_{st} + w_{jk}) \tag{13}$$

Having obtained the distribution of the costs $Y_s$ of departing at all time slices *s*, a MNP choice function calculation method can then be applied to determine the departure time choice proportions and, finally, the overall departure pattern.

Alternatively, and performing analogous calculations, the departure pattern can be obtained assuming that the time is chosen conditional on the route to follow.

The advantage of this last method is that it breaks the original problem into two smaller ones therefore limiting the negative effect of the size of the choice set on the Clark approximation. The disadvantage is that it imposes a hierarchy of choice which is not necessarily justified except by computational convenience and requires the use of the approximation method of Clark.

Whether any of the three alternative methods described above stands out for precision and is suitable for use in a dynamic stochastic network loading algorithm can be determined by comparing the results they return on a set of test cases with those from a calculation method that can be assumed as precise, such as the numerical integration method of Genz (1992, 1993) which we already used for this purpose in previous work.

A number of examples, each replicating a small network with 3 alternative paths between the only OD pair and 6 departure time slices have been designed. We have considered independent and partially correlated routes as well as departure times perceived as independent and as partially correlated. Since we are investigating a network loading, we have assumed the values for the options' costs rather than obtaining them from the traffic model as will be the case in the final DTA model. The route cost patterns considered include a flat cost pattern as well as patterns with variations of the costs over time slices and paths.

The results obtained so far are only preliminary both because of the limited scope of the networks investigated and because, due to the structure of the covariance matrix, we have not been able to carry out the calculations with the numerical integration method to obtain the reference figures. In fact, the covariance matrix for the whole path-departure time set as defined above results positive semi-definite rather than positive definite as should be for input in the numerical integration program we are using.

Since the semi-definiteness of the matrix indicates that its rank is smaller than its dimension, the problem should be reformulated so as to be defined with a full rank covariance matrix. A method to perform this in relation to the numerical integration calculations has been suggested by Genz and Kwong (2000).

While we need to address this issue before obtaining conclusive results, it should be noted that the approximation methods of Clark and of Mendell and Elston do not seem to be affected, at least in their mechanism, by the semi-definiteness of the covariance matrix.

The data obtained so far show that the two ways of finding the departure pattern by using only conditional probabilities give consistent results on flat cost profiles with all the methods investigated: Clark, Mendell-Elston and numerical integration. Inconsistencies, however, arise when systematic costs vary by option and it is interesting to note that they are most often of similar percentage importance when using the three alternative MNP calculation methods, thus suggesting that they are due to the departure pattern calculation method rather than to the approximations. While the inconsistencies observed suggest that this method is likely to return imprecise results, its final assessment requires evaluating the magnitude of the inaccuracies.

Similarly, the third method suggested above, which imposes a choice-hierarchy and requires using the approximation of Clark, returns consistent results from either possible hierarchy as long as the systematic cost profiles are flat while different results are obtained for other cost profiles.

We also noted from the results obtained so far that, when the cost profile is not flat, there are also disagreements between the choice probabilities returned from the approximation of Mendell and Elston applied to the whole choice set and those from its application to the method using only conditional probabilities.

# 7 Summary and further model development

The work reported in this paper sets the basis for the development of a Dynamic Stochastic User Equilibrium model which builds on previous results on static traffic assignment. The aim to capture options' similarity correctly when considering a tripmaker's choice between alternative paths and departure time combinations has lead us to specify a multinomial probit model for the determination of the departure pattern from each origin. As a result of this specification the first building block of the model, the Dynamic Stochastic Network loading, requires the determination of probit choice probabilities for large choice sets. In view of developing an efficient algorithm for finding the equilibrium flow pattern, we considered obtaining such

probit probabilities with analytical approximation methods. However, besides applying the approximations to the whole choice set –which may prove impractical- we have considered ways of dividing the choice sets into smaller ones for the purpose of improving the precision of the calculations. As, at the current stage of work, the merits of the alternative methods could not be assessed, further work will aim at producing reference values and carry out investigations on the accuracy of the methods discussed. In parallel to characterising a suitable network loading method, we plan to develop a formulation for the overall DTA model described in the paper and a related solution algorithm.

# 8 References

Clark C.E. (1961) The greatest of a finite set of random variables. *Operations Research*, **9**, 145-162.

Daganzo C.F. (1994) The cell transmission model: a simple dynamic representation of highway traffic. *Transportation Research B*, **28**, 269-287.

Daganzo C.F. (1995) The cell transmission model, part II: network traffic. *Transportation Research B*, **29**, 79-93.

Daganzo C.F. and Sheffi Y. (1977) On stochastic models of traffic assignment. *Transportation Science*, **11**(3), 253-274.

Genz A. (1992) Numerical Computation of multivariate Normal probabilities. *Journal of Computational and Graphical Statistics*, **1**, 141-149.

Genz A. (1993) Comparisons of methods for the computation of multivariate normal probabilities. *Computing Science and Statistics*, **25**, 400-405.

Genz A. and Kwong K-S (2000) Numerical evaluation of singular multivariate Normal distributions. *Journal of Statistical Computation and Simulation*, **68**, 1-21.

Han S. (2003) Dynamic traffic modelling and dynamic stochastic user equilibrium assignment for general road networks. *Transportation Research B*, **37**(3), 225-249.

Kamakura W.A. (1989) The estimation of multinomial probit models: a new calibration algorithm. *Transportation Science*, **23**(4), 253-265.

Maher M.J. (1992) SAM - A stochastic assignment model. In: *Mathematics in Transport Planning and Control* (ed. J.D. Griffiths), Oxford University Press, 121-132.

Maher M.J. and Hughes P.C. (1997) A probit-based stochastic user equilibrium assignment model. *Transportation Research B*, **31**(4), 341-355.

Maher M.J. and Rosa A. (2006). A probit model for departure time choice. Submitted for presentation at the 11th Meeting of the EURO Working Group on Transportation.

Mc Fadden D., Talvitie A. and Associates (1977) Urban Travel Demand Forecasting Project. Phase 1 Final Report Series. Volume V, Part III. The Institute of Transportation Studies. University of California, Berkeley and Irvine.

Mendell N.R. and Elston R.C. (1974) Multifactorial qualitative traits: genetic analysis and prediction of recurrence risks. *Biometrics*, **30**, 41-57.

Ran B. and Boyce D. (1996) *Modeling Dynamic Transportation Networks. An Intelligent Transportation System Approach*. Second Revised Edition. Springer.

Rosa A. (2003) *Probit based methods in traffic assignment and discrete choice modelling*. PhD thesis. School of the Built Environment, Napier University, Edinburgh.

Rosa A. and Maher M.J. (2002) Algorithms for solving the probit path-based stochastic user equilibrium traffic assignment problem with one or more user classes. In: *Transportation and Traffic Theory in the 21st Century. Proceedings of the 15th International Symposium on Transportation and Traffic Theory* (ed. M.A.P. Taylor), Pergamon Press. 371-392.

Sheffi Y. (1985) *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. Prentice-Hall, Englewood Cliffs, New Jersey.

Small K.A. (1982) Scheduling of consumer activities: work trips. *The American Economic Review*, **72** (3), 467-479.

Szeto W.Y. and Lo H.K. (2004) A cell-based simultaneous route and departure time choice model with elastic demand. *Transportation Research B*, **38**, 593-612.

Tang L.K. and Melchers R.E. (1987) Improved approximation for multinormal integral. *Structural Safety*, **4**, 81-93.

# ANALYSIS OF DYNAMIC SYSTEM OPTIMUM AND EXTERNALITIES WITH DEPARTURE TIME CHOICE

*Andy H. F. Chow: University College London, England. andy@transport.ucl.ac.uk*

## Abstract

This paper aims to analyse the dynamic system optimal assignment with departure time choice, which is an important, yet underdeveloped area. The main contribution of this paper is the necessary conditions and the sensitivity analysis for dynamic system optimizing flow. Following this, we revisit the issue of dynamic externality in a more plausible way. We showed that how the externality can be derived and interpreted from the control theoretic formulation and the sensitivity analysis of traffic flow. To solve the system optimal assignment, we propose a dynamic programming solution approach. We present numerical calculations and discuss the characteristics of the results. In particular, we contrast the system optimal assignment with its equilibrium counterpart in terms of the amount of travel generated, flow profiles, and travel costs.

## 1 Introduction

This paper aims to analyse the dynamic system optimal assignment with departure time choice, which is an important, yet underdeveloped area. The dynamic system optimal assignment process suggests that there is a central "system manager" to distribute network traffic over time within a fixed horizon. Consequently, the total, rather than individual, travel cost of all travellers through the network is minimised.

The travel cost incurred by each traveller is considered to have three distinct components: time-specific costs associated with the departure time of the traveller from the origin, and the arrival time at the destination respectively; and a cost related to the travel time *en route*. Given the assigned network flow, the associated travel times through the network are determined by a traffic model. This paper uses the linear whole-link traffic model proposed by Friesz et al. (1993), who considered the travel time on each link to be a linear non-decreasing function of whole link traffic. This traffic model satisfies the principles of flow conservation, proper flow propagation (i.e. consistency between flows and travel times), non-negativity of flow, first-in-first-out (FIFO), and causality. Detailed discussion of this traffic model can be referred to Mun (2001).

This paper starts with introducing the formulation of dynamic system optimal assignment, which is an optimal control problem with state-dependent response. Following Friesz et al. (2001), the optimality conditions for this special kind of control problem can be derived using the calculus of variations technique. At optimality, traffic is assigned such that the total system travel cost is minimized. To solve the dynamic system optimization, information on the sensitivity of the value of the objective function with respect to the control variable is necessary. Section three illustrates a novel sensitivity analysis of travel time and travel cost with respect to perturbations in inflow. Section four then shows a dynamic-programme algorithm for solving dynamic system optimal assignment. Numerical calculations are given in section five. Finally, concluding remarks are given in section six.

## 2 Formulation of dynamic system optimal assignment

The system optimal assignment with departure time choice for fixed travel demand can be formulated as the following optimal control problem. The optimization problem (1) minimizes the total system travel cost within the planning horizon given a predefined amount of total throughput:

$$\min_{e_a^*(s)} Z = \sum_{\forall a} \int_0^T C_a(s) e_a(s) ds \tag{1a}$$

subject to:

$$g_a[\tau_a(s)]\frac{d\tau_a(s)}{ds} = e_a(s) \quad ,\forall a,\forall s \tag{1b}$$

$$\frac{dx_a(s)}{ds} = e_a(s) - g_a(s) \quad ,\forall a,\forall s \tag{1c}$$

$$\frac{dE_a(s)}{ds} = e_a(s) \quad ,\forall a\,,\forall s \tag{1d}$$

$$\sum_{\forall a} E_a(T) = J_{od} \tag{1e}$$

$$e_a(s) \geq 0 \quad ,\forall \mathrm{a},\forall s \tag{1f}$$

We consider the total travel cost $C_a(s)$ encountered by each traveller on the travel link has three distinct components. The first component is the time spent on travelling along the link, which is determined by the travel time model embedded. In addition to the travel time, we add a time-specific cost $f[\tau_a(s)]$ associated with arrival time $\tau_a(s)$ at the destination. Finally, we add a time-specific cost $h(s)$ associated with departure from the origin at time *s*. Possible choices of these time-specific cost functions are investigated by Heydecker and Addison (2005). Consequently, the total travel cost $C_a(s)$ associated with departure on link *a* at time $s$ is determined as a linear combination of these costs as

$$C_a(s) = h(s) + [\tau_a(s) - s] + f[\tau_a(s)]. \tag{2}$$

The notation $\tau_a(s)$ denotes the exit time from the link *a* for traffic which enters at time *s*. For Friesz's (1993) linear whole-link traffic model, $\tau_a(s)$ takes the following functional form:

$$\tau_a(s) = s + \phi_a + \frac{x_a(s)}{Q_a}, \tag{3}$$

where the amount of whole link traffic at time *s* is represented by $x_a(s)$. The free flow travel time and the capacity of the travel link are denoted by $\phi_a$ and $Q_a$ respectively.

Equations (1b) ensure the proper flow propagation along each route. Equations (1c) are the state equations that govern the evolution of link traffic. Equations (1d) define the cumulative inflow $E_a(s)$. Equation (1e) specifies the amount of total throughput $J_{od}$ generated in the system within the time horizon *T*. Conditions (1f) ensure the positivity of the control variable. Since Friesz's (1993) traffic model has been shown to satisfy FIFO structurally (Mun, 2001), we do not need to add any explicit constraint for this.

The optimality conditions for the optimization problem (1) can be derived as

$$e_a(s)\begin{cases} > 0 \Rightarrow \left.\dfrac{\partial Z}{\partial u}\right|_s + \lambda_a(s) - \lambda_a[\tau_a(s)] = \mu_a(s) = \nu \\ = 0 \Rightarrow \left.\dfrac{\partial Z}{\partial u}\right|_s + \lambda_a(s) - \lambda_a[\tau_a(s)] \geq \mu_a(s) = \nu \end{cases} ,\forall s \in [0,T], \tag{4}$$

where $\mu_a(s) = \nu$ is a constant of time and its magnitude is determined by the predefined amount of throughput. The derivative $\left.\frac{\partial Z}{\partial u}\right|_s$ represents the sensitivity of the value of the objective function with respect to a perturbation *u* in the profile of inflow at time *s*, where

$$\left.\frac{\partial Z}{\partial u}\right|_s = \left.\frac{\partial}{\partial u}\left[\int_0^T C_a(t)e_a(t)dt\right]\right|_s = C_a(s) + \int_0^T \left.\frac{\partial C_a}{\partial u}\right|_s e_a(t)dt \quad . \tag{5}$$

The quantity $\left.\frac{\partial Z}{\partial u}\right|_s$ indeed can also be interpreted as the marginal contribution of adding an additional traffic to the link to the total travel cost on this link. It is the sum of two components: $C_a(s)$ is the travel time experienced by that additional traveller given the current traffic condition; $\int_0^T \left.\frac{\partial C_a}{\partial u}\right|_s e_a(t)dt$ is the additional travel cost, which is also known as externality, added by this traveller to each of the existing travellers. Understanding the nature of this externality is important in managing dynamic network, and it requires knowing $\left.\frac{\partial C_a}{\partial u}\right|_s$, which is analysed in Section 3.

Furthermore, the costate variable $\lambda_a(s)$ is determined as:

$$\lambda_a(s) = \frac{1}{Q_a}\int_{t=s}^{T}(1 + f'[\tau_a(t)])e_a(t)dt \, . \tag{6}$$

This costate variable $\lambda_a(s)$ represents the sensitivity of the value of the objective function wih respect to the changes in state variable $x_a(s)$. In other words, the costate variable $\lambda_a(s)$ is interpreted as the marginal travel cost of increasing the link traffic volume by one unit. The details of derivation of this set of optimality conditions can be found in the full version (Chow, 2006).

## 3 Sensitivity analysis

In this section, we start with establishing an expression for the derivatives of the time of exit from a link with respect to a parameter of the inflow profile. Following this, the externality with respect to additional traffic can be derived.

Consider the expression of the whole link traffic, $x_a(s)$, it can be written alternatively as

$$x_a(s) = E_a(s) - G_a(s) = E_a(s) - E_a[\sigma_a(s)] = \int_{t=\sigma_a(s)}^{s} e_a(t)dt \, , \tag{7}$$

in which $\sigma_a(s)$ is the time of entry to the link that leads to exit at time *s*. The expression for the time of exit in (3) then becomes

$$\tau_a(s) = s + \phi_a + \frac{1}{Q_a} \int_{t=\sigma_a(s)}^{s} e_a(t)dt \,. \tag{8}$$

A perturbation *u* in the profile of inflow $e_a(s)$ induces a change in the time of exit as

$$\begin{aligned} \left.\frac{d\tau_a}{du}\right|_s &= \frac{d}{du}\left( s + \phi_a + \frac{1}{Q_a} \int_{t=\sigma_a(s)}^{s} e_a(t)dt \right) \\ &= \frac{1}{Q_a}\frac{d}{du}\left( \int_{t=\sigma_a(s)}^{s} e_a(t)dt \right) \\ &= \frac{1}{Q_a}\left\{ \int_{t=\sigma_a(s)}^{s} \frac{de_a(t)}{du}dt - \frac{d\sigma_a(s)}{du} e_a[\sigma_a(s)] \right\} \end{aligned} \,. \tag{9}$$

The first term is the bracket can be calculated directly. To determine the second term in (9), we first apply the definitional relationship,

$$\tau_a[\sigma_a(s)] = s \,, \tag{10}$$

and using chain rule implies

$$\left.\frac{d\tau_a}{du}\right|_{\sigma_a(s)} = \left.\frac{\partial\tau_a}{\partial u}\right|_{\sigma_a(s)} + \frac{\partial\tau_a[\sigma_a(s)]}{\partial\sigma_a(s)}\frac{\partial\sigma_a(s)}{\partial u} \,. \tag{11}$$

However, at the same time we note that

$$\left.\frac{d\tau_a}{du}\right|_{\sigma_a(s)} = \frac{ds}{du} = 0 \,, \tag{12}$$

since *s* is fixed with respect to perturbation *u*.

Hence,

$$\left.\frac{\partial\tau_a}{\partial u}\right|_{\sigma_a(s)} + \frac{\partial\tau_a[\sigma_a(s)]}{\partial\sigma_a(s)}\frac{\partial\sigma_a(s)}{\partial u} = 0 \,. \tag{13}$$

Furthermore,

$$\begin{aligned} &\frac{d\tau_a[\sigma_a(s)]}{ds} = \frac{\partial\tau_a[\sigma_a(s)]}{\partial\sigma_a(s)}\frac{\partial\sigma_a(s)}{\partial s} = \frac{ds}{ds} = 1 \\ \Rightarrow &\frac{\partial\tau_a[\sigma_a(s)]}{\partial\sigma_a(s)} = \frac{1}{\dfrac{\partial\sigma_a(s)}{\partial s}} \end{aligned} \,. \tag{14}$$

Therefore,

$$\frac{\partial \sigma_a(s)}{\partial u} = -\left(\frac{\partial \tau_a[\sigma_a(s)]}{\partial \sigma_a(s)}\right)^{-1} \left.\frac{\partial \tau_a}{\partial u}\right|_{\sigma_a(s)} = -\frac{d\sigma_a(s)}{ds}\left.\frac{\partial \tau_a}{\partial u}\right|_{\sigma_a(s)}. \tag{15}$$

Thus,

$$\begin{aligned}\left.\frac{\partial \tau_a}{\partial u}\right|_s &= \frac{1}{Q_a}\left\{\int_{t=\sigma_a(s)}^{s} \frac{de_a(t)}{du}dt - \frac{d\sigma_a(s)}{du}e_a[\sigma_a(s)]\right\} \\ &= \frac{1}{Q_a}\left\{\int_{t=\sigma_a(s)}^{s} \frac{de_a(t)}{du}dt + e_a[\sigma_a(s)]\frac{d\sigma_a(s)}{ds}\left.\frac{\partial \tau_a}{\partial u}\right|_{\sigma_a(s)}\right\}. \\ &= \frac{1}{Q_a}\left\{\int_{t=\sigma_a(s)}^{s} \frac{de_a(t)}{du}dt + g_a(s)\left.\frac{\partial \tau_a}{\partial u}\right|_{\sigma_a(s)}\right\}\end{aligned} \tag{16}$$

The derivative of exit time with respect to the perturbation *u* is then expressed in terms of the dependence of the inflow profile $e_a(s)$ in which *s* lies between *s* and $\sigma_a(s)$, the current outflow $g_a(s)$, and the derivative of exit time at the time of entry, $\sigma_a(s)$.

When the analysis is implemented in computer, calculating $\frac{d\tau_a(s)}{du}$ requires knowing the value of $\left.\frac{d\tau_a}{du}\right|_{\sigma_a(s)}$, in which $\sigma_a(s)$ usually is not an integer. Therefore, a linear interpolation is needed to determine the value of $\left.\frac{d\tau_a}{du}\right|_{\sigma_a(s)}$ as

$$\begin{aligned}&\frac{\left.\frac{d\tau_a}{du}\right|_{\sigma_a(s)} - \left.\frac{d\tau_a}{du}\right|_{\lfloor\sigma_a(s)\rfloor}}{\sigma_a(s) - \lfloor\sigma_a(s)\rfloor} = \frac{\left.\frac{d\tau_a}{du}\right|_{\lceil\sigma_a(s)\rceil} - \left.\frac{d\tau_a}{du}\right|_{\lfloor\sigma_a(s)\rfloor}}{\lceil\sigma_a(s)\rceil - \lfloor\sigma_a(s)\rfloor}, \\ &\Rightarrow \left.\frac{d\tau_a}{du}\right|_{\sigma_a(s)} = \left.\frac{d\tau_a}{du}\right|_{\lfloor\sigma_a(s)\rfloor} + \left(\left.\frac{d\tau_a}{du}\right|_{\lceil\sigma_a(s)\rceil} - \left.\frac{d\tau_a}{du}\right|_{\lfloor\sigma_a(s)\rfloor}\right)(\sigma_a(s) - \lfloor\sigma_a(s)\rfloor)\end{aligned} \tag{17}$$

where the notation $\lceil\sigma_a(s)\rceil$ represent the smallest integer not smaller than $\sigma_a(s)$, and $\lfloor\sigma_a(s)\rfloor$ is the greatest integer not larger than $\sigma_a(s)$.

After deriving the sensitivity of travel time with respect to inflow, the sensitivity of the objective function with respect to inflow can be deduced as

$$\left.\frac{\partial Z}{\partial u}\right|_s = \frac{\partial}{\partial u}\left[\int_0^T C_a(t)e_a(t)dt\right]\Bigg|_s$$
$$= C_a(s) + \int_0^T \left.\frac{\partial C_a}{\partial u}\right|_s e_a(t)dt \quad . \tag{18}$$
$$= C_a(s) + \int_0^T (1+ f'[\tau_a(t)])\left.\frac{\partial \tau_a}{\partial u}\right|_s e_a(t)dt$$

# 4 Solving dynamic system optimum

We propose the following procedure to solve for the dynamic system optimal assignment with fixed travel demand:

*Step 0: Initialisation*

0.1. Guess an initial equilibrium cost $C_{od}^*$;

0.2 set the overall iteration counter $n := 1$;

0.3 set $e_a(k) := 0$ for all links *a*, $a \in [1, A]$, and all times *k*, $k \in [1, K]$. The notation $e_a(k)$ represents the assigned inflow to link *a* between times $k\Delta s$ and $(k+1)\Delta s$. The total number of simulated time steps is denoted as $K = T/\Delta s$ and the total number of parallel links is denoted by *A*; set time index $k := 0$;

0.4 set costates $\lambda_a(k) := 0$ for all times $k \in [1, K]$;

0.5 set the link index $a := 1$;

0.6 set the time index $k := 0$;

0.7 set the overall iteration counter $n^i := 1$.

*Step 1: network loading*

Find $\tau_a(k+1)$ by loading the travel link using the inflow $e_a(k)$ at the current iteration.

*Step 2: equilibrating*

2.1 Calculate $C_a(k+1) = h(k+1) + [\tau_a(k+1) - (k+1)] + f[\tau_a(k+1)] + \lambda_a(k+1) - \lambda_a[\tau_a(k+1)]$;

2.2 calculate $\Omega = \dfrac{C_a(k+1) - C_a(k)}{\Delta s}$ and $\Omega' = \dfrac{\partial \Omega}{\partial e_a(k)} = (1+ f'[\tau_a(k+1)])\dfrac{1}{Q_a}$, in which $f'[\tau_a(k)] = \dfrac{f[\tau_a(k+1)] - f[\tau_a(k)]}{\tau_a(k+1) - \tau_a(k)}$;

2.3 update $e_a(k) := e_a(k) - \pi d$ with the second-order searching direction $d = \Omega / \Omega'$ and the step size $\pi$, which is interpolated linearly as

$$\pi = \frac{C_{od}^* - C_a^0(k+1)}{C_a^1(k+1) - C_a^0(k+1)},$$

where $C_a^1(k+1)$ and $C_a^0(k+1)$ represent the corresponding values of $C_a(k+1)$ when $e_a(k)$ is being updated with $\pi$ is taken as 1 and 0 respectively. To determine $\pi$, two network loadings are required.

*Step 3: Calculating costate variables*

3.1 Compute $\lambda_a(k) = \lambda_a(k+1) + (1 + f'[\tau_a(t)])\frac{e_a(t)}{Q_a}\Delta s$ ;

3.2 calculate $\lambda_a[\tau_a(k)]$ from $\lambda_a(k)$ and $\tau_a(k)$.

*Step 4: Convergence verification*

4.1. Check if $\left|C_a(s) - C_{od}^*\right| \le \varepsilon$ or $n^i$ is greater than the predefined maximum number of inner iterations, then go to step 3.2; otherwise, set $n^i := n^i + 1$ and go to step 1.1.

4.2. if $k = K$, then go to step 3.3; otherwise *k*:= *k* + 1 and go to step 1.1;

4.3. if $a = A$, then go to step 3.4; otherwise *a*:= *a* + 1 and go to step 0.5;

4.4 define $\xi = \frac{\sum_{k\in K}\sum_{a\in A} e_a(k)\left|C_a(k+1) - C_{od}^*\right|}{\sum_{k\in K}\sum_{a\in A} e_a(k)C_{od}^*}$ as the measure of disequilibrium, which is equal to zero at equilibrium. If $n$ is greater than the predefined maximum number of overall iterations or $\xi$ is sufficiently small, i.e. $\xi \le \varepsilon$ where $\varepsilon$ is an arbitrarily small number, then go to step 3.2; otherwise set *n*:=*n*+1 and go to step 1.2;

4.5. check if the total throughput $E_{od} = \sum_{\forall a}\sum_{\forall k} e_a(k)$ from the system is equal to the predefined total demand $J_{od}$ for the *o*-*d* pair. If yes, then *terminate* the algorithm; otherwise update $C^* := C^* + \left[\frac{J_{od} - E_{od}}{\frac{dE_{od}}{dC^*}}\right]$, and go back to step 0.3. The derivative $\frac{dE_{od}}{dC^*}$ is derived by Heydecker (2002) as

$$\frac{\partial E_{od}}{\partial C^*} = \sum_{a\in A}\left(\frac{\left[h'(s_a^0) + f'[\tau_a(s_a^0)]\right] - \left[h'(s_a^1) + f'[\tau_a(s_a^1)]\right]}{\left[h'(s_a^0) + f'[\tau_a(s_a^0)]\right]\left[h'(s_a^1) + f'[\tau_a(s_a^1)]\right]}\right)Q_a .$$

## 5 Numerical calculations

To illustrate the analyses above, we demonstrate some numerical calculations. We consider a single link, which has a free flow time 3 mins and a capacity 20 vehs/min, connecting a single origin-destination pair. The origin-specific cost is considered to be a monotone linear function of time with a slope -0.4. The destination cost function is piecewise linear, which has no penalty for arrivals before the preferred arrival time $t^* = 50$, and increases with a rate 2 afterwards. The size of discretized time interval $\Delta s$ is taken as 1 min. The length of the planning horizon $[0, T]$, where *T*=100, is set such that that all traffic can be cleared by time *T*. The total amount of traffic $J_{od}$ is taken as 390 vehs. Figure 1 plots the inflow, the outflow, and the total cost at dynamic user equilibrium. The traffic is assigned to the link during times 18 and 49.

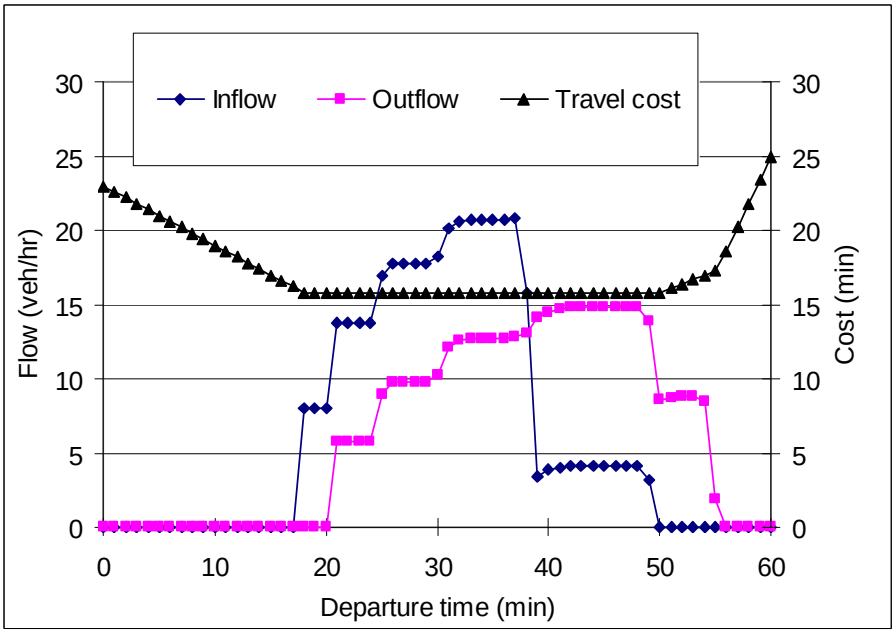



Figure 1 Dynamic user equilibrium assignment

To investigate the accuracy of the sensitivity analysis in Section 3, we suppose the inflow is perturbed at time 18, and plot the associated variations in travel time in Figure 2. The variations are calculated according to (16). In the same figure, we also plot the variations determined by using numerical finite difference method. To calculate the finite difference, we first increase the inflow at time 18 by one unit, and keep the values of other inflows at other times unchanged. The variations in travel times are then calculated by repeated link loading with the original inflow profile versus the perturbed inflow profile. The result shows that the analytical variations given by (16) can represent the true variations in travel time reasonably well. It can be observed that the variations take the value of $1/Q_a$ during time $s$ and $\tau_a(s)$ (i.e during times 19 and 21), and then depend on the profile of outflow and previous variations after $\tau_a(s)$.

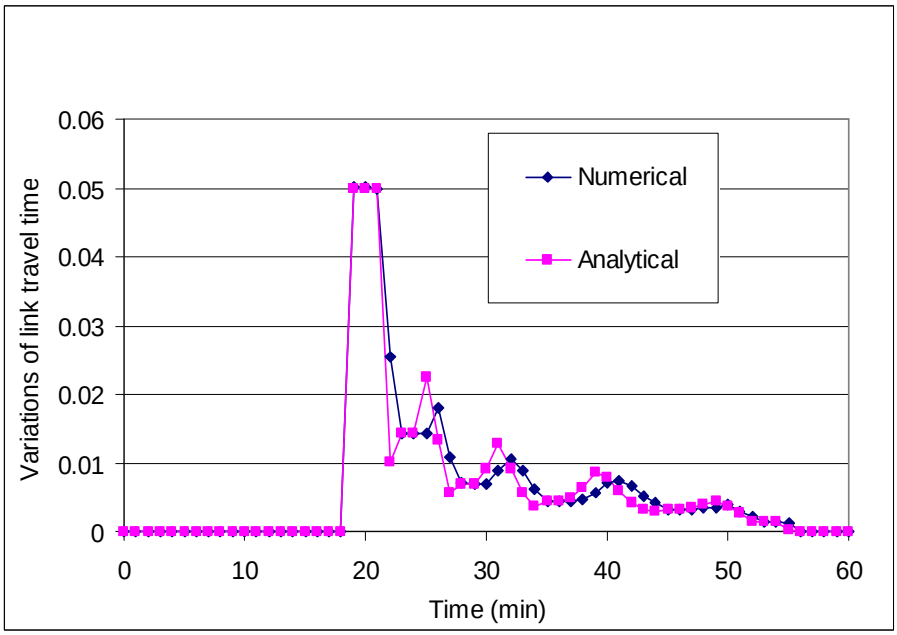



Figure 2 Sensitivity of travel time with respect to a perturbation in inflow

Using the derivative of travel time with respect to inflow, the externality, i.e.

$$\int_0^T \left.\frac{\partial C_a}{\partial u}\right|_s e_a(t)dt = \int_0^T \left(1 + f'[\tau_a(t)]\right)\left.\frac{\partial \tau_a}{\partial u}\right|_s e_a(t)dt,$$

induced by adding an additional inflow at all times can then be calculated accordingly. Figure 3 shows the profile of the externality for inflow under dynamic user equilibrium.

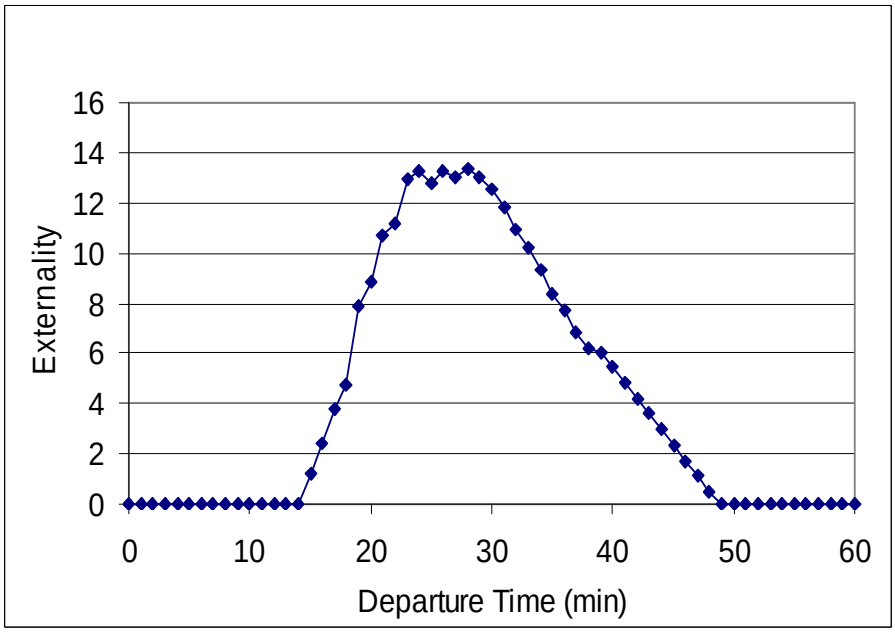



Figure 3 Externality at dynamic user equilibrium

Calculating the dynamic system optimal assignment is still in progress. Figure 4 shows the inflow, the outflow, and the travel cost after one iteration of optimization from the dynamic user equilibrium. With the same total throughput $J_{od}$, the period of assignment shifts from times [18, 49] to times [9, 50]. It is also observed that this assignment, on the one hand, encourages late departures. On the other hand, it also has to maintain a certain amount of early departures to induce a high service rate for the departures at later times. The total system travel cost is decreased from 6,143.45 veh-hr in user equilibrium to 5,777.60 veh-hr. This assignment profile is still subject to further revision.

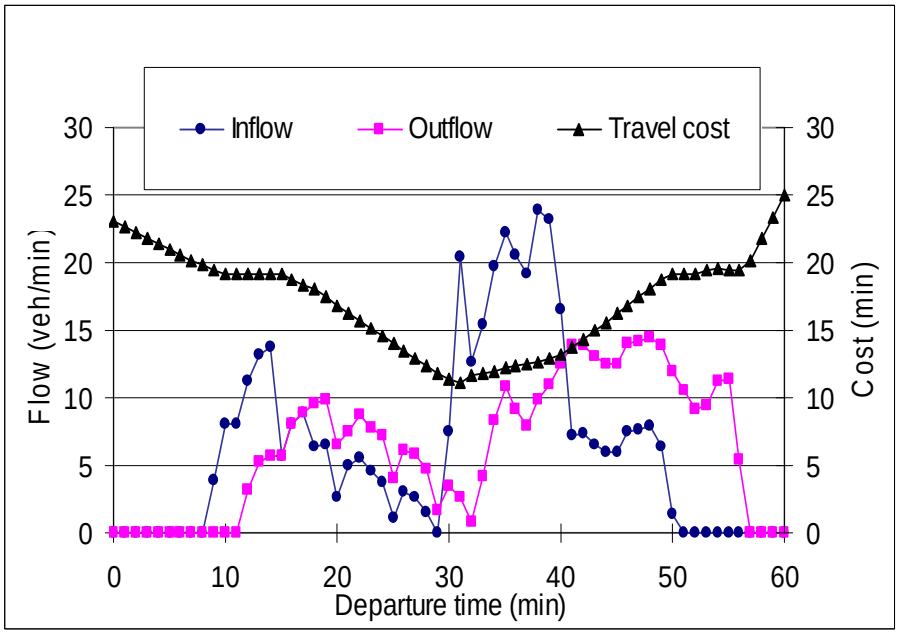



Figure 4 Solving dynamic system optimal assignment

# 6 Concluding remarks

This paper analyses the dynamic system optimizing flow along a single travel link. We propose a novel sensitivity analysis of travel time and travel cost with respect to perturbations in inflow. We also presented a solution method using the dynamic programming approach and applied it to the numerical example. The characteristics of the results were discussed.

The main contribution of this chapter is the necessary conditions and the sensitivity analysis for dynamic system optimizing flow. The investigation also gives us a deeper understanding of the nature of system optimal assignment problems. In addition to analyzing and solving the system optimizing flow, we also note that each additional traveller, who enters the system at a certain time, imposes an additional travel cost on the others who enter the system at that time and thereafter. We regard this additional cost as "externality". Understanding the nature of the externality is important in managing dynamic network. However, previous research is implausible due to the underlying traffic model adopted (Carey and Srinvasan, 1993; Yang and Huang, 1997). This paper revisited the dynamic externality in a more plausible way. We also showed that how the externality can be derived and interpreted from the control theoretic formulation and the sensitivity analysis. This paper considered single-link networks in which only the departure time choices of travellers are considered. We are currently extending the present analysis and discussion to multi-route and multi-destination networks in which the joint choices of departure time and travel route of travellers are investigated.

# Acknowledgement

I would like to thank Professor Ben Heydecker and Dr. JD Addison for their continuing encouragement and supervision of this study. Earlier versions were presented at a series of joint Japan-UK workshops on dynamic traffic assignment, comments from Professor Mike Smith, Professor Richard Allsop, Professor David Watling, and Dr. Simon Shepherd are gratefully appreciated.

# References

Astarita, A (1996) A continuous time link model for dynamic network loading based on travel time functions. J-B Lesort, ed. *Transportation and Traffic Theory*. Pergamon, Oxford, 79-102.

Carey M. and Srinivasan, A. (1993) Externalities, average and marginal costs, and tolls on congested networks with time-varying flows. *Operations Research* **41**, 217-231.

Chow, AHF (2006) Analysis of dynamic system optimum and externalities with departure time choice. *Paper to be presented at the 1st International Symposium on Dynamic Traffic Assignment* (full version), 21 – 23 June. Leeds, England.

Friesz, TL, Bernstein, D, Smith, TE, Tobin, RL and Wie BW (1993) A variational inequality formulation of the dynamic network user equilibrium problem. *Operations Research* **41**(1), 179-91.

Friesz, TL, Bernstein, D, Suo, Z and Tobin, RL (2001) Dynamic network user equilibrium with state-dependent time lags. *Network and Spatial Economics*, **1**(3/4), 319-347.

Heydecker, BG (2002) Dynamic equilibrium network design. MAP Taylor, ed. *Transportation and Traffic Theory*. Pergamon, Oxford, 349-70.

Heydecker, BG and Addison, JD (2005) Analysis of dynamic traffic equilibrium with departure time choice. *Transportation Science,* **39**(1), 39–57.

Mun, JS (2001) A divided linear travel time model for dynamic traffic assignment. *9th World Conference on Transport Research*, Seoul, S.Korea.

Yang, H. and Huang, H.J. (1997) Analysis of the time-varying pricing of a bottleneck with elastic demand using optimal control theory. *Transportation Research* **31B**, 425-440.

# THE EXISTENCE, UNIQUENESS, STABILITY AND BILEVEL OPTIMISATION OF DYNAMIC TRAFFIC EQUILIBRIA

**Yann Golanski, Richard Mounce and Michael Smith**
**Mathematics Dept, University of York, Heslington, York, YO10 5DD, UK**
**Abu Dhabi University, Abu Dhabi, United Arab Emirates**
**E-mails: yg2@york.ac.uk , richardmounce@hotmail.com, mjs7@york.ac.uk**

## ABSTRACT

The paper considers day-to-day dynamic traffic assignment models with deterministic queueing and inelastic demands, and no departure-time choice. Drivers swap (in day to day time) from more to less costly routes at each within day time.

Part A of the paper proves that under natural conditions this day-to-day model has a nonempty convex set of equilibria; and also, by using a Lyapunov function, that the day-to-day dynamical system is stable when certain strong conditions hold.

Part B considers a similar day-to-day dynamic model but on a very simple two-route network. It is again shown that the day-to-day dynamical system is stable. Finally we consider optimisation subject to dynamic equilibrium: *dynamic bilevel optimisation*. We show how in our very simple case gradients of the natural disequilibrium function may be estimated efficiently; such estimates are vital if we wish to optimise control parameters subject to dynamic equilibrium.

## PART A: GENERAL THEORY OF THE DYNAMIC QUEUEING MODEL

**The model.** In our bottleneck model, queueing occurs (first-in first-out) vertically at the exit of links when flow exceeds exit capacity. Initially within-day time is considered to be a continuous variable, varying throughout the time period [0,1] representing a single day. During each day, the rate at which flow enters route $r$ is represented by a non-negative, measurable real-valued function $X_r$ of within-day time. If there are $N$ routes in the network, then all route inflows will be represented by a vector $X$ with $N$ component functions $X_r$. We suppose that demand is inelastic; so we suppose that, for each OD pair and each within-day time, there is a given rigid (but time varying) rate at which traffic leaves each origin. We suppose that this rate for OD pair $k$ is given by $\rho_k$, a given bounded measurable function of within-day time taking only non-negative values. We suppose that a route inflow vector $X = (X_1, X_2, \ldots X_N)$ always belongs to the feasible set

$$D = \{X : X_r \geq 0, measurable, \sum_{r \in R_k} X_r = \rho_k\}$$

where $R_k$ is the set of routes connecting OD pair $k$.

The cost (in time) of traversing any link in the network is the sum of a constant (congestion-free or free-flow) travel time and a bottleneck delay at the link exit; this delay depends on the whole inflow function $X$. The delay may be zero. To represent the queueing delay which arises on saturated links, we suppose that the bottleneck delay $d_i$ on link $i$ is connected to the bottleneck capacity $s_i$ and the bottleneck inflow $x_i$ by the following integral equation:

$$\int_{\tau_0}^{\tau} x_i(u)du = \int_{\tau_0}^{\tau + d_i^X(\tau)} s_i(u)du \quad \text{for all } \tau \in [\tau_0, \tau_1]$$

where the bottleneck is congested throughout the interval $[\tau_0, \tau_1]$. Route cost functions $C_r$ are found by summing the link costs along the route; but at the time that each bottleneck is reached:

$$C_r(X)(\tau) = \sum_{i:i \in r} c_i^x(P_{ir}^X(\tau))$$

where $P_{ir}^{X}(\tau)$ is the time that bottleneck *i* is reached if route *r* is entered at time $\tau$ and the route inflow vector is *X*. Mounce(2006) shows that such route cost functions are Lipschitz continuous functions of time.

The day-to-day dynamical system is derived naturally from the usual dynamic user equilibrium condition, which is that (for each OD pair and each within-day time) more costly routes are unused. A natural day-to-day swap vector is given by $\phi(X)$ where (for all within day times $\tau$ ):

$$\phi(X)(\tau) = \sum_{r,s:r\sim s} X_r(\tau)[C_r(X)(\tau) - C_s(X)(\tau)]_+ \delta_{rs}$$

where $r \sim s$ means that routes *r* and *s* connect the same OD pair, $[x]_+ = \max\{x, 0\}$ and $\delta_{rs}$ is the swap from route *r* to route *s* vector. We let *t* represent day-to-day time and view this as a continuous variable. We then consider the dynamical system.

$$\frac{dX(t)}{dt} = \phi(X)(t), \quad X(0) = X_0 \qquad (1)$$

where $t \geq 0$ and $X_0$ is any initial route inflow vector in *D*. This dynamical system evolves continuously over day-to-day time with each element being a within-day inflow function giving all inflow rates to all routes at all within-day times.

**Existence of equilibria and convexity of the set of equilibria.** At equilibrium, more costly routes are not used, i.e. $C_r(X)(\tau) > C_s(X)(\tau) \Rightarrow X_r(\tau) = 0$. Therefore *X* is at equilibrium if and only if $\phi(X) = 0$. Smith and Wisten (1995) used Schauder's fixed point theorem to prove existence of equilibrium of dynamical system (1) provided (a) that route cost is continuous (as a function of route inflow), and (b) that the feasible set *D* is convex and compact. It is clear that *D* is convex. Mounce (2006) establishes that the feasible set *D* is compact and Mounce (2005) uses an implicit function theorem to show that the route cost vector is indeed a continuous function of the route flow vector; these results fill gaps in Smith and Wisten (1995).

Here we say that the route cost function $C(X)$ is a monotone function of $X$ if and only if

$$\sum_r \int_0^1 (C_r(X)(\tau) - C_r(Y)(\tau))(X_r(\tau) - Y_r(\tau))d\tau \geq 0$$

for all route inflow vectors *X* and *Y*. In the single bottleneck per route (SBPR) case, each route passes through at most one bottleneck. Smith and Ghali (1990b) showed that, in this SBPR case, route cost is a monotone function of route inflow if link cost is a monotone function of link inflow for every link. Mounce (2006) shows that this is true if and only if the link capacities are all non-decreasing functions of within-day time $\tau$ . Provided that *C(X)* is a monotone function of *X*, the set of equilibria is convex; this is shown in Mounce (2005).

**Stability; or convergence to equilibrium.** The dynamical system (1) is globally convergent if for any starting vector $X_0$ the dynamical system converges to the set of equilibria as $\tau \to \infty$. We can show that this occurs here by giving a Lyapunov function *V(X)* for the dynamical system (1), i.e. a function *V(X)* satisfying:

1. $V(X) \geq 0$
2. $V(X) = 0$ if and only if *X* is at equilibrium.
3. $\frac{dV(X)}{dt} < f(X)$ for all non-equilibrium *X*, where *f* is a continuous function of *X* that is negative for all non-equilibrium *X*. (*t* is omitted from *X* here.)

A proof of this can be found in Mounce (2003).

In our case route cost is a monotone function of route inflow and it follows that if we let, for all $X \in D$:

$$V(X) = \sum_{r,s:r\sim s} \int_0^1 X_r(\tau)(C_r(X)(\tau) - C_s(X)(\tau))_+^2 d\tau ,$$

then $V$ is a Lyapunov function for the dynamical system (1). It is not difficult to see that $V(X)$ is nonnegative and zero if and only if $X$ is at equilibrium. We also have (see Mounce (2006)):

$$\frac{dV(X)}{dt} < -\int_0^1 \sum_{r,s:r\sim s} X_r(\tau)(-C(X)(\tau)\cdot\delta_{rs})_+^3 d\tau .$$

Hence, in this case, the dynamical system converges to equilibrium as $t \to \infty$ and we have stability of our day-to-day dynamical system. ($t$ is omitted from $X$ here.)

# PART B: A VERY SIMPLE TWO-ROUTE NETWORK

*Stability in the steady state, with evolution in day-to-day time "t" is studied first. Then we add within day time "τ" to deal with the dynamical state. Finally bilevel optimisation is considered.*

## STABILITY OF EQUILIBRIA IN THE STEADY STATE

**The network.** 1 OD pair connected by 2 non-overlapping routes. Demand is to be rigid.
**The main variables.**

$X_1$ = flow along route 1;
$X_2$ = flow along route 2; and
$X_1 + X_2$ = 1 and $X_1$, $X_2$ are both to be non-negative.

**Cost functions.**

$C_1(X_1)$ = cost on route 1 if flow is $X_1$
$C_2(X_2)$ = cost on route 2 if flow is $X_2$.

**The rigid demand user-equilibrium condition.**

$$X_1[C_1(X_1) - C_2(X_2)]_+ = 0 \text{ and } X_2[C_2(X_2) - C_1(X_1)]_+ = 0$$

or $X_1[C_1(X_1) - C_2(X_2)]_+ + X_2[C_2(X_2) - C_1(X_1)]_+ = 0.$

**Definition of monotonicity.** A real-valued function f of a real variable is *monotone* if and only if

$[f(x + \delta) - f(x)]\delta \geq 0$ for all real numbers x and δ.

This is the case if:

$f'(x) \geq 0$; or $\delta\cdot[f'(x)\delta] \geq 0$ or $\delta\cdot[f'(x, \delta)] \geq 0$ for all real numbers x, δ.

Here: $f'(x)$ is the derivative of f at x; $f'(x; \delta)$ is the directional derivative of f at x in direction δ.

Thus $C_1$ is monotone if and only if:

$\Delta_1[C_1'(X_1)]\Delta_1 \geq 0$ for all real numbers $X_1$, $\Delta_1$ or: $\Delta_1 C_1'(X_1; \Delta_1) \geq 0$ for all real numbers $X_1$, $\Delta_1$ where $C_1'(X_1)$ is the Jacobian of $C_1$ (or the derivative of $C_1$) at $X_1$ and $C_1'(X_1; \Delta_1)$ is the directional derivative of $C_1$ in direction $\Delta_1$, at $X_1$. Directional derivatives allow greater generality.

**Monotonicity assumption.** We assume that $C_1$ and $C_2$ are both monotone.

**A possible natural day-to-day evolution.** Let $\Delta(X) = (\Delta_1(X), \Delta_2(X))$ where:

$$\Delta_1(X) = (-X_1^2[C_1(X_1) - C_2(X_2)]_+ + X_2^2[C_2(X_2) - C_1(X_1)]_+),$$
$$\Delta_2(X) = (-X_2^2[C_2(X_2) - C_1(X_1)]_+ + X_1^2[C_1(X_1) - C_2(X_2)]_+).$$

$\Delta(X) = 0$ if and only if X is an equilibrium. $\Delta(X)$ swaps flow from more to less costly routes and preserves total flow. These swap rates are similar to natural swap rates introduced in Smith (1984).

We now suppose that $X^0 = (X^0_1, X^0_2)$ is any feasible start point. So that

$$X^0_1 + X^0_2 = 1,\ X^0_1 \geq 0 \text{ and } X^0_2 \geq 0.$$

We also suppose that the route flow vector X(t) evolves as day-to-day time passes by obeying:

$X(0) = X^0$ and $dX(t)/dt = \Delta(X(t))$ for all $t \geq 0$.

**A candidate Lyapunov function or objective function**. Changing Smith (1984) slightly, we let

$$V(X) = X_1^2[C_1(X_1) - C_2(X_2)]_+^2 + X_2^2[C_2(X_2) - C_1(X_1)]_+^2 \text{ for all } X \geq 0.$$

Then V(X) measures the departure of any feasible X from equilibrium; or

$V(X) \geq 0$ and $V(X) = 0$ if and only if X is an equilibrium.

Our X will continually follow $\Delta(X) = (\Delta_1(X), \Delta_2(X))$ as shown above. It is natural to find grad V(X) to see if V(X(t)) declines to zero as time becomes large. So let $J_1(X_1) = C_1'(X_1)$ be the derivative (one-dimensional Jacobian) of $C_1(.)$ at $X_1$, and let $J_2(X_2) = C_2'(X_2)$. Then:

$$\begin{aligned}
\tfrac{1}{2}\,\text{grad}V(X) &= \{X_1[C_1(X_1)-C_2(X_2)]_+^2 + X_1^2[C_1(X_1) - C_2(X_2)]_+J_1(X_1),\ -X_1^2[C_1(X_1) - C_2(X_2)]_+J_2(X_2)\} \\
&\quad + \{-X_2^2[C_2(X_2)-C_1(X_1)]_+J_1(X_1),\ X_2[C_2(X_2)-C_1(X_1)]_+^2 + X_2^2[C_2(X_2) - C_1(X_1)]_+J_2(X_2)\}. \\
&= (X_1[C_1(X_1)-C_2(X_2)]_+^2 - \{X_2^2[C_2(X_2)-C_1(X_1)]_+ - X_1^2[C_1(X_1) - C_2(X_2)]_+\}J_1(X_1), \\
&\quad X_2[C_2(X_2)-C_1(X_1)]_+^2 - \{X_1^2[C_1(X_1) - C_2(X_2)]_+ - X_2^2[C_2(X_2) - C_1(X_1)]_+\}J_2(X_2)) \\
&= (X_1[C_1(X_1)-C_2(X_2)]_+^2 - J_1(X_1)\Delta_1(X),\ X_2[C_2(X_2)-C_1(X_1)]_+^2 - J_2(X_2)\,\Delta_2(X)) \\
&= (X_1[C_1(X_1)-C_2(X_2)]_+^2 - C_1'(X_1;\Delta_1(X)),\ X_2[C_2(X_2)-C_1(X_1)]_+^2 - C_2'(X_2;\Delta_2(X)).
\end{aligned}$$

The last two lines here give gradV in terms of both derivatives and directional derivatives (but strictly speaking the final line only holds where gradV exists; although we make use of this definition for convenience).

**Descent to equilibrium?** Assuming that $C_1$ and $C_2$ are both monotone; away from equilibrium, $\Delta(X)$ is a descent direction for objective V at X and so V(X(t)) declines to 0 as $t \to \infty$. Here we just verify this descent (and we do not prove convergence of V to 0 as $t \to \infty$).

Of course $\Delta(X)$ is not a descent direction for objective V at X if X is already an equilibrium as then $V(X) = 0$ and since $V \geq 0$ always there can be no descent direction from X. <u>So let us now suppose that X is not an equilibrium and so $V(X) > 0$.</u> Then

$$X_1[C_1(X_1) - C_2(X_2)]_+ + X_2[C_2(X_2) - C_1(X_1)]_+ > 0$$

and so $$X_1^3[C_1(X_1) - C_2(X_2)]_+^3 + X_2^3[C_2(X_2) - C_1(X_1)]_+^3 > 0.$$

At any X, as we have seen:

$$\tfrac{1}{2}\,\text{grad}\,V(X) = (X_1[C_1(X_1)-C_2(X_2)]_+^2 - C_1'(X_1;\Delta_1(X)),\ X_2[C_2(X_2)-C_1(X_1)]_+^2 - C_2'(X_2;\Delta_2(X))$$

Now

$$\begin{aligned}
&\Delta_1(X)X_1[C_1(X_1)-C_2(X_2)]_+^2 \\
&\quad = (-X_1^2[C_1(X_1) - C_2(X_2)]_+ + X_2^2[C_2(X_2) - C_1(X_1)]_+)\,X_1[C_1(X_1)-C_2(X_2)]_+^2 \\
&\quad = -X_1^2[C_1(X_1) - C_2(X_2)]_+\,X_1[C_1(X_1)-C_2(X_2)]_+^2 \\
&\qquad\qquad (\text{as } X_2^2[C_2(X_2) - C_1(X_1)]_+\,X_1[C_1(X_1)-C_2(X_2)]_+^2 = 0). \\
&\quad = -X_1^3[C_1(X_1)-C_2(X_2)]_+^3.
\end{aligned}$$

Similarly:

$$\Delta_2(X)X_2[C_2(X_2)-C_1(X_1)]_+^2 = -X_2^3[C_2(X_2)-C_1(X_1)]_+^3.$$

Hence:

$$\begin{aligned}
&\tfrac{1}{2}\,V'(X;\Delta(X)) = \tfrac{1}{2}\,\Delta(X)\cdot\text{grad}V(X) = \tfrac{1}{2}\,(\Delta_1(X), \Delta_2(X))\cdot\text{grad}\,V(X) \\
&= (\Delta_1(X), \Delta_2(X))\cdot(X_1[C_1(X_1)-C_2(X_2)]_+^2 - C_1'(X_1;\Delta_1(X)),\ X_2[C_2(X_2)-C_1(X_1)]_+^2 - C_2'(X_2;\Delta_2(X)) \\
&= \\
&\Delta_1(X)X_1[C_1(X_1)-C_2(X_2)]_+^2 - \Delta_1(X)C_1'(X_1;\Delta_1(X)) + \Delta_2(X)X_2[C_2(X_2)-C_1(X_1)]_+^2 - \Delta_2(X)C_2'(X_2;\Delta_2(X)) \\
&= -X_1^3[C_1(X_1) - C_2(X_2)]_+^3 - X_2^3[C_2(X_2) - C_1(X_1)]_+^3 \quad < 0 \quad (\text{as X is a non-equilibrium}) \\
&\quad - \Delta_1(X)C_1'(X_1;\Delta_1(X)) - \Delta_2(X)C_2'(X_2;\Delta_2(X)) \quad \leq 0 \quad (\text{as } C_1 \text{ and } C_2 \text{ are monotone}) \\
&\quad < 0.
\end{aligned}$$

Thus $\Delta(X)\cdot\text{gradV}(X) < 0$ and $\Delta(X)$ is a descent direction for V at X, provided X is not an equilibrium; and we have shown using the chain rule that (for our dynamical system):

$$dV(X(t))/dt = \text{gradV}(X(t)) \cdot dX(t)/dt = \text{gradV}(X(t)) \cdot \Delta(X(t)) = V'(X;\Delta(X)) < 0 \text{ for all } t \geq 0.$$

It is easy now to show (provided some further continuity holds) that $V(X(t)) \to 0$ as $t \to \infty$ and it is then easy to show that $\text{dist}[X(t), E] \to 0$ as $t \to \infty$, where E is the non-empty set of equilibria. This work follows Smith (1984) with a slightly changed swap rate.

## STABILITY OF THE EQUILIBRIUM SET IN THE DYNAMIC STATE WITH CONTINUOUS WITHIN-DAY TIME

Here we just add "(τ)", representing within-day time, to the previous analysis. The within-day time period will be represented by the interval [0, 1], representing a whole day of 24 hours.

**The network.** There will still be just 1 OD pair connected by just 2 non-overlapping routes. There will still also be a given rigid time-varying demand. But now each of the two routes will have a fixed "free-flow" travel time and also a single bottleneck of fixed capacity at the route exit. Any flow at the route-exit exceeding the capacity at the bottleneck exit is stored in a queue just upstream of the bottleneck, and the queue discharge rate equals the capacity.

**The main variables.**

$X_1$ and $X_2$ are now measurable functions of within-day time, both going from $[0, 1] \rightarrow R_+$.

$X_1(\tau)$ = inflow rate into route 1 at time $\tau$ ($\geq 0$ always);

$X_2(\tau)$ = inflow rate into route 2 at time $\tau$ ($\geq 0$ always) ; and

$[X_1 + X_2](\tau)$ = $\rho(\tau)$ where $\rho$: $[0, 1] \rightarrow R_+$ is a given bounded measurable function.

**The feasible set F.** Given $\rho$: $[0, 1] \rightarrow R_+$ the feasible set F of route inflow vectors X is now:

$F = \{ X = (X_1, X_2)$: $X_1, X_2$ are non-negative measurable functions of within-day time $\tau$
satisfying: $[X_1 + X_2](\tau) = \rho(\tau)$ for all $\tau$ in $[0, 1].\}$

**Cost functions.**

$C_1(X_1)(\tau)$ = travel cost experienced on route 1 if entered at time $\tau$ (if inflow is $X_1$).

$C_2(X_2)(\tau)$ = travel cost experienced on route 2 if entered at time $\tau$ (if inflow is $X_2$).

$C_1(.)$ and $C_2(.)$ will both be assumed to be continuous with relevant metrics and directionally differentiable at each $X_1, X_2$.

In the earlier part of the paper and in Smith and Ghali (1990a) and Mounce (2006) it is shown that provided the capacities are, for each bottleneck, non-decreasing functions of time and provided that queues at link exits store those vehicles which cannot emerge from the bottleneck in the right order then $C_1(X_1)$ is a monotone function of the inflow rate $X_1$ and $C_2(X_2)$ is a monotone function of the inflow rate $X_2$. Here costs are defined to be the sum of a free-flow travel time and a bottleneck queueing delay.

**The rigid demand user-equilibrium condition.** Here we say that X is an equilibrium if:

$$X_1[C_1(X_1) - C_2(X_2)]_+(\tau) = 0 \text{ and } X_2[C_2(X_2) - C_1(X_1)]_+(\tau) = 0 \text{ for all } \tau$$

or: $$\{X_1[C_1(X_1) - C_2(X_2)]_+ + X_2[C_2(X_2) - C_1(X_1)]_+\}(\tau) = 0 \text{ for all } \tau.$$

[Here $X_1[C_1(X_1) - C_2(X_2)]_+(\tau) = 0$ is shorthand for $X_1(\tau)[C_1(X_1)(\tau) - C_2(X_2)(\tau)]_+ = 0$. To reduce the number of times $\tau$ appears we will usually utilise this and similar shorthands.]

**Definition of monotonicity in function space.** A function f from one set of bounded measurable functions on $[0, 1] \rightarrow$ another set of bounded measurable functions on $[0, 1]$ is *monotone* iff

$$[f(x + \delta) - f(x)]\cdot\delta \geq 0$$

for all measurable functions x and $\delta$. For directionally differentiable f, f is monotone if and only if

$$\delta\cdot[f'(x; \delta)] \geq 0 \text{ for all } x, \delta;$$

$f'(x; \delta)$ is the directional derivative of f at x (now a function) in direction $\delta$ (now a function).

Thus $C_1$ is monotone if and only if $\Delta_1\cdot C_1'(X_1; \Delta_1) \geq 0$ for all functions $X_1, \Delta_1$, where $C_1'(X_1; \Delta_1)$ is the directional derivative of $C_1$, at $X_1$, in direction $\Delta_1$.

**Monotonicity of the cost functions.** If $C_1$ is the sum of an uncongested travel time and a single queueing delay then it is monotone and $\Delta_1\cdot C_1'(X_1; \Delta_1) \geq 0$ for all $X_1, \Delta_1$. The same applies to $C_2$.

**A possible natural day-to-day evolution in F.** Following the steady state case above, let

$$\Delta_1(X)(\tau) = \{-X_1^2[C_1(X_1) - C_2(X_2)]_+ + X_2^2[C_2(X_2) - C_1(X_1)]_+\}(\tau)$$

$$\Delta_2(X)(\tau) = \{-X_2^2[C_2(X_2) - C_1(X_1)]_+ + X_1^2[C_1(X_1) - C_2(X_2)]_+\}(\tau)$$

and $$\Delta(X)(\tau) = (\Delta_1(X)(\tau), \Delta_2(X)(\tau)) = (\Delta_1(X), \Delta_2(X))(\tau)$$

Here we have used our shorthand by having only one $\tau$ on each of the first two right hand sides here instead of six. This direction is obtained from the steady-state direction by simply adding "$(\tau)$".

Then $\Delta(X)(\tau) = 0$ for all within day times $\tau$ if and only if X is an equilibrium inflow vector in F. As before moving in direction $\Delta(X)$ preserves the total of the two route-inflows as before, but now at each within-day time $\tau$; it swaps flow from the more costly to the less costly route at $\tau$.

We suppose that $X^0 = (X^0_1, X^0_2)$ is any start point in F. So $X^0$ is measurable, is non-negative,

$$[X^0_1 + X^0_2](\tau) = \rho(\tau),\ X^0_1 \geq 0 \text{ and } X^0_2 \geq 0.$$

We also suppose that the route inflow vector $X(\tau)$ evolves with day-to-day time "t" by obeying:

$$X(\tau)(0) = X^0(\tau) \text{ and } dX(\tau)(t)/dt = \Delta(X(\tau)(t)) \text{ for all } t \geq 0 \text{ and all within day times } \tau \in [0, 1].$$

**A candidate Lyapunov function or objective function.** Again following the steady state case, let

$$V(X) = \int \{X_1^2[C_1(X_1) - C_2(X_2)]_+^2 + X_2^2[C_2(X_2) - C_1(X_1)]_+^2\}(\tau)d\tau$$

for all X in F. Then V(X) measures departure from equilibrium at X; or

$$V(X) \geq 0 \text{ and } V(X) = 0 \text{ if and only if X is an equilibrium.}$$

<u>All integrals in this section of the paper will go from 0 to 1; they go over the whole 24-hour day.</u>

**Descent to equilibrium?** In this case now, where $C_1(.)$ and $C_2(.)$ are both monotone, it follows that, away from equilibrium, $\Delta(X)$ is a descent direction for objective V at X. Here as in the steady state case we just verify this descent (and we do not prove convergence of V to 0 as $t \to \infty$; but this does happen under natural conditions just as in the steady state case; see part A of the paper).

Of course $\Delta(X)$ is not a descent direction for objective V at X if X is already an equilibrium in F as then $V(X) = 0$ and since $V \geq 0$ always there can be no descent direction from X.

So let us now suppose that X is not an equilibrium and so $V(X) > 0$. Then:

$$V(X) = \int \{X_1^2[C_1(X_1) - C_2(X_2)]_+^2 + X_2^2[C_2(X_2) - C_1(X_1)]_+^2\}(\tau)d\tau > 0 \text{ and so also}$$
$$\int \{X_1^3[C_1(X_1) - C_2(X_2)]_+^3 + X_2^3[C_2(X_2) - C_1(X_1)]_+^3\}(\tau)d\tau > 0.$$

Using this, at any feasible non-equilibrium X, just as above in the steady state:

$V'(X; \Delta(X))$ $[= \Delta(X) \cdot \text{grad}V(X) = (\Delta_1(X), \Delta_2(X)) \cdot \text{grad } V(X)$ when gradV exists]

$= -2\int \{X_1^3[C_1(X_1) - C_2(X_2)]_+^3 + X_2^3[C_2(X_2) - C_1(X_1)]_+^3\}(\tau)d\tau$ ($< 0$ as X is a non-equilibrium)

$-2\int \Delta_1(X)(\tau)C_1'(X_1;\ \Delta_1(X))(\tau)d\tau - 2\int \Delta_2(X)(\tau)C_2'(X_2;\ \Delta_2(X))(\tau)d\tau$ ($\leq 0$ by monotonicity)

$< 0$.

Thus we have again used the chain rule and our dynamical system to obtain:

$$dV(X(t))/dt = V'(X(t); \Delta(X(t)) < 0 \text{ for all } t \geq 0.$$

And so $\Delta(X)$ is away from equilibrium a descent direction for V. It now follows, under natural and weak conditions, that $V(X(t)) \to 0$ as $t \to \infty$ and it is then easy to show that $\text{dist}[X(t), E] \to 0$ as $t \to \infty$, where E is the non-empty set of equilibria and dist[X(t), E] is the distance between X(t) and E. (the set E of equilibria is non-empty from part A.) Thus in the simple two-route two-bottleneck case we are considering, the set of dynamic equilibria is stable: non-equilibrium inflow functions X(t) in F naturally converge to the equilibrium set E.

## STABILITY OF THE EQUILIBRIUM SET IN THE DYNAMIC STATE WITH TIME SLICES AND STEP FUNCTION INFLOW RATES

Here we let $\tau$ represent an integer. So $\tau = 1, 2, 3, 4, \ldots, M$; where the 24-hour day is divided into M equal intervals or time-slices. So as to be as specific as possible we think of M as being $1440 = 24 \times 60$ so that each interval or time-slice has a duration of 1 minute. We let W denote the set $\{1, 2, 3, \ldots, M\}$ of all possible within-day times $\tau$. These are the names of the intervals or time-slices into which we have divided the whole day. So there are M time-slices:

$$\text{slice } 1 = [0, 1], \text{ slice } 2 = [1, 2], \ldots, \text{ slice } \tau = [\tau - 1, \tau], \ldots, \text{ slice } M = [M-1, M].$$

representing 24 hours in total. We don't worry about the overlap here. Inflows will now be in vehicles per minute. We imagine all route-entry flow rates being constant over a single time slice.

Therefore the total flow entering a route in time slice $\tau$ = the (constant) rate at which flow enters during time-slice $\tau$.

As previously we deal again with a very special network with just two non-overlapping routes having just one bottleneck.

So as to make this analysis as close to that above as possible we will write an M-vector $(X_1, X_2, X_3, \ldots, X_M)$ as $(X(1), X(2), X(3), \ldots, X(M))$ and so on.

**The network.** The network still has just 1 OD pair connected by 2 non-overlapping routes; and each route has just one bottleneck on it. Again the demand is a rigid time-varying demand.

As before each of the two routes will have a fixed "free-flow" travel time and also a single bottleneck of fixed capacity at the route exit. On each route, any flow greater than the capacity at the route exit is stored in a queue just upstream of the bottleneck.

**The main variables.** $X_1$ and $X_2$ are still now functions of within-day time; but now they are step functions which are constant over each time slice.

$X_1(\tau)$ = inflow into route 1 during time-slice $\tau$ (assumed constant);
$X_2(\tau)$ = inflow into route 2 during time-slice $\tau$ (assumed constant); and
$[X_1 + X_2](\tau)$ = $\rho(\tau)$ where each $\rho(\tau)$ is non-negative and given (for all $\tau$ in W).
$X_1, X_2$ are of course both to be non-negative at all $\tau$ in W.

**The feasible set F.** Given the M-vector $\rho$ whose co-ordinates $\rho(\tau)$ are all non-negative the feasible set F of route inflows is now as follows:

$$F = \{ X = (X_1, X_2);\ X_1, X_2 \text{ are M-vectors with non-negative co-ordinates satisfying:}$$
$$X_1(\tau) + X_2(\tau) = \rho(\tau) \text{ for all } \tau \text{ in W.}\}$$

**Cost functions**

$C_1(X_1)(\tau)$ = average travel cost experienced on route 1 if entered at time $\tau$ (if inflow is $X_1$)
$C_2(X_2)(\tau)$ = average travel cost experienced on route 2 if entered at time $\tau$ (if inflow is $X_2$)

Here "at time $\tau$" means "during time interval numbered $\tau$". These cost functions will both be assumed to be generated as previously as the sum of a constant uncongested travel time and a queueing delay at the route exit.

**The rigid demand user-equilibrium condition.** Following closely the two previous cases we say that X is an equilibrium if:

$$X_1[C_1(X_1) - C_2(X_2)]_+(\tau) = 0 \text{ and } X_2[C_2(X_2) - C_1(X_1)]_+(\tau) = 0 \text{ for all } \tau \text{ in W.}$$

or: $\{X_1[C_1(X_1) - C_2(X_2)]_+ + X_2[C_2(X_2) - C_1(X_1)]_+\}(\tau) = 0$ for all $\tau$ in W.

**Definition of monotonicity now.** A function f from one set of M-vectors x $\rightarrow$ another set of M-vectors is *monotone* if and only if

$$[f(x + \delta) - f(x)]\cdot\delta \geq 0$$

for all M-vectors x and $\delta$. If f is also directionally differentiable then f is monotone if and only if

$$\delta\cdot[f'(x; \delta)] \geq 0 \text{ for all x, } \delta$$

where $f'(x; \delta)$ is the directional derivative of f at x (now a M-vector) in direction $\delta$ (now a M-vector). Thus $C_1$ is monotone if and only if:

$$\Delta_1\cdot C_1'(X_1; \Delta_1) \geq 0 \text{ for all N-vectors } X_1, \Delta_1$$

where $C_1'(X_1; \Delta_1)$ is the directional derivative of $C_1$, at $X_1$, in direction $\Delta_1$.

**Monotonicity.** In the earlier part of the paper and in Smith and Ghali (1990a, b) and Mounce (2006) we have shown that $C_1(X_1)$ is a monotone function of inflow rate $X_1$ and $C_2(X_2)$ is a monotone function of inflow rate $X_2$ provided that the bottleneck capacity is a non-decreasing function of within-day time $\tau$. Here we retain the assumption that $C_1$ is monotone. Or that: $\Delta_1\cdot C_1'(X_1; \Delta_1) \geq 0$ for all $X_1, \Delta_1$. We also assume that $C_2$ is monotone.

**A possible natural day-to-day evolution in F.** For each within day time $\tau \in$ W, let

$$\Delta_1(X)(\tau) = \{-X_1^2[C_1(X_1) - C_2(X_2)]_+ + X_2^2[C_2(X_2) - C_1(X_1)]_+\}(\tau),$$
$$\Delta_2(X)(\tau) = \{-X_2^2[C_2(X_2) - C_1(X_1)]_+ + X_1^2[C_1(X_1) - C_2(X_2)]_+\}(\tau), \text{ and}$$
$$\Delta(X)(\tau) = (\Delta_1(X)(\tau),\ \Delta_2(X)(\tau)) = (\Delta_1(X),\ \Delta_2(X))(\tau).$$

Then $\Delta(X)(\tau) = 0$ for all within day times $\tau$ in W if and only if X is an equilibrium in F. As before moving X in F in direction $\Delta(X)$ preserves the total of the route-inflow at each within day time; it is a "swap" function, merely swapping inflow from the more costly to the less costly route at each $\tau$.

We suppose that $X^0 = (X^0_1, X^0_2)$ is any start point in F. So $X^0$ is measurable and non-negative and

$$[X^0_1 + X^0_2](\tau) = \rho(\tau),\ X^0_1 \geq 0 \text{ and } X^0_2 \geq 0.$$

We also suppose that the route inflow vector $X(\tau)$ evolves with day-to-day time "t" by obeying:

$$X(\tau)(0) = X^0(\tau) \text{ and } dX(\tau)(t)/dt = \Delta(X(\tau)(t)) \text{ for all } t \geq 0 \text{ and all within day times } \tau \in W.$$

**A candidate Lyapunov function or objective function**. Rather as before, let

$$V(X) = \sum_\tau \{X_1^2[C_1(X_1) - C_2(X_2)]_+^2 + X_2^2[C_2(X_2) - C_1(X_1)]_+^2\}(\tau)$$

for all X in F. Then V measures departure from equilibrium; or

$$V(X) \geq 0 \text{ and } V(X) = 0 \text{ if and only if X is an equilibrium.}$$

All sums in this section of the paper will go from 1 to M; so they go over all the time-slices.

**Descent to equilibrium?** Provided $C_1(.)$ and $C_2(.)$ are both monotone, it now follows that, away from equilibrium, $\Delta(X)$ is a descent direction for objective V at X and so, again under further natural assumptions, V(X(t)) declines to 0 as $t \to \infty$. Here as above we verify just descent under restricted conditions (and do not prove convergence of V to 0 as $t \to \infty$; this does as before happen under natural conditions).

Of course $\Delta(X)$ is not a descent direction for objective V at X if X is already an equilibrium in F as then $V(X) = 0$ and since $V \geq 0$ always there can be no descent direction from X.

So let us now suppose that X is not an equilibrium and so $V(X) > 0$. Then:

$$\sum_\tau \{X_1^2[C_1(X_1) - C_2(X_2)]_+^2 + X_2^2[C_2(X_2) - C_1(X_1)]_+^2\}(\tau) > 0$$

and so:

$$\sum_\tau \{X_1^3[C_1(X_1) - C_2(X_2)]_+^3 + X_2^3[C_2(X_2) - C_1(X_1)]_+^3\}(\tau) > 0.$$

Using this, at any feasible X, just as above in the steady state:

$V'(X; \Delta(X))$

$= -2 \sum_\tau \{X_1^3[C_1(X_1) - C_2(X_2)]_+^3 + X_2^3[C_2(X_2) - C_1(X_1)]_+^3\}(\tau)$ ($< 0$ as X is a non-equilibrium)

$-2 \sum_\tau \Delta_1(X)(\tau) C_1'(X_1;\ \Delta_1(X))(\tau)$ ($\leq 0$ by monotonicity of $C_1$)

$-2 \sum_\tau \Delta_2(X)(\tau) C_2'(X_2;\ \Delta_2(X)))(\tau)$ ($\leq 0$ by monotonicity of $C_2$)

$< 0$.

Thus $\Delta(X)$ is a descent direction for V away from equilibrium. So again by the chain rule we have shown that:

$$dV(X(t))/dt = V'(X(t); \Delta(X(t)) < 0 \text{ for all } t \geq 0.$$

It now follows, under natural and weak conditions, that $V(X(t)) \to 0$ as $t \to \infty$ and it is then easy to show that dist[X(t), E] $\to 0$ as $t \to \infty$, where E is the non-empty set of equilibria and dist[X(t), E] is the distance between X(t) and E.

Thus in this discrete time-slice model of our simple two-route two-bottleneck case we are considering, the set of equilibria is stable: non-equilibrium inflows in F naturally evolve in such a way as to converge to E as $t \to \infty$.

## BILEVEL OPTIMISATION OF DYNAMIC EQUILIBRIA ON THE SIMPLE NETWORK

Suppose now that there is a two-vector $p = (p_1, p_2)$ of prices; one price for each route. Suppose that we seek the optimal value of these subject to a dynamic equilibrium traffic distribution. Then it is natural to put

$$V(X, p) = \sum_\tau \{X_1^2[C_1(X_1) + p_1 - (C_2(X_2) + p_2)]_+^2 + X_2^2[C_2(X_2) + p_2 - (C_1(X_1) + p_1)]_+^2\}(\tau)$$

and to seek a minimum of an objective function Z(X, p) subject to V(X, p) = 0.

Several methods have been proposed. Most of these methods require the estimation of the steepest descent of the disequilibrium function V in (X, p). This is challenging computationally and the most challenging part of that is to estimate the gradient of V in X-space (with p fixed). Here we show that under *suitable natural but severe restrictions* there is a natural way of estimating this X-gradient.

*The hope here is that these restrictions may be progressively relaxed so that this method leads to an efficient algorithm for solving dynamic bilevel optimisation problems.*

**Steepest descent of V in X-space.**

Above we have utilised directional derivatives as we wished to follow a natural evolution equation which reduces V. But now we may with a little difficulty return to derivatives or Jacobians, partly because we are in Euclidean M-space. We do this as we need steepest descent directions of V in X-space for fixed prices.

Suppose that

$$C_1'(X_1;\ \Delta_1(X)) = C_1'(X_1)\Delta_1(X_1) \text{ for all } \Delta_1(X_1)$$

or that: $C_1'(X_1;\ \Delta_1(X))(\tau) = C_1'(X_1)\Delta_1(X_1)(\tau)$ for all $\tau$ in W.

Here $C_1'(X_1) = J_1(X_1)$ is to be the Jacobian of $C_1(X_1)$ and says how $C_1(X_1)$ changes as $X_1$ changes. If $X_1$ suffers a small displacement $h\Delta_1(X_1)$ then $C_1(X_1)$ will suffer a displacement of about $hJ_1(X_1)\Delta_1(X_1)$. (Here h is a small positive real number.)

This Jacobian approach suffers from the difficulty that $J_1(X_1)$ might not be defined everywhere and might be discontinuous at points where queueing delays vanish or appear. Partly in order to deal with this difficulty *we will now make two very strong assumptions (aiming to relax these eventually).*

ASSUMPTION 1:

Suppose now that the uncongested travel times for the two routes are equal.

ASSUMPTION 2:

Suppose also that there are two within-day integer times A and B with the following property: throughout the time interval [A-1, B] or all time-slices A, A+1, A+2, . . . , B (or [A-1, A], [A, A+1], [A+1, A+2], . . . , [B-1, B]) there is a queue at the exits of both route 1 and route 2; and outside this time interval neither bottleneck is saturated (which implies that there are no queues outside [A-1, B]).

*Using assumption 1 we now think of inflows as being bottleneck inflows rather than route inflows.*

For illustrative purposes suppose finally that B-A = 4 so that we are dealing with just five congested time-slices: [A-1, A], [A, A+1], [A+1, A+2], [A+2, A+3], [A+3, A+4].

Our assumptions allow us to focus on just those five $\tau$ values: A, A+1, A+2, A+3, A+4 = B. For these $\tau$ values the restricted cost function Jacobians $J_1(X_1)$ and $J_2(X_2)$ are given by the equations

$$J_1(X_1) = u_1 J \text{ and } J_2(X_2) = u_2 J$$

where $u_1$ = 1/(the exit capacity of route 1), $u_2$ = 1/(the exit capacity of route 2) and

$$J = \begin{matrix} ½ & 0 & 0 & 0 & 0 \\ 1 & ½ & 0 & 0 & 0 \\ 1 & 1 & ½ & 0 & 0 \\ 1 & 1 & 1 & ½ & 0 \\ 1 & 1 & 1 & 1 & ½. \end{matrix}$$

J has this form as early arrivals delay later arrivals but not conversely. Under our assumptions both the Jacobians $J_1(X_1)$ and $J_2(X_2)$ have this (flow-independent) form uJ between A and B. It now follows from our major assumptions that the network is equilibrated outside [A-1, B] and so the contribution to V arising outside [A-1, B] is zero. Hence, assuming further but with no real loss of generality that the price vector p is zero,

$$V(X) = \sum_{A \le \tau \le B} \{X_1^2(\tau)[C_1(X_1)(\tau) - C_2(X_2)(\tau)]_+^2 + X_2^2(\tau)[C_2(X_2)(\tau) - C_1(X_1)(\tau)]_+\}^2.$$

It follows that:

$$½\ \mathrm{grad}_X V(X) =$$

$$(\sum_{A \le \tau \le B} X_1[C_1(X_1) - C_2(X_2)]_+^2(\tau)e(\tau) - J_1^T(X_1)\Delta_1(X),$$

$$\sum_{A \le \tau \le B} X_2[C_2(X_2) - C_1(X_1)]_+^2(\tau)e(\tau) - J_2^T(X_2)\Delta_2(X))$$

where $e(\tau)$ is the M-vector with 1 in the $\tau^{th}$ place and 0 elsewhere.

Using the formulae $J_1(X_1) = u_1 J$ and $J_2(X_2) = u_2 J$ and the structure of J above:

$$
J^T = \begin{matrix} ½ & 1 & 1 & 1 & 1 \\ 0 & ½ & 1 & 1 & 1 \\ 0 & 0 & ½ & 1 & 1 \\ 0 & 0 & 0 & ½ & 1 \\ 0 & 0 & 0 & 0 & ½ \end{matrix}
$$

$$
= \begin{matrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{matrix} - \begin{matrix} ½ & 0 & 0 & 0 & 0 \\ 1 & ½ & 0 & 0 & 0 \\ 1 & 1 & ½ & 0 & 0 \\ 1 & 1 & 1 & ½ & 0 \\ 1 & 1 & 1 & 1 & ½ \end{matrix}
$$

and hence: $J_1(X_1)^T = u_1(\mathbf{1} - J)$ and $J_2(X_2)^T = u_2(\mathbf{1} - J)$.

It now follows that $J_1(X_1)^T \Delta_1(X) = u_1[\mathbf{1}\Delta_1(X) - J\Delta_1(X)]$ may be readily estimated since both $\mathbf{1}\Delta_1(X)$ and $u_1 J\Delta_1(X)$ may be readily estimated.

*Firstly*:

$\mathbf{1}\Delta_1(X) = (\sum_{A \le \tau \le B} \Delta_1(X)(\tau), \sum_{A \le \tau \le B} \Delta_1(X)(\tau), \sum_{A \le \tau \le B} \Delta_1(X)(\tau), \ldots, \sum_{A \le \tau \le B} \Delta_1(X)(\tau))^T$
and the sums involved here are available as all the $\Delta_1(X)(\tau)$ are available.

*Secondly*:

the vector $u_1 J\Delta_1(X)$ is approximately $[C_1(X_1 + h\Delta_1(X)) - C_1(X_1)]/h$ if h is small and positive and this may be estimated by evaluating both $C_1(X_1)$ and $C_1(X_1 + h\Delta_1(X))$.

Thus $J_1(X_1)^T \Delta_1(X) = u_1 \mathbf{1}\Delta_1(X) - u_1 J\Delta_1(X)$ may be readily estimated and similarly $J_2(X_2)^T \Delta_2(X) = u_2 \mathbf{1}\Delta_2(X) - u_2 J\Delta_2(X)$ may also be readily estimated.

As these derivatives are readily estimated, the bilevel method in Smith (2005, 2006) (and other methods) may be applied to this simple dynamical optimisation problem. Earlier work along these lines, optimising controls in dynamical networks at equilibrium, is given in Smith et al (1998).

## REFERENCES


R Mounce (2003), "Convergence in a continuous day-to-day dynamic traffic assignment model", Ph.D. Uni. of York, UK.

R Mounce (2005), "Dynamics and equilibrium in a continuous queueing model for traffic networks". Proceedings of the 4th IMA International Conference on Mathematics in Transport, University College London, UK.

R Mounce (2006), "Convergence in a continuous dynamic queueing model for traffic networks", to appear: *Transpn. Res.*

M J Smith (1984), "The stability of a dynamic model of traffic assignment - an application of a method of Lyapunov". *Transpn. Sci*, 18, 245 - 252.

M J Smith and M O Ghali (1990a), "The dynamics of traffic assignment and traffic control; a theoretical study", *Transpn. Res.* 24B, 409 - 422.

M J Smith and M O Ghali (1990b), "Dynamic traffic equilibrium and dynamic traffic control", Proc. 11th ISTTT (Ed: Koshi), 273 - 290.

M J Smith and M B Wisten (1995), "A continuous day-to-day traffic assignment model and the existence of a continuous dynamic user equilibrium". *Annals of Operations Research*, 60, 59-79.

M J Smith, Y Xiang, R Yarrow (1998), "Descent Methods of Calculating Locally Optimal Signal Controls and Prices in Multi-Modal and Dynamic Transportation Networks." Proc. of the 4th EURO Trans Meeting (Ed: Bell), 9 – 34.

M J Smith (2005), "Bilevel optimisation of prices in a variety of transportation models", Proc. 16th ISTTT (Ed: Mahmassami), 1-21.

M J Smith (2006) "Bilevel Optimisation of Prices and Signals in Transportation Models". In: Mathematical and Computational Models for Congestion Charging (Eds: Lawphongpanich, Hearn and Smith), 159 – 199.

# Dynamic Non-cooperative Games as a Foundation for Modeling Dynamic User Equilibrium*

*Terry L. Friesz: Penn State University, USA, tfriesz@psu.edu*
*Reetabrata Mookherjee: Penn State University, USA, reeto@psu.edu*
*Changhyun Kwon: Penn State University, USA, chkwon@psu.edu*

**Abstract**

In this paper we take the point of view that there is a formalism for modeling a within-day dynamic user equilibrium (DUE) that is an extension of traditional differential game theory to accomodate the natural formulation of DUE as an infinite dimensional differential variational inequality (DVI) involving explicit state-dependent time shifts, explicit control variables and explicit equations of state. We also show how a second time scale (day-to-day) may be included to model the learning process behind the formation of demand. An example based on both time scales is included.

*Keywords:* Dynamic User Equilibrium; Differential Variational Inequality; Optimal Control

## 1 Introduction

In this paper we take the point of view that there is a formalism for modeling dynamic user equilibrium (DUE) that is not widely understood or applied. That formalism is the extension of traditional differential game theory to accomodate the natural formulation of DUE as an infinite dimensional variational inequality involving explicit state-dependent time shifts, explicit control variables and explicit equations of state. We call this the differential variational inequality (DVI) formalism (Friesz and Mookherjee (2006)). We begin with some foundation material from the theory of deterministic optimal control, and mathematical programming in function spaces. From there we show how time shifts may be considerd by appeal to the notion of G-differentiability. Next we show how dynamic Cournot-Nash-Bertrand games may be formulated as differential variational inequalities, leading to necessary conditions for such dynamic games that are static variational inequalities. We then discuss how functional fixed point algorithms whose subproblems are tractable optimal control problems — without time shifts even when the original dynamic game has time shifts — may be constructed and implemented.

We then show how a well-known DUE model, proposed by Friesz, Bernstein, Smith, Tobin and Wie (1993), may be treated using the apparatus of differential variational inequalities (DVIs). In particular, the DVI formalism is shown to accomodate both path-based and arc-based formulations of DUE, as well as alternative models of delay and explicit queue spill-back constraints. We observe that the DVI formalism allows a direct and quite simple treatment of the first-in-first-out queue discipline. We also observe that the formalism may be extended to account for stochastic phenomena, including both imperfect and incomplete information. We conclude this paper by applying the formalism to create two entirely new formulations of dynamic user equilibrium when: (1) there are dual time scales (day-to-day and within-day); and (2) demand information is uncertain.

## 2 Differential Variational Inequality with State Dependent Time Shifts

Dynamic systems comprised of game-theoretic agents having control of their own (but not necessarily anyone else's) strategic variables are self-organizing if observable, persistent behavioral patterns and hierarchies emerge with the passage of time. Moreover, time-shifted variational inequalities with explicit state dynamics and explicit controls are

*This paper supplies the mathematical background and a more rigorous theoretical development missing from Friesz and Mookherjee (2006), which is mainly concerned with computation and is largely based on intuitive arguments. The present paper purposely subsumes the previous Friesz and Mookherjee (2006) paper to provide a self-contained reference.

known to arise in the modeling of such systems if the game-theoretic agents have a forward-looking or anticipatory perspective and the emergent behavior is some variety of Cournot-Nash-Bertrand equilibrium, be it static or moving in nature.

Here we take the point of view that infinite dimensional variational inequalities with state dynamics among their constraints and having explicit control variables are direct generalizations of optimal control problems. Because such problems contain ordinary differential equations of state among their constraints, they are one variety of differential variational inequality (DVI) that we refer to as a differential variational inequality with controls (DVIC). It stands to reason that the study of DVICs should involve the derivation of a generalized version of the Pontryagin maximum principal as well as other necessary conditions similar to those encountered in optimal control theory – as we do in Section 2.2. We know of no other manuscripts that use the optimal control perspective taken herein for the study of time-shifted infinite dimensional (dynamic) variational inequalities with state dynamics and explicit controls.

In particular, we will consider the notion of a variational inequality in Hilbert space that includes state dynamics as constraints in the form of a two-point boundary value problem depending parametrically on control variables. Both the principal operator of the variational inequality and the dynamics themselves will involve time shifts that are state-dependent. In fact we consider the following operator

$$x\left(u,u_D,t\right)=\arg\left\{\frac{dx}{dt}=f\left(x,u,u_D,t\right),\ x\left(t_0\right)=x^0,\Gamma\left[x\left(t_f\right),t_f\right]=0\right\}\in\left(\mathcal{H}^1\left[t_0,t_f\right]\right)^n \tag{1}$$

where $t_0$ and $t_f$ are given and

$$\left[t_0,t_f\right]\subseteq\Re^1_+$$

Furthermore $u_D\left(t\right)$ is a shorthand for the shifted control vector

$$u_D\left(t\right)=\left(\begin{array}{c}u_1\left(t+D_1\left(x_1\right)\right)\\ \vdots\\ u_m\left(t+D_m\left(x_m\right)\right)\end{array}\right)$$

where $D_i:\left(\mathcal{H}^1\left[t_0,t_f\right]\right)^n\longrightarrow\mathcal{H}^1\left[t_0,t_f\right]$ for each $i\in[1,m]$. The other relevant mappings are

$$\begin{array}{rcl} f & : & \left(\mathcal{H}^1\left[t_0,t_f\right]\right)^n\times\left(L^2\left[t_0,\tau\right]\right)^m\times\left(L^2\left[t_0,t_f\right]\right)^m\times\Re^1_+\longrightarrow\left(L^2\left[t_0,t_f\right]\right)^n\\ \Gamma & : & \left(\mathcal{H}^1\left[t_0,t_f\right]\right)^n\times\Re^1_+\longrightarrow\left(\mathcal{H}^1\left[t_0,t_f\right]\right)^r\\ u & \in & U\subseteq\left(L^2\left[t_0,t_f\right]\right)^m;\ u_D:\left(\mathcal{H}^1\left[t_0,t_f\right]\right)^n\times\Re^1_+\longrightarrow\left(L^2\left[t_0,t_1\right]\right)^m\end{array}$$

where

$$t_1=t_f+\max\left\{D_i\left[x\left(t_f\right)\right]:i\in[1,m]\right\} \tag{2}$$

In the above $\left(L^2\left[t_0,t_f\right]\right)^m$ is the $m$-fold product of the space of square-integrable functions $L^2\left[t_0,t_f\right]$, while $\left(\mathcal{H}^1\left[t_0,t_f\right]\right)^n$is the $n$-fold product of the Sobolev space $\mathcal{H}^1\left[t_0,t_f\right]$.

Additionally we invoke the following regularity condition for the two-point boundary value problem (1):

**Definition 1** *Regular Dynamics. We shall say the state dynamics operator $x\left(u,u_D,t\right)$ given by (1) is $\left(x^0,U,\Gamma\right)$-regular if the terminal state constraint $\Gamma\left[x\left(t_f\right),t_f\right]=0$ is reachable from the given initial state $x^0$ for all $u\in U$.*

The notation $x\left(u,u_D,t\right)$ is a direct generalization of that used by Minoux (1986) to describe how the Pontryagin minimum principle of optimal control theory may be derived using notions from infinite dimensional mathematical programming; it denotes an operator which determines the state vector for any pair of shifted and un-shifted control vectors. In order to use the operator notation $x\left(u,u_D,t\right)$, we will invoke $\left(x^0,U,\Gamma\right)$-regularity to ensure that the parametric boundary value problem (1) is well posed. Such a regularity condition should not be interpreted as an *a priori* stipulation that the variational inequality to be introduced below has a solution; rather it is a stipulation that the constrained dynamics of (1) have a solution for all controls that are considered pertinent to the problem of interest.

## 2.1 A Related Optimal Control Problem

Before studying differential variational inequalities with state-dependent time shifts, we need to derive necessary conditions for a related optimal control problem. That derivation relies on the notion of a so-called *G-derivative*:

**Definition 2** *(G-differentiable, Minoux (1986)) Let $V$ be a normed vector space and $J$ be a functional on $V$. If for all $\varphi \in V$ the limit $\delta J(v,\varphi)$ defined by*

$$\delta J(v,\varphi) \equiv \lim_{\theta \to 0} \frac{J(v+\theta\varphi) - J(v)}{\theta}$$

*exists, then $J$ is said to be differentiable in the sense of Gateaux (G-differentiable) at $v \in V$.*

With the preceding background, we consider the following optimal control problem:

$$\min \Gamma\left[x(t_f), t_f\right] + \int_{t_0}^{t_f} G(x,u,u_D,t)dt \tag{3}$$

subject to

$$\frac{dx}{dt} = f(x,u,u_D,t);\ x(t_0) = x^0 \tag{4}$$

$$u \in U \tag{5}$$

This is a non-standard optimal control problem, and we will need its necessary conditions. In fact we will state and prove the following result:

**Theorem 3** *(Necessary Conditions for Optimal Control with State-Dependent Time Shifts) If (i) $u \in U \subseteq \left(L^2[t_0,\tau]\right)^m$; (ii) $u_D \in \left(L^2[t_0,t_f]\right)^m$; (iii) the operator $x(u,u_D,t) : \left(L^2[t_0,t_f]\right)^m \times \left(L^2[t_0,\tau]\right)^m \longrightarrow \left(\mathcal{H}_\infty{}^1[t_0,t_f]\right)^n$ is $\left(x^0,U,\Gamma\right)$-regular, continuous and G-differentiable with respect to $u$ and $u_D$; (iv) $D_i(x_i) : \left(\mathcal{H}^1[t_0,t_f]\right)^n \longrightarrow \mathcal{H}^1[t_0,t_f]$ is continuously differentiable with respect to $x_i$ for each $i \in [1,m]$; (v) $\Gamma[x,t] : \left(\mathcal{H}^1[t_0,t_f]\right)^n \times \Re^1_+ \longrightarrow \mathcal{H}^1[t_0,t_f]$ is continuously differentiable with respect to $x$; (vi) $G(x,u,u_D,t) : \left(\mathcal{H}^1[t_0,t_f]\right)^n \times \left(L^2[t_0,\tau]\right)^m \times \left(L^2[t_0,\tau]\right)^m \times \Re^1_+ \longrightarrow L^2[t_0,t_f]$ is continuously differentiable with respect to $x$, $u$ and $u_D$; (vii) $f(x,u,u_D,t) : \left(\mathcal{H}^1[t_0,t_f]\right)^n \times \left(L^2[t_0,\tau]\right)^m \times \left(L^2[t_0,\tau]\right)^m \times \Re^1_+ \longrightarrow \left(L^2[t_0,t_f]\right)^n$ is continuously differentiable with respect to $x$, $u$ and $u_D$; (viii) $U \subseteq \left(L^2[t_0,\tau]\right)^m$ is convex and compact; and (ix) $x^0 \in \Re^n$*

*then any solution $u^* \in U$ of the optimal control problem (3) through (5) obeys the following necessary conditions:*

**1.** *the finite dimensional variational inequality principle:*

$$\sum_{i=1}^m \frac{\partial H_1^*}{\partial u_i}(u_i - u_i^*) \geq 0 \quad \forall t \in [t_0, D_i(x(t_0))], u \in U$$

$$\sum_{i=1}^m \left\{ \frac{\partial H_1^*}{\partial u_i} + \left[ \frac{\partial H_1^*}{\partial (u_D)_i} \frac{1}{1 + \sum_{j=1}^m \frac{\partial D_i(x^*)}{\partial x_j}\frac{dx_j^*}{dt}} \right]_{s_i(t)} \right\} (u_i - u_i^*) \geq 0 \ \ \forall t \in [D_i(x^*(t_0)), t_f + D_i(x^*(t_f))], u \in U$$

*where* $s_i(t) = \arg\left[s = t - D_i(x(s))\right] \ \ \forall t \in [D_i(x^*(t_0)), t_f + D_i(x^*(t_f))], i \in [1,m]$ and $H_1^* = H_1(x^*,u^*,u_D^*,\lambda^*,t) = G(x^*,u^*,u_D^*,t) + (\lambda^*)^T f(x^*,u^*,u_D^*,t) \ \ \forall t \in [t_0,t_f]$;

**2.** *the state dynamics*

$$\frac{dx^*}{dt} = f(x^*,u^*,u_D^*,t);\ x^*(t_0) = x^0; \quad \text{and}$$

**3.** *the adjoint dynamics*

$$(-1)\frac{d\lambda^*}{dt} = \nabla_x (\lambda^*)^T f(x^*,u^*,u_D^*,t);\ \lambda^*(t_f) = \frac{\partial \Gamma[x^*(t_f),t_f]}{\partial x}.$$

**Proof.** The below proof extends the fixed time shift analysis of **?** to state-dependent time shifts. Note that

$$x\left(u,u_D,t\right)=x\left(t_0\right)+\int_{t_0}^{t}f\left[x\left(u,u_D,t\right),u,u_D,t\right]dt$$

It is immediate that

$$x\left(u+\theta\rho,u_D+\theta\rho_D\right)=x\left(t_0\right)+\int_{t_0}^{t}f\left[x\left(u+\theta\rho,u_D+\theta\rho_D\right),u+\theta\rho,u_D+\theta\rho_D,t\right]dt$$

Consequently,

$$\begin{aligned}\delta x\left(u,\rho;u_D,\rho_D\right) &= \int_{t_0}^{t}\left\{\frac{\partial f\left[x\left(u,u_D,t\right),u,u_D,t\right]}{\partial x}\delta x\left(u,\rho;u_D,\rho_D\right)+\frac{\partial f\left[x\left(u\right),u,u_D,t\right]}{\partial u}\delta u\left(\rho\right)\right.\\ &\left.+\frac{\partial f\left[x\left(u\right),u,u_D,t\right]}{\partial u_D}\delta u_D\left(\rho_D\right)\right\}dt\end{aligned}$$

where the G-derivatives of $u$ and $u_D$ obey

$$\delta u\left(\rho\right)=\lim_{\theta\longrightarrow 0}\frac{\left(u+\theta\rho\right)-u}{\theta}=\rho;\ \delta u_D\left(\rho_D\right)=\lim_{\theta\longrightarrow 0}\frac{\left(u_D+\theta\rho_D\right)-u_D}{\theta}=\rho_D$$

Employing the shorthand $y=\delta x\left(u,\rho;u_D,\rho_D\right)$, we have the integral equation

$$y=\int_{t_0}^{t}\left[\frac{\partial f}{\partial x}y+\frac{\partial f}{\partial u}\rho+\frac{\partial f}{\partial u_D}\rho_D\right]dt \tag{6}$$

It is of course immediate from this integral equation that $y$ obeys

$$\frac{dy}{dt}=\frac{\partial f}{\partial x}y+\frac{\partial f}{\partial u}\rho+\frac{\partial f}{\partial u_D}\rho_D;\quad y\left(t_0\right)=0 \tag{7}$$

which is recognized as an initial value problem, verifying that the G-derivative of $x$ is well defined. The G-derivative of $J$ obeys

$$\begin{aligned}\delta J\left(u,\rho;u_D,\rho_D\right) &= \left[\frac{\partial\Gamma\left[x\left(t\right),t\right]}{\partial x}\delta x\left(u,\rho;u_D,\rho_D\right)\right]_{t_0}^{t_f}+\int_{t_0}^{t_f}\left[\frac{\partial G}{\partial x}\delta x\left(u,\rho;u_D,\rho_D\right)+\frac{\partial G}{\partial u}\delta u\left(\rho\right)+\frac{\partial G}{\partial u_D}\delta u\left(\rho_D\right)\right]\\ &= \frac{\partial\Gamma\left[x\left(t_f\right),t_f\right]}{\partial x}y\left(t_f\right)+\int_{t_0}^{t_f}\left[\frac{\partial G}{\partial x}y+\frac{\partial G}{\partial u}\rho+\frac{\partial G}{\partial u_D}\rho_D\right]dt\end{aligned}$$

We introduce *adjoint variables* $\lambda$ defined by the final value problem

$$-\frac{d\lambda}{dt}=\left(\frac{\partial f}{\partial x}\right)^T\lambda+\left(\frac{\partial G}{\partial x}\right)^T;\quad\lambda\left(t_f\right)=\frac{\partial\Gamma\left[x\left(t_f\right),t_f\right]}{\partial x} \tag{8}$$

so that

$$\delta J\left(u,\rho;u_D,\rho_D\right)=\int_{t_0}^{t_f}\left[-\left(\frac{d\lambda}{dt}\right)^T y-\lambda^T\frac{\partial f}{\partial x}y+\frac{\partial G}{\partial u}\rho+\frac{\partial G}{\partial u_D}\rho_D\right]dt \tag{9}$$

Note that

$$\begin{aligned}\left[\lambda^T y\right]_{t_0}^{t_f} &= \left[\lambda\left(t_f\right)\right]^T y\left(t_f\right)-\left[\lambda\left(t_0\right)\right]^T y\left(t_0\right)\\ &= \frac{\partial\Gamma\left[x\left(t_f\right),t_f\right]}{\partial x}y\left(t_f\right)\end{aligned}$$

due to (8) and the fact that $y(t_0) = 0$, so an integration by parts yields

$$\begin{aligned}\int_{t_0}^{t_f} -\left(\frac{d\lambda}{dt}\right)^T y dt &= \int_{t_0}^{t_f} \lambda^T \frac{dy}{dt} dt - \left[\lambda^T y\right]_{t_0}^{t_f} \\ &= \int_{t_0}^{t_f} \lambda^T \frac{dy}{dt} dt - \frac{\partial \Gamma\left[x(t_f), t_f\right]}{\partial x} y(t_f) \\ &= \int_{t_0}^{t_f} \lambda^T \left[\frac{\partial f}{\partial x} \cdot y + \frac{\partial f}{\partial u} \cdot \rho + \frac{\partial f}{\partial u_D} \rho_D\right] dt - \frac{\partial \Gamma\left[x(t_f), t_f\right]}{\partial x} y(t_f) \end{aligned} \tag{10}$$

It follows that

$$\begin{aligned}\delta J(u, \rho; u_D, \rho_D) &= \frac{\partial \Gamma\left[x(t_f), t_f\right]}{\partial x} y(t_f) + \int_{t_0}^{t_f} \left\{\lambda^T \left[\frac{\partial f}{\partial x} \cdot y + \frac{\partial f}{\partial u} \cdot \rho + \frac{\partial f}{\partial u_D} \rho_D\right]\right. \\ &\qquad \left. -\lambda^T \frac{\partial f}{\partial x} y + \frac{\partial G}{\partial u} \rho + \frac{\partial G}{\partial u_D} \rho_D\right\} dt - \frac{\partial \Gamma\left[x(t_f), t_f\right]}{\partial x} y(t_f) \\ &= \int_{t_0}^{t_f} \left[\lambda^T \frac{\partial f}{\partial u} + \frac{\partial G}{\partial u}\right] \rho dt + \int_{t_0}^{t_f} \left[\lambda^T \frac{\partial f}{\partial u_D} + \frac{\partial G}{\partial u_D}\right] \rho_D dt \end{aligned}$$

Defining $H_1(x, u, u_D, \lambda, t) = G(x, u, u_D, t) + \lambda^T f(x, u, u_D, t)$, we have

$$\delta J(u, \rho; u_D, \rho_D) = \int_{t_0}^{t_f} \left[\frac{\partial H_1}{\partial u} \rho + \frac{\partial H_1}{\partial u_D} \rho_D\right] dt \tag{11}$$

as an expression for the G-derivative of the criterion with respect to both $u$ and $u_D$. Moreover, terms of the form

$$\int_{t_0}^{t_f} \frac{\partial H_1}{\partial (u_D)_i} (\rho_D)_i \, dt = \int_{t_0}^{t_f} \frac{\partial H_1}{\partial (u_D)_i} \delta u_i \left(t + D_i(x_i)\right) dt$$

may be re-expressed by making the change of variables

$$\Delta_i = t + D_i(x(t)) \Longleftrightarrow t = \Delta_i - D_i(x(t))$$

Because the $D_i(x)$ are differentiable with respect to $x_i$, the implicit function theorem gives

$$\frac{dt}{d\Delta_i} = -\frac{\partial \left[t - \Delta_i + D_i(x)\right] / \partial \Delta_i}{\partial \left[t - \Delta_i + D_i(x)\right] / \partial t} = \frac{1}{1 + \sum_{j=1}^{m} \frac{\partial D_i(x)}{\partial x_j} \dot{x_j}}$$

or,

$$dt = \frac{1}{1 + \sum_{j=1}^{m} \frac{\partial D_i(x)}{\partial x_j} \dot{x_j}} d\Delta_i \tag{12}$$

Note that

$$t = t_0 \Longrightarrow \Delta_i = t_0 + D_i(x(t_0)); \text{ Putting } t = t_f \Longrightarrow \Delta_i = t_f + D_i(x(t_f))$$

Furthermore, without loss of generality, we may take $\delta (u_D)_i = 0$ for any time $t < D_i(x(t_0))$ and $\delta (u)_i = 0$ for any time $t > D_i(x(t_0))$. A change of variables based on (12) leads to

$$\begin{aligned}\int_{t_0}^{t_f} \frac{\partial H_1}{\partial (u_D)_i} (\rho_D)_i \, dt &= \int_{D_i(x_i(t_0))}^{t_f + D_i(x_i(t_f))} \frac{\partial H_1}{\partial (u_D)_i} \delta (u_D)_i \, dt \\ &= \int_{D_i(x_i(t_0))}^{t_f + D_i(x_i(t_f))} \left[\frac{\partial H_1}{\partial (u_D)_i} \frac{1}{1 + \sum_{j=1}^{m} \frac{\partial D_i(x)}{\partial x_j} \dot{x_j}}\right]_{s_i(t)} \delta (u)_i \, dt \\ &= \int_{D_i(x_i(t_0))}^{t_f + D_i(x_i(t_f))} \left[\frac{\partial H_1}{\partial (u_D)_i} \frac{1}{1 + \sum_{j=1}^{m} \frac{\partial D_i(x)}{\partial x_j} \dot{x_j}}\right]_{s_i(t)} \rho_i dt \end{aligned} \tag{13}$$

where $s_i(t)$ obeys $s_i(t) = \arg[s = t - D_i(x(s))]$ for any given instant of time $t$ at which the term

$$\frac{\partial H_1}{\partial (u_D)_i} \frac{1}{1 + \sum_{j=1}^{m} \frac{\partial D_i(x)}{\partial x_j} \dot{x_j}}$$

must be evaluated. Note that the change of variables in (13) has re-expressed the G-derivative of $u_D$ as a derivative of $u$. We next note that

$$\int_{t_0}^{t_f} \frac{\partial H_1}{\partial u_i} \rho_i dt = \int_{t_0}^{D_i(x_i(t_0))} \frac{\partial H_1}{\partial u_i} \rho_i dt + \int_{D_i(x_i(t_0))}^{t_f + D_i(x_i(t_f))} \frac{\partial H_1}{\partial u_i} \rho_i dt \tag{14}$$

This last result means that for the change of variables introduced above the G-derivative is expressible in terms of $\rho$; that is

$$\delta J(u, \rho; u_D, \rho_D) = \delta J(u, \rho; u, \rho) \equiv \delta J(u, \rho)$$

Using (13) and (14) we obtain

$$[\delta J(u, \rho)]_i = \int_{t_0}^{D_i(x(t_0))} \frac{\partial H_1}{\partial u_i} \rho_i dt + \int_{D_i(x(t_0))}^{t_f + D_i(x(t_f))} \left\{ \frac{\partial H_1}{\partial u_i} + \left[ \frac{\partial H_1}{\partial (u_D)_i} \frac{1}{1 + \sum_{j=1}^{m} \frac{\partial D_i(x)}{\partial x_j} \dot{x_j}} \right]_{t=s_i} \right\} \rho_i dt$$

Note that in the above each component of $\delta J(u, \rho)$ has a different upper limit of integration and thereby we cannot give an inner product representation of the G-derivative in terms of a gradiemt and a direction vector. However, without loss of generality we may define $\delta(u)_i = 0$ for any $t > t_f + D_i(x(t_f))$. Since $\delta(u)_i = \rho_i$ we may write

$$[\delta J(u, \rho)]_i = \int_{t_0}^{D_i(x(t_0))} \frac{\partial H_1}{\partial u_i} \rho_i dt + \int_{D_i(x(t_0))}^{t_1} \left\{ \frac{\partial H_1}{\partial u_i} + \left[ \frac{\partial H_1}{\partial (u_D)_i} \frac{1}{1 + \sum_{j=1}^{m} \frac{\partial D_i(x)}{\partial x_j} \dot{x_j}} \right]_{t=s_i} \right\} \rho_i dt$$

where $t_1$ is defined by (2) and the same for all $i \in [1, m]$, which has the effect of defining the G-derivative of the criterion as

$$\delta J u, \rho) = \int_{t_0}^{t_1} \left[ \frac{\partial H_1}{\partial u} \rho \right] dt$$

Optimality requires $u^* \in U$ to obey

$$\delta J(u^*, \rho) \geq 0 \quad \forall \rho \geq 0 \tag{15}$$

which directly yields the desired necessary conditions when it is observed that each direction may be stated as $\rho = (u - u^*)$ for some $u \in U$. ■

The following result, stemming directly from the above proof, is also important:

**Corollary 4** *(Gradient of the Criterion in the Presence of Time Shifts) For regularity in the sense of Definition 5, the gradient of the criterion (3) is defined by*

$$[\nabla J(u)]_i = \begin{cases} \frac{\partial H_1}{\partial u_i} & \text{if } t \in [t_0, D_i(x_i(t_0))] \\ \frac{\partial H_1}{\partial u_i} + \left[ \frac{\partial H_1}{\partial (u_D)_i} \frac{1}{1 + \sum_{j=1}^{m} \frac{\partial D_i(x)}{\partial x_j} \dot{x_j}} \right]_{s_i(t)} & \text{if } t \in [D_i(x^*(t_0)), t_f + D_i(x^*(t_f))] \end{cases}$$

*for* $i = [1, m]$.

**Proof.** By the Riesz representation theorem we know

$$\delta J(u^*, \rho) = \langle \nabla J(u^*), (u - u^*) \rangle \quad \forall u \in U \tag{16}$$

The result is then immediate. ■

## 2.2 Statement of a DVI with State Dependent Time Shifts

With the above background we are now ready to study the following problem:

$$\text{find } u^* \in U \text{ such that}$$

$$\left\langle F\left(x\left(u^*, u_D^*\right), u^*, u_D^*, t\right), u - u^*\right\rangle \geq 0 \text{ for all } u \in U \tag{17}$$

where

$$x\left(u, u_D, t\right) = \arg\left\{\frac{dx}{dt} = f\left(x, u, u_D, t\right),\ x\left(t_0\right) = x^0, u \in U, \Gamma\left[x\left(t_f\right), t_f\right] = 0\right\} \in \left(\mathcal{H}^1\left[t_0, t_f\right]\right)^n \tag{18}$$

We refer to (17) as a differential variational inequality with explicit controls and time shifts, abbreviated $DVIC(F, f, \Gamma, D, U, x^0)$.

### 2.2.1 Necessary Conditions

To develop necessary conditions for solutions of (17) we will rely on the following notion of regularity:

**Definition 5** *[Regularity of $DVIC(F, f, \Gamma, D, U, x^0)$] We call $DVIC(F, f, \Gamma, D, U, x^0)$ regular if: (i) $u \in U \subseteq \left(L^2\left[t_0, \tau\right]\right)^m$; (ii) $u_D \in \left(t_f^2\left[t_0, t_f\right]\right)^m$; (iii) the operator $x\left(u, u_D, t\right) : \left(L^2\left[t_0, t_f\right]\right)^m \times \left(L^2\left[t_0, \tau\right]\right)^m \longrightarrow \left(\mathcal{H}_\infty{}^1\left[t_0, t_f\right]\right)^n$ is $\left(x^0, U, \Gamma\right)$-regular, continuous and G-differentiable with respect to $u$ and $u_D$; (iv) $D_i\left(x\right) : \left(\mathcal{H}^1\left[t_0, t_f\right]\right)^n \longrightarrow \mathcal{H}^1\left[t_0, t_f\right]$ is continuously differentiable with respect to $x_i$, for each $i \in [1, m]$; (v) $\Gamma\left(x, t\right) : \left(\mathcal{H}^1\left[t_0, t_f\right]\right)^n \times \Re_+^1 \longrightarrow \left(\mathcal{H}^1\left[t_0, t_f\right]\right)^r$ is continuously differentiable with respect to $x$; (vi) $F(x, u, u_D, t) : \left(\mathcal{H}^1\left[t_0, t_f\right]\right)^n \times \left(L^2\left[t_0, \tau\right]\right)^m \times \left(L^2\left[t_0, t_f\right]\right)^m \times \Re_+^1 \longrightarrow \left(L^2\left[t_0, t_f\right]\right)^m$ is continuous with respect to $x$ and $u$; (vii) $f\left(x, u, u_D, t\right) : \left(\mathcal{H}^1\left[t_0, t_f\right]\right)^n \times \left(L^2\left[t_0, \tau\right]\right)^m \times \left(L^2\left[t_0, t_f\right]\right)^m \times \Re_+^1 \longrightarrow \left(L^2\left[t_0, t_f\right]\right)^n$ is continuously differentiable with respect to $x$, $u$ and $u_D$; (viii) $U \subseteq \left(L^2\left[t_0, \tau\right]\right)^m$ is convex and compact; and (ix) $x^0 \in \Re^n$.*

We next note that (17) may be restated as the following optimal control problem

$$\min \gamma^T \Gamma\left[x\left(t_f\right), t_f\right] + \int_{t_0}^{t_f} \left[F\left(x^*, u^*, u_D^*, t\right)\right]^T u dt \tag{19}$$

subject to

$$\frac{dx}{dt} = f\left(x, u, u_D, t\right);\ x\left(t_0\right) = x^0 \tag{20}$$

$$u \in U \tag{21}$$

where $x^* = x\left(u^*, u_D^*\right)$ is the optimal state vector and $\gamma \in \Re^r$ is the vector of dual variables for the terminal constraints $\Gamma\left[x\left(t_f\right), t_f\right] = 0$. We point out that this optimal control problem is a mathematical abstraction and of no use for computation, since its criterion depends on knowledge of the variational inequality solution $u^*$. In what follows we will need the Hamiltonian for (19) through (21), namely

$$H_2\left(x, u, u_D, \lambda, t\right) = \left[F\left(x^*, u^*, u_D^*, t\right)\right]^T u + \lambda^T f\left(x, u, u_D, t\right) \tag{22}$$

where $\lambda\left(t\right)$ is the adjoint vector that solves the adjoint equations and transversality conditions for given state variables and controls. It is now a relatively easy matter to derive the necessary conditions stated in the following theorem:

**Theorem 6** *[Necessary Conditions for $DVIC(F, f, \Gamma, D, U, x^0)$] When regularity in the sense of Definition 5 holds, solutions $u^* \in U$ of $DVIC(F, f, \Gamma, D, U, x^0)$ must obey:*

***1.*** *the finite dimensional variational inequality principle:*

$$\sum_{i=1}^m \left[F_i\left(x^*, u^*, u_D^*, t\right) + \sum_{j=1}^m \lambda_j \frac{\partial f_i\left(x^*, u^*, u_D^*, t\right)}{\partial u_i}\right]\left(u_i - u_i^*\right) \geq 0 \quad \forall t \in \left[t_0, D_i\left(x\left(t_0\right)\right)\right], u \in U$$

$$\sum_{i=1}^{m}\left\{F_i\left(x^*,u^*,u_D^*,t\right)+\sum_{j=1}^{m}\lambda_j\frac{\partial f_j\left(x^*,u^*,u_D^*,t\right)}{\partial u_i}\right.$$

$$\left.+\left[\lambda_j\frac{\partial f_j\left(x^*,u^*,u_D^*,t\right)}{\partial\left(u_D\right)_i}\frac{1}{1+\sum_{j=1}^{m}\frac{\partial D_i(x^*)}{\partial x_j}f_j\left(x^*,u^*,u_D^*,t\right)}\right]_{s_i(t)}\right\}\left(u_i-u_i^*\right)\geq 0$$

$$\forall t\in\left[D_i\left(x^*\left(t_0\right)\right),t_f+D_i\left(x^*\left(t_f\right)\right)\right],u\in U$$

***2.*** *the state dynamics*

$$\frac{dx^*}{dt}=f\left(x^*,u^*,u_D^*,t\right);\ x^*\left(t_0\right)=x^0;\quad \textit{and}$$

***3.*** *the adjoint dynamics*

$$(-1)\frac{d\lambda^*}{dt}=\nabla_x\left(\lambda^*\right)^T f\left(x^*,u^*,u_D^*,t\right);\ \lambda^*\left(t_f\right)=\nu^T\frac{\partial\Gamma\left[x^*\left(t_f\right),t_f\right]}{\partial x}$$

*where $\nu\in\Re^r$ is the vector of dual variables for the terminal constraints $\Gamma\left[x\left(t_f\right),t_f\right]=0$.*

**Proof.** $DVIC(F,f,\Gamma,D,U,x^0)$ is equivalent to the optimal control problem

$$\min\nu^T\Gamma\left[x\left(t_f\right),t_f\right]+\int_{t_0}^{t_f}\left[F\left(x^*,u^*,u_D^*,t\right)\right]^T udt$$

subject to

$$\begin{array}{rcl}\dfrac{dx}{dt} & = & f\left(x,u,u_D,t\right);\ \ x\left(t_0\right)=x^0\\ u & \in & U\end{array}$$

with Hamiltonian $H_2\left(x,u,u_D,\lambda,t\right)=\left[F\left(x^*,u^*,u_D^*,t\right)\right]^T u+\lambda^T f\left(x,u,u_D,t\right)$.By virtue of regularity we may apply Theorem 3; the necessary conditions follow immediately. ■

## 2.3 Fixed Point Formulation and Algorithm

Furthermore, there is a fixed point form of $DVIC(F,f,\Gamma,D,U,x^0)$. In particular we state and prove the following result:

**Theorem 7** *(fixed point formulation of $DVIC(F,f,\Gamma,D,U,x^0)$) When regularity in the sense of Definition 5 holds and $f\left(x,u,u_D,t\right):\left(\mathcal{H}^1\left[t_0,t_f\right]\right)^n\times\left(L^2\left[t_0,\tau\right]\right)^m\times\left(L^2\left[t_0,t_f\right]\right)^m\times\Re_+^1\longrightarrow\left(L^2\left[t_0,t_f\right]\right)^n$ is convex, $DVIC(F,f,\Gamma,D,U,x^0)$ is equivalent to the following fixed point problem:*

$$u=P_U\left[u-\alpha F\left(x\left(u,u_D,t\right),u,u_D,t\right)\right]$$

*where $P_U\left[.\right]$ is the minimum norm projection onto $U\subseteq\left(L^2\left[t_0,\tau\right]\right)^m$ and $\alpha\in\Re_{++}^1$.*

**Proof.** The fixed point problem considered requires that

$$u=\arg\min_v\left\{\frac{1}{2}\left\|u-\alpha F\left(x\left(u,u_D,t\right),u,u_D,t\right)-v\right\|^2:v\in U\right\}\tag{23}$$

where $\alpha\in\Re_{++}^1$ is any strictly positive real number. That is, we seek the solution of the optimal control problem

$$\min_v\gamma^T\Gamma\left[x\left(t_f\right),t_f\right]+\int_{t_0}^{t_f}\frac{1}{2}\left[u-\alpha F\left(x,u,u_D,t\right)-v\right]^2dt$$

subject to

$$\begin{aligned} \frac{dx}{dt} &= f(x, v, v_D, t);\ x(t_0) = x^0 \\ u &\in U \end{aligned}$$

where $u$ and $u_D$ are treated as fixed vectors. Because of regularity and the assumed convexity of $f(x, v, v_D, t)$, a necessary and sufficient condition for a solution $v^* \in U$ of this optimal control problem is

$$[\nabla_v H_3(x^*, v^*, v_D^*, \eta^*, t)]^T (v - v^*) \geq 0 \qquad \forall v \in U \tag{24}$$

where $H_3(x, v, v_D, \eta, t) = \frac{1}{2}[u - \alpha F(x, u, u_D, t) - v]^2 + \eta^T f(x, v, v_D, t)$ and for given $x$ and $v$

$$\eta = \arg\left\{(-1)\frac{d\eta}{dt} = \nabla_x H_3(x, v, v_D, \eta, t),\ \eta(t_f) = \gamma^T \frac{\partial \Gamma[x(t_f), t_f]}{\partial x(t_f)}\right\}$$

Note that $\nabla_v H_3(x, v, v_D, \eta, t) = -u + \alpha F(x, u, u_D, t) + v + \nabla_v \eta^T f(x, v, v_D, t)$. Because $u = v$ by virtue of (23) we have

$$\nabla_u H_3(x, v, v_D, \eta, t) = \alpha F(x, u, u_D, t) + \nabla_u \eta^T f(x, u, u_D, t) \tag{25}$$

Now if we set $\lambda = \frac{\eta}{\alpha}$; we have

$$\left[F(x^*, u^*, u_D^*, t) + \nabla_u (\lambda^*)^T f(x^*, u^*, t)\right]^T (u - u^*) \geq 0 \qquad \forall v \in U$$

which is identical to the finite dimensional variational inequality principle of Theorem 6. The other optimality conditions are also identical. This completes the proof. ■

Naturally there is an associated fixed point algorithm based on the iterative scheme

$$u^{k+1} = P_U\left[u^k - \alpha F\left(x\left(u^k, u_D^k\right), u^k, u_D^k, t\right)\right]$$

The detailed structure of the fixed point algorithm is:
**Step 0.** Initialization: identify an initial feasible solution $u^0 \in U$ and set $k = 0$.
**Step 1.** Solve optimal control problem: call the solution of the following optimal control problem $u^{k+1}$.

$$\min_v J^k(v) = \gamma^T \Gamma[x(t_f), t_f] + \int_{t_0}^{t_f} \frac{1}{2}\left[u^k - \alpha F\left(x^k, u^k, u_D^k, t\right) - v\right]^2 dt \tag{26}$$

$$\text{subject to } \frac{dx}{dt} = f(x, v, v_D, t);\ x(t_0) = x^0 \tag{27}$$

$$v \in U \tag{28}$$

**Step 2.** Stopping test: if $\left\|u^{k+1} - u^k\right\| \leq \varepsilon$ where $\varepsilon \in \Re^1_{++}$ is a preset tolerance, stop and declare $u^* \approx u^{k+1}$. Otherwise set $k = k + 1$ and and go to Step 1.
The convergence of this algorithm is guaranteed by the following result:

**Theorem 8** *When $DVIC(F, f, \Gamma, D, U, x^0)$ is regular in the sense of Definition 5 and $f(x, u, u_D, t) : \left(\mathcal{H}^1[t_0, t_f]\right)^n \times \left(L^2[t_0, \tau]\right)^m \times \left(L^2[t_0, t_f]\right)^m \times \Re^1_+ \longrightarrow \left(L^2[t_0, t_f]\right)^n$ is convex, while additionally $F(x, u, u_D, t)$ is strongly monotonic for $u \in U$, the fixed point algorithm presented above converges.*

**Proof.** Consider

$$u^{k+1} - u^* = P_U\left[u^k - \alpha F\left(x\left(u^k, u_D^k\right), u^k, u_D^k, t\right)\right] - P_U\left[u^* - \alpha F\left(x\left(u^*, u_D^*\right), u^*, u_D^*, t\right)\right]$$

and note that $P_U$ is a contraction; that is, the projection of a vector is never greater in length than the length of the vector itself. Thus

$$\|P_U(v)\| \leq \|v\|$$

for any $v \in U \subseteq \left(L^2[t_0, \tau]\right)^m$. Define

$$F^k = F\left(x\left(u^k, u_D^k\right), u^k, u_D^k, t\right);\ F^* = F\left(x\left(u^*, u_D^*\right), u^*, u_D^*, t\right)$$

Because $F$ obeys a strong monotonicity condition, we have

$$\langle F^k - F^*, u^k - u^* \rangle \geq \varepsilon \left\| u^k - u^* \right\|$$

where $\varepsilon \in \Re^1_{++}$. We also know that both $\left\| F^k - F^* \right\|$ and $\left\| u^k - u^* \right\|$ are bounded, by virtue of the boundedness of $U$ and the continuity of $F$. Consequently, there must exist $\beta \in \Re^1_{++}$ such that

$$\left\| F^k - F^* \right\|^2 \leq \beta \left\| u^k - u^* \right\|^2 \tag{29}$$

The contractive property of $P_U$ and the strong monotonicity of $F$ together with property (29) mean

$$\begin{aligned} \left\| u^{k+1} - u^* \right\|^2 &\leq \left\| \left( u^k - u^* \right) - \alpha \left( F^k - F^* \right) \right\|^2 \\ &= \left\| u^k - u^* \right\|^2 + \alpha^2 \left\| F^k - F^* \right\|^2 - 2\alpha \langle F^k - F^*, u^k - u^* \rangle \\ &\leq (1 + \beta - 2\alpha\varepsilon) \left\| u^k - u^* \right\|^2 \end{aligned}$$

Note that we may chose $\alpha > 0$ such that $1 + \beta - 2\alpha\varepsilon < 1$ which is equivalent to $\alpha > \frac{\beta}{2\varepsilon}$ a condition ensuring

$$\left\| u^{k+1} - u^* \right\|^2 < \left\| u^k - u^* \right\|^2$$

Consequently, the algorithm is a strict contraction mapping and convergence is assured. ■

## 2.4 Descent in Hilbert Space for the Projection Sub-Problems

It is important to realize that the fixed point algorithm of Section 2.3 can be carried out in continuous time provided we employ a continuous time representation of the solution of each subproblem (26)-(28) from Step 1 of the fixed point algorithm. This may be done using a continuous time gradient projection method. For our present circumstances, that algorithm may be stated as

*Descent Algorithm in Hilbert Space for the Projection Sub-Problems*

**Step 0.** *Initialization.* Pick $v^{k,0}(t) \in U$ and set $j = 0$.
**Step 1.** *Finding state variables. Solve the state dynamics*

$$\frac{dx}{dt} = f\left( x, v^{k,j}, v_D^{k,j}, t \right) \tag{30}$$

$$x(t_0) = x^0 \tag{31}$$

Call the solution $x^{k,j}(t)$ .In the event a discrete time method is used to solve the state dynamics (30) and (31), curve fitting is used to obtain the continuous time state vector $x^{k,j}(t)$.
**Step 2.** *Finding adjoint variables.* Solve the adjoint dynamics

$$(-1)\frac{d\lambda}{dt} = \nabla_x H^k \left|_{x=x^{k,j}} \right. ; \ \lambda(t_f) = \frac{\partial \Gamma \left[ x^{k,j}(t_f), t_f \right]}{\partial x(t_f)} \tag{32}$$

where

$$H^k = \frac{1}{2}\left[ u^k - \alpha F\left( x^k, u^k, u_D^k, t \right) - v \right]^2 + \lambda^T f\left( x, v^{k,j}, v_D^{k,j}, t \right)$$

Call the solution $\lambda^{k,j}(t)$.In the event a discrete time method is used to solve the adjoint dynamics (**??**) and (32), curve fitting is used to obtain the continuous time adjoint vector $\lambda^{k,j}(t)$.
**Step 3.** *Finding the gradient.* Determine

$$\nabla_v J^{k,j}(t) = \nabla_v H^k$$

**Step 4.** *Stopping test. For a fixed and suitably small fixed step size*

$$\theta_k \in \Re^1_{++}$$

*determine*

$$v^{k,j+1}(t) = P_U\left[v^{k,j}(t) - \theta_k \nabla_v J^{k,j}\right] \tag{33}$$

In the event a discrete time method is used to solve the above projection subproblem, curve fitting is used to obtain the continuous time control vector (33).

**Step 5.** *Stopping test.* For $\varepsilon_2 \in \Re^1_{++}$, a pre-set tolerance, stop if $\left\|v^{k,j+1} - v^{k,j}\right\| < \varepsilon_1$and declare $v^{k*} \approx v^{k,j+1}$. Otherwise set $j = j + 1$ and go to Step 1.

This gradient projection algorithm in Hilbert space has known convergence properties. In fact the following result obtains:

**Theorem 9** *If $DVIC(F, f, \Gamma, D, U, x^0)$ is regular in the sense of Definition 5 while the conditions*

$$\langle v - v' + \lambda^T \left[\nabla_v f(x, v, v_D, t) - \nabla_v f(x, v', v'_D, t)\right], v - v'\rangle \geq \xi \|v - v'\| \tag{34}$$

*and*

$$\left\|v - v' + \lambda^T \left[\nabla_v f(x, v, v_D, t) - \nabla_v f(x, v', v'_D, t)\right]\right\| \leq \delta \|v - v'\| \tag{35}$$

*are satisfied for some $\xi, \delta \in \Re^1_{++}$ and all $v, v' \in U$, then the gradient projection algorithm for the fixed point sub-problem converges.*

**Proof.** Note that

$$\nabla_v J^k(v) = v - u^k + \alpha F\left(x^k, u^k, u^k_D, t\right) + \lambda^T \nabla_v f(x, v, vD, t)$$

From (34) we have

$$\left\langle v - u^k + \alpha F\left(x^k, u^k, u^k_D, t\right) + \lambda^T \nabla_v f(x, v, v_D, t) - \right.$$

$$\left.\left[v' - u^k + \alpha F\left(x^k, u^k, t\right) + \lambda^T \nabla_v f(x, v', v'_D, t)\right], v - v'\right\rangle \geq \xi \|v - v'\|$$

or

$$\langle \nabla_v J^k(v) - \nabla_v J^k(v'), v - v'\rangle \geq \xi \|v - v'\|$$

which is recognized as a coerciveness condition. Also (35) can be similarly re-stated as

$$\left\|\nabla_v J^k(v) - \nabla_v J^k(v')\right\| \leq \delta \|v - v'\|$$

which is recognized as a condition. Of course

$$v^{k,j+1} - v^{k*} = P_U\left[v^{k,j} - \theta_k \nabla_v J^k\left(v^{k,j}\right)\right] - P_U\left[v^{k*} - \theta_k \nabla_v J^k\left(v^{k*}\right)\right]$$

Because of the contractive nature of the projection operator, we have immediately that

$$\begin{aligned}\left\|v^{k,j+1} - v^{k*}\right\|^2 &\leq \left\|v^{k,j} - v^{k*} - \theta_k\left(\nabla_v J^k\left(v^{k,j}\right) - \nabla_v J^k\left(v^{k*}\right)\right)\right\|^2 \\ &= \left\|v^{k,j} - v^{k*}\right\|^2 + (\theta_k)^2 \left\|\nabla_v J^k\left(v^{k,j}\right) - \nabla_v J^k\left(v^{k*}\right)\right\|^2 \\ &\quad -2\theta_k \langle \nabla_v J^k\left(v^{k,j}\right) - \nabla_v J^k\left(v^{k*}\right), v^{k,j} - v^{k*}\rangle\end{aligned}$$

Because of coerciveness and the Lipschitz assumption, we have

$$\begin{aligned}\left\|v^{k,j+1} - v^{k*}\right\|^2 &\leq \left\|v^{k,j} - v^{k*}\right\|^2 + (\theta_k \delta)^2 \left\|v^{k,j} - v^{k*}\right\|^2 - 2\theta_k \xi \left\|v^{k,j} - v^{k*}\right\|^2 \\ &= \left[1 + (\theta_k \delta)^2 - 2\theta_k \xi\right] \left\|v^{k,j} - v^{k*}\right\|^2\end{aligned}$$

We may select $\theta_k$ such that $1 + (\theta_k \delta)^2 - 2\theta_k \xi < 1$ which is equivalent to a non-zero step obeying $\theta_k < \frac{2\xi}{\delta^2}$,a condition ensuring the algorithm is a strict contraction mapping. ■

# 3 Brief Overview of Friesz, Bernstein, Suo and Tobin (2001) DUE Model

Most of the dynamic network user equilibrium (DUE) models proposed to date are comprised of four essential submodels:

1. a model of path delay;
2. flow dynamics;
3. flow propagation constraints; and
4. a route/departure-time choice model.

Peeta and Ziliaskopoulos (2001), in a comprehensive review of DTA and DUE research, note that there are several published models comprised of the four submodels named above.

## 3.1 Choice of Formulation

Recently Friesz and Mookherjee (2006) have shown how the DUE formulations by Friesz et al. (1993) and Friesz et al. (2001) may be numerically solved using infinite dimensional mathematical programming and a fixed point algorithm in Hilbert space. The Friesz et al. (1993) and Friesz et al. (2001) formulations are more computationally demanding than most if not all other DUE models because of the complicated path delay operators, equations of motion and time lags they embody. As such the algorithmic results they report and which are reviewed in this paper should work as well or better when adapted to other DUE models, including those for which path delay is determined by a nonlinear response surface or by simulation for a so-called rolling horizon. In the balance of this subsection, we closely follow Friesz et al. (2001) in presenting the DUE formulation emphasized in this paper.

The network of interest will form a directed graph $G(\mathcal{N}, \mathcal{A})$, where $\mathcal{N}$ denotes the set of nodes and $\mathcal{A}$ denotes the set of arcs; the respective cardianlities of these sets are $|\mathcal{N}|$ and $|\mathcal{A}|$. An arbitrary path $p \in \mathcal{P}$ of the network is

$$p \equiv \left\{a_1, a_2, ..., a_i, ..., a_{m(p)}\right\}$$

where $\mathcal{P}$ is the set of all paths and $m(p)$ is the number of arcs of $p$. We also let $t_e$ denote the time at which flow exists an arc, while $t_d$ is the time of departure from the origin of the same flow. The exit time function $\tau^p_{a_i}$ therefore obeys

$$t_e = \tau^p_{a_i}(t_d)$$

The relevant arc dynamics are

$$\begin{aligned} \frac{dx^p_{a_i}(t)}{dt} &= g^p_{a_{i-1}}(t) - g^p_{a_i}(t) \qquad \forall p \in \mathcal{P}, \quad i \in \{1, 2, ..., m(p)\} \\ x^p_{a_i}(t) &= x^p_{a_{i,0}} \qquad \forall p \in \mathcal{P}, \quad i \in \{1, 2, ..., m(p)\} \end{aligned}$$

where $x^p_{a_i}$ is the traffic volume of arc $a_i$ contributed by path $p$, $g^p_{a_i}$ is flow exiting arc $a_i$ and $g^p_{a_{i-1}}$ is flow entering arc $a_i$ of path $p \in \mathcal{P}$. Also, $g^p_{a_0}$ is the flow exiting the origin of path $p$; by convention we call this the flow of path $p$ and use the symbolic name

$$h_p = g^p_{a_0}$$

Furthermore

$$\delta_{a_i p} = \begin{cases} 1 & \text{if } a_i \in p \\ 0 & \text{if } a_i \notin p \end{cases}$$

so that

$$x_a(t) = \sum_{p \in \mathcal{P}} \delta_{ap} x^p_a(t) \qquad \forall a \in \mathcal{A}$$

is the total arc volume.

Arc unit delay is $D_a(x_a)$ for each arc $a \in \mathcal{A}$. That is, arc delay depends on the number of vehicles in front of an auto as that auto enters an arc. Of course total path traversal time is

$$D_p(t) = \sum_{i=1}^{m(p)} \left[\tau^p_{a_i}(t) - \tau^p_{a_{i-1}}(t)\right] = \tau^p_{a_{m(p)}}(t) - t \qquad \forall p \in \mathcal{P}$$

It is expedient to introduce the following recursive relationships that must hold in light of the above development:

$$\begin{aligned}
\tau^p_{a_1}(t) &= t + D_{a_1}[x_{a_1}(t)] \quad \forall p \in \mathcal{P} \\
\tau^p_{a_i}(t) &= \tau^p_{a_{i-1}}(t) + D_{a_i}\left[x_{a_i}(\tau^p_{a_{i-1}}(t))\right] \quad \forall p \in \mathcal{P}, \quad i \in \{2,3,...,m(p)\}
\end{aligned}$$

from which we have the nested path delay operators first proposed by Friesz et al. (1993):

$$D_p(t,x) \equiv \sum_{i=1}^{m(p)} \delta_{a_i p} \Phi_{a_i}(t,x) \quad \forall p \in \mathcal{P},$$

where

$$x = (x^p_{a_i} : p \in \mathcal{P}, i \in \{1,2,...,m(p)\}$$

and

$$\begin{aligned}
\Phi_{a_1}(t,x) &= D_{a_1}(x_{a_1}(t)) \\
\Phi_{a_2}(t,x) &= D_{a_2}(x_{a_2}(t + \Phi_{a_1})) \\
\Phi_{a_3}(t,x) &= D_{a_3}(x_{a_3}(t + \Phi_{a_1} + \Phi_{a_2})) \\
&\vdots \\
\Phi_{a_i}(t,x) &= D_{a_i}(x_{a_i}(t + \Phi_{a_1} + \cdots + \Phi_{a_{i-1}})) \\
&= D_{a_i}(x_{a_i}(t + \sum_{j=1}^{i-1} \Phi_{a_j})).
\end{aligned}$$

To ensure realistic behavior, we employ asymmetric early/late arrival penalties

$$F[t + D_p(t,x) - t_A]$$

where $t_A$ is the desired arrival time and

$$\begin{aligned}
&t + D_p(t,x) > t_A \Longrightarrow \quad F(t + D_p(t,x) - t_A) = \chi^L(x,t) > 0 \\
&t + D_p(t,x) < t_A \Longrightarrow \quad F(t + D_p(t,x) - t_A) = \chi^E(x,t) > 0 \\
&t + D_p(t,x) = t_A \Longrightarrow \quad F(t + D_p(t,x) - t_A) = 0 \\
&\quad \chi^L(t,x) > \chi^E(t,x)
\end{aligned}$$

We now combine the actual path delays and arrival penalties to obtain the *effective delay operators*

$$\Psi_p(t,x) = D_p(t,x) + F\{t + D_p(x,t) - T_A\} \qquad \forall p \in \mathcal{P} \tag{36}$$

Since the volume which enters and exits an arc should satisfy the conservation law, we must have

$$\int_0^t g^p_{a_{i-1}}(t)\,dt = \int_{D_{a_i}(x_{a_i}(0))}^{t+D_{a_i}(x_{a_i}(t))} g^p_{a_i}(t)\,dt \quad \forall p \in \mathcal{P}, i \in [1, m(p)] \tag{37}$$

where $g^p_{a_0}(t) = h_p(t)$. Differentiating the both sides of (37) with respect to time $t$ and using the chain rule, we have

$$\begin{aligned}
h_p(t) &= g^p_{a_1}(t + D_{a_1}(x_{a_1}(t)))(1 + D'_{a_1}(x_{a_1}(t))\dot{x}_{a_1}) \qquad \forall p \in \mathcal{P} \\
g^p_{a_{i-1}}(t) &= g^p_{a_i}(t + D_{a_i}(x_{a_i}(t)))(1 + D'_{a_i}(x_{a_i}(t))\dot{x}_{a_i}) \qquad \forall p \in \mathcal{P}, \quad i \in [2, m(p)]
\end{aligned}$$

These are *proper flow progression constraints* derived in a fashion that make them completely *consistent with the chosen dynamics and point queue model of arc delay.* These constraints involve a state dependent time lag $D_{a_i}(x_{a_i}(t))$ but make no explicit reference to the exit time functions. These flow propagation constraints describe the expansion and contraction of vehicle platoons; they were first presented by Friesz, Tobin, Bernstein and Suo (1995), Astarita (1995), Astarita (1996) independently proposed flow propagation constraints that may be readily placed in the above form.

### 3.2 Recast of DUE as a DVI with State Dependent Time Shifts

Given the traveling cost $\Theta_p$ for path $p$, the infinite dimensional variational inequality formulation for dynamic network user equilibrium itself is: find $(g^*, h^*) \in \Omega$ such that

$$\langle \Theta\left(t, x\left(h^*\right)\right), \left(h - h^*\right)\rangle = \sum_{p\in\mathcal{P}} \int_{t_0}^{t_f} \Theta_p\left[t, x\left(h^*\right)\right]\left[h_p\left(t\right) - h_p^*\left(t\right)\right] dt \geq 0 \tag{38}$$

for all $(g, h) \in \Omega$, all of whose solutions **?** show are dynamic user equilibria[1]. In particular the solutions of (38) obey

$$\Theta_p\left(t, x^*\right) > \mu_{ij} \Longrightarrow h_p^*\left(t\right) = 0 \tag{39}$$

$$h_p^*\left(t\right) > 0 \Longrightarrow \Theta_p\left(t, x^*\right) = \mu_{ij} \tag{40}$$

for $p \in \mathcal{P}_{ij}$ where $\mu_{ij}$ is the lower bound on achievable costs for any $ij$-traveler, given by

$$\mu_p = \operatorname{ess\,inf}\left\{\Theta_p\left(t, x\right) : t \in \left[t_0, t_f\right]\right\} \geq 0$$

and

$$\mu_{ij} = \min\left\{\mu_p : p \in \mathcal{P}_{ij}\right\} \geq 0$$

We call a flow pattern satisfying (39) and (40) a *dynamic user equilibrium*. The behavior described by (39) and (40) is readily recognized to be a type of Cournot-Nash non-cooperative equilibrium. It is important to note that these conditions do not describe a stationary state, but rather a time varying flow pattern that is a Cournot-Nash equilibrium (or user equilibrium) at each instant of time.

## 4 Extensions

### 4.1 Dual Time Scales (day-to-day and within-day)

Let $\tau \in \Upsilon \equiv \{1, 2, ..., L\}$ be one typical day within the planning horizon, and take the length of each day to be $\Delta$, while the clock time within each day $\tau$ is presented by $t \in [(\tau - 1)\Delta, \tau\Delta]$ for all $\tau \in \{1, 2, ..., L\}$. The planning horizon consists of $L$ consecutive days. We assume the travel demand for each day changes based on the moving average of congestion experienced over previous days. We postulate that the travelling demand $Q_{ij}^{\tau}$ for day $\tau$ between a given O-D pair $(i, j) \in \mathcal{W}$ determined by the following system of difference equations:

$$\begin{aligned} Q_{ij}^{\tau+1} &= \left[ Q_{ij}^{\tau} - \eta_{ij}^{\tau} \left\{ \frac{\sum_{p\in\mathcal{P}_{ij}} \sum_{j=0}^{\tau-1} \int_{j\cdot\Delta}^{(j+1)\cdot\Delta} \Psi_p\left[t, x\left(h^*, g^*\right)\right] dt}{\left|\mathcal{P}_{ij}\right| \cdot \tau \cdot \Delta} - \chi_{ij} \right\} \right]^{+} \qquad \forall \tau \in \{1, 2, ...L-1\} \\ Q_{ij}^{1} &= \tilde{Q}_{ij} \end{aligned} \tag{41}$$

where $\tilde{Q}_{ij} \in \mathbb{R}_+$ is the fixed traveling demand for the O-D pair $(i, j) \in \mathcal{W}$ for the first day. The operator $[x]^+$ is equivalent to $\max[0, x]$.

### 4.2 Uncertain Travel Demand Information

Once again let us assume $\tau \in \Upsilon \equiv \{1, 2, ..., L\}$ be one typical day within the planning horizon, and take the length of each day to be $\Delta$, while the clock time within each day $\tau$ is presented by $t \in [(\tau - 1)\Delta, \tau\Delta]$ for all $\tau \in \{1, 2, ..., L\}$. where the planning horizon consists of $L$ consecutive days. Here we assume that the travel demand for each day is a random variable in the following multiplicative form

$$\hat{Q}_{ij}^{\tau} = Q_{ij}^{\tau} \cdot z_{ij}$$

[1] Although we have purposely supressed the functional analysis subtleties of the formulation, it should be noted that (38) involves an inner product in a Hilbert space, namely $\left(L^2\left[0, T\right]\right)^{|\mathcal{P}|}$.

where $\hat{Q}_{ij}^{\tau}$ is the realized travel demand on day $\tau$ between the OD pair $(i,j)$ where as $z_{ij}$ is the random variable. To keep exposition simple we assume that distribution of $z_{ij}$ is known exactly, however, it can further be generalized to have only partial information (e.g., first and second moments) about $z_{ij}$. The average travel volume, $Q_{ij}^{\tau}$ may be computed from (41).

# 5 Numerical Example

In what follows, we consider a 5 arc, 4 node traffic network shown below. The forward star array and arc delay functions $D_a\left(x_a\left(t\right)\right)$ for all 5 arcs of the network are contained in the following table:

| Arc name | From node | To node | Arc Delay, $D_a\left(x_a\left(t\right)\right)$ |
|---|---|---|---|
| $a_1$ | 1 | 2 | $\frac{1}{2}+\frac{x_{a_1}}{70}$ |
| $a_2$ | 1 | 3 | $1+\frac{x_{a_2}}{150}$ |
| $a_3$ | 2 | 3 | $\frac{1}{2}+\frac{x_{a_3}}{100}$ |
| $a_4$ | 2 | 4 | $1+\frac{x_{a_4}}{150}$ |
| $a_5$ | 3 | 4 | $\frac{1}{2}+\frac{x_{a_5}}{100}$ |

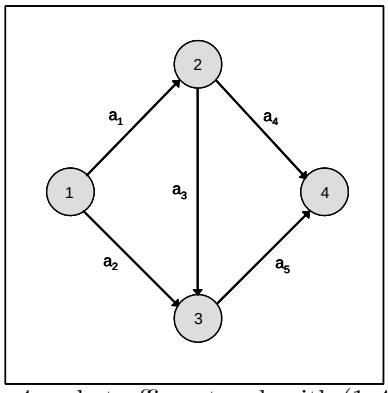



Fig 1 : The 5-arc 4-node traffic network with $(1,4)$ being the OD-pair

There is a travel demand of $Q_{14}^{1}=75$ units from node 1 (origin) to node 4 (destination) on day 1. There are 3 paths connecting nodes 1 through 4, namely

$$\begin{aligned}\mathcal{P}_{14} &= \{p_1,p_2,p_3\}\\ p_1 &= \{a_1,a_4\}\\ p_2 &= \{a_2,a_5\}\\ p_3 &= \{a_1,a_3,a_5\}\end{aligned}$$

We consider the planning horizon to be 4 days (i.e., $L=4$) and the length of each day is $\Delta=24$ hours. The desired arrival time for commuters is $T_A=13$ (1:00 PM of every day).The controls (path flows and arc exit flows) and states (arc traffic volumes) are enumerated in the following table:

| Paths | Path Flows | Arc Exit Flows | Traffic Volume of Arcs |
|---|---|---|---|
| $p_1$ | $h_{p_1}$ | $g_{a_1}^{p_1},g_{a_4}^{p_1}$ | $x_{a_1}^{p_1},x_{a_4}^{p_1}$ |
| $p_2$ | $h_{p_2}$ | $g_{a_2}^{p_2},g_{a_5}^{p_2}$ | $x_{a_2}^{p_2},x_{a_5}^{p_2}$ |
| $p_3$ | $h_{p_2}$ | $g_{a_1}^{p_3},g_{a_3}^{p_3},g_{a_5}^{p_3}$ | $x_{a_1}^{p_3},x_{a_3}^{p_3},x_{a_5}^{p_3}$ |

We consider the symmetric early/late arrival penalty

$$F\left[t+D_p\left(x,t\right)-T_A\right]=\left[t+D_p\left(x,t\right)-T_A\right]^2$$

Furthermore, without any loss of generality, we take the initial traffic volumes on every arc to be zero:

$$x_{a_i}^{p}\left(0\right)=0 \qquad \forall p\in\mathcal{P}, i\in\left[1,m\left(p\right)\right]$$

We forgo the detailed symbolic statement of this example and instead provide numerical results in graphical form for an essentially exact solution achieved after 29 iterations of the fixed point algorithm. Figures 2, 3 and 4 depict departure rates and arc exit flows for paths $p_1, p_2$ and $p_3$ respectively.

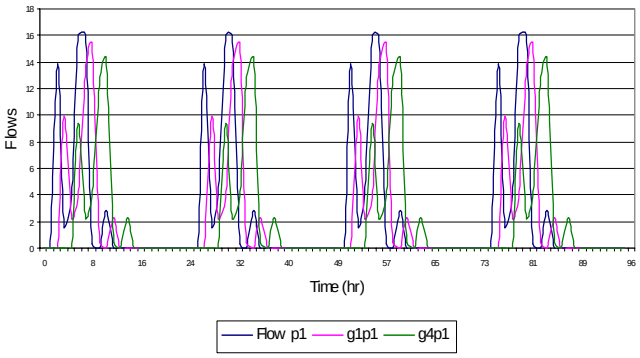



Fig 2 : Path and arc exit flows for path 1

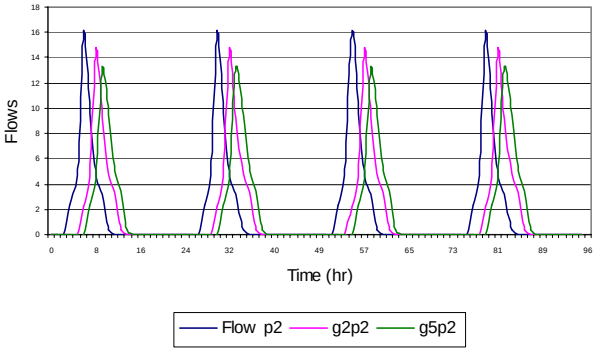



Fig 3 : Path and arc exit flows for path 2

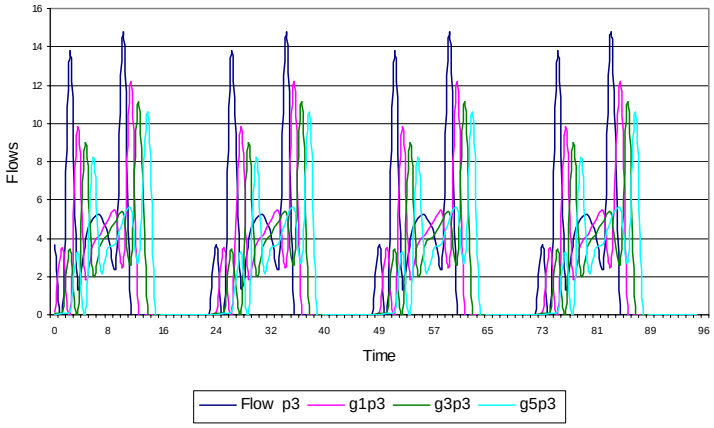



Fig 4 : Path and arc exit flows for path 3

Cumulative traffic volumes on the 5 different arcs are plotted against time in Figure 5 where

$$\begin{aligned} x_{a_1}(t) &= x_{a_1}^{p_1}(t) + x_{a_1}^{p_3}(t) \\ x_{a_2}(t) &= x_{a_2}^{p_2}(t) \\ x_{a_3}(t) &= x_{a_3}^{p_3}(t) \\ x_{a_4}(t) &= x_{a_4}^{p_1}(t) \\ x_{a_5}(t) &= x_{a_5}^{p_2}(t) + x_{a_5}^{p_3}(t) \end{aligned}$$

for all time $t \in [0, L\Delta]$.

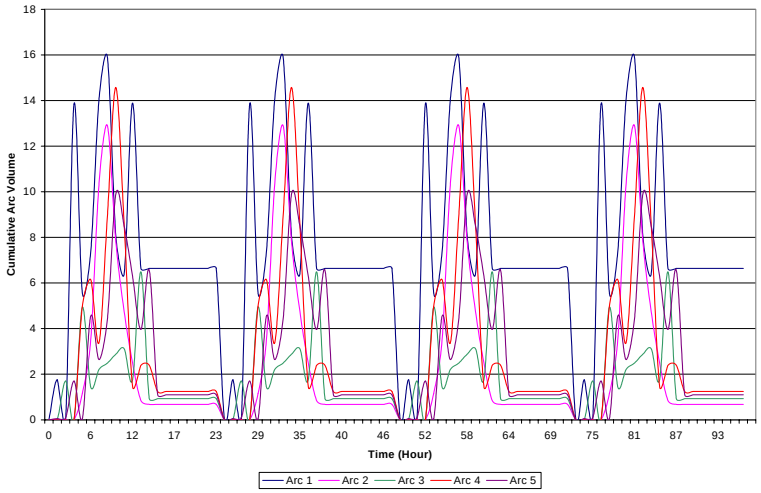



Fig 5 : Cumulative arc volume vs. time

Note that the effective path delay operator in (36) gives the unit travel cost along a path $p$ at time $t$. Figure 6 analyzes the effective delay and flow for path $p_2$ by plotting both for the same time scale which shows that path flow is maximal when the associated unit travel cost (effective path delay) is at its well defined minimum.

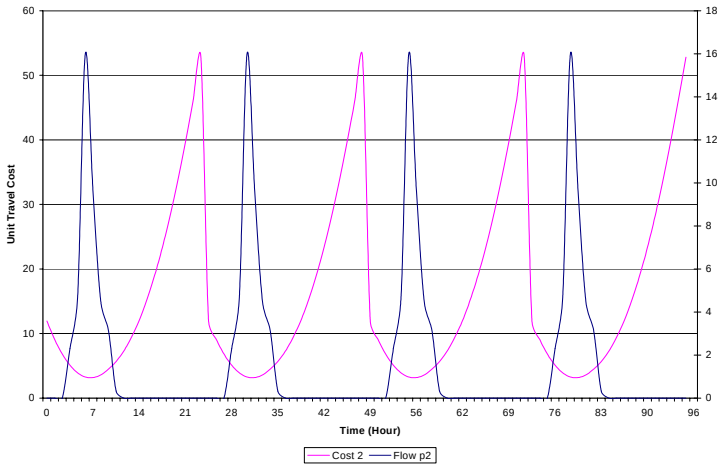



Fig 6 : Comparison of path flows and associated unit travel costs for path $p_2$

Net travel demand and demand reduction are plotted below against the same time scale (day) which clearly demonstrates that more commuters switch to alternative mode (e.g., telecommuting) as their rolling average experience

of congestion increases with passage of time.

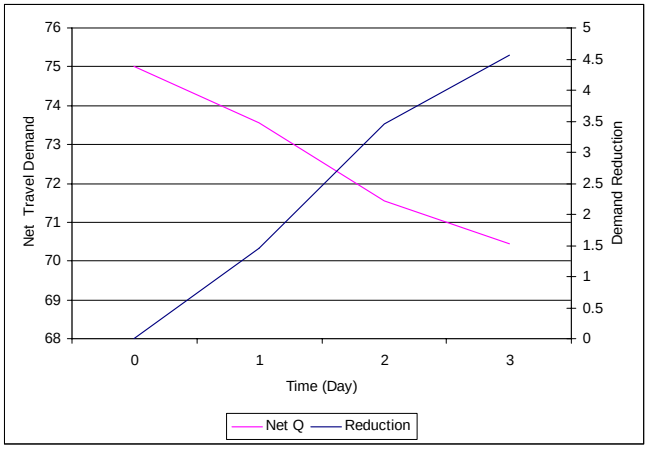



Fig 7 : Net travel demand and demand reduction

# 6 Concluding Remarks

We have explained how traditional non-cooperative differential game theory may be extended to accomodate the natural formulation of DUE as an infinite dimensional variational inequality involving explicit state-dependent time shifts. We show that such a perspective is not only useful for analysis but also leads to simple yet effective algorithms for the computation of DUE solutions. We also apply this formalism to create two entirely new formulations of dynamic user equilibrium when: (1) there are dual time scales (day-to-day and within-day); and (2) demand information is uncertain. Our future DUE reserach will provide a more in-depth analysis of the stochastic DUE problem in the presence of incomplete traffic information.

# References

Astarita, V.: 1995, Flow propagation description in dynamic network loading models, *Proc. of International Conference on Applications of Advanced Technologies in Transportation Engineering*.

Astarita, V.: 1996, A continuous time link model for dynamic network loading based on travel time functions, *Proc. of 13th International Symposium on Theory of Traffic Flow*, pp. 107 – 126.

Friesz, T. L., Bernstein, D., Smith, T. E., Tobin, R. L. and Wie, B. W.: 1993, A variational inequality formulation of the dynamic network user equilibrium problem, *Operations Research* **41**, 179 – 191.

Friesz, T. L., Bernstein, D., Suo, Z. and Tobin, R. L.: 2001, Dynamic network user equilibrium with state-dependent time lags, *Networks and Spatial Economics* **1**, 319 – 347.

Friesz, T. L. and Mookherjee, R.: 2006, Solving the dynamic network user equilibrium problem with state-dependent time shifts, *Transportation Research Part B* **40**(3), 207 – 229.

Friesz, T. L., Tobin, R. L., Bernstein, D. and Suo, Z.: 1995, Proper flow propagation constraints which obviate exit functions in dynamic traffic assignment, *INFORMS Spring National Meeting, Los Angeles, April 23-26* .

Minoux, M.: 1986, *Mathematical Programming*, John Wiley Sons.

Peeta, S. and Ziliaskopoulos, A.: 2001, Foundations of dynamic traffic assignment, *Networks and Spatial Economics* **1**(3), 233 – 265.

# A Link-Node Complementarity Model and Solution Algorithm for Dynamic User Equilibria with Exact Flow Propagations

**Jeff X. Ban***
California Center for Innovative Transportation (CCIT)
Institute of Transportation Studies (ITS)
University of California
2105 Bancroft Way, Suite 300
Berkeley, CA 94720
Tel: (510) 642-5112, Fax: (510) 642-0910
Email: xban@berkeley.edu
* **Corresponding Author**

**Henry X. Liu**
Department of Civil Engineering
University of Minnesota
122 Civil Engineering Building
500 Pillsbury Drive S.E.
Minneapolis, MN 55455
Tel: (612) 625-6347, Fax: (612) 626-7750
Email: henryliu@umn.edu

**Michael C. Ferris**
Computer Sciences Department
University of Wisconsin at Madison
Tel: (608) 262-4281, Fax: (608) 262-9777
Email: ferris@cs.wisc.edu

**Bin Ran**
Department of Civil and Environmental Engineering
University of Wisconsin at Madison
Tel: (608) 262-0052, Fax: (608) 262-5199
Email: bran@engr.wisc.edu

**Submittal to 2006 DTA Symposium**

# A link-node complementarity model and solution algorithm for dynamic user equilibria with exact flow propagations

**Jeff X. Ban[1], Henry X. Liu[2], Michael C. Ferris[3], and Bin Ran[4]**

[1] California Center for Innovative Transportation, ITS, University of California, 2105 Bancroft Way, Suite 300, Berkeley, CA 94720
[2] Department of Civil Engineering, University of Minnesota, Twin-Cities
[3] Computer Sciences Department, University of Wisconsin at Madison
[4] Department of Civil and Environmental Engineering, University of Wisconsin at Madison

## Abstract

We propose a link-node based complementarity formulation for the basic deterministic dynamic user equilibrium problem with single-user-class and fixed demands. In particular, the proposed model captures the exact flow propagation constraints that were usually approximated by previous studies. The solution existence and compactness condition for the proposed model is established under mild assumptions. The model is solved by an iterative algorithm with a relaxed NCP solved accurately and efficiently at each iteration. Therefore, the required number of relaxed NCP solves is fairly small. Numerical examples are also provided in this paper to illustrate the proposed model and solution approach.

## 1. Introduction

Dynamic traffic assignment (DTA) which can predict, in a short-term fashion, the future dynamic traffic states, has been extensively studied for decades, particularly accelerated in the last fifteen years since the emergence of the intelligent transportation systems (ITS). In this paper, we are interested in the so-called dynamic user equilibrium (DUE) which is the fundamental yet most challenging problem of DTA. Two distinct approaches have dominated the methodologies applied to the DTA research: the simulation-based (microscopic/mesoscopic) and the analytical (macroscopic) approach. In this paper, we focus on the analytical DTA models, especially the variational inequality (VI) method which is more capable of computing dynamic network equilibria than the constrained optimization approach (Ran and Boyce, 1996; Chen, 1999; Peeta and Ziliouskopolous, 2001).

VI has been applied for long for modeling various traffic interactions for static traffic assignment problems (Smith, 1979; Dafermos, 1981). It had not been used to model DTA until Friesz et al. (1993). Later on, Ran and Boyce (1994, 1996) extensively studied the issues of applying VI to formulate and solve DTA problems. The VI approach has also been used for DTA study by Lo et al. (1996), Ran et al. (1996), Heydecker and Verlander (1999), to name just a few. In particular, Friesz et al. (2001) and Friesz and Mookherjee (2005) developed the differential variational inequality (DVI) technique to model and solve DUE in the continuous time domain.Although formulated continuously in the temporal domain, most DUE models, e.g. those by Friesz et al. (1993) and Ran and Boyce (1996), were solved by the time discretization since to date solving the continuous-time DUE model directly for practical transportation networks is still not feasible. However, the discretized models were only treated as part of the solution procedure without rigorous investigations on their mathematical properties such as the solution existence conditions. Chen and Heush (1998) were among the first to investigate explicitly on the discrete-time VI model for DTA. Bliemer (2001) and Bliemer and Bovy (2003) further improved the model by Chen and Heush (1998) and proposed the link-route based quasi-variational inequality (QVI) formulation. Lo and Szeto (2002) integrated the cell transmission model (CTM) into DUE which was formulated as a route based VI problem. Due to the nature of CTM (Daganzo, 1994, 1995a), the model by Lo and Szeto (2002) is discretized in both temporal and spatial domains.

As one special case of VI, the nonlinear complementarity problem (NCP) has been fully studied in the mathematical programming community (Facchinei and Pang, 2003 and references therein). Efficient

solution approaches have been developed during the last decade for solving large scale NCPs (Cao and Ferris, 1996; Billups et al., 1997; Ferris et al., 1999). Generally, solving an NCP is much easier than a regular VI. However to date, NCP has not been widely applied in modeling and solving DUE. The only study, to the best of the author's knowledge, is Wie et al. (2002) which formulated the discrete-time DUE with departure time choice as an NCP. It was also pointed out in Wie et al. (2002) that continuous-time and discrete-time DUE models are significantly different. Generally, the former are infinite dimensional mathematical programming problems while the latter are finite dimensional mathematical programming problems. The NCP based DUE model by Wie et al. (2002), nevertheless, projected the link exit flows to two neighboring time grids. In this sense, the exact flow propagation of DUE (Astarita, 1996) was not fully respected. Further, the linear programming based solution approach in Wie et al. (2002) is similar to the Frank-Wolfe (FW) algorithm. Due to the well-known convergence problem of FW, such a method may not be effective for solving DUE, especially for producing accurate solutions.

In this paper, we formulate the discrete-time DUE problem with exact flow propagations as a link-node based NCP. We consider the basic discrete-time DUE problem which is deterministic, single-user-class with fixed travel demands. We first start with the continuous-time DUE and formulate it as an infinite-dimensional mixed complementarity problem (MiCP) with side constraints. Based on this MiCP model, we adopt the discretization scheme by Astarita (1996) and prove that an NCP formulation exists for discrete-time DUE with exact flow propagations. We further prove that the solution set of the proposed NCP is nonempty and compact, a sharper result compared with that in Wie et al. (2002). To solve the NCP, we develop an iterative algorithm with a relaxed NCP solved at each iteration. Due to the exact solve of each relaxed NCP, the solution process requires much less iterations than previous methods, which is demonstrated by the case study conducted in this paper.

This paper is organized as follows. The continuous time DUE formulation is first discussed in Section 2, with a MiCP formulation provided. In Section 3, the method for deriving the discrete-time DUE model is presented with exact flow propagations. The derivation of the NCP model and its solution existence condition is established. Section 4 mainly presents the solution algorithm of the proposed model, including the network loading process and two gap functions to monitor the convergence of the algorithm. Numerical examples are provided in Section 5, followed by the concluding remarks and future research directions in Section 6.

## 2. Continuous time DUE model

In this section, we introduce the link-node based continuous-time formulation for DUE. Friesz et al. (1993), Ran and Boyce (1996), and Bliemer and Bovy (2003) have extensively studied the path-based and link-based continuous-time DUE models.

Assume a given transportation network can be represented as a connected and directed graph, denoted as $G(N, A)$, where $N$ is the set of nodes and $A$ is the set of links (arcs). Since we are dealing with dynamic (or time-varying) traffic flows, we denote $t \in [0, T']$ as the continuous time and $T'$ is the total study period. Also denote $R$ and $S$ as the origin and destination node set, respectively. Throughout this paper, we will use index $a \in A$ to denote a link, index $i \in N$ or $j \in N$ to denote a node, and index $s \in S$ to denote a destination. Moreover, we only consider the basic DUE problem, i.e., the deterministic and single-user-class DUE with fixed demands, for which $d_{is}(t)$ denotes the (fixed) travel demand rate from node $i$ to destination $s$ at time instant $t$. We conventionally set $d_{ss}(t) = 0, \forall s \in S, t \in [0, T']$.

### *2.1 DUE Condition*

The link-node model is derived from the so-called (link-node based) DUE condition which describes the optimality condition of the DUE problem. The DUE condition is a dynamic extension to the Wardrop's first principle (Wardrop, 1952) for the static case and can be stated as follows:

*If, from each decision node to every destination node at each instant of time, the actual travel times for all the routes that are being used are equal and minimal, then the dynamic traffic flow over the network is in a travel-time based dynamic user equilibrium (DUE) state.*

In this condition, a "decision node" with respect to a given destination node means any node in the network which either generates OD trips (i.e., origin nodes) or is traversed by flows heading to the destination (i.e., the intermediate nodes). Hence, we require that a destination node must not be a decision node of itself. Mathematically, the DUE condition can be expressed as follows:

$$0 \le u_{as}(t) \perp \{\tau_a(t) + \pi_{h_a s}[t + \tau_a(t)] - \pi_{l_a s}(t)\} \ge 0, \forall a \in A, s \in S, t \in [0, T']. \quad (1)$$

Here $u_{as}(t)$ denotes the inflow rate to link *a* with respect to destination *s* at time instant *t*, $\pi_{l_a s}(t)$ and $\pi_{h_a s}(t)$ denote, respectively, the minimum travel time from the tail node ($l_a$) and head node ($h_a$) of link *a* to destination *s* at time *t* and $\tau_a(t)$ is the link travel time for link *a* at time instant *t*. We conventionally set $\pi_{ss}(t) = 0, \forall s \in S, t \in [0, T']$. In addition, "$\perp$" in (1) means "perpendicular" so that $x \perp y \Leftrightarrow x^T y = 0$. Note that the DUE condition can also be conveniently expressed in a Dynamic Programming (DP) manner. For details, one can refer to Han and Heydecker (2006).

### *2.2 Dynamic Network Constraints*

The dynamic network constraints describe the defining set of the DUE problem, which must be satisfied by any feasible solution. Five types of constraints have been identified in the literature (Ran and Boyce, 1996; Bliemer and Bovy, 2003), namely the mass balance constraints, the flow conservation constraints, the flow propagation constraints, First-in-first-out (FIFO) constraints, and other definitional constraints. In the following, we simply list these constraints without further discussions. For detailed descriptions, one can refer to Ran and Boyce (1996) and Bliemer and Bovy (2003).

#### *2.2.1 Mass Balance Constraints*

Mass balance constraints, as shown in equation (2), define the relationship between the link flow on a given link *a* with respect to destination *s* at time *t*, denoted as $x_{as}(t)$, and the inflow rate and exit flow rate for the same link to the same destination at the same time instant (denoted as $u_{as}(t)$ and $v_{as}(t)$, respectively).

$$\frac{dx_{as}(t)}{dt} = u_{as}(t) - v_{as}(t), \forall a \in A, s \in S, t \in [0, T'] \quad (2)$$

Equation (2) implies that $x_{as}(t) = \int_0^t [u_{as}(w) - v_{as}(w)]dw + x_{as}(0), \forall a \in A, s \in S, t \in [0, T']$. We further assume the initial condition as

$$x_{as}(0) = 0, \forall a \in A, s \in S. \quad (3)$$

Then, the mass balance constraints (2) can be rewritten as:

$$x_{as}(t) = \int_0^t [u_{as}(w) - v_{as}(w)]dw, \forall a \in A, s \in S, t \in [0, T']. \quad (4)$$

#### *2.2.2 Flow Conservation Constraints*

The flow conservation constraints require that for a given destination at any given time instant, the flows entering any node, together with the demands generated at this node, must all exit from this node unless this node is the destination itself. Mathematically, they can be expressed as:

$$\sum_{a\in A(i)} u_{as}(t) = d_{is}(t) + \sum_{a\in B(i)} v_{as}(t), \forall i \in N, s \in S, i \neq s, t \in [0, T'], \tag{5}$$

where $A(i)$ is the set of links whose tail nodes are *i* and $B(i)$ is the set of links whose head nodes are *i*.

*2.2.3 Flow Propagation Constraints*

These constraints describe the consistent evolvement of traffic flows in both temporal and spatial domains. It has been proved that in a continuous time fashion, the flow propagation constraints can be represented as (Astarita, 1996):

$$v_{as}[t'+\tau_a(t')] = \frac{u_{as}(t')}{1+d\tau_a(t')/dt'}, \forall a \in A, s \in S, t' \in [0, T']. \tag{6}$$

Note that equation (6) can be easily obtained by differentiating both sides of the Type I flow propagation constraints in Ran and Boyce (1996, page 74, equation (4.22)). Equation (6) can be rewritten as

$$v_{as}(t) = \int_{t'\in[0,T']} \lambda_a(t', t) u_{as}(t') d\mu(t'), \forall a \in A, s \in S, t \in [0, T']. \tag{7}$$

Here $\lambda_a(t', t)$ is denoted as the indicator function defined as follows:

$$\lambda_a(t', t) = \begin{cases} \dfrac{1}{1+d\tau_a(t')/dt'}, & \text{if } t = t'+\tau_a(t'). \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

From (8), it is obvious that $\lambda_a(t', t)$ is not continuous. Therefore, the integral in (7) is a Lebesgue integral and $\mu(t')$ is the corresponding Lebesgue measure. The indicator function defined in (8) is similar to the "dynamic effective flow rate factor" proposed by Bliemer (2000). However, our definition here is link-based and thus much simpler compared with that in Bliemer (2000). Also, since we assume vehicles to different destinations experience the same travel time as long as they enter the link at the same time, $\lambda_a(t', t)$ is independent of individual destinations.

*2.2.4 FIFO Constraints*

FIFO constraints were first introduced into the DTA study by Carey (1987). Since then, it has been assumed to be a "discipline" respected by the dynamic traffic flows. FIFO requires that any vehicle entering into a link earlier must also exit from the link earlier. Ran and Boyce (1996) have shown that in order for FIFO to hold, an extra restriction should be imposed on $\tau_a(t)$ as $d\tau_a(t)/dt > -1$. However, explicitly imposing such a constraint will dramatically increase the complexity of the resulting model. Therefore, most of the DUE models tend to implicitly guarantee FIFO by choosing a proper link performance function. How to design such a function, however, is beyond the scope of this paper. Here we conventionally assume $\tau_a(t)$ is a function of the link flow at time *t* on link *a*:

$$\tau_a(t) = g(x_a(t)). \tag{9}$$

Here $x_a(t)$ is the total link flow on link *a* at time *t* (see equation (10) below). The function in (9) satisfies FIFO when *g* is linear, or if the gradient of *g* with respect to *x* is bounded from above when *g* is nonlinear. For more discussions, one can refer to Nie and Zhang (2005), Carey et al. (2003), and Xu et al. (1999). Note that such a function is only suitable for moderately congested networks (Daganzo, 1995). More sophisticated functional forms should be adopted for networks with heavy congestion.

*2.2.5 Other Definitional Constraints:*

Other definitional constraints are listed in equations (10) and (11) as follows.

$$\begin{cases} u_a(t) = \sum_{s\in S} u_{as}(t) \\ v_a(t) = \sum_{s\in S} v_{as}(t), \forall a \in A, t \in [0,T'] \\ x_a(t) = \sum_{s\in S} x_{as}(t) \end{cases}, \tag{10}$$

$$\begin{cases} u_{as}(t) \geq 0, v_{as}(t) \geq 0, x_{as}(t) \geq 0, \forall a \in A, s \in S, t \in [0,T'] \\ \pi_{is}(t) \geq 0, \forall i \in N, s \in S, i \neq s, t \in [0,T'] \end{cases}, \tag{11}$$

Here $u_a(t), v_a(t)$ denote, respectively, the total inflow rate and exit flow rate of link *a* at time *t*. Together with $x_a(t)$, they are referred as "aggregated" variables, while $u_{as}(t)$, $v_{as}(t)$ and $x_{as}(t)$ are called the "disaggregated" variables. For aggregated and disaggregated variables that satisfy (10), we call them "corresponding" to each other.

*2.3 MiCP Formulation*

Equations (1), (4), (5), (7) and (10) – (11) constitute the DUE model defined on disaggregated variables $x = (x_{as}(t))_{\forall a,s,t}, u = (u_{as}(t))_{\forall a,s,t}$, $v = (v_{as}(t))_{\forall a,s,t}$ and $\pi = (\pi_{is}(t))_{\forall i,s,t;i\neq s}$; whereas $\lambda = (\lambda_a(t',t))_{\forall a,t,t'}$ and $\tau = (\tau_a(t))_{\forall a,t}$ can be treated as functions of these defining variables. However, such a DUE model can be further simplified. Firstly, from equations (4) and (7), *x* and *v* can be readily represented by *u* and $\lambda$. Then we can model DUE using *u* and $\pi$ only, while both $\lambda$ and $\tau$ are functions of *u*. In this manner, the only significant equations are (1), (5), the nonnegativity constraints on *u* and $\pi$ in (11), and the definitions of $\lambda$ and $\tau$ in (8) and (9). As a result, we have the following model for DUE: trying to find ($u$, $\pi$) such that the following is satisfied

$$\begin{cases} 0 \leq u_{as}(t) \perp \{\tau_a(t) + \pi_{h_a s}[t+\tau_a(t)] - \pi_{l_a s}(t)\} \geq 0, \forall a \in A, s \in S, t \in [0,T'], & \text{(12a)} \\ \sum_{a\in A(i)} u_{as}(t) = d_{is}(t) + \sum_{a\in B(i)} \int_0^t \lambda_a(t',t) u_{as}(t') d\mu(t'), \forall i \in N, s \in S, i \neq s, t \in [0,T'], & \text{(12b)} \\ \pi_{is}(t) \geq 0, \forall i \in N, s \in S, i \neq s, t \in [0,T'] & \text{(12c)} \end{cases}$$

, whereas $\lambda$ and $\tau$ are defined in equations (8) and (9) respectively. Variables *x* and *v* are indeed intermediate and can be expressed by *u* through equations (4) and (7). Clearly, (12a) just repeats (1), and (12b) and (12c) correspond to (5) and the nonnegativity on $\pi$ in (10). Note that (12a) and (12b) define an infinite dimensional MiCP (Ulbrich, 1999), while (12c) imposes a side constraint which requires the minimum travel time from node *i* to destination *s* at time *t* must be nonnegative. Solving the infinite dimensional MiCP (12) with side constraints is generally difficult and we thus study the discretized problem starting from the next section.

## 3. Discrete time DUE model

*3.1 Discrete Time DUE with Exact Flow Propagation*

In order to obtain the discrete-time model, we can evenly divide the entire study period into $K'$ time intervals by introducing the length of each time interval $\Delta$ such that $K'\Delta = T'$. The notation at discrete-time is first listed as follows:

$u_{as}^k = u_{as}(k\Delta)$: the inflow rate to link *a* towards destination *s* at time interval *k*, $u = (u_{as}^k)_{\forall a,s,k}$

$v_{as}^k = v_{as}(k\Delta)$: the exit flow rate from link *a* towards destination *s* at time interval *k*, $v = (v_{as}^k)_{\forall a,s,k}$

$u_{as}^k\Delta$: the inflows to link *a* towards destination *s* during time interval *k*

$v_{as}^k\Delta$: the exit flows from link *a* towards destination *s* during time interval *k*

$x_{as}^k = x_{as}(k\Delta)$ : the flows of link $a$ towards destination $s$ at time interval $k$, $x = (x_{as}^k)_{\forall a,s,k}$

$u_s^k = \sum_{s\in S} u_{as}^k$ : the aggregated inflow rate to link $a$ at time interval $k$, $u^A = (u_a^k)_{\forall a,k}$

$v_a^k = \sum_{s\in S} v_{as}^k$ : the aggregated exit flow rate from link $a$ at time interval $k$, $v^A = (v_a^k)_{\forall a,k}$

$x_a^k = \sum_{s\in S} x_{as}^k$ : the aggregated flows of link $a$ towards destination $s$ at time interval $k$, $x^A = (x_a^k)_{\forall a,k}$

$\tau_a^k(u) = \tau_a(k\Delta)$ : the travel time of link $a$ at time interval $k$, a function of $u$, $\tau = (\tau_a^k)_{\forall a,k}$

$\pi_{is}^k = \pi_{is}(k\Delta)$ : the minimum travel time from node $i$ to destination $s$ at time interval $k$, $\pi = (\pi_{is}^k)_{\forall i,s,k,i\neq s}$

$d_{is}^k = d_{is}(k\Delta)$ : the demand rate generated from node $i$ to destination $s$ at time interval $k$, $d = (d_{is}^k)_{\forall i,s,k,i\neq s}$

$d_{is}^k\Delta$ : the demands generated from node $i$ to destination $s$ during time interval $k$

$e_a^k(u) \equiv (k-1)\Delta + \tau_a^k(u)$ : the exit time for vehicles *e*ntering $a$ at the start of time interval $k$, a function of $u$, $e = (e_a^k)_{\forall a,k}$

Using this simple discretization scheme, one should be cautious about how to discretize (12a) and 12(b). First of all, (12b) is related to the flow conservation (5) and flow propagation (6). Assuming $(k-1)\Delta = t$, (5) can be easily discretized as

$$\sum_{a\in A(i)} u_{as}^k = d_{is}^k + \sum_{a\in B(i)} v_{as}^k, \forall i\in N, s\in S, i\neq s, k = 1,2,\cdots,K' \cdot \tag{13}$$

According to (6), the exit flow rate $v_{as}^k$ is related to inflow rate $u_{as}^{k'}$ if $e_a^{k'}(u) = (k'-1)\Delta + \tau_a^{k'} = k\Delta$. However, such an integer $k'$ may not exist since travel time $\tau_a^{k'}$ is real valued. In this paper, we adopt the discretization method in Astarita (1996) to address this problem. The method first constructs the pair of $(e_a^{k'}, v_{as}(e_a^{k'}))$ for any inflow at $k'$ and we then have, by discretizing (6),

$$v_{as}(e_a^{k'}) = u_{as}^{k'} \cdot \lambda_a^{1,k'}(u), \forall a\in A, s\in S, k' = 1,2,\cdots,K', \tag{14}$$

and

$$\lambda_a^{1,k'}(u) = \frac{\Delta}{\tau_a^{k'+1}(u) - \tau_a^{k'}(u) + \Delta} \cdot \tag{15}$$

Note that $\lambda_a^{1,k'}$ is a function of $u$ since $\tau_a^{k'}$ is so. Then for any $k$, we try to find $k'$ such that $e_a^{k'} \le k\Delta < e_a^{k'+1}$, as shown in Figure 1.

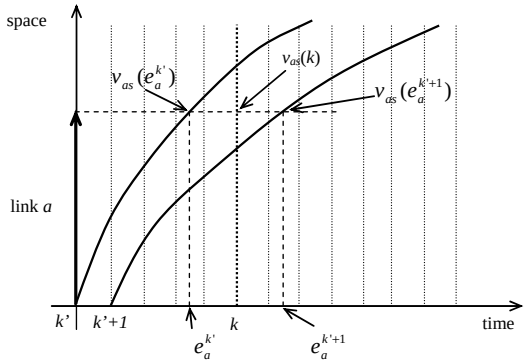



**Figure 1 Discretization of flow propagation constraints**

In Figure 1, the solid line represents the trajectory of traffic flows entering link $a$ at time $k'$ or $k'+1$. The exit flow rate at time interval $k$ can be computed using linear interpolation of $v_{as}(e_a^{k'})$ and $v_{as}(e_a^{k'+1})$ as:

$$v_{as}^{k} = \sum_{k';e_a^{k'} \le k\Delta < e_a^{k'+1}} \lambda_a^{2,k',k}(u) \cdot v_{as}(e_a^{k'}) + (1 - \lambda_a^{2,k',k}(u)) \cdot v_{as}(e_a^{k'+1}), \tag{16}$$

where $\lambda_a^{2,k',k}(u)$ is defined as:

$$\lambda_a^{2,k',k}(u) = \frac{e_a^{k'+1} - k\Delta}{e_a^{k'+1} - e_a^{k'}} = \frac{\tau_a^{k'+1}(u) + (k'-k)\Delta}{\tau_a^{k'+1}(u) - \tau_a^{k'}(u) + \Delta}. \tag{17}$$

Substitute equations (14) – (17) to (13), we obtain the discretized version of (12b) as

$$\sum_{a \in A(i)} u_{as}^{k} = d_{is}^{k} + \sum_{a \in B(i)} \sum_{k';e_a^{k'} \le k\Delta < e_a^{k'+1}} [\lambda_a^{1,k'}(u) \cdot \lambda_a^{2,k',k}(u) \cdot u_{as}^{k'} + \lambda_a^{1,k'+1}(u) \cdot (1 - \lambda_a^{2,k',k}(u)) \cdot u_{as}^{k'+1}], \forall i \in N, s \in S, i \ne s, k = 1, \cdots, K'. \tag{18}$$

Similarly, to discretized $\pi_{h_a s}[t + \tau_a(t)]$ in (12a), we project it to the two neighboring grids of $e_a^k = (k-1)\Delta + \tau_a^k$, denoted as $l$ and $l$+1, such that $l\Delta \le e_a^k < (l+1)\Delta$. In other words,

$$\pi_{h_a s}[t + \tau_a(t)] = \pi_{h_a s}(e_a^k) = \lambda_a^{3,k,l}(u) \cdot \pi_{h_a s}^{l} + [1 - \lambda_a^{3,k,l}(u)] \cdot \pi_{h_a s}^{l+1}. \tag{19}$$

Here we define

$$\lambda_a^{3,k,l}(u) = (l + 2 - k - \frac{\tau_a^k(u)}{\Delta}). \tag{20}$$

Some observations thus follow. First, all $\lambda^1 = (\lambda_a^{1,k'})_{\forall a,k'}, \lambda^2 = (\lambda_a^{2,k',k})_{\forall a,k',k;e_a^{k'} \le k\Delta < e_a^{k'+1}}, \lambda^3 = (\lambda_a^{3,k,l})_{a,k,l;l \le e_a^k/\Delta < l+1}$ and $e$ are functions of $\tau$ which is itself a function of the aggregated link inflow rate vector $u^A$. Due to the relation between aggregated and disaggregated variables in (10), the above four functions are also defined on the disaggregated link inflow vector $u$. Furthermore, since we assume the link travel time defined in (9) satisfies FIFO, $\lambda_a^{1,k'} > 0, \forall a, k'$. Meanwhile, from the definition of $\lambda^2$ and $\lambda^3$, we can easily observe $0 < \lambda_a^{2,k',k} \le 1$, $\forall a, k', k; e_a^{k'} \le k\Delta < e_a^{k'+1}$ and $0 < \lambda_a^{3,k,l} \le 1, \forall a, k, l; l \le e_a^k/\Delta < l+1$.

Finally substituting (19) to (12a), we have the following discretized DUE model with exact flow propagations:

$$\begin{cases} 0 \le u_{as}^{k} \perp \{\tau_a^k(u) + \sum_{l \le e_a^k(u)/\Delta < l+1} \lambda_a^{3,k,l}(u) \cdot \pi_{h_a s}^{l} + [1 - \lambda_a^{3,k,l}(u)] \cdot \pi_{h_a s}^{l+1} - \pi_{l_a s}^{k}\} \ge 0, \forall a \in A, s \in S, k = 1, \cdots K' & \text{(21a)} \\ \sum_{a \in A(i)} u_{as}^{k} = d_{is}^{k} + \sum_{a \in B(i)} \sum_{k';e_a^{k'}(u) \le k\Delta < e_a^{k'+1}(u)} [\lambda_a^{1,k'}(u) \cdot \lambda_a^{2,k',k}(u) \cdot u_{as}^{k'} + \lambda_a^{1,k'+1}(u) \cdot (1 - \lambda_a^{2,k',k}(u)) \cdot u_{as}^{k'+1}], \forall i \in N, s \in S, i \ne s, k = 1, \cdots, K' & \text{(21b)} \\ \pi_{is}^{k} \ge 0, \forall i \in N, s \in S, i \ne s, k = 1, \cdots, K' & \text{(21c)} \end{cases}$$

Note that (21) is a MiCP with side constraints (21c). Also since $u$ is the defining variable, $e_a^{k'}(u)$ and $e_a^{k}(u)$ will both change as $u$ does. This implies that the summation terms in the right hand sides of both (21a) and (21b) are not fixed; rather they change as $u$ does. In this sense, model (21) is not close form.

We next show that (21) has an equivalent NCP formulation.

**Theorem 1** If the link travel time function $\tau_a^k(u)$ is positive for any $a \in A, k = 1, \cdots K'$ and $u \ge 0$, then model (21) is equivalent to the following NCP model: *find* $(u, \pi)$ *such that*

$$\begin{cases} 0 \le u_{as}^{k} \perp \{\tau_a^k(u) + \sum_{l \le e_a^k(u)/\Delta < l+1} \lambda_a^{3,k,l}(u) \cdot \pi_{h_a s}^{l} + [1 - \lambda_a^{3,k,l}(u)] \cdot \pi_{h_a s}^{l+1} - \pi_{l_a s}^{k}\} \ge 0, \forall a \in A, s \in S, k = 1, \cdots K', & \text{(22a)} \\ 0 \le \pi_{is}^{k} \perp (\sum_{a \in A(i)} u_{as}^{k} - d_{is}^{k} - \sum_{a \in B(i)} \sum_{k';e_a^{k'}(u) \le k\Delta < e_a^{k'+1}(u)} [\lambda_a^{1,k'}(u) \cdot \lambda_a^{2,k',k}(u) \cdot u_{as}^{k'} + \lambda_a^{1,k'+1}(u) \cdot (1 - \lambda_a^{2,k',k}(u)) \cdot u_{as}^{k'+1}], \forall i \in N, s \in S, i \ne s, k = 1, \cdots K'. & \text{(22b)} \end{cases}$$

**Proof**. We need to prove (***a***) if $(u,\pi)$ solves (21), then it is also a solution to (22); and (***b***) if $(u,\pi)$ solves (22), then it must also solves (21).

(***a***) is straightforward. If $(u,\pi)$ solves (21), then it is obvious that $(u,\pi)$ also solves (22).

In order to prove (***b***), suppose $(u,\pi)$ solves (22). Since $(u,\pi)$ is given here, $e_a^{k'}(u)$ and $e_a^{k}(u)$ are fixed for any $a\in A, k', k=1,\cdots K'$. This implies that both (21) and (22) become close form. For the sake of contradiction, assume $(u,\pi)$ is not a solution of (21). Then we must have $i\in N, s\in S, i\neq s$ at some time interval *k*, such that $\sum_{a\in A(i)} u_{as}^k - d_{is}^k - \sum_{a\in B(i)}\sum_{k':e_a^{k'}(u)\le k\Delta<e_a^{k'+1}(u)} [\lambda_a^{1,k'}(u)\cdot\lambda_a^{2,k',k}(u)\cdot u_{as}^{k'} + \lambda_a^{1,k'+1}(u)\cdot(1-\lambda_a^{2,k',k}(u))\cdot u_{as}^{k'+1}] > 0$. Hence we have $\pi_{is}^k = 0$ due to (22b). This implies $\pi_{l_a s}^k = \pi_{is}^k = 0$ for any link $a\in A(i)$. On the other hand, Since $\lambda^1 > 0, 0<\lambda^2\le 1$ and $u_{as}^{k'}\ge 0$, $u_{as}^{k'+1}\ge 0$, $d_{is}^k\ge 0$, we have $\sum_{a\in A(i)} u_{as}^k > d_{is}^k \ge 0$. Thus, we must have at least one link $a\in A(i)$ at time interval *k*, such that $u_{as}^k > 0$. Then, we have $\tau_a^k(u) + \sum_{l\le e_a^k(u)/\Delta<l+1} \lambda_a^{3,k,l}(u)\cdot\pi_{h_a s}^l + [1-\lambda_a^{3,k,l}(u)]\cdot\pi_{h_a s}^{l+1} - \pi_{l_a s}^k = 0$ due to (22a). This means $\tau_a^k(u) + \sum_{l\le e_a^k(u)/\Delta<l+1} \lambda_a^{3,k,l}(u)\cdot\pi_{h_a s}^l + [1-\lambda_a^{3,k,l}(u)]\cdot\pi_{h_a s}^{l+1} = 0$ since $\pi_{l_a s}^k = 0$. Nevertheless, $\tau_a^k(u) > 0$, $\pi_{h_a s}^l \ge 0$, $\pi_{h_a s}^{l+1}\ge 0$, and $0<\lambda^3\le 1$, this is a contradiction. □

Note that in NCP (22), *u* is the defining variable, and $\lambda^1, \lambda^2, \lambda^3$, *e* and $\tau$ are functions defined on *u*. However, (22) is not close form due to the same reason for model (21). We further define two vector functions

$$F_u(u,\pi) = \left( \tau_a^k(u) + \sum_{l\le e_a^k(u)/\Delta<l+1} \lambda_a^{3,k,l}(u)\cdot\pi_{h_a s}^l + [1-\lambda_a^{3,k,l}(u)]\cdot\pi_{h_a s}^{l+1} - \pi_{l_a s}^k \right)_{\forall a\in A, s\in S, k=1,\cdots K'}, \tag{23}$$

$$F_\pi(u,\pi) = \left( \sum_{a\in A(i)} u_{as}^k - d_{is}^k - \sum_{a\in B(i)}\sum_{k':e_a^{k'}(u)\le k\Delta<e_a^{k'+1}(u)} [\lambda_a^{1,k'}(u)\cdot\lambda_a^{2,k',k}(u)\cdot u_{as}^{k'} + \lambda_a^{1,k'+1}(u)\cdot(1-\lambda_a^{2,k',k}(u))\cdot u_{as}^{k'+1}] \right)_{\forall i\in N, s\in S, i\neq s, k=1,\cdots K'} \tag{24}$$

We can then compactly express (22) as follows.

$$\begin{cases} 0\le u \perp F_u(u,\pi)\ge 0 \\ 0\le \pi \perp F_\pi(u,\pi)\ge 0 \end{cases}. \tag{25}$$

We denote model (22) and (25) as *NCPDUE* thereafter in this paper.

*3.2 Solution Existence and Compactness Condition*

This section establishes the solution existence condition for *NCPDUE*. For this purpose, we need a solution existence condition for VIs, as stated in Lemma 1.

**Lemma 1** Let $K\subseteq R^n$ be compact (closed and bounded) and convex and let $F: K\to R^n$ be continuous. Then the solution set of VI defined by *K* and *F* is nonempty and compact.

**Proof**. The proof can be found in Facchinei and Pang (2003) in Corollary 2.2.5. □

The solution existence result discussed here can then be summarized as follows.

**Theorem 2** Suppose *a)* the link travel time function $\tau_a^k(u)$ is positive for any $a\in A, k=1,\cdots K'$ and bounded from above for any finite *u*, *b)* the given OD demand $d_{is}^k$ is nonnegative and bounded from above for any

$i \in N, s \in S, k = 1, \cdots K'$, *c)* $\lambda^1$ is bounded from above, and *d)* $F_u(u,\pi)$ and $F_\pi(u,\pi)$ are continuous with respect to $(u,\pi)$. Then the solution set of *NCPDUE* is nonempty and compact.

**Proof**. First of all, choose three scalars as follows.

$$\alpha_d > \max_{s\in S} \max_{i\in N, i\neq s} \max_{k=1,\cdots,K'} d_{is}^k \tag{26a}$$

$$\alpha_u > \max_{s\in S} \max_{i\in N, i\neq s} \max_{k=1,\cdots,K'} \max_{u\geq 0} (d_{is}^k + \sum_{a\in B(i)} \sum_{k';e_a^{k'}(u)\leq k\Delta < e_a^{k'+1}(u)} [\lambda_a^{1,k'}(u)\cdot\lambda_a^{2,k',k}(u)\cdot u_{as}^{k'} + \lambda_a^{1,k'+1}(u)\cdot(1-\lambda_a^{2,k',k}(u))\cdot u_{as}^{k'+1}]) \tag{26b}$$

$$\alpha_\pi > \max_{s\in S} \max_{a\in A} \max_{k=1,\cdots,K'} \max_{u\geq 0,\pi\geq 0} (\tau_a^k(u) + \sum_{l\leq e_a^k(u)/\Delta < l+1} \lambda_a^{3,k,l}(u)\cdot\pi_{h_a s}^l + [1-\lambda_a^{3,k,l}(u)]\cdot\pi_{h_a s}^{l+1}) \tag{26c}$$

These three scalars exist because of the three boundedness assumptions and the fact that $0 < \lambda_a^{2,k',k} \leq 1$, $\forall a, k', k; e_a^{k'}(u) \leq k\Delta < e_a^{k'+1}(u)$ and $0 < \lambda_a^{3,k,l} \leq 1, \forall a, k, l; l \leq e_a^k(u)/\Delta < l+1$. For the same reason, they are all positive. We further define a set $E = \{y = (u,\pi) \geq 0 \,|\, u \leq \alpha_u 1,\ \pi \leq \alpha_\pi 1\}$ and a function $F = (F_u(u,\pi),\ F_\pi(u,\pi))$. Here **1** is a vector with all components being 1 with the proper dimension. Set *E* and function *F* constitute a VI which *tries to find* $y^* = (u^*, \pi^*) \in E$ *such that*

$$F(y^*)^T (y - y^*) \geq 0, \forall y \in E. \tag{27}$$

Since *E* is compact and convex and further *F* is continuous with respect to $(u,\pi)$, according to Lemma 1, the above VI has a nonempty and compact solution set, denoted as $\Gamma = \{y^* = (u^*, \pi^*)\}$. We next show that the solution set of *NCPDUE*, denoted as $\Xi$, coincides with $\Gamma$.

First of all, for each solution $y^* = (u^*, \pi^*) \in \Gamma$, since *E* is a polyhedral set, there must exist multipliers $\gamma$ and $\eta$ such that

$$\begin{cases} 0 \leq u^* \perp F_u(u^*, \pi^*) + \gamma \geq 0 & \text{(28a)} \\ 0 \leq \pi^* \perp F_\pi(u^*, \pi^*) + \eta \geq 0 & \text{(28b)} \\ 0 \leq \gamma \perp \alpha_u 1 - u^* \geq 0 & \text{(28c)} \\ 0 \leq \eta \perp \alpha_\pi 1 - \pi^* \geq 0 & \text{(28d)} \end{cases}$$

In order to prove $y^* = (u^*, \pi^*)$ solves *NCPDUE*, we need to show that $\gamma = 0$ and $\eta = 0$. Suppose $\gamma_{bs}^k > 0$ for some *b*, *s*, *k*. We then must have $u_{bs}^{k\,*} = \alpha_u > 0$ from (28c). This implies $\rho_1 = \tau_b^k(u^*) + \sum_{l\leq e_b^k(u^*)/\Delta < l+1} \lambda_b^{3,k,l}(u^*)\cdot\pi_{h_b s}^{l\,*} + [1-\lambda_b^{3,k,l}(u^*)]\cdot\pi_{h_b s}^{l+1\,*} - \pi_{l_b s}^{k\,*} + \gamma_{bs}^k = 0$ due to (28a). This also means that for the tail node of link *b*, denoted as node *i*, $\sum_{a\in A(i)} u_{as}^{k\,*} - d_{is}^k - \sum_{a\in B(i)} \sum_{k';e_a^{k'}(u^*)\leq k\Delta < e_a^{k'+1}(u^*)} [\lambda_a^{1,k'}(u^*)\cdot\lambda_a^{2,k',k}(u^*)\cdot u_{as}^{k'\,*} + \lambda_a^{1,k'+1}(u^*)\cdot(1-\lambda_a^{2,k',k}(u^*))\cdot u_{as}^{k'+1\,*}] + \eta_{is}^k \geq$

$\alpha_u - (d_{is}^k + \sum_{a\in B(i)} \sum_{k';e_a^{k'}(u^*)\leq k\Delta < e_a^{k'+1}(u^*)} [\lambda_a^{1,k'}(u^*)\cdot\lambda_a^{2,k',k}(u^*)\cdot u_{as}^{k'\,*} + \lambda_a^{1,k'+1}(u^*)\cdot(1-\lambda_a^{2,k',k}(u^*))\cdot u_{as}^{k'+1\,*}]) + \eta_{is}^k > 0$ based on (26b). Hence $\pi_{is}^{k\,*} = \pi_{l_b s}^{k\,*} = 0$ according to (28b). Therefore, we have $0 < \tau_b^k(u^*) \leq \rho_1 = \tau_b^k(u^*) + \sum_{l\leq e_b^k(u^*)/\Delta < l+1} \lambda_b^{3,k,l}(u^*)\cdot\pi_{h_b s}^l + [1-\lambda_b^{3,k,l}(u^*)]\cdot\pi_{h_b s}^{l+1} + \gamma_{bs}^k = 0$ which is a contradiction.

Similarly if $\eta_{is}^k > 0$ for some *i*, *s*, *k*, we have $\pi_{is}^{k\,*} = \alpha_\pi > 0$ based on (28d). According to (28a), we must have $\rho_2 = \tau_a^k(u^*) + \sum_{l\leq e_a^{k'}(u^*)/\Delta < l+1} \lambda_a^{3,k,l}(u^*)\cdot\pi_{h_a s}^{l\,*} + [1-\lambda_a^{3,k,l}(u^*)]\cdot\pi_{h_a s}^{l+1\,*} - \pi_{l_a s}^{k\,*} + \gamma_{as}^k \geq 0$ for any link $a \in A(i)$. However

since $\pi_{l_a s}^{k\,*} = \pi_{is}^{k^*} = \alpha_\pi$ and $\gamma_{as}^k = 0$ , $\rho_2 = \tau_a^k(u^*) + \sum_{l \le e_a^k(u^*)/\Delta < l+1} \lambda_a^{3,k,l}(u^*) \cdot \pi_{h_a s}^{l\,*} + [1 - \lambda_a^{3,k,l}(u^*)] \cdot \pi_{h_a s}^{l+1\,*} - \alpha_\pi < 0$ due the definition in (26c). This is a contradiction.

So far we have proved that $\Gamma \subseteq \Xi$. In order to prove $\Xi \subseteq \Gamma$, we notice that for any solution $y^* = (u^*, \pi^*) \in \Xi$, (28) has to be satisfied with the two set of multipliers $\gamma = 0$ and $\eta = 0$. Due to the equivalence between (28) and VI (27), $y^*$ must also be a solution to VI (27). This implies $y^* \in \Gamma$ and further $\Xi \subseteq \Gamma$. Therefore, $\Xi = \Gamma$ and since $\Gamma$ is nonempty and compact, so is the solution set of *NCPDUE*.

□

## 4. Solution algorithm

The solution algorithm in this section is based on the fact that for a given flow $\bar{u}$ (denoted as base inflow), the link travel time vector $\tau$ will be fixed by a network loading procedure. Hence, all $\lambda^1, \lambda^2, \lambda^3$ and *e* will be fixed as well. As a result, the originally non-close-form NCP model (22) will become close-form as in (29):

$$\begin{cases} 0 \le u_{as}^k \perp \{\tau_a^k(\bar{u}) + \sum_{l \le e_a^k(\bar{u})/\Delta < l+1} \lambda_a^{3,k,l}(\bar{u}) \cdot \pi_{h_a s}^l + [1 - \lambda_a^{3,k,l}(\bar{u})] \cdot \pi_{h_a s}^{l+1} - \pi_{l_a s}^k\} \ge 0, \forall a \in A, s \in S, k = 1, \cdots K', & (29a) \\ 0 \le \pi_{is}^k \perp (\sum_{a \in A(i)} u_{as}^k - d_{is}^k - \sum_{a \in B(i)} \sum_{k';\, e_a^{k'}(\bar{u}) \le k\Delta < e_a^{k'+1}(\bar{u})} [\lambda_a^{1,k'}(\bar{u}) \cdot \lambda_a^{2,k',k}(\bar{u}) \cdot u_{as}^{k'} + \lambda_a^{1,k'+1}(\bar{u}) \cdot (1 - \lambda_a^{2,k',k}(\bar{u})) \cdot u_{as}^{k'+1}], \forall i \in N, s \in S, i \ne s, k = 1, \cdots K'. & (29b) \end{cases}$$

Note the only difference between model (22) and (29) is that $\lambda^1, \lambda^2, \lambda^3$, *e* and $\tau$ in (29) are evaluated at the base inflow $\bar{u}$, while in (22) they are functions of *u*. In other words, (29) is a close-form NCP which is much easier to solve. We denote (29) as the "relaxed" NCP of (22).

The above observation outlines an iterative algorithm to solve model (22). It is heuristic in the sense that the convergence can not be established using regularity conditions. This is mainly due to the fact that (22) can not be expressed in a close form. The algorithm is listed as below.

**Algorithm DUE**

*Step 1*. Initialization. Assign an initial feasible base inflow $(\bar{u})^0$.

*Step 2*. Main Loop. Set *n=0*.

*Step 2.1* Construct current relaxed NCP at $(\bar{u})^n$ by a network loading procedure.

*Step 2.2* Solve the relaxed NCP and denote its solution $(\tilde{u})^n, (\tilde{\pi})^n$ as a "candidate" solution.

*Step 2.3* Convergence Test. If certain convergence criterion is satisfied at the candidate solution, go to Step 3; else, go to Step 2.4.

*Step 2.4* Update and Move. Set $(\bar{u})^{n+1} = (\bar{u})^n + \theta \cdot ((\tilde{u})^n - (\bar{u})^n)$, *n=n+1* and go to Step 2.1.

*Step 3.* Find an optimal solution $(\tilde{u})^n, (\tilde{\pi})^n$.

In the above algorithm, $0 \le \theta \le 1$ is a pre-defined step size. In addition, there are several options for the convergence test in Step 2.3. The most commonly used and also simplest way is to check whether the base inflows stabilize between two consecutive iterations, i.e.,

$$Gap\_u = |(\bar{u})^n - (\bar{u})^{n+1}| \le \varepsilon_1. \quad (30)$$

A more rigorous way is to check whether the complementarity condition in equation (1) holds:

$$Gap\_DUE = u^T F_u(u, \pi) \le \varepsilon_2. \quad (31)$$

In (30) and (31), $\varepsilon_1$ and $\varepsilon_2$ are chosen as small positive scalars.

The next two sections will further discuss the network loading procedure and the method for solving the relaxed NCP model (29).

*4.1 Link-Based Network Loading Procedure*

The network loading procedure in Algorithm DUE is link-based, i.e., for a given based inflow $\bar{u}$, it is to load/propagate $\bar{u}_{as}^k$ to link *a* for each link *a* at any time interval *k* towards a given destination *s*. This will generate other traffic measurements such as $\bar{v}, \bar{x}, \tau(\bar{u}), e(\bar{u})$. In turn, $\lambda^1(\bar{u}), \lambda^2(\bar{u}), \lambda^3(\bar{u})$ can also be determined. In Carey and Ge (2004) and Nie and Zhang (2005), this loading procedure is also called the "solution algorithm for link travel time model." In this paper, we intend to call it a "loading procedure" since it indeed generates not only travel times but also link flows and exit flows.

To be consistent with the discretization scheme in Section 3.1, we need to adopt the loading process in Astarita (1996). However, since the original algorithm by Astarita (1996) may have numerical problems (Nie and Zhang, 2005), in this paper, we apply the improved algorithm by Nie and Zhang (2005) which performs the loading based on cumulative departure curves (Algorithm D2). Furthermore, in order to construct the relaxed NCP model (29), one needs to track the relation between $\bar{u}_{as}^{k'}$ and $\bar{v}_{as}^k$ for any *a* and *s* and appropriate $(k', k)$ pairs. Due to the fact that inflows to any destination *s* will experience the same travel time at each entrance time *k*, we can first use a three dimensional matrix $V(a, k', k)$ to represent this relation. In particular, $V(a, k', k) = \rho$ means that a proportion of $\rho$ $(0 \le \rho \le 1)$ of the inflows $\bar{u}_{as}^{k'}\Delta$ will exit the link *a* at time *k*, i.e., become part of $\bar{v}_{as}^k\Delta$. It turns out that we can introduce a "stack" for this tracking purpose. The revised loading procedure, denoted as Algorithm DL, is listed as follows. Note that since the loading procedure is the same for each link, we only show it for link *a*. Also, we omit the "bar" symbol on each variable to simplify the notation.

**Algorithm DL**

*Step 0* Initialization. Set $l = \lfloor \alpha_a / \Delta \rfloor, x_{as}^1 = 0, v_{as}^k = 0, \forall s \in S, k = 1, \cdots, l; e_a^1 = \alpha_a; R_{as}^u = 0, \forall s \in S$. Create an empty stack *SR*. Set *k*=1.

*Step 1* Move. Set *k*=*k*+1. Compute $x_{as}^k = x_{as}^{k-1} + (u_{as}^{k-1} - v_{as}^{k-1})\Delta, \forall s \in S$ and $x_a^k = \sum_{\forall s \in S} x_{as}^k, u_a^k = \sum_{\forall s \in S} u_{as}^k, v_a^k = \sum_{\forall s \in S} v_{as}^k$.

Then compute $\tau_a^k$. Set $e_a^k = (k-1)\Delta + \tau_a^k$ and $n_l = \lfloor e_a^k / \Delta - l \rfloor$. If $n_l < 1$, go to Step 1.1; otherwise, go to Step 1.2.

*Step 1.1*: Update $R_{as}^u = R_{as}^u + u_{as}^{k-1}, \forall s \in R$. Push pair $(k-1,1)$ into stack *SR*.

*Step 1.2*: Set *l*=*l*+1; $\rho_k = (l\Delta - e_a^{k-1})/(e_a^k - e_a^{k-1})$; $v_{as}^l = R_{as}^u + \rho_k u_{as}^{k-1}, \forall s \in S$. Push pair $(k-1, \rho_k)$ to stack *SR*. Then pop each entry $(i, \rho)$ from stack *SR* and set $V(a, i, l) = \rho$.

For *j*=2 to $n_l$: set *l*=*l*+1, $\rho_k = \Delta/(e_a^k - e_a^{k-1})$, $v_{as}^l = \rho_k u_{as}^{k-1}, \forall s \in S$, and $V(a, k-1, l) = \rho_k$.

Set $\rho_k = (e_a^k - l\Delta)/(e_a^k - e_a^{k-1})$ and update $R_{as}^u = \rho_k u_{as}^{k-1}, \forall s \in S$. Push the pair $(k-1, \rho_k)$ to stack SR.

*Step 2*. If $k < T'$, go to Step 1; otherwise, stop.

In Algorithm DL, $\lfloor x \rfloor$ denotes the integral part of a real value *x*, $\alpha_a$ is the free flow travel time for link *a*, and $R_{as}^u\Delta$ is the undistributed inflows for destination *s*. The term "undistributed" here represents inflows that have not yet been used to determine exit flows (Nie and Zhang, 2005). The stack *SR* is used to track the time interval and proportion of those inflows that are undistributed. Therefore, each entry in SR is a pair of ($k$, $\rho$) where *k* is the time interval and $\rho$ is the proportion, both for inflows. Besides tracking the relation between *u* and *v*, Algorithm DL above also extends Algorithm D2 in Nie and Zhang (2005) by performing the loading for each individual destination. It also worth noting that the matrix *V* represents $\lambda^1$ and $\lambda^2$ in (29) and $\lambda^3$ can be computed using equation (20). Consequently, NCP (29) can be constructed readily using *V* and $\lambda^3$.

*4.2 Solving the Relaxed NCP*

Since the relaxed model (29) is a well-defined NCP with continuous and close-form defining functions, it can be readily solved using existing solution techniques. Facchinei and Pang (2003) provides a comprehensive review of solution methods for NCPs. In particular, the projection-based methods play a central role in solving NCPs because calculating the projection on the nonnegative orthant, the defining set of an NCP, is extremely easy and efficient compared with that on a general convex set. Based on this observation, Dirkse and Ferris (1995) developed a path search algorithm which, under certain regularity conditions, is proved to be globally convergent with quadratic convergence rate (near the solution). The algorithm was later evolved to the PATH solver which is now available in GAMS (General Algebraic Modeling System, see Brooke et al., 1998). In this paper, we directly adopt the PATH solver which has been shown to be effective to solve NCP (29). For detailed descriptions of the solver, one can refer to Ferris and Munson (1998). Using the PATH solver requires developing GAMS codes for the relaxed NCP (29) which is straightforward and hence the details are omitted here.

## 5. Numerical examples

In this section, a case study is provided to demonstrate the model and solution algorithm proposed in the paper. We start with the link travel time function that is actually used.

*5.1 Link Travel Time Function*

In this paper, we choose the following linear form for the link travel time function:

$$\tau_a^k = \alpha_a(1+\beta_a^u u_a^k + \beta_a^x x_a^k), \tag{32}$$

where $\alpha_a$ is the free flow travel time for link *a*, $\beta_a^u$ and $\beta_a^x$ are constants that are shown in Table 1. Here we introduce the aggregated link inflow into the link travel time function by the constant $\beta_a^u$. The purpose is to make the solution algorithm for relaxed NCP (29) numerically stable. The reason is as follows. First, (29) can be rewritten in a matrix notation as

$$\begin{cases} 0 \le u_s \perp [\tau(u) + \Omega_s \pi_s] \ge 0 \\ 0 \le \pi_s \perp [\Lambda_s u_s - d_s] \ge 0 \end{cases}, \forall s \in S \tag{33}$$

with $\Omega_s$ and $\Lambda_s$ are fixed matrices computed using the base inflow. Here $u_s = (u_{as}^k)_{\forall a,k}$, $\pi_s = (\pi_{is}^k)_{\forall i,k;i\neq s}$, and $d_s = (d_{is}^k)_{\forall i,k;i\neq s}$ denote destination-based variables.

Model (33) has a very special structure such that it can be easily decomposed to individual destinations. The interactions between variables of different destinations only exist in the link travel time vector $\tau(u)$. If we compute the Jacobian matrix of (33), denoted as *M*, we will have

$$M = \begin{array}{c} \begin{matrix} \pi_1 & \cdots & \pi_{|S|} & u_1 & \cdots & u_{|S|} \end{matrix} \\ \begin{bmatrix} 0 & \cdots & 0 & \Lambda_1 & 0 & 0 \\ \vdots & \ddots & \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \Lambda_{|S|} \\ \Omega_1 & 0 & 0 & \partial\tau/\partial u_1 & \cdots & \partial\tau/\partial u_{|S|} \\ 0 & \ddots & 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \Omega_{|S|} & \partial\tau/\partial u_1 & \cdots & \partial\tau/\partial u_{|S|} \end{bmatrix} \end{array} \begin{matrix} \pi_1 \\ \cdots \\ \pi_{|S|} \\ u_1 \\ \cdots \\ u_{|S|} \end{matrix}. \tag{34}$$

In (34), each variable on the right side of the matrix indicates the corresponding row is computed by taking the partial derivative of the function perpendicular to the particular variable over all variables; while each variable on the top indicates that the column is computed by taking partial derivative of each

function to the particular variable. If $\beta_a^u$=0, the diagonal of $\partial\tau/\partial u_s$ will be all zero since $x_a^k$ only includes inflows up to time ($k$-1) and not $u_a^k$. This implies that the diagonals of the $M$ are all zero. Such a matrix can be proved to be not positive definite and thus may cause problems to solve the relaxed NCP. Adding a small positive $\beta_a^u$ to the diagonal turns out to be helpful to stabilize the solution process. It worth noting that $\beta_a^u << \beta_a^x$, therefore, the possibility of FIFO violations due to $\beta_a^u$ is expected to be small which will be verified in later numerical studies.

*5.2 Case Study*

The case study is a small example tested on the hypothetical network depicted in Fig. 2, denoted as the D3 network. In the DUE literature, this network was first used by Chen and Hsuen (1998). In this paper, we use slightly different specifications for the D3 network, as shown in Table 1. The network has two origins: node 1 and 2, and one destination: node 3. Further, the length of each time interval is set as $\Delta = 0.25$ minutes (15 seconds).

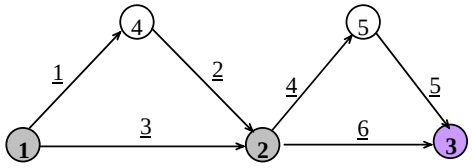



**Fig. 2 Test network**

**Table 1 Configuration of the test network**

| **Link** | $\alpha_a$ **(min)** | $\beta_a^u$ **(min$^2$/vehicle)** | $\beta_a^x$ **(min/vehicle)** |
|---|---|---|---|
| 1 | 1.2 | 0.00125 | 0.01 |
| 2 | 1.2 | 0.00125 | 0.01 |
| 3 | 2.16 | 0.00125 | 0.0056 |
| 4 | 1.2 | 0.00125 | 0.01 |
| 5 | 1.2 | 0.00125 | 0.01 |
| 6 | 2.4 | 0.00125 | 0.005 |

To simulate the fluctuation of traffic volumes during peak hours, we employ a parabolic-shaped curve to represent the OD demands between each OD pair. In particular, the demand rate between any OD pair *rs* is assumed to be calculated through equation (35):

$$d^{rs}(k) = 40 + 120 * (1 - \left(\frac{k - K/2}{K/2}\right)^2), \forall 1 \le k \le K, \forall r \in R, s \in S, \quad (35)$$

where $K$ denotes the total number of intervals during which OD trips will be generated.

For the case study, we set $K$=120 which is equivalent to 30 minutes. Algorithm DUE can solve successfully the proposed model. Fig. 3 first depicts the convergence performance of the algorithm for the first 25 iterations. In this figure, although the units for the two gaps are different as indicated, we plot them together in order to show their slight different performances. We can easily observe that both gaps decrease monotonically, whereas *Gap_u* decreases faster. When close to $10^{-4}$, *Gap_DUE* is stabilized. Furthermore, after 25 iterations, the absolute difference of inflow rates between two consecutive iterations is close to $10^{-5}$ vehicles/minute. This should be accurate enough for most transportation related applications. It is also worth noting that due to the exact solve of each relaxed NCP by the PATH solver (usually to $10^{-6}$), the proposed Algorithm DUE requires much less number of major iterations than previous DUE solution methods based on FW.

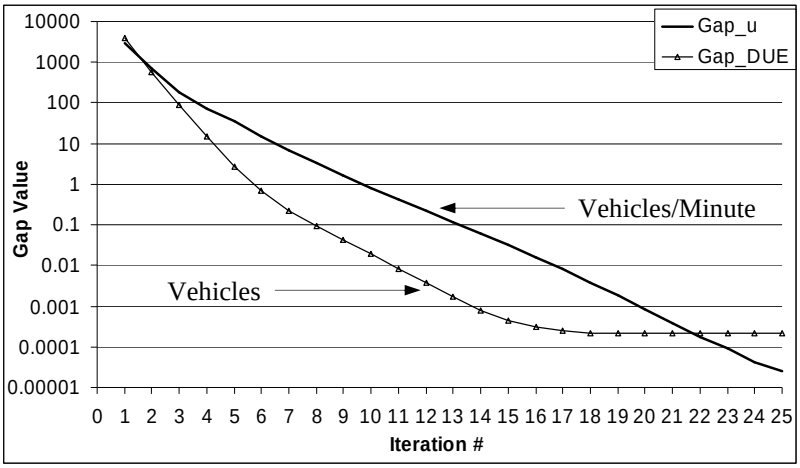



**Fig. 3 Convergence of the algorithm**

Fig. 4 below shows the aggregated inflow rate on each link. We can see that the inflow rates to link 1 and 3 change more abruptly than those to link 4 and 6. This may be due to the fact that the choice of links at node 1 will be impacted by the inflows at node 2, and not vice versa. Notice that the inflows to link 2 and 5 are exactly the exit flows from link 1 and 4, respectively. We can therefore observe that the shapes of exit flows are similar to their corresponding inflows, but are smoother. We further plots in Fig. 5 the travel times from node 1 and 2 respectively to destination 3 via different links. This figure shows that at the beginning ($k$=1 to 15), the travel time from node 1 to 3 via link 1 is higher than that via link 3. Consequently, all vehicles choose link 3 during this period as depicted in Figure 4. This is also true for the period of $k$>=115. While from node 2 to 3, choosing either link 4 or link 6 will have almost equal travel times, hence both links will be selected. The time-dependent link flows are also presented in Fig. 6.

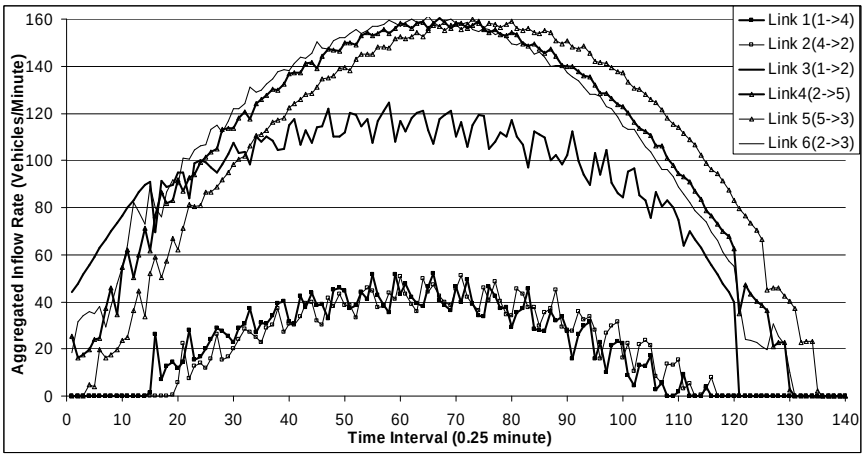



**Fig. 4 Inflow Rates**

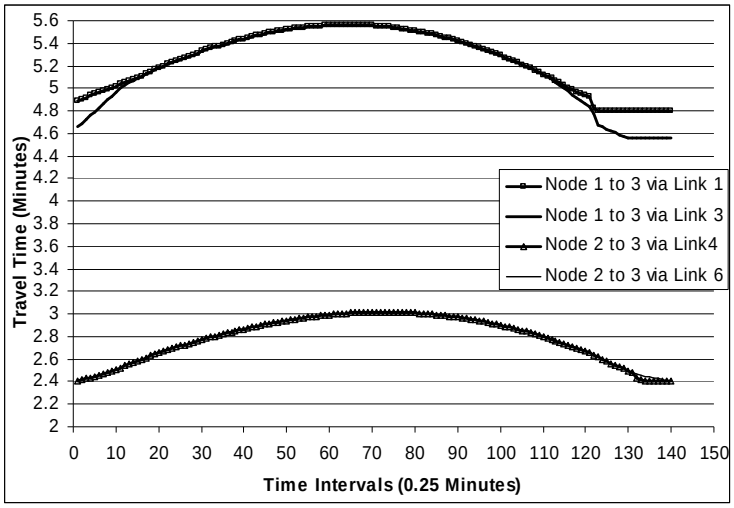



**Fig. 5 Travel Times**

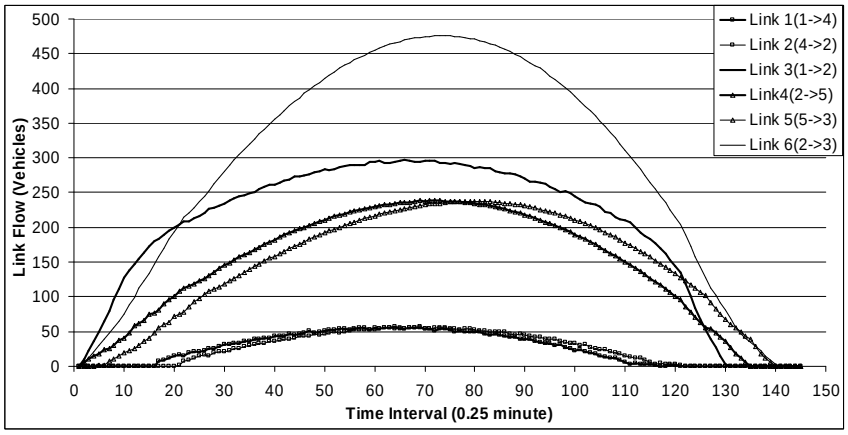



**Fig. 6 Link Flows**

We illustrate the discretized version of the derivative of link travel time (over time), i.e. $(\tau_a^{k+1} - \tau_a^k)/\Delta$, in Fig. 7. Clearly link travel times change rather abruptly, but the derivatives are within the range of (-0.1, 0.1). Therefore, FIFO is not violated for our case study since $(\tau_a^{k+1} - \tau_a^k)/\Delta > -1$ is held for all links at all time intervals.

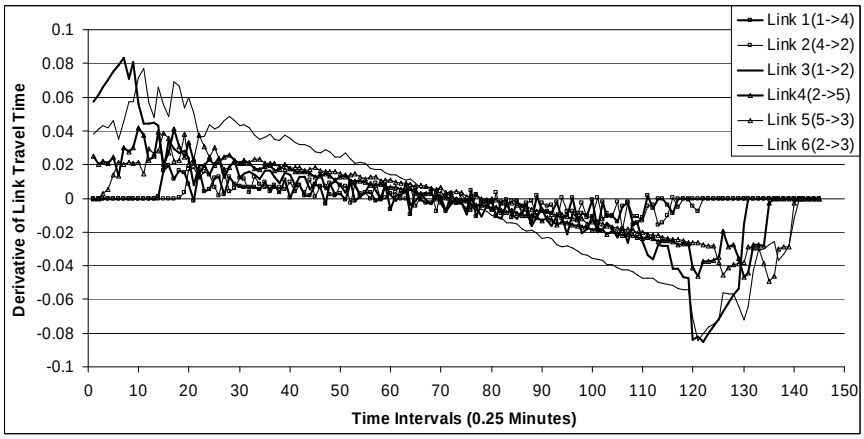



**Fig. 7 Link Travel Time Changes**

## 6. Conclusions and future research

We presented a link-node based NCP model for basic DUE problem and explicitly captured the exact flow propagations. The solution existence and compactness condition was established under mild assumptions. We also developed an iterative solution algorithm for the proposed model by solving a relaxed NCP in each iteration. The relaxed NCP can be solved very efficiently using existing solution technique and thus the entire algorithm only required fairly small number of iterations. The case study demonstrated that the proposed model and solution approach are effective for solving DUE problems.

For future studies, the proposed NCP model, especially its solution approach, merits further investigations. Especially, certain solution convergence condition needs to be established. Secondly, the travel time function used in this paper may be suitable only for mild congestion scenarios. In this regard, more sophisticated functional forms need to be developed to capture heavy traffic congestion. Lastly, the model and solution approach proposed in this paper need to be further tested on large scale DUE problems. Due to the special structure of the proposed NCP model, certain decomposition scheme based on individual destinations may be applied. The authors have developed such schemes for solving asymmetric and static user equilibrium problems (Ban et al., 2006), as well as some preliminary extensions to DUE (Ban, 2005). More rigorous extensions to DUE are currently under investigation and results will be reported in subsequent papers.

## References


1. Astarita, V., 1996. A continuous time link model for dynamic network loading based on travel time function. In: Lesort, J.-B. (Ed.), *Transportation and Traffic Theory*. Pergamon, Oxford, 79-102.
2. Ban, X., 2005. *Quasi-variational Inequality Formulations and Solution Approaches for Dynamic User Equilibria*. Ph.D Thesis, University of Wisconsin-Madison.
3. Ban, **X**., Liu, H., and Ferris, M.C., 2006. A link-node based complementarity model and its solution algorithm for asymmetric user equilibria. In *Proceedings of the 85th Transportation Research Board Annual Meeting (CD-ROM)*.
4. Billups, S.C., Dirkse, S. P. and Ferris, M. C., 1997. A comparison of large scale mixed complementarity problem solvers. *Computational Optimization and Applications* 7, 3-25
5. Bliemer, M.C.J., 2001. *Analytical dynamic traffic assignment with interacting user-classes: Theoretical advances and applications using a variational inequality approach*. Ph.D. Thesis, Delft University of Technology, The Netherlands.
6. Bliemer , M.C.J. and Bovy, P.H.L., 2003. Quasi-variational inequality formulation of the multiclass dynamic traffic assignment problem. *Transportation Research* 37B, 501-519.
7. Brooke, A., Kendrick, D. and Meeraus, A., et al., 1998. *GAMS, a user's guide*. GAMS Development Corporation.
8. Cao, M., and Ferris, M. C., 1996. A pivotal method for affine variational inequalities. M*athematics of Operations Research* 21, 44-64.
9. Carey, M., 1987. Optimal Time-Varying Flows on Congested Networks. *Operations Research* 35, 58-69.
10. Carey, M., Ge, Y.E. and McCartney, M., 2003. A whole-travel time model with desirable properties. *Transportation Science* 37(1), 83-96.
11. Carey, M. and Ge, Y.E. (2004). Efficient discretisation for link travel time models. *Networks and Spatial Economics* 4, 269-290.
12. Chen, H.K. 1999. *Dynamic Travel Choice Models*. Springer-Verlag, New York.
13. Chen, H.K. and Hsuen, C.F., 1998. A model and an algorithm for the dynamic user-optimal route choice problem. *Transportation Research* 32B(3), 219-234.
14. Dafermos, S.C., 1980. Traffic equilibrium and variational inequality. *Transportation Science*, 14, 42-54.
15. Daganzo, C.F., 1994. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research* 28B(4), 269-287.
16. Daganzo, C.F., 1995a. The cell transmission model part II: network traffic. *Transportation Research* 29B(2), 79-93.
17. Daganzo, C.F., 1995b. Properties of link travel time functions under dynamic loads. *Transportation Research*, 29B, 95-98.

18. Dirkse, S.P. and Ferris, M.C., 1995. The PATH solver: A non-monotone stabilization scheme for mixed complementarity problems. *Optimization Methods and Software* 5, 123-156.
19. Facchinei, F. and Pang J.S., 2003a. *Finite-Dimensional Variational Inequalities and Complementarity Problems (Vol. I and II)*. Springer-Verlag, New York
20. Ferris, M.C. and Munson, T.S., 1998. *PATH 4.6 User Manual*. University of Wisconsin-Madison.
21. Ferris, M. C. Fourer, R., and Gay, D. M., 1999. Expressing complementarity problems and communicating them to solvers. *SIAM Journal on Optimization* 9, 991-1009.
22. Friesz, T.L., Bernstein, D. and Smith, T.E., et al., 1993. A Variational Inequality Formulation of the Dynamic Network User Equilibrium Problem. *Operations Research* 41(1), 179-191.
23. Friesz, T.L., Bernstein, D., Suo, Z. and Tobin, R.L. (2001) Dynamic network user equilibrium with time-dependent time lags. *Networks and Spatial Economics* 1, 319-347.
24. Friesz, T.L. and Mookherjee, R., 2005 Solving the dynamic network user equilibrium problem with state-dependent time shifts. *Transportation Research, Part B*. In Press.
25. Han, S. and Heydecker, B.G., 2006. Consistent objectives and solution of dynamic user equilibrium models. *Transportation Research* 40B, 16-34.
26. Heydecker, B.G. and Verlander, N., 1999. Calculation of dynamic traffic equilibrium assignments. Proceedings of the European Transport Conferences, Seminar F, 79-91.
27. Lo, H.K., Ran, B. and Hongola, B., 1996. Multiclass Dynamic Traffic Assignment Model: Formulation and Computational Experiences. *Transportation Research Record 1537*, 74-82.
28. Lo, H. and Szeto, W.Y., 2002. A cell-based variational inequality formulation of the dynamic user optimal assignment problem. *Transportation Research* 36B, 421-443.
29. Nie, X.J. and Zhang, H.M., 2005. Delay-function-based link models: their properties and computational issues. *Transportation Research*, Part B, in press.
30. Peeta, S. and Ziliaskopoulos, A.K., 2001. Foundations of Dynamic Traffic Assignment: the Past, the Present and the Future. *Networks and Spatial Economics* 1, 233–265.
31. Ran, B. and Boyce, D.E., 1994. *Dynamic Urban Transportation Network Models: Theory and Implications for Intelligent Vehicle Highway Systems, Lecture Notes in Economics and Mathematical Systems* 417. Springer-Verlag, New York.
32. Ran, B. and Boyce, D.E, 1996. *Modeling Dynamic Transportation Networks: An Intelligent Transportation Systems Oriented Approach*, 2nd revised edn. Springer-Verlag, New York.
33. Ran, B., Lo, H. and Boyce, D.E., 1996. A Formulation and Solution Algorithm for A Multi- class Dynamic Traffic Assignment Problem *Transportation and Traffic Theory*. Edited by Lesort, J.B., Elsevier Science, UK, 195-216.
34. Smith, M.J., 1979. The Existence, Uniqueness and Stability of Traffic Equilibria. *Transportation Research* 13B, 295-304.
35. Ulbrich, M., 2002. *Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. Ph.D thesis, Technische Universitat Munchen, Germany.
36. Wardrop, J.G., 1952. Some theoretical aspects of road traffic research. In *Proceedings of the Institution of Civil Engineers*, Part II, 1, 325-378.
37. Wie, B.W., Tobin, R.L. and Carey, M., 2002. The existence, uniqueness and computation of an arc-based dynamic network user equilibrium formulation. *Transportation Research* 36B, 897-918.
38. Xu, Y.W., Wu, J.H., Florian, M., Marcotte, P., and Zhu, D.L., 1999. Advances in the continuous dynamic network loading problem. *Transportation Science* 33, 341-353.

# DESCENT DIRECTION BASED ALGORITHM FOR DUE ASSIGNMENT

*Agachai Sumalee: Institute for Transport Studies, University of Leeds, UK, asumalee@its.leeds.ac.uk*
*Masao Kuwahara: Institute of Industrial Science, University of Tokyo, Japan, kuwahara@iis.u-tokyo.ac.jp*

## Abstract

The paper proposes a solution algorithm for solving dynamic user equilibrium (DUE) assignment without departure time choice. The DUE problem is firstly formulated as a variational inequality (VI) and then transformed to an equivalent optimization problem using an associated gap function of the VI. The objective function of the problem is the gap function which is a result of a minimization problem. Thus, the paper employs the theorem for deriving the derivative of the optimal value function to define the derivative of the gap function. This requires derivatives information of the path travel time at different departure times with respect to different path inflow profiles which were readily defined in the literature. The paper then describes the descent direction algorithm for solving the DUE assignment based on Wolfe's Reduced Gradient method. The algorithm is then tested with a small example.

## 1 Introduction

Congestion occurring in the network is dynamic by its nature. This observation provides a strong support for the development of a dynamic traffic assignment (DTA) model. With its simplest form, there are two major elements of the DTA model: the dynamic loading model and route choice model. The dynamic loading model is responsible for evaluating the dynamic movement and resulting congestion of the given traffic volume in the network (see e.g. Chen and Florian, 1998, Friesz et al., 1993). Several researchers have attempted in defining an appropriate and plausible dynamic link model (see e.g. Carey et al., 2003, Friesz et al., 1993, Danganzo, 1994, Merchant and Nemhauser, 1978, Newell, 1993). For the route choice model, most of the assumptions made for the DTA model are similar to those for the static version. The main assumptions of route choice include deterministic user equilibrium and stochastic equilibrium. However, applying these same principles to the DTA framework causes a major difficulty in finding the modelling solution (equilibrated route choice inflows). This is mainly due to a complex structure of the dynamic loading model and the inter-relationship of the travel cost from one point in time to another. This represents the drawback of the DTA model as compared to its counterpart, the static model, whose modelling solution can be found efficiently.

Solving a general case of the dynamic user equilibrium (DUE) is a challenging task. Several researchers have tackled this problem with a rather restricted assumption. Wie et al (1995) proposed a heuristic route-departure swapping algorithm for a discrete time DUE comparing the route and departure time cost with the minimum OD travel cost. Lo and Szeto (2002) applied the projection method for solving variational inequality to the DUE assignment with cell-transmission model which requires the information about the set of routes *as priori*. Kuwahara and Akamatsu (1997) offered a convergence guaranteed solution algorithm with a restricted problem of the DUE under the case of one-to-many origin-destination case. With this restricted problem, they showed that the assignment can be decomposed by departure time. However, this property may not hold for a general network. Akamatsu (2001) later re-formulated the case with one-to-many OD without using path variable and developed an efficient algorithm which does not need to generate the path set. The decomposition of the assignment by departure time is also applicable for the case with many-to-one OD. Friesz and Mookherjee (2006) revisited the problem by reformulating it as a fixed point problem based on the differential variational inequality with state dependent time shifts. They also proposed a solution algorithm based on optimal control problem for the continuous-time DUE assignment.

This paper is motivated by the works of Han and Heydecker (2006) which provides an interesting discussion on the decomposition of DUE assignment and its equivalent formulation as a minimization problem (using a gap function) and Balijepalli and Watling (2005) which defines the derivative of the path travel cost with respect to inflow rates. In this paper, the problem of DUE assignment is first formulated as a VI and then

transformed to a minimization problem using its equivalent gap function. Then, through the optimal value function theorem (Dem'yanov and Malozemov, 1990) the descend direction in the space of path inflow vectors to minimize the gap function is defined using the derivative of the nested path cost operator as defined in Balijepalli and Watling (2005). The paper then applies Wolfe's reduced gradient projection method (Wolfe, 1963) to solve the DUE assignment problem.

The paper is structured into four further sections. The next section describes the definition of the dynamic link model and dynamic user equilibrium adopted in this paper. Then, section 3 explains the algorithm to find the descend direction inflow vector and the algorithm for solving the DUE problem. Section 4 tests the algorithm proposed with a small five link example. The last section summarises the paper and discussed future research issues.

# 2 Preliminary

This section presents the notations and assumptions of traffic model adopted in this paper. For the dynamic link, the paper adopts the model suggested by Friesz et al (1993) which exhibits several desired properties for a dynamic link model. A brief description of this model is given below.

Let $t$ be the indices of a discrete time instance and $\tau(t)$ denotes the travel time for traffic entering the link at time $t$ to traverse the entire link. We assume that the time interval is discretized to create a finite set of time periods $\{t = 0, 1, \ldots, T\}$ and the length of each time period is unity for simplicity. Let $u(t)$ and $w(t)$ define the inflow rate and outflow rate at time $t$ of the link considered. The amount of traffic on the link at any given time, $x(t)$, can be defined as:

$$x(t) = \sum_{s=0}^{t} \left( u(s) - w(s) \right) \tag{1}$$

Then, following the definition of the dynamic link model: $\tau(t) = f(x(t))$, where $f$ is a linear function with the argument of $x(t)$. We now discuss the formulation of the deterministic user equilibrium condition (DUE) without departure time choice for the route choice model under the DTA framework.

For each origin-destination pair, denoted $(r,s)$, let $\Pi(r,s)$ represents the set of feasible paths between this OD pair and each path is indicated by a subscript $p$. For each path, $C_p(t)$ denotes the path travel cost for the traffic departing its origin at departure time $t$ and $h_p(t)$ defines the inflow rate for path $p$ at time $t$.

**Definition 1**: The path inflow rate pattern, $h_p^*(t)$, is a discrete time dynamic user equilibrium if and only if:

$$\left.\begin{aligned} & h_p^*(t)\left[ c_p\left(t, h^*\right) - \mu_{(r,s)}\left(t, h^*\right) \right] = 0 \\ & c_p\left(t, h^*\right) - \mu_{(r,s)}\left(t, h^*\right) \geq 0 \\ & h_p^*(t) \geq 0 \end{aligned}\right\} \quad \forall p \in \Pi(r,s), r \in R, s \in S, t = 0, \ldots, T$$

$$\sum_{p \in \Pi(r,s)} h_p^*(t) = Q_{(r,s)}(t) \qquad \forall r \in R, s \in S, t = 0, \ldots, T \tag{2}$$

where $\mu_{(r,s)}\left(t, h^*\right)$ is the minimum travel cost for traffic departing origin $r$ at departure time $t$ and destined to destination $s$. This condition can be described as a generalisation of the Wardop's user equilibrium condition for the static model. As pointed out in Wie et al (1995), the drawback of this formulation is the it requires the set of all used paths at the beginning of the algorithm. We can also formulate (2) as a variational inequality.

**Theorem 1**: The discrete time DUE problem of definition 1 is equivalent to the following variational inequality: find $h^* \in \Omega$ such that

$$\sum_{t=0}^{T}\sum_{r\in R}\sum_{s\in S}\sum_{p\in\Pi(r,s)} C_p\left(t,h^*\right)\left[h_p^*\left(t\right)-h_p\left(t\right)\right]\geq 0 \quad \forall h\in\Omega$$

*where* (3)

$$\Omega=\left\{h:\begin{array}{ll}\sum_{p\in\Pi(r,s)} h_p\left(t\right)=Q_{(r,s)}\left(t\right) & \forall r\in R,s\in S,t=0,\ldots,T\\ h_p\left(t\right)\geq 0 & \forall r\in R,s\in S,\forall p\in\Pi\left(r,s\right),t=0,\ldots,T\end{array}\right\}$$

The VI expression in (3) can also be formulated as an equivalent optimization problem using its associated gap function. This formulation will then allow us to develop an algorithm to find the descent direction in the space of path inflow vector to minimize the equivalent gap function.

**Lemma 1**: Let $\Psi$ be a set of solution to a given VI condition of $f\left(x^*\right)\left(x^*-x\right)\leq 0 \ \ \forall x\in\Omega$, a function $\Theta\left(x\right)$ from $x$ to $\Re$ is a gap function for this VI if:

(i) $\Theta\left(x\right)$ is restricted in sign;

(ii) $\Theta\left(x\right)=0\Leftrightarrow x\in\Psi$

The first condition means that $\Theta\left(x\right)$ should have the same sign (either positive or negative) for all $x\in\Omega$. Following Hearn et al (1984) the gap function for $f\left(x^*\right)\left(x^*-x\right)\leq 0 \ \ \forall x\in\Omega$ satisfying the conditions stated above can be defined as:

$$\Theta\left(x\right)=\min_{\tilde{x}\in\Omega}-f\left(x\right)\left(x-\tilde{x}\right)=\min_{\tilde{x}\in\Omega} f\left(x\right)\tilde{x}-f\left(x\right)x$$

$\Theta\left(x\right)$ will always have a negative value and when $\Theta\left(x\right)=0$, $x$ is the solution to its associated VI.

Using Lemma 2, the gap function for the VI condition for the discrete time dynamic DUE in (3) can be defined as:

$$\Theta\left(h\right)=\min_{\tilde{h}\in\Omega}\left\{\sum_{t=0}^{T}\sum_{r\in R}\sum_{s\in S}\sum_{p\in\Pi(r,s)} C_p\left(t,h\right)\tilde{h}_p\left(t\right)\right\}-\sum_{t=0}^{T}\sum_{r\in R}\sum_{s\in S}\sum_{p\in\Pi(r,s)} C_p\left(t,h\right)h_p\left(t\right) \quad (4)$$

This formulation has also been proposed in Han and Heydecker (2006) for the development of their algorithm for solving the dynamic DUE. However, the key difference is that in this paper we do not decompose the assignment and the departure time at the origin as proposed in their paper. Consider the optimization problem:

$$\min_{h\in\Omega}\Theta\left(h\right)=\min_{h\in\Omega}\sum_{t}\Theta\left(h,t\right)$$

*where* (5)

$$\Theta\left(h,t\right)=\min_{\tilde{h}(t)\in\Omega(t)}\left\{\sum_{r\in R}\sum_{s\in S}\sum_{p\in\Pi(r,s)} C_p\left(t,h\right)\tilde{h}_p\left(t\right)\right\}-\sum_{r\in R}\sum_{s\in S}\sum_{p\in\Pi(r,s)} C_p\left(t,h\right)h_p\left(t\right)$$

From (5), for each of $\Theta\left(h,t\right)$, which is the dis-aggregated gap function by the departure time from the origins, there exists the contribution from the inflow from the path and future time period as $C_p(t,h)$ is a function of a vector of $h$ not just $h(t)$. Thus, it is unlikely that we can decompose the assignment by incremental of departure time using (5). However, the formulation of (5) provides some possibility in applying an optimization method to solve the DUE. This will be discussed in the next section.

## 3 Descent direction finding and solution algorithm

The problem stated in (5) can be considered as a minimization problem of an optimal value function, $\Theta(h)$. The strategy adopted in this paper is a straightforward application of the theorem related to the directional derivative of the optimal value function. At this stage, several assumptions are required to simplify the development of the algorithm. These assumptions are mainly relevant to the differentiability of the optimal value function.

*Assumption 1*: The dynamic travel time function, $f(x(t))$ is continuously differentiable with respect to the path inflow profile vector, $h$.

*Assumption 2*: For a given vector of path inflow profile, $h$, the solution of (4) is unique, i.e. there exists only one shortest path for each OD pair (unique shortest path solution).

Assumption 1 is generally guaranteed by the definition of the dynamic link travel time model as discussed earlier. Assumption 2 is required for the following lemma to establish the differentiability of the optimal value function. This may not generally be true in different network. When this is not true, the gap function becomes non-differentiable at some point and one need to employ non-smooth optimization algorithm. However, for simplicity we assume that Assumption 2 is true in this paper.

**Lemma 2** (Dem'yanov and Malozemov, 1990): Let $\varphi(x)$ be an optimal value function of the problem $\varphi(x) = \min_{y \in \Omega} f(x, y)$ and assume that $f$ is continuously differentiable everywhere and there exists a unique solution, $y$, to this problem for a given $x$. Then, $\varphi(x)$ is continuously differentiable everywhere with respect to $x$ and can be defined as:

$$\nabla \varphi(x) = \nabla_x f\left(x, y^*(x)\right)$$

where $y^*(x)$ is the solution of $\min_{y \in \Omega} f(x, y)$.

This lemma provides us a possible formulation of the derivative of $\Theta(h)$ in (5) which can be defined as:

$$\frac{\partial \Theta(h)}{\partial h_{\tilde{p}}(\tilde{t})} = \sum_{t=0}^{T} \sum_{r \in R} \sum_{s \in S} \sum_{p \in \Pi(r,s)} \frac{\partial C_p(t,h)}{\partial h_{\tilde{p}}(\tilde{t})} \tilde{h}_p^*(t) - \left\{ \sum_{t=0}^{T} \sum_{r \in R} \sum_{s \in S} \sum_{p \in \Pi(r,s)} \left[ \frac{\partial C_p(t,h)}{\partial h_{\tilde{p}}(\tilde{t})} h_p(t) \right] \right\} - C_{\tilde{p}}(\tilde{t}) \quad (6)$$

where $\tilde{h}_p^*(t)$ is the solution of $\min_{\tilde{h}(t) \in \Omega(t)} \left\{ \sum_{r \in R} \sum_{s \in S} \sum_{p \in \Pi(r,s)} C_p(t,h) \tilde{h}_p(t) \right\}$ which is simply a shortest-path problem. Finding $\tilde{h}_p^*(t)$ will also generate the new path which will then be included in the assignment process. From (6), we can notice the requirement of the calculation of the derivative of path travel cost with respect to the path inflow profile, $\frac{\partial C_p(t,h)}{\partial h_{\tilde{p}}(\tilde{t})}$. Balijepalli and Watling (2005) have already defined the exact formulation of this derivative for the general network using the same dynamic link model as adopted in this paper. Given the structure of (5) which involves only linear equality constraints, the method of Wolfe's Reduced Gradient Projection (Bazaraa et al., 1993, Wolfe, 1963) will be employed for solving the problem. We need some notations before introduction the algorithm. Define the flow conservation constraint in the form of **Ah** = **Q** in which **h** is a stacked vector of path inflow for all departure time periods and the elements of **A**, $a_{i,j}$ if path $j$ is related to OD pair $i$ ($q_i(t)$ is the demand for OD pair $i$ at departure time $t$) and $\mathbf{a}_j$ denote the column $j$ of **A**:

$$\mathbf{h}=\begin{bmatrix} h_1(1) \\ \vdots \\ h_p(1) \\ h_1(2) \\ \vdots \\ h_p(2) \\ \vdots \\ h_1(T) \\ \vdots \\ h_p(T) \end{bmatrix} \quad \mathbf{A}=\begin{bmatrix} a_{1,1}(1) & \cdots & a_{1,p}(1) & & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ a_{i,1}(1) & \cdots & a_{i,p}(1) & & 0 & \cdots & 0 \\ & \vdots & & \ddots & & \vdots & \\ 0 & \cdots & 0 & & a_{1,1}(T) & \cdots & a_{1,p}(T) \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & & a_{i,1}(T) & \cdots & a_{i,p}(T) \end{bmatrix}$$

The algorithm can be summarised as follows:

*Initialisation Step*: Choose a feasible path inflow profile, $\mathbf{h}_k$, satisfying the flow conservation constraint. Let $k = 1$ and go to Step 1

*Step 1*: Given $\mathbf{h}_k$, conduct the dynamic network loading to evaluate $C_p(t,h)$. Solve the shortest path problem to find $\tilde{\mathbf{h}}$ and then evaluate $\Theta(h)$, see equation (4)

*Step 2*: Calculate $\nabla_{\mathbf{h}}\Theta$ following equation (6)

*Step 3*: Let $\mathbf{d}_k = (\mathbf{d}_B, \mathbf{d}_N)$ where $\mathbf{d}_B$ and $\mathbf{d}_N$ are obtained as below from (10) and (11), respectively. If $\mathbf{d}_k = 0$, stop and $\mathbf{h}_k$ is the solution. Otherwise, go to step 2.

$I_k$ = index set of the path with the highest inflow for each OD pair at each time period from Step 1 (7)

$$\mathbf{B}=\{\mathbf{a}_j : j \in I_k\} \quad \mathbf{N}=\{\mathbf{a}_j : j \notin I_k\} \tag{8}$$

$$\mathbf{r}^{\mathrm{T}} = \nabla_{\mathbf{h}}\Theta^{\mathrm{T}} - \nabla_B\Theta^{\mathrm{T}} \cdot \mathbf{I} \cdot \mathbf{A} \tag{9}$$

$$d_j = \begin{cases} -r_j & \text{if } j \notin I_k \text{ and } r_j \le 0 \\ -h_j r_j & \text{if } j \notin I_k \text{ and } r_j > 0 \end{cases} \tag{10}$$

$$\mathbf{d}_B = -\mathbf{I} \cdot \mathbf{N} \cdot \mathbf{d}_N \tag{11}$$

*Step 4*: Solve the following line-search problem:

$$\min_{0 \le \lambda \le \lambda_{\max}} \Theta(\mathbf{h}_k + \lambda \mathbf{d}_k)$$

$$\text{where } \lambda_{\max} = \begin{cases} \min_{1 \le j \le n} \left\{ \dfrac{-h_{jk}}{d_{jk}} : d_{jk} \le 0 \right\} & \text{if } \mathbf{d}_k < \mathbf{0} \\ \infty & \text{if } \mathbf{d}_k \ge \mathbf{0} \end{cases}$$

and $h_{jk}$ and $d_{jk}$ are the $j$th components of $\mathbf{h}_k$ and $\mathbf{d}_k$, respectively. Let $\lambda_k$ be an optimal solution, and let $\mathbf{h}_{k+1} = \mathbf{h}_k + \lambda_k \mathbf{d}_k$. Replace $k$ by $k + 1$ and repeat Step 1.

Note that in our numerical experiment the Golden-section algorithm is adopted for solving the line-search sub-problem in Step 4. Noteworthy, other line-search algorithm (e.g. Armijo-line search) can also be applied to this step. The stopping criteria can also be set against the acceptable value of the gap function in Step 1 (apart from the value of $\mathbf{d}_k$).

## 4 Illustrative example

This section illustrates the application of the proposed algorithm for the DUE assignment with a small five-links network. Figure 1 depicts the test network. We assumes the linear-exit time function for all link in the form of $a+bx(t)$. Table 1 shows the parameters for the link exit time functions adopted in the test. There is

one OD pair in this network from node 1 to node 4 with three possible paths including path 1 (link1-> link 3), path 2 (link2 -> link4), and path3 (link1 -> link5 -> link4). Figure 2 shows the given inflow demand profile over the period of 25 units of time. The planning horizon is $t_0 = 0$ and $t_f = 40$ with the discretization of time at 1 unit time.

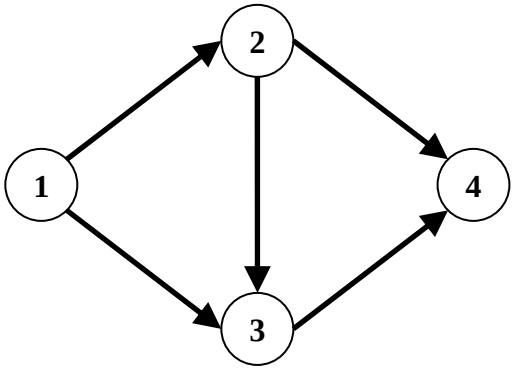



Figure 1: Test network

| Link no | From node | To Node | a | b |
|---|---|---|---|---|
| 1 | 1 | 2 | 2.0 | 0.025 |
| 2 | 1 | 3 | 2.5 | 0.010 |
| 3 | 2 | 4 | 2.5 | 0.010 |
| 4 | 3 | 4 | 2.0 | 0.020 |
| 5 | 2 | 3 | 2.0 | 0.020 |

Table 1: Link exit time parameters for the test network

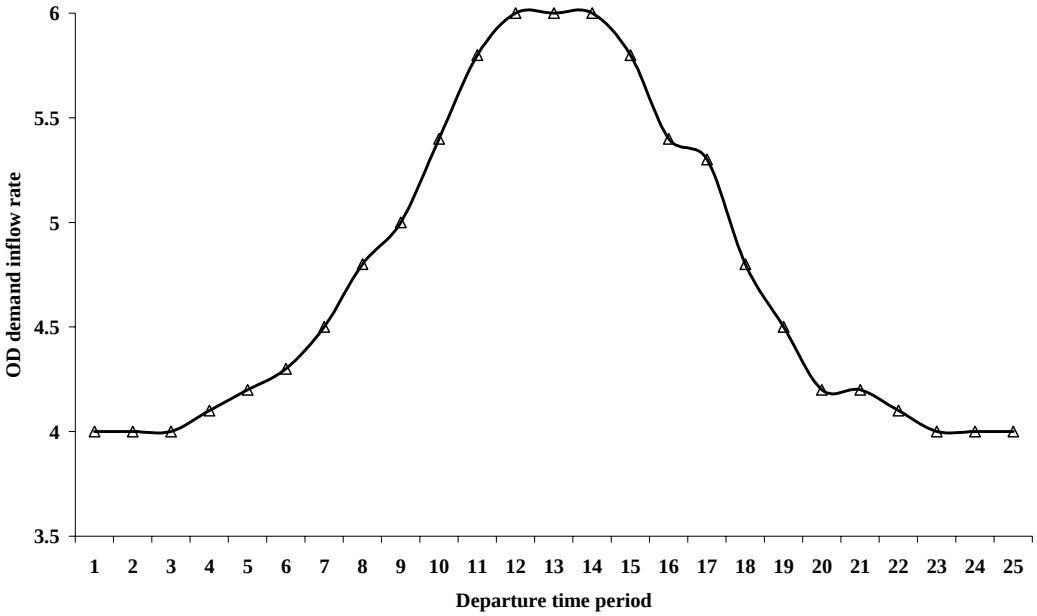



Figure 2: OD inflow rate profile

The algorithm proposed in the previous section is then applied to the problem with the stopping criteria on the gap function of 1.5. To demonstrate the output from the algorithm, figure 3 show the equilibrated inflow profiles for the three paths in the network. Figure 4 shows an example of inflow and outflow profile for link 2. Figure 5 shows the path travel costs at the initial condition adopted. Figure 6 then presents the path travel costs from the solution as found by the algorithm. To demonstrate the convergence of the algorithm, figure 7 plots the gap function at each iteration of the algorithm (the algorithm terminated after six iterations). The final value of the gap function is 1.5 comparing to the value of 45.8 at the initial point. The algorithm was

written in Visual Basics and tested on a 1.7GHz laptop with 256 MB-RAM. The computational time for this test was approximately 10 minutes.

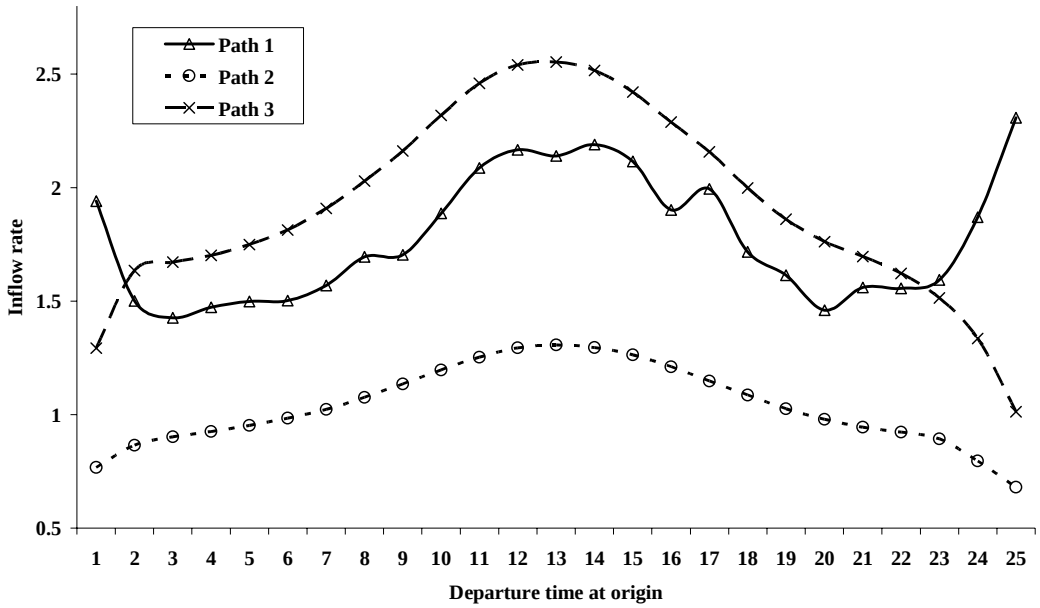



Figure 3: Inflow and outflow profile for path 1

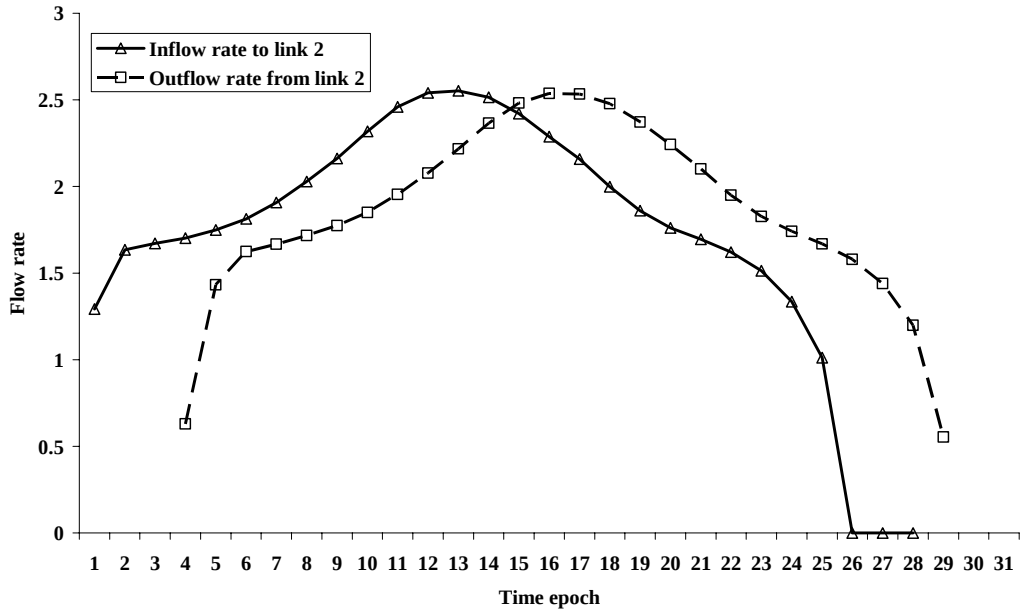



Figure 4: Inflow and outflow profile for path 3

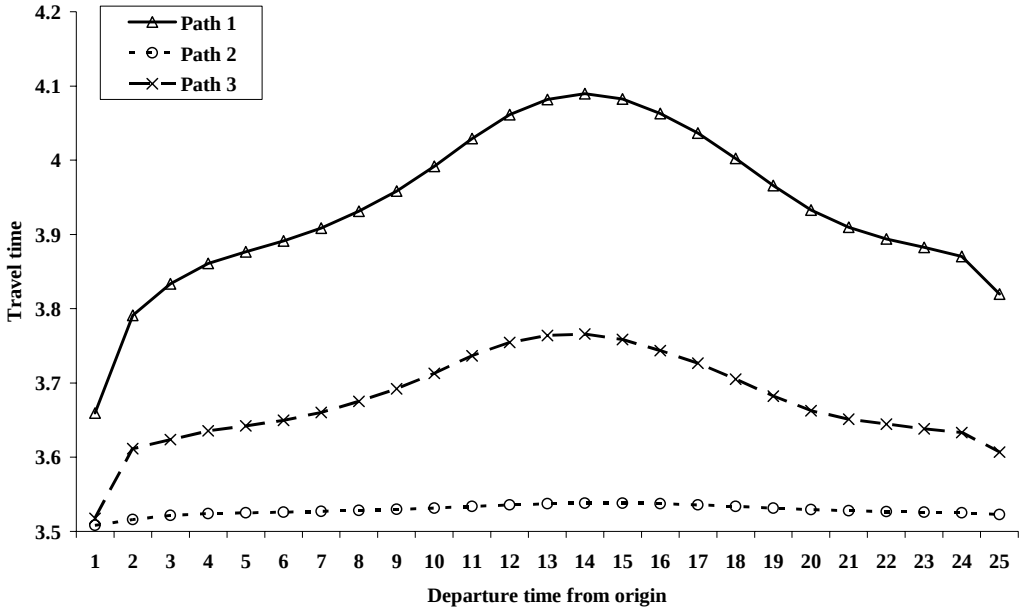



Figure 5: Travel cost at different departure time based on the initial condition of the inflow profile

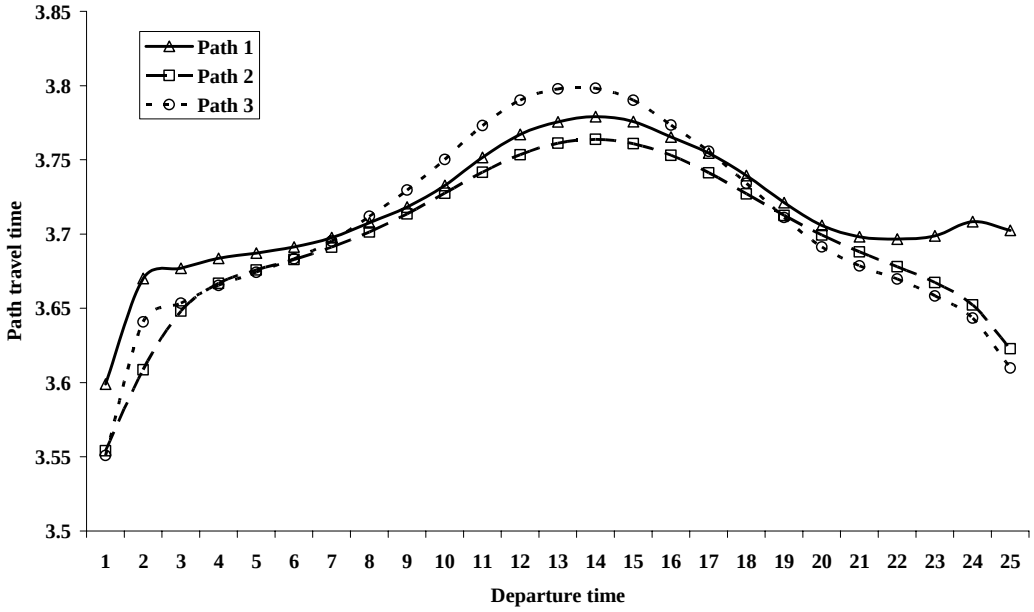



Figure 6: Equilibrated travel cost profile for all paths

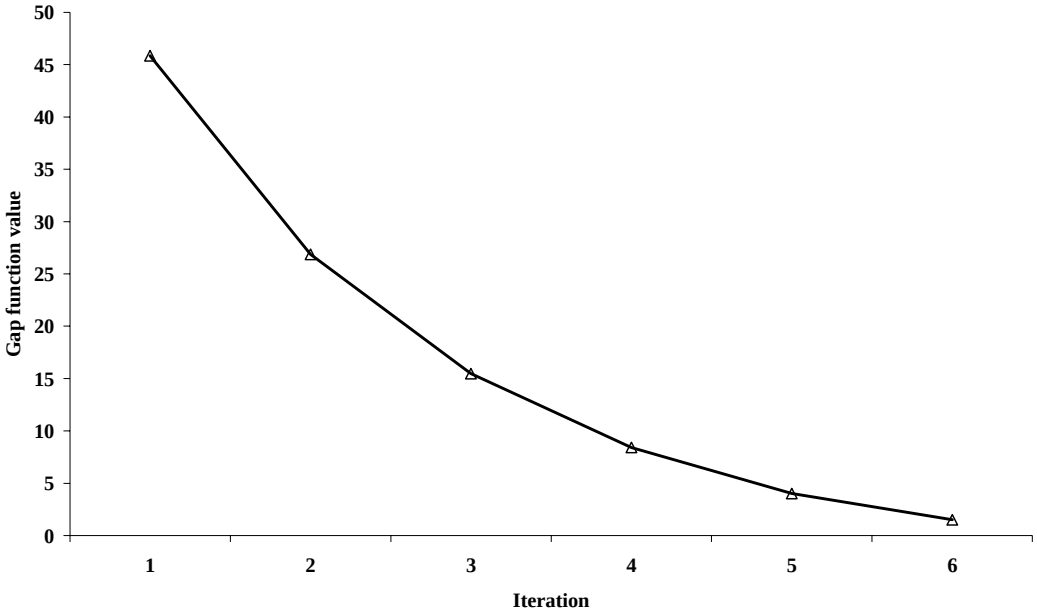



Figure 7: Gap function at different iterations of the algorithm

# 5 Conclusions

Presenting the dynamical phenomena of congestion in traffic network modelling is a key advance step toward a more realistic modelling framework. Nevertheless, solving the assignment problem of dynamic user equilibrium is still a real challenging task. This paper proposed an alternative solution algorithm for the DUE. The main idea behind the algorithm was the reformulation of the DUE problem as a minimization problem using its equivalent gap function formulation. Then, we utilised the optimal value function theorem to define the derivative of the gap function (which is a minimization problem itself) and applied the formulation of the derivative of dynamic nested cost operator with respect to path inflow rate to complete the numerical algorithm. The derivative information of the gap function was then fed into the Wolfe's reduced gradient projection algorithm to search for the descent direction in the space of path inflow rate (for all departure times). The paper described the detail of the algorithm and then tested it with the small five-link network. The result obtained is promising in which the algorithm converged to an acceptable solution with the criteria set on the gap function. Although, this initial test shows a rather encouraging result several issues need to be tackled in the future research. Firstly, the computational effort for the jacobian of path travel time is very intensive (accounted for more than 70% of the computational time). Future research will look into a way to reduce this computational burden using some approximation methods. Secondly, it is still not theoretical guaranteed concerning the stationary point of the gap function which may involve a saddle point. Further research will investigate the property of the gap function of DUE and test the performance of other line-search techniques (e.g. Armijo line search). Thirdly, several assumptions made for the differentiability property of the gap function may not be valid in general cases (e.g. unique shortest path). These assumptions should be relaxed in the subsequent work.

# Acknowledgements

This research is funded by the UK EPSRC under the Platform Grant project, "Towards a unified theoretical framework for transport network and choice modelling". However, note that this paper should not represent the conclusive view of the project. The research was conducted partly while the first author visiting Centre for Collaborative Research, University of Tokyo, Japan. The authors would like to sincerely thank David Watling and Chandra Balijepalli for a fruitful discussion on DTA and the cost derivatives which led to the initial idea of this work.

## References

AKAMATSU, T. (2001) An Efficient Algorithm for Dynamic Traffic Equilibrium Assignment with Queues. *Transportation Science,* 35**,** 389-404.

BALIJEPALLI, N. C. & WATLING, D. P. (2005) Doubly Dynamic Equilibrium Approximation Model for Dynamic Traffic Assignment. IN MAHMASSANI, H. S. (Ed.) *Transportation and Traffic Theory: Flow, Dynamics, and Human Interaction.* Oxford, UK, Elsevier.

BAZARAA, M. S., SHERALI, H. D. & SHETTY, C. M. (1993) *Nonlinear programming: theory and algorithms,* New York, John Wiley & Sons, Inc.

CAREY, M., GE, Y. E. & MCCARTNEY, M. (2003) A Whole-Link Travel-Time Model with Desirable Properties. *Transportation Science,* 37**,** 83-96.

CHEN, J. H. W. & FLORIAN, M. (1998) The continuous dynamic network loading problem: a mathematical formulation and solution method. *Transportation Research Part B,* 32**,** 173-187.

DANGANZO, C. F. (1994) The Cell Transmission Model: A Simple Dynamic Representation of Highway Traffic Consistent with the Hydrodynamic Theory. *Transportation Research Part B,* 28**,** 269-287.

DEM'YANOV, V. F. & MALOZEMOV, V. N. (1990) *Introduction to minmax,* New York, Dover.

FRIESZ, T. L., BERNSTEIN, D., SMITH, T. E., TOBIN, R. L. & WIE, B. W. (1993) A variational inequality formulation of the dynamic network user equilibrium problem. *Operations Research,* 41**,** 179-191.

FRIESZ, T. L. & MOOKHERJEE, R. (2006) Solving the dynamic network user equilibrium problem with state-dependent time shifts. *Transportation Research Part B,* 40**,** 207-229.

HAN, S. & HEYDECKER, B. G. (2006) Consistent objectives and solution of dynamic user equilibrium models. *Transportation Research Part B: Methodological,* 40**,** 16.

HEARN, D. W., LAWPHONGPANICH, S. & NGUYEN, S. (1984) Convex programming formulations of the asymmetric traffic assignment problem. *Transportation Research Part B: Methodological,* 18**,** 357-365.

KUWAHARA, M. & AKAMATSU, T. (1997) Decomposition of the dynamic assignments with queues: DUO and DUE. *Transportation Research Part B,* 31**,** 1-10.

LO, H. K. & SZETO, W. Y. (2002) A cell-based variational inequality formulation of the dynamic user optimal assignment problem. *Transportation Research Part B,* 36**,** 421-443.

MERCHANT, D. K. & NEMHAUSER, G. L. (1978) A model and an algorithm for the dynamic traffic assignment problems. *Transportation Science,* 12**,** 183-199.

NEWELL, G. F. (1993) A simplified theory of kinematic waves in highway traffic, part I: general theory; part II: queuing at freeway bottlenecks part III: multi-destination flows. *Transportation Research Part B,* 27**,** 281-313.

WIE, B. W., TOBIN, R. L., FRIESZ, T. L. & BERNSTEIN, D. (1995) A Discrete Time, Nested Cost Operator Approach to the Dynamic Network User Equilibrium Problem. *Transportation Science,* 29**,** 79-92.

WOLFE, P. (1963) Methods of Nonlinear Programming. IN GRAVES, R. L. & WOLFE, P. (Eds.) *Recent Advances in Mathematical Programming.*

# OBSERVABILITY IN ESTIMATING TIME DEPENDENT ORIGIN-DESTINATION FLOWS FROM TRAFFIC COUNTS

*Ramachandran Balakrishna: Massachusetts Institute of Technology, USA* rama@mit.edu
*Moshe Ben-Akiva: Massachusetts Institute of Technology, USA* mba@mit.edu
*Yang Wen: Massachusetts Institute of Technology, USA* wenyang@mit.edu

## Abstract
Time-dependent Origin-Destination (OD) flows are fundamental inputs to any Dynamic traffic Assignment (DTA). However, the true flows are generally unobserved. DTA systems often employ standard OD estimation methodologies to efficiently extract this information from link counts. Practical sensor coverage levels, however, dictate that the classical OD estimation approaches rely on target (or seed) OD flows to remove the indeterminacy caused by using fewer independent counts than OD variables, and to provide structural information about the unknown flows. In this paper, we evaluate the property of *observability* in the context of sequential time-dependent OD estimation for large-scale DTA systems. We hypothesize that under certain conditions, and given sufficient prior intervals of count observations, the OD flows for a particular time interval can be uniquely identified. A framework for the validation of observability in OD estimation is outlined. Guidelines for determining the number of "warm-up" intervals are suggested. Practical issues concerning the calculation of assignment matrices, key inputs for OD estimation, are also detailed. We present an empirical case study that verifies observability on a network from Los Angeles, California.

## 1 Introduction
Ever growing traffic congestion and the externalities associated with it has forced transportation engineers and researchers around the world to better manage existing supply and demand. Intelligent Transportation Systems (ITS) have emerged as a key field in this direction, providing the tools and technologies to ease travel by "intelligently" managing all components of transportation (including planning, design and operations). One of the applications of ITS is the provision of pre-trip and en-route travel information that enables informed driver decision making, lower delays and more reliable travel times. A necessary requirement of such systems is the ability to predict future traffic, so as to proactively respond to drivers' concerns and avoid potentially undesirable network conditions such as long delays, congestions, queuing and excessive pollution. Dynamic Traffic Assignment (DTA) systems simulate and forecast network conditions under varying traffic demand and influencing factors such as weather, special events, construction activities and accidents.

DTA systems model complex and dynamic interactions between transportation demand and supply, to estimate current conditions, anticipate future network performance and generate consistent, anticipatory route guidance. A key factor influencing the reliable deployment of such systems is appropriate model calibration. Effective calibration enables DTA systems to capture the demand-supply interactions in the study area, through their numerous model parameters and inputs. A critical component of such an exercise is the set of time-dependent matrices of OD flows that capture network travel demand. True demand patterns are generally unobserved, and must be inferred from indirect measurements of the same. OD estimation methods are typically employed for this purpose.

Classical OD estimation methods solve a system of linear equations that map the unknown OD flows to observed sensor count measurements. However, the number of independent sensors is typically much lower than the dimension of the OD matrix, thus mandating the need for a starting matrix of OD flows that will render the approach feasible. Starting flows are often selected arbitrarily, and impact the final OD estimates. The dependence on the seed matrix is an undesirable

property of real-time DTA systems. In this paper, we hypothesize that OD estimation for real traffic networks is an *observable* phenomenon: the OD flows for a given time interval may largely be identified uniquely, given repeated sensor measurements from sufficiently many past intervals. We provide a methodology for testing this hypothesis, and present empirical results that validate the same.

The rest of this paper is organized as follows. We begin with a review of research focusing on issues related to observability in the context of traffic network modeling. We then discuss the mathematical basis for analyzing dynamic systems for observability, and describe the implications for the OD estimation problem. A framework for the validation of observability in OD estimation is outlined, and demonstrated through a case study on a real network from Los Angeles, California. We conclude with a summary of our main findings and future research directions.

## 2 Literature Review

Literature on the observability of the OD estimation procedure is limited. Ashok and Ben-Akiva (2000) propose a Kalman Filter based real-time OD estimation methodology, and mention observability as a desirable property of dynamic systems. The authors state that under conditions of observability, the influence of the initial value of the state vector in their state-space model would disappear with time. They also list several factors affecting observability, primary being the ratio of the number of sensors to the number of OD pairs. The authors also emphasize the importance of two linkages: between the OD flows and counts through an assignment matrix, and between the OD flows over time in the form of a transition matrix.

Several papers report on the optimal sensor location problem, which is closely related to observability. The basic sensor location problem aims to identify the minimum number of sensors (and their locations) required in order to efficiently estimate all unknown OD flows. A restricted variant of this formulation incorporates resource constraints that limit the number of sensors that may be deployed. Alternatively, one may also determine the best OD estimation accuracy level that may be achieved with a given pattern of sensors.

Yang et al. (1991) propose the Maximum Possible Relative Error (MPRE) index as an upper bound on OD estimation error. They provide a quadratic formulation that estimates the relative OD error (with the unknown true flows as the reference) by maximizing the sum of squared relative errors, using the observed link counts as constraints. The approach assumes perfect knowledge of the assignment matrix (and hence route choice), and focuses on static OD estimation.

Yang et al. use their theoretical framework to develop rules that impact the observability of OD flows. Their OD covering rule dictates that all OD flows must be counted at least once. Yang and Zhou (1998) extend the MPRE results to propose additional guidelines to increase the reliability of the estimated flows. For example, they recommend that sensors be placed so that high fractions of all OD flows are measured, while ensuring that each sensor provides information about multiple OD pairs. The mutual independence of all sensor measurements is also a requirement.

Bianco et al. (1997) propose a two-stage procedure that derives the complete network traffic flow vector before producing a reliable estimate of the (static) OD matrix based on a minimal-cost set of sensor counts. The authors use a heuristic algorithm based on the combined cutset principle of graph theory, and claim that the OD estimation error is always bounded, even when the OD covering rule is violated. The perfect knowledge of turning fractions is assumed.

Other papers report on the optimal location of Automatic Vehicle Identification (AVI) readers (Anthony et al., 2004), image sensors (Gentili and Mirchandani, 2004) and variable message signs (Huynh et al., 2002).

Most of the sensor location literature makes restrictive assumptions, such as the network-wide availability of turning fractions, knowledge of the true assignment matrix, and error-free counts. In addition, the focus on static flows effectively ignores traffic dynamics. This paper furthers the state of the art along several dimensions. Observability in the context of the *dynamic* OD estimation problem is discussed, and the mathematical requirements for the unique identification of OD flows are outlined. An empirical methodology to test for observability is presented, and validated on a real and large traffic network with actual sensor data.

## 3 Methodology

We begin with general background on the estimation of dynamic OD flows for a network with $n_{OD}$ OD pairs and $n_l$ instrumented links. The time period of interest, such as the AM peak, is assumed to be sub-divided into H uniform intervals, $h$=1,2,…,H. The variables of interest are denoted as $\mathbf{x}_h$, a matrix of OD demand rates departing their origins during time interval $h$. The classical OD estimation problem can then be stated as the solution for unknown OD flows $\mathbf{x}_h \quad \forall h$, given vectors of aggregate sensor counts $\mathbf{y}_h$ and some starting (target) OD flows $\mathbf{x}^0$. The problem is generally formulated as a sequence of optimizations of a function of two error terms, as defined by the following equations (Cascetta et al, 1993):

$$\mathbf{y}_h = \mathbf{a}_h^h \mathbf{x}_h + \mathbf{v}_h \tag{1}$$

$$\mathbf{x}_h^a = \mathbf{x}_h + \mathbf{w}_h \tag{2}$$

Equation (1) captures the fit between the observed counts $\mathbf{y}_h$ and the fitted values $\hat{\mathbf{y}}_h = \mathbf{a}_h^h \mathbf{x}_h$, where $\mathbf{a}_h^h$ is an $n_l$ x $n_{OD}$ matrix of assignment fractions mapping the OD flows in interval $h$ to the counts measured at the end of the same interval. The counts on the left hand side of equation (1) have been adjusted where relevant by $\sum_{p=h-p'}^{h-1} \mathbf{a}_h^p \hat{\mathbf{x}}_p$, the contributions from vehicles departing in past intervals ($p$' denotes the number of intervals required to span the longest trip on the network). Equation (2) models the dependence of the estimated OD flows on the target values $\mathbf{x}_h^a$. The terms $\mathbf{v}_h$ and $\mathbf{w}_h$ are errors that must be minimized after applying a (non-linear) transformation such as least squares. The starting OD flow matrix is used as the *a priori* OD matrix for the first interval ($h$=1). For subsequent intervals ($h$>1), a transition matrix is applied, that captures dependence of the flows in interval $h$ on those estimated for $h$-1:

$$\mathbf{x}_h^a = \mathbf{\Phi}_h \hat{\mathbf{x}}_{h-1} \tag{3}$$

The simplest definition of the transition matrix $\mathbf{\Phi}_h$ would be the identity matrix, thus setting $\mathbf{x}_h^a = \hat{\mathbf{x}}_{h-1}$.

*Observability*

In order to mathematically define observability in OD estimation, we start with the dynamic system encapsulated in equations (1), (2) and (3), but assume noise-free measurements. The evolution of the system over $n$ consecutive intervals can then be represented by the following set of equations:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} a_1^1 \\ a_2^2 \Phi_2 \\ a_3^3 \Phi_3 \Phi_2 \\ \vdots \\ a_n^n \prod_{i=2}^{n} \Phi_i \end{bmatrix} \mathbf{x}^0 = \Gamma \mathbf{x}^0$$

The dynamic system is observable in $n$ measurement intervals if the initial state $\mathbf{x}^0$ can be determined uniquely from the $n$ sets of count measurements. It is easily verified that observability requires that the matrix $\Gamma$ be of full rank $n$ so as to be non-singular.

Other basic requirements for observability may be derived from analyzing the structure of the assignment matrix. Clearly, if an entire column of assignment fractions were zero, the corresponding OD flow will never be measured, and hence cannot be estimated. Such a flow would be unobserved, unless it is structurally related to other measured OD flows in some known way (through a non-diagonal transition matrix). Similarly, when the entries of a column of $\mathbf{a}_h^h$ are all negligibly small, the OD estimate will be subject to a high variance. These observations are essentially the OD covering and maximum flow interception rules proposed by Yang and Zhou (1998).

*Testing for observability*

The mathematical definition of a system's observability, while simple, is not easy to test. The matrix $\Gamma$ is a complex, non-linear function of network travel times and drivers' route choice decisions. Travel times in turn are manifestations of travel demand levels, which are yet to be estimated. Noise in the various measurement processes introduces further complexities that render a theoretical observability analysis intractable.

We propose an empirical methodology to test for observability. For this purpose, we revert to the implications of an unobservable system on the outcome of the sequential OD estimation process. Ashok (1996) states that under conditions of non-observability, the effects of the starting matrix $\mathbf{x}^0$ do not disappear with time. The OD estimates will thus depend on the seed matrix, even when $h$ is far from 1. Consequently, the use of different starting matrices should yield significantly different flow estimates as the sequential procedure progresses across $h$.

The validity of the observability property may thus be tested using the following algorithm:

1. Generate D separate starting OD matrices $\mathbf{x}^{0,d}$, $d$ = 1,2,…,D.
2. Obtain sensor count measurements for all H time intervals in the analysis period, $h$ = 1,2,…,H.
3. $d$ = 1
4. Perform sequential OD estimation for all H intervals, using $\mathbf{x}^{0,d}$ as the target matrix for $h$ = 1. Archive the output flows $\mathbf{x}_h^d \quad \forall h$.
5. $d = d + 1$
6. Go to step 4 until $d$ = D + 1.
7. Compute statistics to compare estimates across $h$ and $d$.

The generation of starting matrices must reflect the fact that the true underlying demand patterns on the network are unobserved. Two dimensions are relevant in this context. The structure of the starting OD matrix must be varied to capture a range of possible *relative* magnitudes between the $n_{OD}$ flows. In addition, the matrices must be drawn with different mean flow levels, to determine any scale effects.

*Determining* H

For an observable system, the effect of $\mathbf{x}^0$ should diminish with *h*. The number of intervals before this effect falls below a pre-specified threshold should ideally be small. However, several factors may influence the rate of this process. A strong indicator of the magnitude of H is the ratio $H^* = \left\lceil \left( n_{OD} / n_l \right) \right\rceil$. Intuitively, each application of the transition equation carries over structural OD flow information into all future time intervals, implicitly generating additional measurement equations. After $H^*$ consecutive periods, there would therefore be sufficient equations to estimate the OD flows uniquely. It should be noted, however, that the above hypothesis relies on the existence and *a priori* knowledge of the temporal (interval-over-interval) relationships between the underlying OD flows, typically encapsulated in the transition matrix $\Phi_h$. When such relationships are either absent or hard to determine, each OD flow must be estimated from exactly one count measurement. The seed flows could then play a significant role in determining the final solution, thus preventing observability.

The speed with which an OD flow estimate converges towards its "true" value may also be influenced by the ratio of the maximum travel time between the OD pair and the length of the estimation interval. Vehicles that stay on the network for many intervals will be counted multiple times during their trips. Counts from future intervals would thus provide useful measurements on past OD flows, thereby accelerating the convergence. In order to exploit this characteristic, however, one would have to *simultaneously* estimate the flows departing in more than one time interval, so that past flows are periodically re-estimated as new counts data become available. Ashok (1996) describes such a state augmentation procedure that displays attractive computational properties while efficiently utilizing the available sensor information.

*Note on the calculation of assignment matrices*

The assignment matrix $\mathbf{a}_h^h$ is critical to the OD estimation process, and may be obtained in several ways. The most frequently adopted approach involves the simulation of a demand matrix using a network loading mechanism (together with a route choice model), and recording vehicle crossings at each sensor location. A simple post-processing of the resulting database yields the fraction of each OD pair counted at every sensor. Such a simulated assignment matrix is adequate when the starting OD matrix is close to their true levels, and when the route choice model and network travel times are good approximations of reality.

Often, however, all of the above components are unknowns that must be calibrated in the absence of *a priori* values. In such cases, the simulation of an arbitrary demand matrix may result in the estimation of stochastic and biased assignment factions, the bias arising from artificial capacity bottlenecks caused by unrealistic starting demand structure (Jha et al., 2004). The use of low demand levels to prevent artificial congestion could further increase the stochasticity, since fewer vehicles are used in the computation of the assignment fractions. To circumvent these drawbacks, a natural approach is to compute the assignment fractions analytically from simulated travel times.

The theoretical basis and equations for the calculation of analytical assignment matrices may be found in Ashok (1996). The approach essentially combines route choice fractions with the origin-to-sensor travel times for the first and last departures in interval *p*, to evaluate the fraction of an OD pair's demand that crosses each sensor in every time interval. Route choice fractions may be obtained using a discrete choice model along with OD travel times by path.

An analytical assignment matrix is expected to be more reliable than its simulated counterpart, since the travel times on which it is based are averaged across vehicles from multiple OD pairs and hence less stochastic. Also, such a matrix explicitly captures the contributions of every OD pair to the observed sensor counts, in particular the pairs that have zero starting flows (an OD pair with a zero starting flow would never contribute to the simulated assignment matrix, and will thus be ignored during the current and future OD estimation steps).

## 4 Case study

A dataset from the South Park region of downtown Los Angeles was used to empirically test for observability. The network (Figure 1) was coded using 243 nodes connected by a total of 606 directed freeway and arterial links. Count data from 203 loop detector stations were obtained and aggregated into uniform, 15-minute time intervals. In addition, arterial detector occupancy and freeway speed records were available. Occupancies were converted into density estimates, and used in conjunction with counts to infer speeds on arterial links.

*Experimental setup*

The beginning of the analysis period was chosen as 3:00 AM, when traffic conditions were ascertained to be clearly in the free-flow regime. Sequential OD estimation was carried out, using equation (3) to calculate the target matrix for each interval $h>1$. The assignment matrix, a crucial input for OD estimation, was calculated analytically from simulated network travel times. The necessary travel time estimates between each origin node and sensor were obtained from the DynaMIT (Ben-Akiva et al., 2001) mesoscopic traffic simulator.


**Figure 1. The Study Network (source: Google Maps)**

In the absence of any *a priori* information regarding the structure of the OD matrix, $n_{OD}$=3908 pairs were identified to cover every feasible origin-destination combination. A maximum of $H^* = 20$ was thus assumed.

*Supply and route choice model calibration*

DynaMIT's supply and route choice models were calibrated using the archived sensor data, in order to replicate the traffic dynamics observed in the field. Several network links were sub-divided into segments based on physical attributes (number of lanes and location on the network). The resulting 740 segments were further grouped based on their similarity with segments with sensors. Speed-density relationships were fitted for each of the 15 segment groups, using techniques outlined in Kunde (2002). Three route choice model parameters (a travel time coefficient, freeway bias factor and a penalty for switching between facility types) were estimated through a grid search.

*Starting (seed) OD flows*

Three starting OD matrices were randomly generated, by drawing from various distributions:

- Case D: $x_r^0 = 1 \quad \forall r$
- Case U: $x_r^0 = U[70,130] \quad \forall r$
- Case N: $x_r^0 = N(20,5) \quad \forall r$

Cases D, U and N represent draws from constant, uniform and normal distributions respectively. The means of the various distributions (in vehicle departures per 15 minutes) were selected so as to introduce an additional scale effect. The index *r* denotes an OD pair.

*Measures of performance*

Two comparisons were used to study the impact of the starting OD matrix. Since the true OD flows on the network are unobserved, the OD estimates from the three cases were compared in pairs (for each time interval *h*) through the Root Mean Square Normalized (RMSN) and "scale" statistics:

$$RMSN_h = \sqrt{\frac{\sum_{r=1}^{R}\left(x_{r,h}^i - x_{r,h}^j\right)^2}{\sum_{r=1}^{R}\left(x_{r,h}^i\right)^2}} \qquad Scale_h = \sqrt{\frac{\sum_{r=1}^{R}\left(x_{r,h}^i\right)^2}{\sum_{r=1}^{R}\left(x_{r,h}^j\right)^2}}$$

where *i* and *j* belong to the set {D, U, N} and $i \neq j$. In addition, the fit between the observed sensor counts and the corresponding estimated values was compared both graphically and through the RMSN statistic.

*Numerical results*

Figure 2 provides a graphical summary of the three starting OD matrices, outlining both the structural and scale variations between them. Interval-specific RMSN and scale statistics computed between cases D-N and N-U indicate that the estimated OD flows from the three starting points have converged after *h*=14 intervals, with the RMSN values consistently below the 7% level. The scales, evaluated as 0.996 for D-N and 0.999 for N-U, further validate the conclusion that the OD estimates after 14 intervals are comparable across starting points.

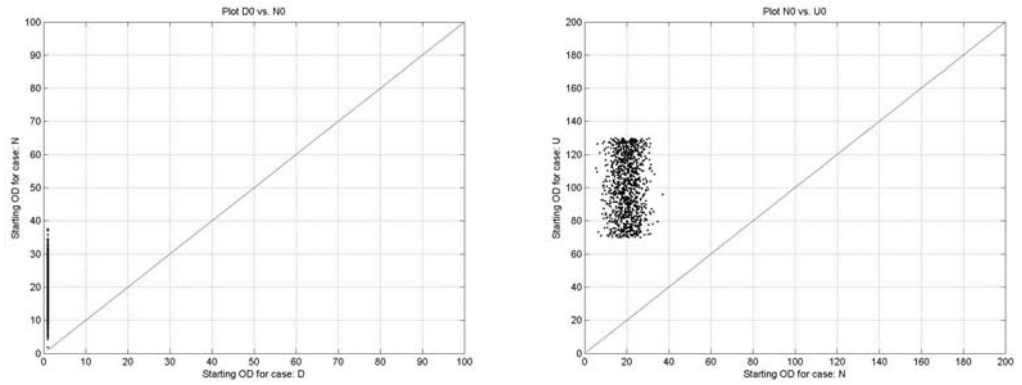



**Figure 2. Comparison of Starting OD Flows: (a) N vs. D (b) U vs. N**

Figure 3 graphically summarizes the three OD estimates for *h*=14. It is observed that case D overestimates the flow for a single OD pair. Further analysis revealed this pair's origin and destination nodes to lie along a freeway mainline, while no sensor intercepted flow along the primary mainline path. The OD flow was therefore estimated based on very small contributions to parallel arterial streets, and was hence unreliable.

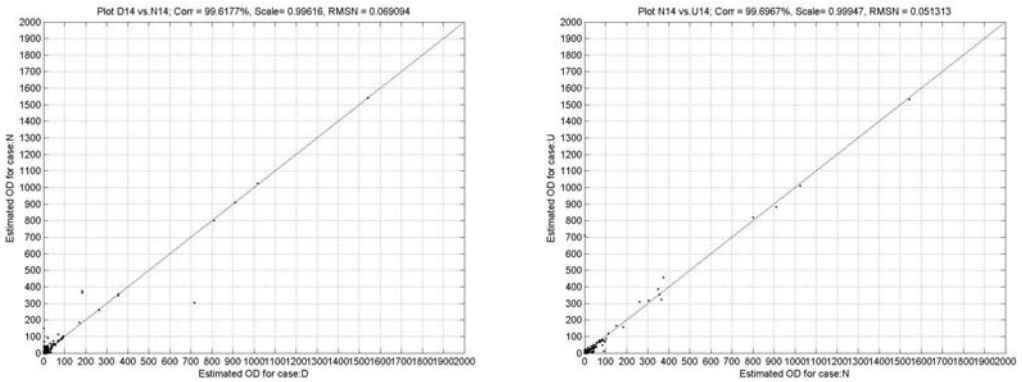



**Figure 3. Comparison of OD Flows for *h*=14: (a) N vs. D (b) U vs. N**

Having established that the OD flows are indeed observable, dynamic OD flows were estimated for the remainder of the 24-hour period. Graphical comparisons of the fit between observed and fitted counts (simulated by DynaMIT) for two intervals in the AM peak are presented in Figure 4. Stable RMSN statistics in the range of 0.15-0.19 were obtained for the combined AM+PM peak period.

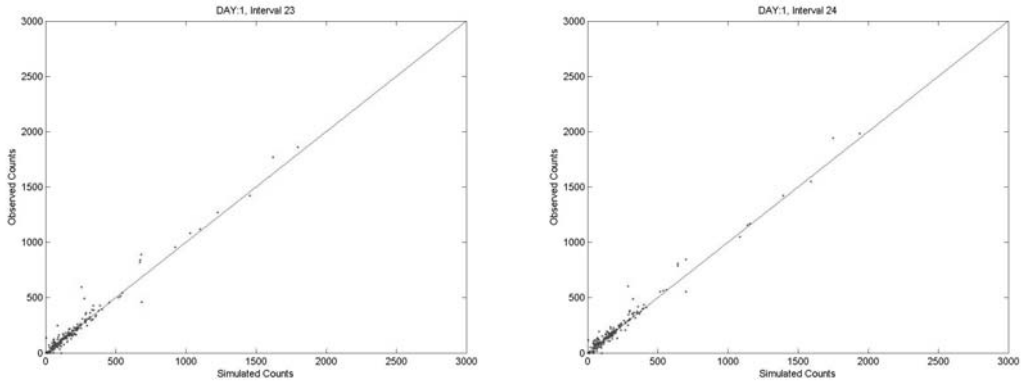



**Figure 4. Simulated vs. Observed Sensor Counts: (a) 8:00-8:15 (b) 8:15-8:30**

Figure 5 further illustrates the precision with which the estimated OD flows replicate temporal traffic dynamics. The plots compare the observed and simulated sensor count variability across 15-minute intervals for sample arterial and freeway detectors, by time of day (3 AM-midnight).

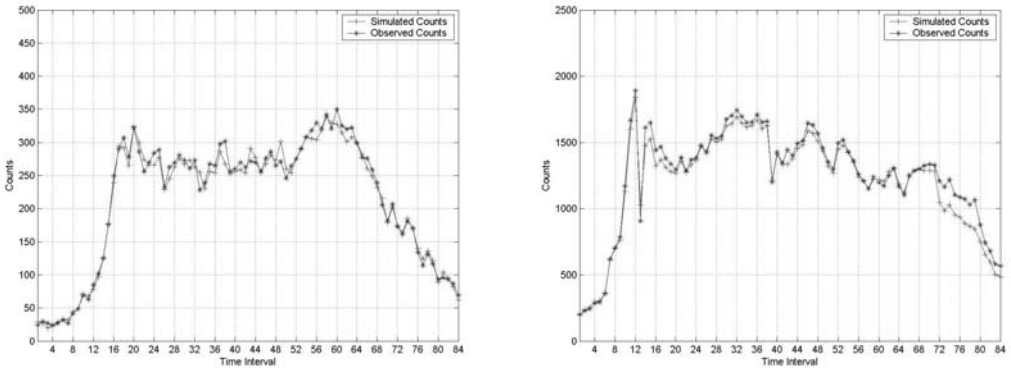



**Figure 5. Temporal Sensor Count Variability: (a) Arterial and (b) Freeway**

## 6 Conclusion

Classical OD estimation procedures, used for calibrating dynamic traffic assignment systems, typically update an *a priori* target (seed) matrix of OD flows with limited aggregate sensor count information. The choice of seed flows is known to impact the final estimates, which is undesirable from transportation planning and operations perspectives. In this paper, we present the hypothesis that dynamic OD flows are observable when sensor measurements from sufficiently many past time intervals, along with known interval-over-interval OD transition relationships, are available. We outline a framework for the numerical testing of this hypothesis, and present empirical evidence of its validity. Three randomly generated seed OD matrices (drawn from different distributions) were employed together with a sequential OD estimation approach. The least squares based estimates were found to converge after several intervals. We present some intuition for the number of consecutive warm-up intervals needed for convergence and observability.

The case study illustrated the proposed test for observability using a large and real traffic dataset. However, the true underlying OD flows were unobserved. Ongoing research focuses on the use of known OD inputs and simulated sensor count data from the MITSIMLab (Yang and Koutsopoulos, 1996 and Yang et al., 2000) microscopic traffic simulator, to demonstrate both the uniqueness and accuracy of the estimated flows. Various sensor configurations are also being tested. Directions for future research include analyses of the impact of assignment and transition matrix specification errors on observability.

## Acknowledgements

This research was partially supported by the National Science Foundation under projects CMS-0339005 and CMS-0339108.

## References

1. Anthony, C., C. Piya and P. Surachet (2004) “A Multi-Objective Model for Locating Automatic Vehicle Identification Readers.” Presented at the 83rd annual meeting of the Transportation Research Board.
2. Ashok, K. (1996) “Estimation and Prediction of Time-Dependent Origin-Destination Flows.” PhD thesis, Massachusetts Institute of Technology.
3. Ashok, K. and M. Ben-Akiva (2000) “Alternative Approaches for Real-Time Estimation and Prediction of Time-Dependent Origin-Destination Flows.” *Transportation Science*, Vol. 34(1), pp. 21-36.

4. Balakrishna, R. (2002) "Calibration of the Demand Simulator within a Dynamic Traffic Assignment System." Masters thesis, Massachusetts Institute of Technology.
5. Balakrishna, R., H. N. Koutsopoulos and M. Ben-Akiva (2005) "Calibration and Validation of a Dynamic Traffic Assignment System." Accepted for presentation and publication at the 16th International Symposium on Transportation and Traffic Theory, 18-21 July.
6. Ben-Akiva, M., M. Bierlaire, D. Burton, H. N. Koutsopoulos and R. Mishalani (2001) "Network State Estimation and Prediction for Real-Time Transportation Management Applications." *Networks and Spatial Economics*, Vol. 1(3/4), pp. 293-318.
7. Bianco, L., G. Confessore and P. Reverberi (1997) "Optimal Location of Traffic Counting Points for Transport Network Control." IFAC Transportation Systems, Chania, Greece.
8. Brogan, W. L. (1991) "Modern Control Theory." Prentice Hall, 3rd edition.
9. Cascetta, E., D. Inaudi and G. Marquis (1993) "Dynamic Estimation of Origin-Destination Matrices using Traffic Counts." *Transportation Science*, Vol. 27(4), pp. 363-373.
10. Gentili, M. and P. Mirchandani (2004) "Locating Image Sensors on Traffic Networks." Presented at the fifth triennial symposium in transportation analysis, TRISTAN V.
11. Huynh, N., Y. Chiu and H. S. Mahmassani (2002) "Finding Near Optimal Locations for Variable Message Signs for Real-Time Network Traffic Management." Presented at the 82nd annual meeting of the Transportation Research Board.
12. Jha, M., G. Gopalan, A. Garms, B. P. Mahanti, T. Toledo and M. Ben-Akiva (2004) "Development and Calibration of a Large-Scale Microscopic Traffic Simulation Model." *Transportation Research Record*, No. 1876, pp. 121-131.
13. Kunde, K. (2002) "Calibration of Mesoscopic Traffic Simulation Models for Dynamic Traffic Assignment." Masters thesis, Massachusetts Institute of Technology.
14. Yang, H., Y. Iida and T. Sasaki (1991) "An Analysis of the Reliability of an Origin-Destination Trip Matrix Estimated from Traffic Counts." *Transportation Research* B, Vol. 25, pp. 351-363.
15. Yang, H. and J. Zhou (1998) "Optimal Traffic Counting Locations for Origin-Destination Matrix Estimation." *Transportation Research* B, Vol. 33(2), pp. 109-126.
16. Yang, Q. and H. N. Koutsopoulos (1996) "A Microscopic Traffic Simulator for Evaluation of Dynamic Traffic Management Systems." *Transportation Research* C, Vol. 4, No. 3, pp. 113-129.
17. Yang, Q., H. N. Koutsopoulos and M. Ben-Akiva (2000) "A Simulation Model for Evaluating Dynamic traffic Management Systems." *Transportation Research Record*, No. 1710, pp. 122-130.

# DYNAMIC ORIGIN-DESTINATION MATRIX ESTIMATION FROM LINK COUNTS: AN APPROACH COHERENT WITH DYNAMIC TRAFFIC ASSIGNMENT

*Thomas Durlin: LICIT- ENTPE / INRETS, France, thomas.durlin@entpe.fr*
*Vincent Henn: LICIT- ENTPE / INRETS, France, vincent.henn@entpe.fr*

## Abstract
A wide range of models has been developed to solve the problem of estimating a dynamic OD matrix from link counts. We propose here a simple approach based on both traffic flow considerations and on assignment equilibrium assumptions. Our two step process consists in: 1. the propagation and calculation of flows on links and through intersection in the network, and 2. the double tracking of composition changes from origins to destinations and from destinations to origins to calculate the OD matrix coefficients. We focus on the study of elementary networks that allows an illustration of some assignment phenomenon. It could also be useful in an approach based on the decomposition of complex networks into elementary ones.

## 1. Introduction

In an operational context, the objective of dynamic traffic assignment (DTA) models is to represent traffic evolutions on a road network when traffic conditions change. More precisely they seek to describe the assignment of origin-destination (OD) flows on the different paths connecting every OD couple corresponding to an equilibrium state. Considering the Wardrop definition (Wardrop, 1952) this equilibrium corresponds to the state such that no user can decrease its travel time by changing its path.
To account for the effects of the demand and supply variations on assignment, DTA models are classically composed of two sub-models:

- a Dynamic Network Loading (DNL) model, providing the links travel times for a given case of given case of demand and supply variation ;
- an assignment model that determines the assignment coefficients from the calculated travel times, i.e. the flows assignment between all possible paths for each origin-destination pair.

An iterative process is classically used to calculate the traffic assignment equilibrium: the DNL model calculates the travel times for a given assignment case. The assignment model then uses those travel times to determine a new assignment. While the equilibrium is not reached, the process “travel times calculation / assignment calculation” goes on.
The quality of the results of the DTA model depends on the ability of the DNL model and of the assignment model to describe properly the traffic evolution and the users path choice behaviour respectively, i.e. on the intrinsic quality of the two sub-models. Another criterion for quality is the coherence between the operational goal of the DTA model, the model itself and the used data, in particular the dynamic OD matrix.

Link counts are the dynamic data sources the most easily available and the less costly, even if the use of other types of data is possible, such as number plates (Watling, Maher, 1992) or parking survey (Bierlaire, Toint, 1995). A wide range of models has been developed to solve the complex problem of estimating a dynamic OD matrix from link counts. This problem is twofold whether you work in the overdetermined case or in the underdetermined one, i.e. whether you dispose or not of a sufficient set of observation equations compared to the number of variables (the dynamic OD path flows).
For overdetermined problems some models use the times series of traffic counts to track the variations in the OD matrix (see for example (Cremer, Keller, 1981)). Those simple but efficient models take into account propagation effects, such as travel time or platoon dispersion (Bell, 1991b). They do not need prior information (a prior OD matrix) but are relevant only for overdetermined cases.
Two classes of models respond to the underdetermined case. They use assumption on the structure of the observed OD matrix to increase the number of equations. The first one, generally derived from static OD matrix estimation models, are based on the minimization of the distance between a prior OD matrix and the estimated matrix. See for instance (Van Zuylen, Willumsen, 1980), (Bell, 1991a). They suffer from the

complex determination of the required prior matrix and from the artificial aspect of the minimization process that can lead to unphysical results.
Another class of models are DTA-based, i.e. they use relations between paths derived from DTA concepts to increase the number of observation equations. The bilevel models directly search an OD matrix such that its assignment on the network gives flows that minimize the distance with the observed flows ((Yang et al, 1992), (Yang, 1995), (Bell et al, 1997)). This approach is more coherent with the rest of the global process of DTA: it ensures that the assignment of the estimated OD matrix with the DTA model gives the real observed situation, at least on the links where measurements are available.
(Wu, Chang, 1996) develops the idea of dynamic screenline flows. Those screenlines are built from the link counts so that they cut the network into smaller and isolated parts. Flows between a given sub-network and the others are then known. The restraint OD matrix inside this sub-network can be calculated independently from the global one for the rest of the network.

We propose here a hybrid model between the time series and the assignment based models. We do not intend to create a general model for dynamic OD matrix estimation. We rather focus on theoretical considerations on some elementary networks that highlight some specific behaviours. The study of such elementary networks can be of some interest in a dynamic screenline flows approach.
Our approach is a two steps process. First we calculate dynamic flows on all links on the network thanks to the given link counts. As a consequence this means that now we know flows for each origin and destination and the splitting rates in the intersections. The second step consists in the determination of the OD flows: every change in flows in an origin (respectively in a destination) can be tracked to the destinations (respectively to the exit). Our assumptions on the network configuration are strong. It has no cycle to allow this second step, and the problem must not be "too underdetermined": considerations on assignment equilibrium must give enough new equations to reach the number of variables.
Part 2 explains the general dynamic OD matrix process whose elements are detailed in the rest of the paper. Parts 3 and 4 respectively deal with the propagation of flows on links and through intersections. Part 5 introduces travel time notions that are used in part 6 to present the assignment equilibrium assumptions. Two illustrations are presented in part 7.

## 2. Dynamic OD matrix estimation process

The proposed method for dynamic OD matrix estimation from link counts is a two step process:
- flows estimation on all links,
- downwards and forwards propagation of composition variations.

The first step is the calculation of flows in the network. Parts 3 and 4 detail how we estimate flows on all links from a partial set of link counts. We use a traffic flow model to propagate information on a link downwards and upwards a count location (part 3). Flows in intersections are estimated according to the flow conservation principle (part 4). For underdetermined cases (i.e. cases where some flows are still unknown) we use assignment equilibrium assumptions that give some new relations between flows (or more precisely between path flows). Those assumptions are detailed in part 6. Once the unknown flows in an intersection are estimated, they can be propagated upwards or downwards to the next intersection to increase the number of observation equations in it. We assume that this simple process is sufficient to know flows on each link entry and exit for simple networks.

The second step consists in the determination of the OD flows: every change in OD flows in an origin has consequences on flows of origin or destinations links. This change can be:
- a simple change in composition of the origin flow imperceptible with the simple information on the flow value on the entry link, but it can be seen in a change in flows on the destinations links;
- a simple change in the flow value on the origin link (but not in its composition) that can be seen both on the entry and on the exits;
- a complex change in both the flow value and its composition that can be seen on the entry and/or on the exits.

We first track the change in the origins flows to destination to obtain an origin decomposition of link flows (downwards tracking). The origin-destination decomposition is done with a second track of flow variations from the destinations to the origins (upwards tracking).
We use the cumulative flow method proposed by (Newell, 1992) to track the composition variations from the entry of a link to its exit (or from its exit to its entry).
The downwards and upwards tracking of composition variations through intersections (detailed in part 5) may need more information than the simple flow conservation principle. Therefore we may also use information on path flows given from the previous assignment equilibrium assumptions.
Once each origin and destination flow variation has been taken into account, the dynamic origin flow compositions give the dynamic OD matrix coefficients.

# 3. Flows on links

The study of directional flows in the intersections is the core of time series method or assignment based models because it gives information on turning rates. These rates directly results from the path choices and the path flows, that depends themselves on the OD matrix. However a link count provides dynamic information on the traffic state on the link, basically flows and concentrations (or more precisely occupation rates). This information is temporally and spatially limited. Therefore it must be propagated downwards and upwards on the whole link to intersections.
Like assignment based models we choose to describe this propagation with the same DNL model as we use in the DTA model (Durlin, Henn, 2006). This DNL model is a first order macroscopic model: the Lighthill-Whitham-Richards model. We will first describe the model and its numerical scheme for a downwards propagation, and then we'll see how it can be adapt to upstream propagation.

## 3.1. The LWR model

We choose the Lighthill-Whitham-Richards (LWR) model ((Lighthill, Whitham, 1955) and (Richards, 1956)) as our DNL model because it is a good compromise between our two opposite needs of calculation simplicity and accuracy in traffic phenomena description. It represents traffic as a homogenous and continuous flow described with three variables: the flow $Q(x,t)$, the density $K(x,t)$ and the flow speed $V(x,t)$. The LWR model is based on three equations (the conservation equation, the flow definition and the equilibrium fundamental relation). The LWR model needs few parameters for its fundamental relation. In our application, we restraint the traffic to free flow states and we use a parabolic expression. Therefore we only need two parameters: $Kc$ the critical concentration (the maximum concentration in free flow state) and $Qx$ the maximum flow.

The LWR model is generally numerically solved by using the Godunov scheme (Daganzo, 1995) and (Lebacque, 1996), based on a spatial and time discretization. We prefer here another approach based on an event discretization: the Wave Tracking (WT) method. Events are treated with a time increasing approach that avoids the use of spatial and temporal steps.
The WT method has been applied to the LWR model for a single road (Henn, 2005b) with some boundary conditions (Henn, 2005a). Some specific developments for a DTA use have also been made to take into account the intersection delays and flow restriction from an average point of view (Durlin, Henn, 2005). Note that the WT method gives piecewise constant flows.

## 3.2. Upwards propagation

Flow models are classically used for a downwards propagation[1]: for given initial conditions (initial concentrations on the link) and given boundaries conditions (concentration variations at the entry of the link due to the upstream demand variation) they calculate the evolution of the traffic state for increasing event

[1] The expression « time increasing propagation » would be more adequate than « downwards propagation » because congestion shock waves can propagate upwards but always with a propagation towards increasing times. However the second expression is more explicit: we want to know what is happening upwards.

times (see figure 2.a). On the contrary an upstream propagation seems unusual because it means working with a decreasing event time (see figure 2.b).
(Newell, 1993) shows that the traffic state in a point ($x$,$t$) depends both on the demand $\Delta$ in the point ($x$-$dx$ , $t$-$dt$) and on the supply $\Sigma$ in the point ($x$+$dx$ , $t$-$dt$): the demand propagates downstream and the supply upstream. In free flow state (when the supply exceeds the demand) the flow in $x$ is then completely determined by the demand. Therefore under the assumption of a free flow state flows can be propagated upwards (and towards negative times).
To solve this specific upstream propagation we use the same WT method as for the classical downstream propagation. However the WT scheme is a time increasing process. We then adopt the variables transformation $x \rightarrow -x$ and $t \rightarrow -t$. We also inverse the role of $t$ and $x$ in our scheme so that even congestion shock waves propagates with ($-x$) increase: all event are ($-x$) increasing (instead of time increasing with the classical downwards propagation). We work now in ($-t$,$-x$) – diagram, where shockwave speeds are the inverse of the shockwave speeds for the downwards propagation (see figure 2.c). The resulting calculation of shock wave propagation and intersection follows the same process as in the previous case.

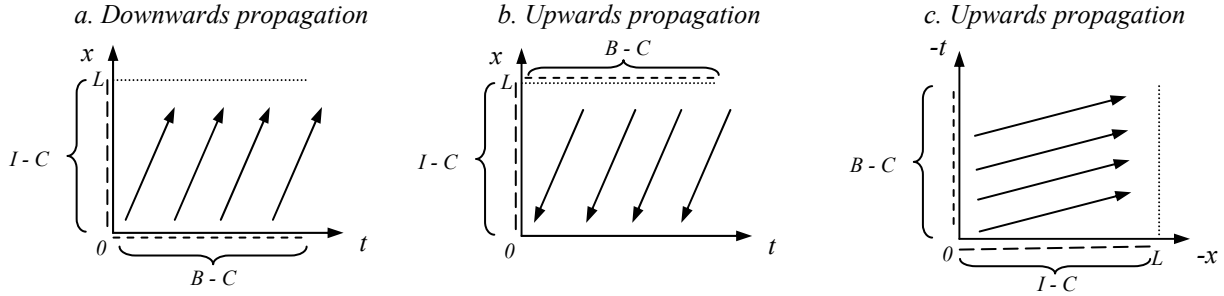



**Figure 2. Illustration of downwards and upwards propagation: principles and variables transformations (B - C and I - C: Boundary and Initial Conditions)**

This upwards propagation resolution scheme can also handle congestion shock waves. However the propagation of those waves is theoretically uncertain. A congestion shock wave can be detected with a link count and modelized in the vicinity of the link count abscissa. However the evolution of the shock wave upstream this abscissa can not be precisely described because it depends on the upstream demand variations that are inaccessible with a unique count. In congestion the link count can only estimate the supply and not the demand. Therefore we limit our study in this paper to free flow cases.
Note that the same phenomenon of uncertainty happens for free flow cases but at a lower scale so that it can be neglected. The traffic state downstream the count depends both on the upstream demand (that is known) and the downstream supply. Considering cases where the network parameters are set to (variable but) known values the supply is perfectly known and can be taken into account. For example a queue spill back upstream an intersection can be theoretically modelized even if no count is available in the vicinity of the queue.

Thanks to the downwards and upwards propagations we are now able to propagate the temporally and spatially limited count information anywhere on the link, in particular to its entry and to its exit. From now on we will no more mention this local aspect of the link count information and assume that having a count on a link means knowing flow both at its entry and its exit.

# 4. Flow propagation on the network

We have seen in part 3 how we propagate information on traffic (flows and concentrations) on a link upstream and downstream a count. We explain here how to handle simple nodes (diverge and merge) for the flow propagation step (first step of the OD matrix estimation process) and for the composition tracking (second step of the process).

## 4.1. Flow propagation

We consider here only simple nodes: one entry - n exits (diverge) or n entries – one exit (merge). These intersection configurations ensure us that every change in the OD composition at the entries or at the exits directly modifies the splitting rates and is then observable if all flows are known.

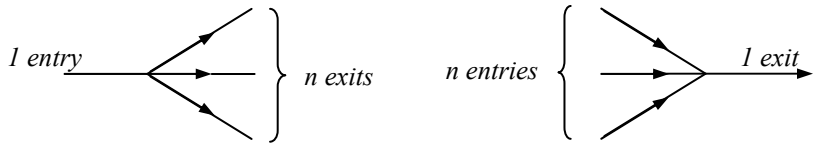



**Figure 3. Simple intersection configurations: diverge and merge**

The problem of flow determination is linear. The number of flow variables for each time interval[2] is then n+1 (the flows on each entry and exit). The conservation of flows in the intersection gives one equation:

$$Q_{entry} = \sum_{j=1}^{n} Q_{exit\,j} \text{ (diverge)} \qquad \text{or} \qquad \sum_{i=1}^{n} Q_{entry\,i} = Q_{exit} \text{ (merge)}$$

Different cases are possible depending on the number of known flows which gives the number of observation equations:

- Every flow at each entry and each exit is known (n+1 observation equations). The system is then overdeterminated. If counts and flow propagation are not perfect, there might be no solution to the problem. A data correction process could be used to avoid violation of the flow conservation law.
- Only one flow is unknown. It could be easily calculated (n+1 variables, n observation equations + 1 conservation equation).
- More than one flow is unknown. Other equations are needed to solve the problem. Therefore we try to use some assumption based on assignment equilibrium considerations. For example, flows at two exits can be interdependent at the equilibrium if the two exits belong to two paths connecting the same OD (flows are such that travel times on the two paths are equals). See part 6 for more details.

Once the unknown flows in an intersection are estimated, they can be propagated upwards or downwards to the next intersection to increase the number of observation equations in it. We assume that this simple process is sufficient to know flows on each link entry and exit for simple networks.

### 4.2. Composition tracking

The composition tracking is the second step of the process. We first track changes in the origins flows to destinations (downwards tracking) and then we track flow variations from the destinations to the origin (upwards tracking).
The previous flow conservation equation for diverges can be extended considering OD flows:

$$Q_{entry}^{OD_k} = \sum_{j=1}^{n} Q_{exit_j}^{OD_k} \qquad \text{for all OD flows } OD_k \text{ on the entry and exit links}$$

Any change in one exit flow composition is obviously attributable to a change in the entry composition and this new composition can be calculated. Therefore the upwards tracking is relevant in the case of diverge intersections. However a change in the entry composition does not induce calculable exit flows: if one exit has a fix constraint flow, such a composition modification induces changes only on non constraint exits. The downwards tracking in diverge needs more information. That is why we must also use information on path flows given from the assignment assumptions.
Composition tracking in merge follows the same principle but behaves symmetrically: the downwards tracking is efficient and the upwards tracking needs assignment assumptions.

## 5. Travel times functions

In the next part we will explain assumptions on assignment equilibrium and how we use them to estimate some path flows. But before this, we need to introduce some notions on travel time functions that are needed to present those assignment considerations.

[2] The WT scheme ensures us that flows are piecewise constant. Time can then be discretized into intervals during when all flows are constant and whose duration are variable. Therefore we prefer use « time interval » instead of « time step».

### 5.1. Link travel time function

Considering the LWR model in free flow state with a parabolic fundamental relation the vehicle speed $V(k)$ is:

$$V(k) = \frac{Q(k)}{k} = \frac{Q_x}{K_c^{\,2}}.(2.K_c - k) \qquad \text{with } k \in [0, K_c]$$

The travel time function $T(k)$ for a link whose length is $L$ is then:

$$T(k) = \frac{2.K_c^{\,2}.L}{(2.K_c - k).Q_x} \qquad \text{with } k \in [0, K_c]$$

Using $q$ instead of $k$ as the variable, $T$ becomes:

$$T(q) = \frac{K_c.L}{q}\left(1 - \sqrt{1 - \frac{q}{Q_x}}\right) \qquad \text{with } q \in [0, Q_x]$$

Those expressions of the link travel time are true only in the stationary case: the vehicle encounters the same traffic condition on the whole link. In the general dynamic case concentration and flow are variable functions of time and space $k(x,t)$ and $q(x,t)$. However as explain in the part 6 we will use the stationary case expression from now on.

### 5.2. Link travel time function with constraint flows

In a network some OD are connected with a single path. This is particularly true with small networks. Therefore these unassignable OD flows do not depend on travel times. We call then constraint flows. If such flows are known to use a link the travel time function of the link can be modified according to this information. The travel time function $T'$ for the unconstraint flow $q'$ (i.e. the flow that has chosen to use the link) is obtained with a translation $T'(q') = T(q'+q_c)$:

$$T'(q') = \frac{K_c.L}{q - q_c}\left(1 - \sqrt{1 - \frac{q - q_c}{Q_x}}\right) \qquad \text{with } q' \in [0, Q_x - q_c]$$

with $q_c$ the total constraint flow. The maximum unconstraint flow is then reduced to $Q_x' = Q_x - q_c$. Note that $q_c$ can be composed of different OD flows, but only its total value and not its precise composition is important. The constraint flow is not necessarily constant but can be dynamic: $q_c(t)$.

### 5.3. Sub-path travel time function

We call sub-path a chain of successive links that have the same assignment role. This means that the unconstraint flow entering the first link must travel on the other link of the sub-path.
The figure 4 shows such a sub-path $A - B$ composed of three links. All the vehicles of the flow $q'$ entering the sub-path in $A$ go on until $B$. In this example there is a known constraint flow $q_c$ from $C$ to $D$ using the link 2. We could also have a slightly more complex case with a constraint flow from $A$ to $D$ or from $C$ to $B$, but the method will be the same.

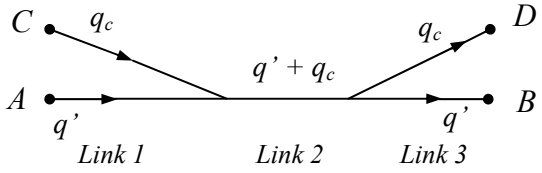



**Figure 4. A small sub-path with a constraint flow**

We search an expression for the travel time function of such a sub-path for the stationary case. Travel times are evidently cumulative, so the travel time function for the sub-path will be the sum of each travel time function. We use here the corrected functions integrating the constraint flows. Each link $i$ has its own travel time function $T'^{\,i}$ depending on its own parameters ($L^i$, $Q_x^{\,i}$, $K_c^{\,i}$ and $q_c^{\,i}$). The maximum unconstraint flow on the sub-path is then the minimum of each $Q_x'^{\,i} = Q_x^{\,i} - q_c^{\,i}$. The sub-path travel time function is then:

$$T_{A-B}(q') = \sum_i T'^{\,i}(q') \quad \text{with } q' \in \left[0, \min_i (Q_x^{\,i} - q_c^{\,i})\right]$$

For the sake of simplicity the three links have here the same parameters ($L$ = *1000 m*, $Q_x$ = *2 veh/s*, $K_c$ = *0.3 veh/m*, $q_c$=*1 veh/s*). As a consequence they all have the same basic travel time function but the link 2 have a specific travel time function integrating the constraint flow $q_c$. Figure 5 shows those different functions.

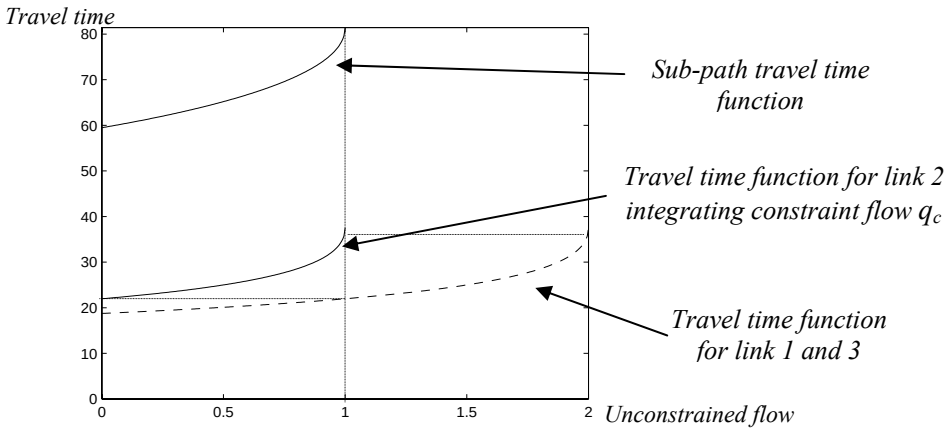



**Figure 5. Link travel time functions without and with constraint flow and sub-path travel time function**

We are now able to estimate path travel time functions depending on the parameters of the links composing this path and possibly on some known constraint flows on these links. Note that those functions are bijections so that we could express their inverse functions that give flows depending on travel times:

$$q'(t) = T'^{-1}(t)$$

# 6. Assignment equilibrium considerations

If link counts are not numerous enough or/and unfortunately badly located, some underdetermination may appear in the flow conservation problem in intersections. Therefore we need to find other relations between flows. The OD matrix estimation problem is usually solved to obtain the OD information for a DTA model application. We logically choose to use the classical DTA assumptions to obtain our new (and consistent) relations.

We assume that the network follows an assignment equilibrium principle. Considering the Wardrop definition (Wardrop, 1952) the equilibrium state corresponds to the state such that no user can decrease its travel time by changing its path. Considering paths connecting a same OD couple, the travel times of all used paths (paths with a nonzero path flow) are then equal to the same value $T_{path}$ and the travel times of unused paths (paths with a zero path flow) are greater that $T_{path}$. Link flows are then such that the path travel times satify this constraint.
Consider for example the network of figure 6 with two paths between A and B but only one available link count on link 2. According to part 5, the travel time on the first path can be estimated and the inverse travel time function of the second path gives the flow on link 3 corresponding to the assignment equilibrium state. As we have seen in part 5, the two paths can be composed of successive links with their own parameters.

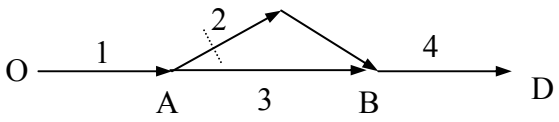



**Figure 6. The one OD-two path flows network**

Our second assumption concerns the dynamics properties of the assignment equilibrium. For sake of simplicity we consider that the dynamic equilibrium is in fact a succession of static equilibriums: at each instant users choose their path as if the demand was constant. A simple change in the flow entering the intersection A induces a simple change in the assignment between the two paths. This means that we neglect the transition phase between the two stationary states.
This assumption holds if the time scale of OD variations is greater than the travel times on the network because we consider that links stay on free flow regime. This is expected to be true with small elementary

networks. Bigger networks with great travel times are to be cut in elementary networks so that this is still true at the smaller network level.

Note that only one link count is needed to know all dynamic link flows. If the count is located on link 2 or 3 you can easily find the equilibrium flow on link 3 and so calculate flows on links 1 and 4. If it is located on link 1 or 4 you can use the same principle to calculate the assignment between the two paths which will give you all missing link flows.

# 7. Illustrations

Two cases are treated here to illustrate our approach. We first analyze the problem: number of variables, equations. We also use a measure *TDS* of quality of the OD matrix estimation problem based on the total demand scale proposed by (Bierlaire, 2002). It evaluates the underdetermination of the problem and consequently it gives an estimation of the dependence of the supplementary assumptions that are needed (here our assignment equilibrium assumptions). It consists in a calculation of the width of the interval [$Q_{min}$ , $Q_{max}$] of all admissible values of the total demand on the network, with $Q_{min}$ and $Q_{max}$ the minimum and maximum values. The characteristic values are:

- *TDS = 0*: the total demand is known and the underdetermination only concerns the repartition of this demand between OD flows,
- *TDS >0*: the underdetermination concerns both the total demand and its repartition between OD flows,
- *TDS = infinity*: an OD flow can take any value because this is not captured by any count this flow[3] (link counts locations do not allow to create a screen line that isolates this origin or/and this destination and can capture the total OD flow[4])

After this problem analysis we briefly explain the different steps of the process.

### 7.1. Case 1: the two OD-three path flows network

Figure 7 shows the network configuration. The two dot lines represent the link counts on links 2 and 3.

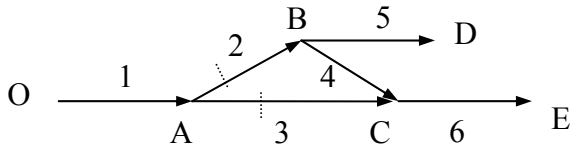



**Figure 7. The two OD-three path flows network**

There are three paths: I: 1-2-5 (OD), II: 1-2-4-6 (OE) and III: 1-3-6 (OE). The flow towards D is constrained to use the path I.

However this network presents some underdetermination:

- Intersection A: 3 variables – 2 observation equations + 1 conservation equation: no underdetermination,
- Intersection B and C: 3 variables – 1 observation equation + 1 conservation equation: one unconstraint variable.

The two counts on links 2 and 3 constitute a screen line that captures all flows from the origin. The total demand scale is then: *TDS = 0*. The underdetermination only concerns the assignment of flows towards E between path II and path III. Using some assignment assumptions to treat this underdetermination is then relevant.

We know the dynamic travel times from A to C on the link 3 $T_3(q_3)$ and from A to B on the link 2 $T_2(q_2)$. The assignment equilibrium assumption gives us the flow on link 4 $q_4$:

$$q_4 = T_4^{-1}(T_3(q_3) - T_2(q_2))$$

[3] In fact as we work with capacity constraint links all flows are obviously bounded but we still use the *infinity* value because it is easily identifiable as a non captured OD flow configuration.

[4] See (Wu, Chang, 1996) for further details on screen lines.

Tracking upwards this flow $q_4$ gives us the composition variations in the intersection A and consequently the dynamic OD matrix.

If link counts are now on links 3 and 4, our estimation process is still efficient. However, the total demand scale is now: *TDS = infinity*. The flow towards D is not captured but the assignment of flows towards E is known (path flows II and III are measured). As the underdetermination consists now in a total demand evaluation and not an assignment problem, our assignment assumptions seem less relevant even if they are sufficient to estimate the OD matrix.
Note that wherever the two link counts are located on the network, our process is efficient. This gives some flexibility to the use of this elementary network inside any bigger network.

### 7.2. Case 2: The two OD-four path flows network

This network (see figure 8) presents a more complex configuration:

- Four paths: I: 1-2-5-8 (OF), II: 1-2-5-7-9 (OG), II: 1-2-4-6-9 (OG) and III: 1-3-6-9 (OG),
- Intersection A: 3 variables – 2 observation equations + 1 conservation equation: no underdetermination,
- Intersection B and C: 3 variables – 1 observation equation + 1 conservation equation: one unconstraint variable,
- Intersection B and C: 3 variables – 1 observation equation: two unconstraint variables.
- The total demand scale is still: *TDS = 0*: the problem is an assignment problem.

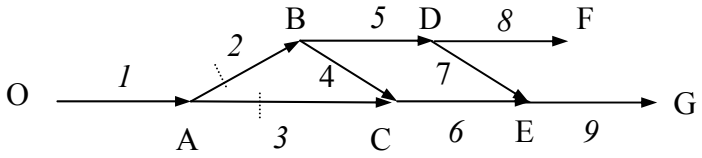



**Figure 8. The two OD-four path flows network**

We decompose this complex network into to two simple networks (see figure 9). The first is the one we have studied in the previous subpart. We can use the same process to solve the problem. We now know flows on links 1 to 6, as well as the OD matrix for the two OD couples OD and OE.
The second network is also the same, but the previous link 3 is now a subpath 4+6, with a known constraint flow on the link 6. As we know flows on links 4 and 6 and consequently travel times from B to E the flow on link 7 is easily calculated. The OD matrix for the two OD AF and AG couples is then calculated.
The global OD matrix is calculated by mixing the two sub OD matrixes. Travel times corrections are made to take into account the flow propagation.

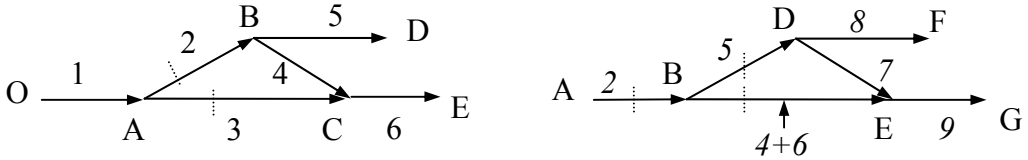



**Figure 9. Decomposition into two elementary networks**

## 8. Comments and conclusion

The paper presents a method for the dynamic OD matrix estimation from link counts. As dynamic OD matrices are needed as data for DTA models, their estimation process must be relevant and consistent with the DTA assumptions.
The approach is then based on considerations on both traffic flow model and assignment equilibrium. We first propagate flows on links and through intersection in the whole network. We use some assignment equilibrium considerations to cope with the problem of underdetermination that may appear in the intersection. We then track the OD flows variations thanks to the flow variations on the origin and destination links. Two applications on small networks illustrate the process and highlight some interesting facts, such as the relevancy of assumptions depending on the network configuration.
Our method is hardly directly applicable to general and complex networks because it makes strong assumptions on the network configuration and on the level of underdetermination: assignment equilibrium

considerations must be sufficient to avoid underdetermination in the flow and composition propagation through intersections. Nevertheless we propose a decomposition approach: complex networks could be cut to some extent into smaller elementary networks that can be treated with our method. The results on the restricted OD matrices related to those subnetworks are then mixed to obtain a global OD matrix. A typology of interesting and relevant elementary networks is still to be made.

## References


Bell M.G.H., The estimation of origin-destination matrices by constrained generalised least squares, *Transportation Research – Part B,* volume 23, pages 13-22, 1991.

Bell M.G.H., The real time estimation of origin-destination flows in the presence of platoon dispersion, *Transportation Research – Part B,* volume 23, pages 115-125, 1991.

Bell M.G.H., Shield C.M., Busch F., Kruse G., A stochastic user equilibrium path flow estimator, *Transportation Research – Part C,* volume 5, pages 197-210, 1997.

Bierlaire M., The total demand scale: a new measure of quality for static and dynamic origin-destination trip tables, *Transportation Research – Part B,* volume 36, pages 837-850, 2002.

Bierlaire M., Toint P.L., MEUSE: an origin-destination matrix estimator that exploits structure, *Transportation Research – Part B,* volume 29, pages 47-60, 1995.

Cremer M., Keller H., Dynamic identification of flows from traffic counts at complex intersections, *Proc. of the 8th International Symposium of Transportation and Traffic Theory*, University of Toronto, Toronto, Canada, pp 121-142, 1981.

Daganzo C. F., A finite difference approximation of the kinematic wave model of traffic flow, *Transportation Research – Part B,* volume 29, n°4, pages 261-276, 1995.

Durlin T., Henn V., A delayed flow intersection model for dynamic traffic assignment. In: Proceedings of the Joint Conference-10th EWGT Meeting and 16th Mini-Euro Conference, Poznan, Poland, 2005.

Durlin T., Henn V., Tracking waves in the LWR model for a progressive dynamic traffic assignment calculation, *Submitted for presentation to the 11th EWGT Conference, Bari, Italia, September 2006.*

Henn V., Tracking waves through boundaries: boundary conditions in the wave tracking resolution of the LWR, *Proceeding of the 4th IMA Conference on Mathematics in Transport studies*, London, United Kingdom, September 2005.

Henn V., A wave-based resolution scheme for the hydrodynamic LWR traffic flow model. In S.P. Hoogendoorn, S. Luding, P.H.L. Bovy, M. Schreckenberg, and D.E. Wolf, editors, *Traffic and granular flows'03*, pages 105-124. Springer, October 2005.

Lebacque J.P., The Godunov scheme and what it means for first order traffic models. In J.B. Lesort, editor, *Proceedings of the 13th International Symposium on the Transportation and Traffic Theory,* Lyon, France, pages 647-678, 1996. Pergamon, Oxford.

Lighthill M.L., Whitham G.B., On kinematic waves II - A theory of traffic flow in long crowded roads. *Proceedings of the Royal Society,* volume A, n°229, pages 317-345, 1955.

Newell G.F., A simplified theory of kinematic waves in highway traffic, Part 1: General theory, *Transportation Research – Part B*, volume 27B, n°4, pages 281-287, 1993.

Richards P.I., Shockwaves on the highway, *Operation Research*, volume 4, pages 42-51, 1956.

Van Zuylen H.J., Willumsen L.G., The most likely trip matrix estimated from traffic counts, *Transportation Research - Part B*, volume 14, pages 281-293, 1980.

Wardrop J.G., Some theoretical aspects of road traffic research, *Proceedings of the Institute of Civil Engineering,* Part 2, 1, pages 325-378, 1952.

Watling D.P., Maher M.J., A statistical procedure for estimating a mean origin-destination matrix from a partial registration plate survey, *Transportation Research – Part B*, volume 26, n°3, pages 171-193, 1992.

Wu J., Chang G.-L., Estimation of time-varying origin-destination distributions with dynamic screenline flows, *Transportation Research – Part B*, volume 30, n°4, pages 277-290, 1996.

Yang H., Heuristic algorithms for the bilevel origin destination matrices from link traffic counts in congested networks, *Transportation Research - Part B 29*, pages 231-242, 1995.

Yang H., Sasaki T., Iida Y., Asakura Y., Estimation of origin-destination matrices from link traffic counts on congested networks, *Transportation Research - Part B*, volume 26, n°6, pages 417-434, 1992.

# Determination of Behaviour-Consistent Information-based Control Strategies using Fuzzy Modeling

*Alexander Paz: Purdue University, USA apaz@purdue.edu*
*Srinivas Peeta: Purdue University ,USA peeta@purdue.edu*

## Abstract

This study proposes a fuzzy control based methodology to determine behaviour-consistent information-based control strategies based on the controller's estimation of driver behaviour. It is the core of the broader problem where the objective is to effectively control/influence the performance of a vehicular traffic system by providing information to the drivers. Experiments are performed to evaluate the effectiveness of the proposed methodology. The experiment results suggest the importance of using a behaviour-consistent approach to determine the information-based control strategies; that is, the effects of the level of responsiveness of drivers to information provision may require more meaningful strategies than those provided under the traditional DTA models to have a reliable estimation/control of system performance.

## 1 Introduction

The primary functional capabilities of dynamic traffic assignment (DTA) for advanced traveler information systems (ATIS) operations are to predict network states over time and to provide routing information to drivers consistent with some control objectives. Existing DTA models focus primarily on ensuring an adequate representation of the traffic flow dynamics while determining the vehicular routes that satisfy some system-level and/or driver class objectives (for example, Papageorgiou, 1990; Ben-Akiva et al., 1991; Peeta and Mahmassani, 1995). However, traditional DTA models are behaviourally restrictive, and limited in their ability to model driver response to information. This motivates the development of a DTA paradigm integrating traffic control/information provision decisions and realistic driver behaviour representation. A comprehensive exposition of the state-of-the-art on DTA is provided by Peeta and Ziliaskopoulos (2001).

Driver behaviour is a fundamental factor and a key source of complexity in predicting traffic network states unfolding over time. However, DTA models are based on a rigid framework, they either assume homogeneous drivers or pre-specify the behaviour of a few driver behaviour classes assumed in conjunction with the information characteristics. Few DTA models consider heterogeneity among drivers. Even these models assume that driver behaviour classes can be pre-specified. Further, they assume that driver behaviour class fractions in the ambient traffic stream are known deterministically *a priori*. This rigidity precludes the seamless incorporation of driver behaviour characteristics and raises issues vis-à-vis the realistic modeling of driver response to information provision.

Incorrect prediction of traffic system states based on the aforementioned assumptions can negatively impact the validity and effectiveness of the information-based control strategies and potentially deteriorate system performance. In reality, the natural mechanism for driver route choice, even under information provision, is based on the driver's innate behavioural tendencies, past experience, situational factors (such as time-of-day, weather conditions, and trip purpose), and the ambient traffic conditions encountered. This is true irrespective of whether drivers receive personalized, generic, or no information (Peeta and Yu, 2004).

While information provision and content can be used as control variables to influence system performance, they cannot imply perfect compliance by the drivers to the supplied information, as is predominantly done in the DTA arena. From the traffic controller perspective, providing personalized or class-specific information based on a better understanding of driver response tendencies and ambient traffic conditions could generate a more effective control paradigm. The explicit consideration of driver behaviour leads to a new dimension of complexity in predicting traffic states. The proposed research focuses on developing an approach that enables theoretical consistency with the underlying behavioural process by estimating drivers' likely response behaviour to information-based control strategies. In summary, this research focuses on developing information-based control strategies in which the controller factors the likely driver behaviour while determining the information to disseminate to drivers.

## 2 Problem Definition

The behaviour-consistent information-based control strategies problem is defined as follows. A central controller seeks to determine information-based control strategies that are consistent with driver behaviour and address its objectives of enhancing system performance. The approach used by the controller is to influence driver route choice decisions (by providing route guidance information) in such a way that the percentages of drivers taking specific routes are close to the corresponding percentages under a system-wide objective, the system optimal (SO) solution. To achieve this consistency, the controller estimates the driver route choice behaviour using a model, and uses it to determine the appropriate information provision strategies that result in driver decisions which lead to route choice percentages that are close to those under the SO solution. Hence, the broader problem being addressed here is to determine the information-based route guidance strategies that minimize the difference between the controller-desired and actual percentages of drivers taking each route.

This study adopts a perspective that by directing the system, to the extent possible, to a time-dependent SO state, the objective of the controller to enhance system performance can be achieved in a behaviourally more realistic manner that that under the traditional DTA approaches. The validity of this perspective has been successfully tested in Peeta and Paz (2006), where the authors use this approach to control a vehicular traffic network in a within-day context. The results showed the importance of incorporating driver behaviour realism in the determination of the information provision strategies. Significant differences in terms of total system travel time were obtained when the behaviour-consistent approach was compared to the traditional approaches. Figure 1 illustrates the overall logic of the proposed approach for the broader problem in the context of real-world deployment. In this paper, we focus on the key sub-problem of this approach, represented by the non-shaded box in the middle of the flowchart in Figure 1. This sub-problem, addressed by the controller, is *the determination of the percentages of drivers that should be recommended specific routes* so that when drivers make their decisions according to the controller-estimated driver behaviour model, close to SO percentages are obtained.

## 2.1 Definition of Terms

*Desired Routes* (*DK*)*:* Desired routes are routes that the controller would like the drivers to take. These are the time-dependent system optimal routes obtained using current demand and network conditions.

*Preferred Routes* (*PK*)*:* Preferred routes are routes that are likely to be accepted by drivers. These routes represent the set of routes that are consistent with driver behaviour and can be obtained using historical data, travel surveys and/or two-way communication systems.

*Controllable Routes* (*CK*)*:* Controllable routes are defined as routes that belong to both the set of desired routes and the set of preferred routes. A controllable route is one where the controller is likely to be able to influence the percentage of drivers taking the route.

*Behaviour-Consistency Gap*: Difference between the desired percentage of drivers taking a route and the percentage of drivers that must be recommended that route in order to achieve the desired percentage.

## 2.2 Problem Statement

Consider a traffic network represented by a directed graph *G(N,A)* where *N* is the set of nodes and *A* the set of directed arcs. A node can represent a trip origin, a destination and/or a junction of physical links. A network with multiple origins $i \in I$ and destinations $j \in J$ is considered for generality. Given the set of vehicle trip desires expressed as the number of vehicle trips $R_{ij}$ from node *i* to node *j*, $\forall\, i \in I$ and $j \in J$, the system optimal traffic assignments expressed as the system optimal percentage of drivers assigned to each (desired) route *k* in the network $SO_k$, the estimated set of preferred routes for the drivers $PK_{ij}$, $\forall\, i \in I$ and $j \in J$, the average experienced travel times by drivers for the set of preferred routes $T_{rk}$, $\forall\, k \in PK_{ij}$, $r \in R_{ij}$; determine the behaviour-consistent information-based control strategies ($\theta_k, \phi_k$) for the set of controllable routes $CK_{ij}$, $\forall\, i \in I$ and $j \in J$, that minimize the difference between the system optimal traffic assignment percentages

$SO_k$ and the controller-estimated percentage of drivers taking routes $E_k$, $\forall k \in CK_{ij}$. Here, the information strategies are defined as personalized information $\theta_k$ for route $k$ and generic information $\phi_k$ for route $k$.

# 3 Determination of the Behavior-Consistent Information Strategies

A fuzzy control approach is used to determine the behaviour-consistent information-based control strategies and can be conceptually defined as follows. It consists of an input stage, a processing stage, and an output stage. The input stage maps the inputs to the appropriate membership functions according to the control rules. The processing stage invokes each appropriate rule and generates a result for each, then combines the results of all fired (used) rules. Finally, the output stage converts the combined result to a specific control output value. It should be noted here that the fuzzy control approach is employed within a stage as illustrated in Figure 1; for simplicity, the time dimension is ignored in the discussion hereafter in Section 3.

## 3.1 Input Stage

For a given origin-destination (OD) pair *ij*, the control inputs are a function of the difference between the system optimal percentages and the controller-estimated percentages of drivers taking routes. The control inputs for iteration $q$ of the control process are the vectors of error ($\boldsymbol{e_k^q}$) and change in error ($\Delta \boldsymbol{e_k^q}$) defined by:

$$e_k^q = SO_k - E_k^q \quad \text{and} \quad \Delta e_k^q = e_k^q - e_k^{q-1}, \qquad \forall k \in CK_{ij} \tag{1}$$

where $SO_k$ is the system optimal traffic assignment percentage along route $k$, $E_k^q$ is the controller-estimated percentage of drivers choosing route $k$ given the information strategies, and $e_k^{q-1}$ is the error in the previous iteration of the control process. The estimated percentage of drivers choosing routes is calculated using a rule-based route choice model where travel time, route complexity and information are the factors defining the route choices.

## 3.2 Processing Stage

The processing stage can be summarized as follows. For a given OD pair (*ij*), for each route $k$ and iteration $q$ of the control process, the approach uses a fuzzy logic framework (Tsoukalas and Uhrig, 1997). The current inputs, $\boldsymbol{e_k^q}$ and $\Delta \boldsymbol{e_k^q}$ are matched against all the control rules resulting in $M$ values of $\Delta \theta_{kh}^{q^*}$, $h$=1,…,$M$ and $M$ values of $\Delta \boldsymbol{\phi}_{kh}^{q^*}$, $h$=M+1,…,2$M$. The degree at which rule $h$ is activated, denoted by $\gamma_{kh}^q$, depends on the relevant components of $\boldsymbol{e_k^q}$ and $\Delta \boldsymbol{e_k^q}$. An aggregation scheme is used to aggregate the outcomes from all rules. Once the fuzzy aggregate outcomes $\Delta \theta_h^{q^*}$ and $\Delta \boldsymbol{\phi}_h^{q^*}$ are defined, they are defuzzified using the center of gravity method (CGM) described in Tsoukalas and Uhrig (1997). The results, $\Delta \theta_k$ and $\Delta \boldsymbol{\phi}_k$, are then used to update the control variables as follows:

$$\theta_k^q = \theta_k^{q-1} + \Delta \theta_k^q \qquad \forall k \in CK_{ij} \tag{2}$$

and

$$\phi_k^q = \phi\left(min_h\left(\Delta \phi_k^q, \overline{\phi}_h\right)\right) \qquad \forall k \in CK_{ij},\ h=M+1,\ldots,2M \tag{3}$$

where $\theta_k^{q-1}$ is the numeric control values of the previous iteration.

## 3.3 Output Stage

There are two types of control outputs, personalized and generic information. Personalized information corresponds to route recommendations while generic information corresponds to qualitative descriptions of

routes (linguistic variables). While in both cases the information given to the drivers is discrete (e.g. a route is recommended or not), intermediate continuous variables are used to achieve a smooth convergence and reduce jumps in the objective function that can result from large variations in the control strategy.

In this approach $\theta_k$ is directly used as a control variable and represents the fraction of drivers that must be recommended to take route *k*, and $\Delta\theta_k$ represents its fraction of change or adjustment. $\phi_k$ is the linguistic message to display or provide to the drivers; it influences the fractions of drivers taking route *k* or switching to other routes. Given that $\phi_k$ is a linguistic variable, an additional step given by Equation 3 is required. The approach used is to select the fuzzy set (linguistic message) with the closest centroid to $\Delta\phi_k$. The rationality here is that the resulting chosen fuzzy set has the largest degree of membership (or mapping) among all the possible fuzzy sets. In addition, the center of gravity method used to obtain (defuzzify) the outcome can be viewed as a measure of central tendency or weighted average where the weights are the centroids of the fuzzy sets. Therefore, the fuzzy set with the closest centroid to the final outcome can be used as a representative set of the fuzzy sets used to calculate the outcome.

As indicated in the decision process, both types of control strategies are calculated simultaneously and for all routes. This is necessary and adds several dimensions of complexity to the problem. The information-based control strategies must be calculated simultaneously and for all routes under control because they are mutually dependent. Some drivers may have access to both types of information and use them to make their routes choice decisions. Therefore, the effect of one strategy affects the effect of the other on the entire set of drivers choosing routes.

Information on a route directly affects the fraction of drivers taking that route as well as the other routes because the information results in driver switching from some routes to others. For example, recommending route *k* makes it more likely to be chosen by drivers, and some of these drivers are switching from other routes, possibly reducing the fraction of drivers taking those routes. Hence,, it is necessary to simultaneously determine the information-based control strategies for all routes given their interdependencies. These interdependencies illustrated through the experimental results presented in the next section.

# 4 Experiments

Experiments are designed to evaluate the performance of the fuzzy controller and to illustrate the significance of behaviourally-consistent approaches to determine the information-based control strategies. Two sets of experiments are conducted to evaluate the performance of the controller for various driver classes (in terms of information type, and their level of responsiveness to information).

## 4.1 General Experimental Setup

The study network is the Borman expressway corridor in northwest Indiana, USA. It consists of a sixteen-mile section of interstate I-80/94 (called the Borman expressway), I-90 toll freeway, I-65, and the surrounding arterials and streets. The Borman network has 197 nodes and 458 links. While the procedure discussed in Section 3 can be used to determine the information strategies for multiple OD pairs, a single OD pair has been chosen here to illustrate insights for the proposed approach. There are four preferred routes and four desired routes, but only three of them fully overlap. The controller tries to achieve the SO percentages only on the set of controllable routes (the three routes that fully overlap). Thus, desired routes 1, 2 and 3 are defined as the controllable routes in this experimental setup. The SO percentages are 56%, 33% and 11% for routes 1, 2 and 3, respectively.

In the estimated behaviour model used by the controller, two types of drivers are considered based on their level of responsiveness to information provided. The first type of drivers, labelled as “more responsive” to the information strategies, are very likely to behave consistently with the information provided. The second type of drivers, categorized as “less responsive” to information strategies, are less likely to behave consistently with the information provided. These drivers are not as influenced by the information as the first type of drivers; they rely more on their past experience and perceptions to make route choice decisions.

Drivers that are not influenced at all by the information are viewed here as drivers without access to information.

In the figures illustrating the experiment results, the dotted lines are used to show the results for "more responsive" drivers and the continuous lines indicate "less responsive" drivers. The initial points (iteration 0) represent the case when no information is given to the drivers.

## 4.2 Experiments: Personalized Information

### *Specific Objectives and Design*

The objective of these experiments is to evaluate the performance of the controller to determine personalized information under the two classes of driver responsiveness to information. It is assumed that all drivers can receive personalized information, but only a subset of them is provided route recommendations. The remaining drivers do not receive information, and hence their route choice decisions are assumed to be without the influence of information. The control variable is the vector $\theta$ that represents the percentages of drivers that are recommended to take specific routes.

### *Experiment Results and Analysis*

Figures 2 and 3 present the results of these experiments. Figure 2 shows the estimated fraction (percentage) of drivers choosing routes in each iteration of the fuzzy controller under the currently calculated vector of information strategies. It can be noticed that the controller can achieve close to the desired percentages (SO; shown by the three horizontal lines in the figure for the three routes) through its information provision strategies. However, it achieves a faster convergence when all drivers are more responsive to the information strategies. This is because when drivers are more likely to make route choice decisions consistent with the recommendation, the controller can achieve its objective with fewer recommendations.

Figure 3 shows the percentage of drivers that must be recommended to take each route in order to achieve the desired percentages. The control values at convergence indicate that more recommendations are required for two of the three routes under the more responsive behaviour scenario when compared to the less responsive behaviour scenario. This may seem counterintuitive since it is expected that fewer recommendations are necessary to achieve the desired percentages under more responsive behaviour. For example, as can be seen in Figure 2, under this type of behaviour the controller achieves its objective in fewer (about 5) iterations. As can be seen after 5 iterations in Figure 2, the estimated percentages are almost constant, but the recommended percentages in Figure 3 still have a lot of variability. This implies the existence of multiple solutions. For example, in Figure 3, in the neighborhood of 6 iterations, the controller still needs to use less information under the more responsive case compared to the less responsive case.

The results from Figure 3 indicate that there are significant behaviour-consistency gaps in all cases. That is, there are significant differences between the desired percentage of drivers choosing routes and the percentages of drivers that must be recommended to take the routes in order to achieve the desired percentages. Some of the behaviour-consistency gaps are negative, while others are positive. Hence, the experiment results highlight the importance of using a behaviour-consistent approach to determine the information-based control strategies.

## 4.3 Experiments: Generic Information

### *Specific Objectives and Design*

The objective of these experiments is to evaluate the performance of the controller to determine generic information under the two classes of driver responsiveness to information. Here, all drivers are assumed to have access to generic information. Therefore, drivers route choice decisions are always influenced by generic information. The control variable here is the vector $\phi$ that represents linguistic messages describing route conditions.

The messages that can be provided are: “Very Light Traffic”, “Light Traffic”, “Moderate Traffic”, “Heavy Traffic”, and “Very Heavy Traffic”. Numbers from 1 to 5 are use to represent these messages in the results (Figure 5) where 1 corresponds to “Very Light Traffic”, 2 corresponds to “Light Traffic”, 3 corresponds to “Moderate Traffic”, and so on.

### *Experiment Results and Analysis*

Figures 4 and 5 show the experiment results. Figure 4 shows the controller-estimated percentage of drivers choosing routes for each iteration of the fuzzy controller under the currently calculated vector of information strategies. As shown by Figure 4, the controller can achieve close to the desired percentages. Though the controller achieves a faster convergence when all drivers have a more responsive behaviour, it achieves slightly better convergence when all drivers are less responsive. This is because when all drivers have access to the information, the change from one message to another produces a larger effect in the percentage of drivers choosing routes under the more responsive case. Therefore, while the controller achieves convergence faster, it has reduced ability to get closer to the desired percentage due to the large effects of information provision under more responsive behaviour.

Figure 5 shows the results for the vector of control values $\phi$, the set of messages that the controller provides to the drivers. As illustrated, the procedure converges to a stable set of messages. The control values at convergence indicate that in general, stronger messages are require under less responsive behaviour compared to those under the more responsive case. This result is intuitive because stronger messages are needed to compensate the fact that drivers are less influenced by the messages in the less responsive behaviour case.

It is not possible to define behaviour-consistency gaps for linguistic information because each message represents an unknown percentage of drivers choosing routes. This is another reason to use a behaviour-consistent approach to determine the information-based control strategies. Traditional approaches cannot incorporate the linguistic nature of some information strategies.

## 5 Summary and Conclusions

This study proposes a fuzzy control based methodology to determine behaviour-consistent information-based control strategies based on the controller’s estimation of driver behaviour. It has key features that enable realism and more efficient field deployment. The estimation of the driver behaviour to the control strategies enables the determination of behaviour-consistent information provision strategies, and can lead to faster improvements in system performance. The fuzzy control approach is consistent with the proposed controller-estimated driver behaviour model, which is also based on simple if-then rules. The use of a fuzzy control approach simplifies the controller design and reduces the need for tuning or calibration of control parameters. The control framework is independent of the model structure used for the controller-estimated driver behaviour. The knowledge and experience of traffic control personnel can be seamlessly incorporated into the construction and calibration of the control rules. Also, the information strategies involve linguistic or qualitative information (fuzzy variables) which makes the problem amenable to a fuzzy logic-based approach.

The experiment results highlight the complexity of the problem faced by the controller and show the effectiveness and robustness of the fuzzy controller approach to address the multidimensionality of the problem. They also indicate the importance of using a behavior-consistent approach to determine the information-based control strategies. That is, the effects of the level of responsiveness of drivers to information provision may require more meaningful strategies than those provided under the traditional DTA models to have a reliable estimation/control of system performance.

A single within-day and day-to-day framework that uses the proposed methodology is currently under development. The resulting framework will be able to update model parameters based on both on-line and off-line data measurements to further improve system performance. The objective is to calibrate both behavioral and control parameters to better predict driver behavior.

## References

Ben-Akiva, M., DePalma, A., and Kaysi, I. (1991) Dynamic Network Models and Driver Information Systems. *Transportation Research A*, 25(5).

Papageorgiou, M. (1990) Dynamic Modelling, Assignment and Route Guidance in Traffic Networks. *Transportation Research B*, 24**,** 471–495.

Peeta, S. and Mahmassani, H.S. (1995). Multiple User Classes Real-Time Traffic Assignment for On-Line Operations: A Rolling Horizon Solution Framework. *Transportation Research C*, 3(2), 83-98.

Peeta, S. and Ziliaskopoulos, A. (2001) Foundations of Dynamic Traffic Assignment: The Past, the Present and the Future. *Networks and Spatial Economics*, 1(3/4), 233-266.

Peeta, S. and Yu, J.W. (2004) Adaptability of a Hybrid Route Choice Model to Incorporating Driver Behaviour Dynamics under Information Provision*. IEEE Transactions on Systems, Man, and Cybernetics Part A*: *Systems and Humans*, 34(2), 243-256.

Peeta, S. and Paz, A. (2006) Behaviour-Consistent Within-day Traffic Routing under Information Provision. *In*: *9th International IEEE Conference on Intelligent Transportation Systems*, Toronto, Canada.

Tsoukalas, L. H. and Uhrig, R. E. (1997) *Fuzzy and Neural Approaches in Engineering*. New York: John Wiley and Sons, Inc.

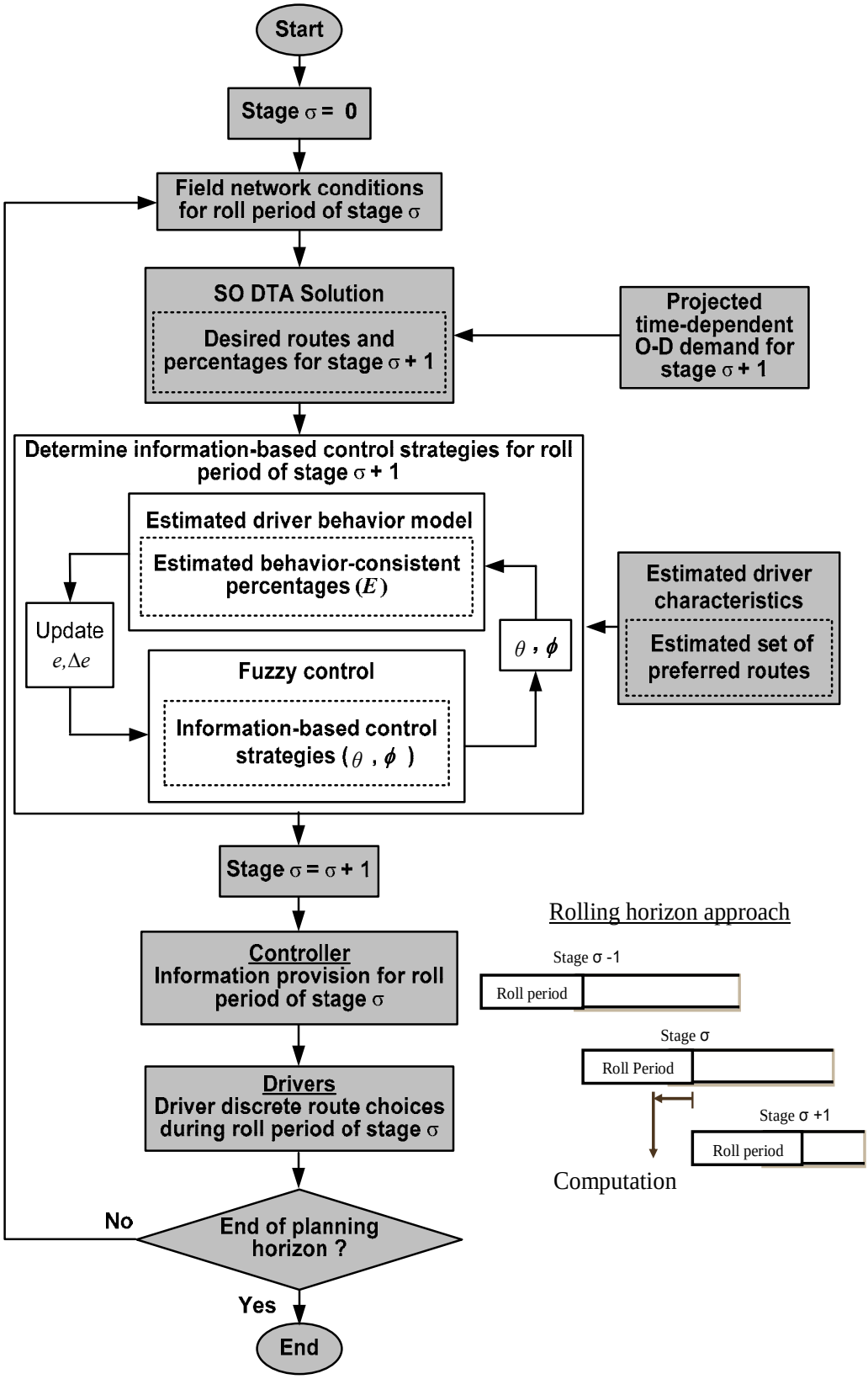



Figure 1. Conceptual Framework for the Behaviour-Consistent Within-day Traffic Routing under Information Provision Problem.

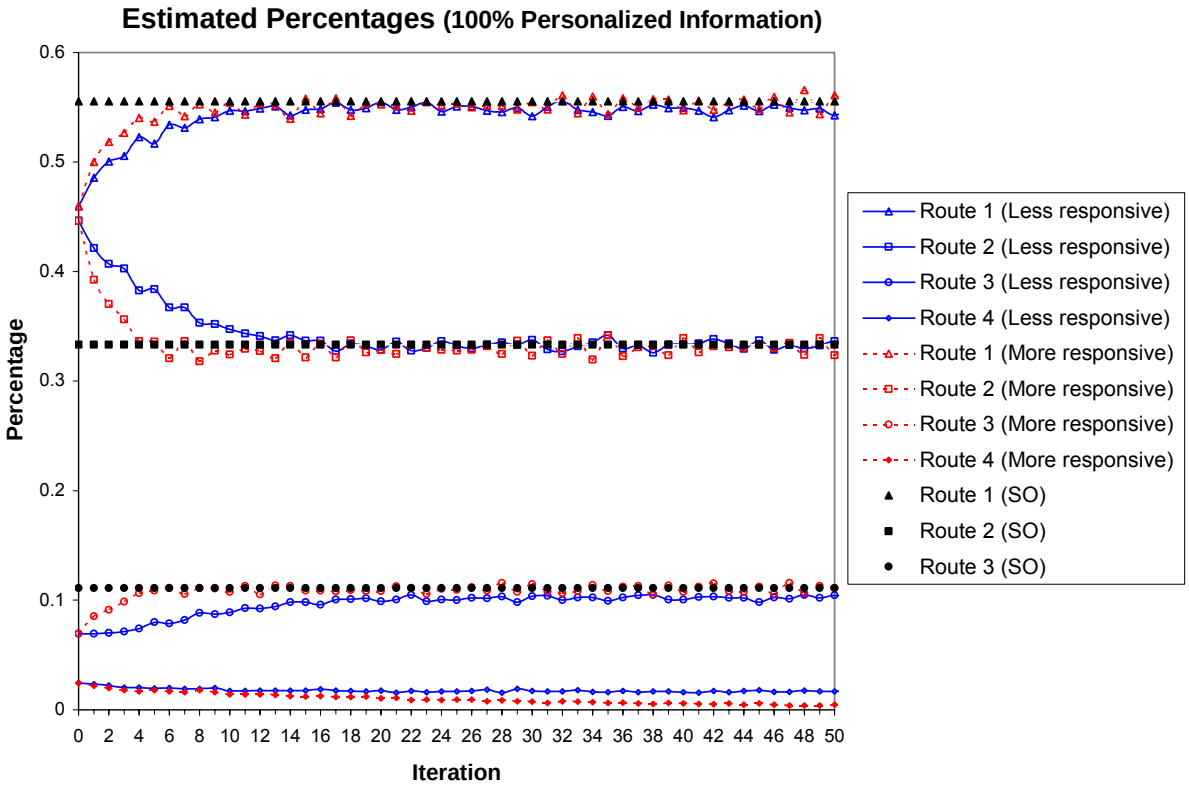



Figure 2. Estimated Percentage of Drivers Choosing Routes under 100% Personalized Information.

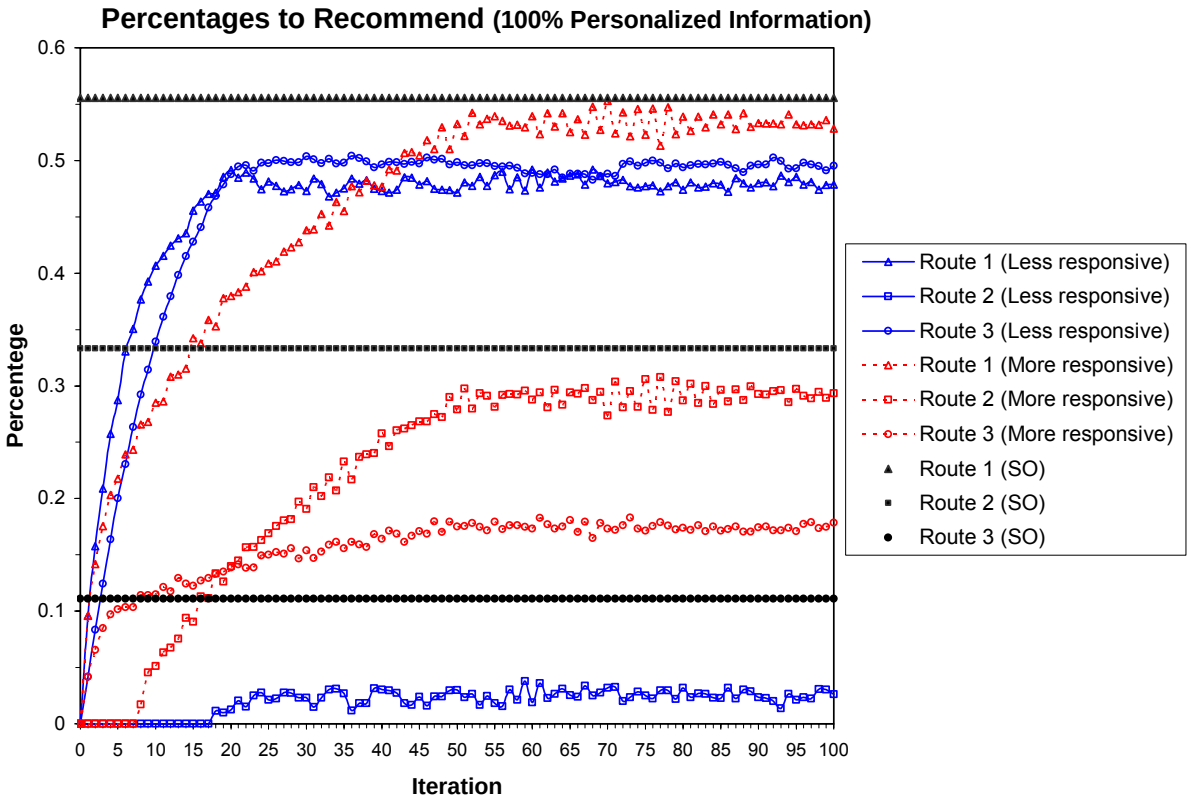



Figure 3. Percentage of Drivers that must be Recommended to Take Specific Routes under 100% Personalized Information.

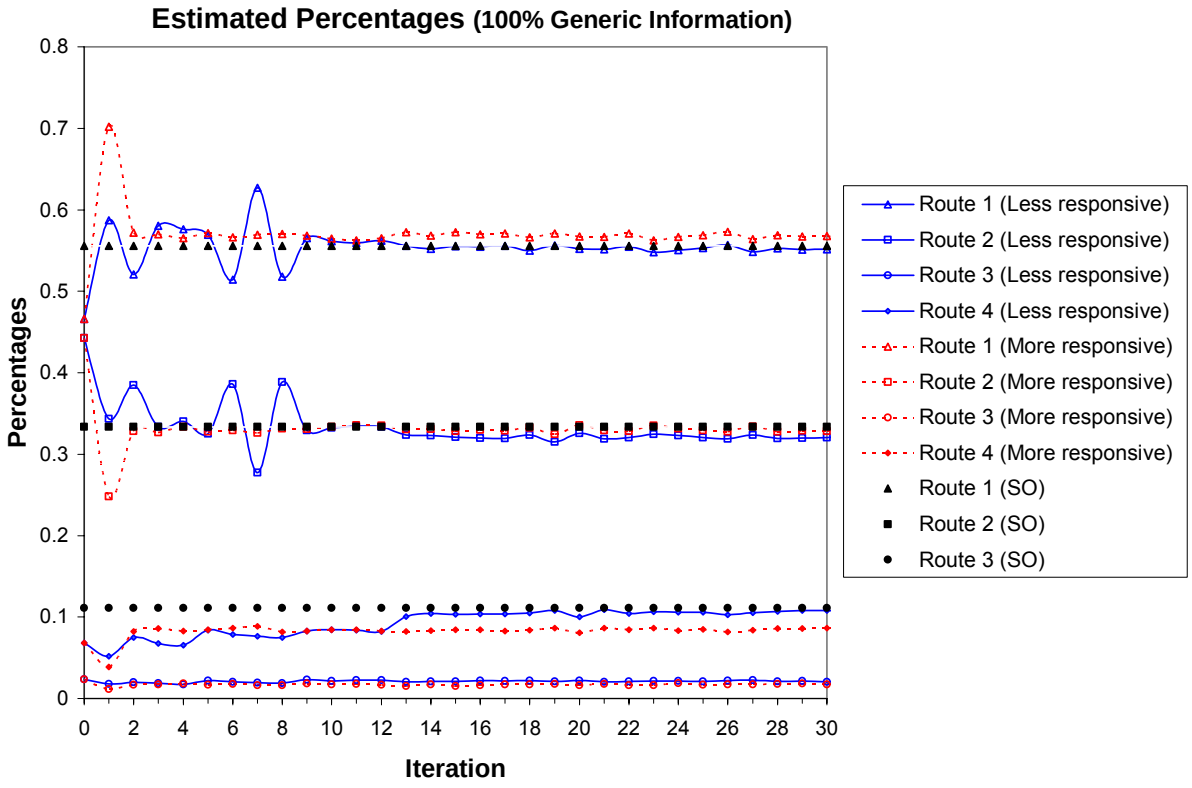



Figure 4. Estimated Percentage of Drivers Choosing Routes under 100% Generic Information.

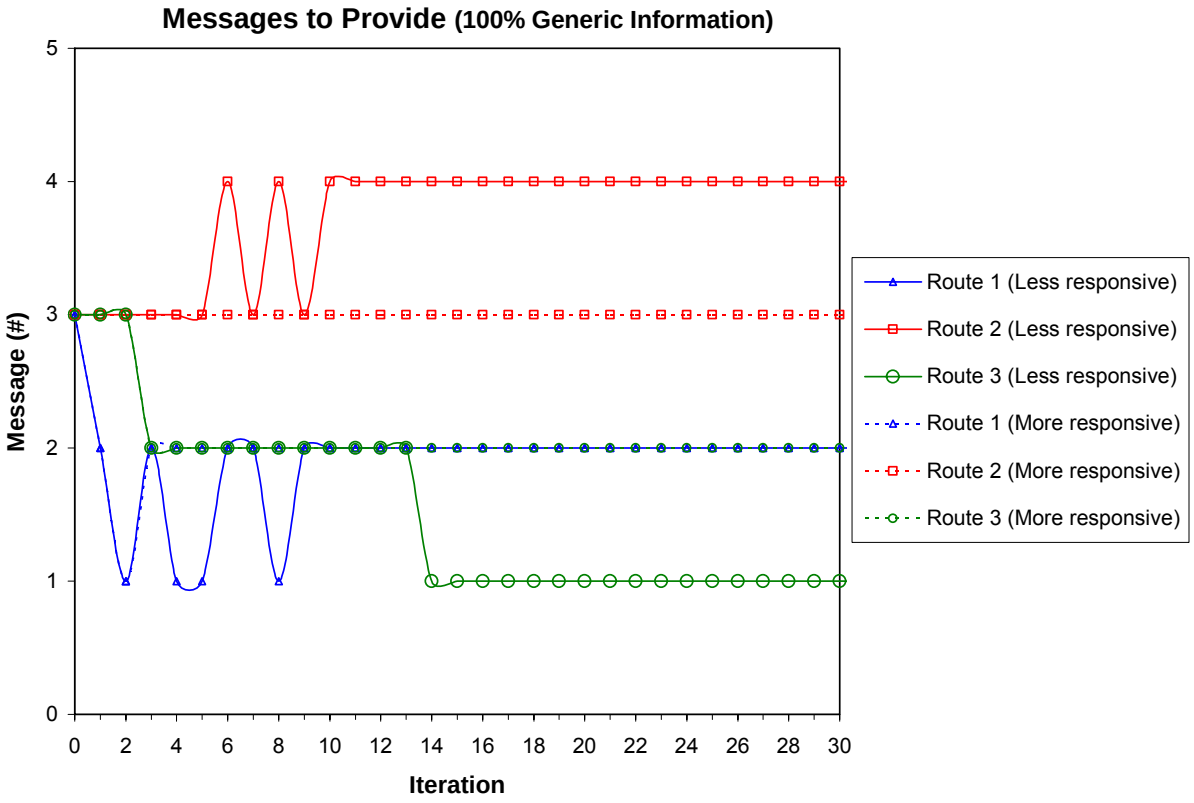



Figure 5. Messages to Provide under 100% Generic Information.

# LINEAR PROGRAMMING DYNAMIC TRAFFIC ASSIGNMENT MODELS: FORMULATIONS, EXTENSIONS AND COMPUTATIONAL RESULTS

*Satish V Ukkusuri: Rensselaer Polytechnic Institute,USA. ukkuss@rpi.edu*
*S. Travis Waller: University of Texas at Austin,USA* stw@mail.utexas.edu
*Ampol Karoonsoontawong: University of Texas at Austin,USA* ampolk@mail.utexas.edu

## Abstract
The opportunities for new technologies and improved computational resources have encouraged considerable research in the development of new methodologies and solution algorithms in dynamic transportation networks. This paper reviews the authors work on dynamic traffic assignment (DTA) model based on a linear programming approach where the embedded traffic flow model satisfies the Cell Transmission Model (CTM). The advantage of the developed models apart from realistic traffic flow modeling is that they lend very well to numerous extensions and analysis which are difficult with traditional formulations. This review begins with the formulation of a single destination based DTA model which satisfies user optimal conditions (Waller and Ukkusuri, 2003). This model is extended to a single level network design (NDP) model which accounts for user optimal conditions. This formulation is further extended to incorporate long term demand uncertainty. Bilevel network design problems accounting for DTA conditions are analyzed based on the initial formulations (Karoonsoontawong and Waller, 2006). Results from all these formulations are briefly discussed and compared.

## 1 Introduction
The ability to capture traffic dynamics using dynamic traffic assignment (DTA) models have important consequences on the ability to understand traffic congestion, travel behavior and system management. The governing relationships to model traffic dynamics form the core of any DTA model and are usually classified into two broad categories: (i) the travel choice principle which governs the behavioral rule for traffic assignment and (ii) the network loading component which captures the interaction and movement of vehicles under the behavioral assumptions in (i). Usually, in travel choice decisions, the most commonly adapted principle is the dynamic extension of Wardrop's principle called the dynamic user optimal principle and its stochastic extension, Stochastic dynamic user optimal. In the literature most of the models, capture both (i) and (ii) as either nonlinear complementary problem (e.g. Boyce et al., 2001), variational inequality problem (Friesz et al., 1993), fixed point problem (e.g. Perakis and Kachani, 2004) or a non-linear mathematical program (e.g. Lo and Szeto, 2002). Almost all of these formulations use complex non-linear relationships to capture traffic dynamics and require the route travel times to be strictly monotone with respect to route flows. However, it is clear that this assumption is clearly problematic in bottleneck networks (Waller and Ukkusuri, 2003).

The traffic flow component depicts how traffic propogates on a transport network and hence governs the network performance in terms of travel time. In most of the previous formulations (see for e.g. Boyce et al., 2002), this is modeled as a set of side constraints which render the resulting formulations to be difficult to solve. Further, most of the studies in the past have aspired to address the following traffic behavior properties in DTA models:

1. First-in first-out (FIFO) (e.g. Tong and Wong, 2000; Heydecker and Addison, 1998): FIFO can be defined at different levels but at a OD level it means that travelers who leave the origin earlier, reach the destination earlier.
2. Queue spillback and Shockwave propogation (Kuwahara and Akamatsu, 2001, Lo and Szeto, 2004): This refers to the end of the queue spilling backwards into the network when the flow exceeds capacity
3. Consistency (Carey, 2001): This refers that the DTA models satisfies what it intends to perform for all instances of the problem

In this study, we synthesize the authors work on linear programming based DTA models (Ukkusuri and Waller, 2004; Ukkusuri et al. (2004); Karoonsoontawong and Waller (2005, 2006); Kyunghwi et al. (2006)). The developed models address (1), (2) and (3) explicitly by using the Cell Transmission Model (CTM) as the

embedded traffic flow model. Initially, a linear programming based dynamic user optimal (DUO) formulation is presented. This formulation will be extended to model the network design problem, account for demand uncertainities in a single level NDP model and develop a bi-level dynamic network design problem.

## 2 CTM based Dynamic User Optimal (DUO) Formulation

The basic relationships of the cell transmission model that describe the evolution of traffic flow are extensively discussed in Daganzo (1994, 1995) and in Ziliaskopoulos (2000). Here, we briefly review the basic relationships. A summary of the notation used in the LP formulation is given in Table 1.

$$x_i^t = x_i^{t-1} + y_{ki}^{t-1} - y_{ij}^{t-1} \quad k \in \Gamma^{-1}(i),\ j \in \Gamma(i),\ \forall i \in C_O,\ \forall t \in T \tag{1}$$

$$y_{ki}^t = min\{ x_k^t ,\ min[ Q_i^t ,\ Q_k^t ],\ \delta_i^t ( N_i^t - x_i^t )\} \quad \forall (k,i) \in E_O,\ \forall t \in T \tag{2}$$

The first relationship constrains the cell mass conservation and is shown in Equation (1) for the Ordinary cell type, where $x_i^t$ is the number of vehicles in cell $i$ at time $t$ and $y_{ij}^t$ is the number of vehicles moving from cell $i$ to cell $j$ at time $t$.

Equation (2) states that the flow between two cells is constrained by the number of vehicles occupying the beginning cell, the remaining capacity at the ending cell, and the minimum of the maximum flow that can get out of the beginning cell and into the ending cell. These relationships are shown for the ordinary cell type but can be extended to Diverging, Merging, Sink, and Source cell type. Note that equation (2) is modeled with a relaxation within the linear program as in Ziliaskopoulos (2000). The implication of this is that even though the capacity and flow constraints are not violated, the entire allowable traffic will not necessarily exit cell $i$. This results in the potential of traffic being "held" in a cell even though it could move as has been observed in the SO version of this model. While this is clearly problematic and may suggest the need for a more constrained, and more difficult, formulation, it can be observed that in the single-destination UO version of this model, no unit of flow will be held if it causes any delay in that units arrival time. The holding, therefore, only occurs in the case of a non-unique solution and correct flows can be found in post-processing as demonstrated in Waller and Ukkusuri (2003).

By incorporating these constraints along with the minimization of $x_i^t$, a linear program for system optimal dynamic assignment is achieved. In the next section, we develop an equivalent user optimal model from the system optimal (SO) objective function (Waller and Ukkusuri, 2003).

**Table 1.** LP-DTA Notation

| | |
|---|---|
| $D$ : | *Sum of the total demand within the network* |
| $M_t$: | *Cost per time interval that will yield user optimal flows* |
| $C$ : | *The set of cells--ordinary cells ($C_O$), diverging cells ($C_D$), merging cells ($C_M$), source cells ($C_R$) and sink cells ($C_S$).* |
| $T$ : | *The set of discrete time intervals* |
| $S$ : | *The maximum time interval in the set T* |
| $x_i^t$ : | *Number of vehicles in cell i at time interval t* |
| $N_i^t$ : | *The maximum number of vehicles in cell i at time interval t,* |
| $y_{ij}^t$ : | *The number of vehicles moving from cell i to cell j at time interval t* |
| $E$ : | *The set of cell connectors—ordinary cell connectors ($E_O$), merging cell connectors ($E_M$), diverging cell connectors ($E_D$), source cell connectors ($E_R$), and sink cell connectors ($E_S$)* |
| $Q_i^t$ : | *The maximum number of vehicles that can flow into or out of cell i during time interval t* |
| $\delta_i^t$ : | *The ratio of forward to backward propagation speed for each cell and time interval* |
| $\Gamma(i)$ : | *The set of successor cells to i* |
| $\Gamma^{-1}(i)$ | *The set of predecessor cells to cell i* |
| $\tau$ : | *Discretization time interval* |
| $d_i^t$ | *The demand (inflow) at cell i in time interval t* |
| $TAB$ | *total available budget* |
| $\tilde{\chi}$ | *increase in $\tilde{N}_i^t$ per one unit of $b_i$* |
| $\tilde{\phi}$ | *increase in $\tilde{Q}_i^t$ per one unit of $b_i$* |
| $\tilde{\xi}$ | *vec( $\tilde{d}_i^t \ \forall i \in C, t \in T$ ; $\tilde{N}_i^t \ \forall i \in C, t \in T$ ; $\tilde{Q}_i^t \ \forall i \in C, t \in T$ ; $\tilde{\chi}$ ; $\tilde{\phi}$ )* |
| | *vector of stochastic parameters* |
| $b_i$ | *amount of budget allocated to cell i* |

The complete UO-DTA formulation is given below (Waller and Ukkusuri, 2003; Ukkusuri, 2002). Note, the constraint set is identical to the SO version of the problem

$$\text{Minimize} \quad \sum_{\forall t \in T} \sum_{\forall (i,j) \in E_s} M_t y_{ij}^t \tag{3-1}$$

Subject to:

$$x_i^t - x_i^{t-1} - \sum_{k \in \Gamma^{-1}(i)} y_{ki}^{t-1} + \sum_{j \in \Gamma(i)} y_{ij}^{t-1} = 0, \ \ \forall i \in C \setminus \{C_R, C_S\}, \forall t \in T \tag{3-2}$$

$$y_{ij}^t - x_i^t \le 0, y_{ij}^t \le Q_j^t, y_{ij}^t \le Q_i^t, y_{ij}^t + \delta_j^t x_j^t \le \delta_j^t N_j^t, \ \forall (i,j) \in E_o \cup E_R, \forall t \in T \tag{3-3}$$

$$y_{ij}^t - x_i^t \le 0, y_{ij}^t \le Q_i^t, \forall (i,j) \in E_s, \forall t \in T \tag{3-4}$$

$$y_{ij}^t \le Q_j^t, y_{ij}^t + \delta_j^t x_j^t \le \delta_j^t N_j^t, \forall (i,j) \in E_D, \forall t \in T \tag{3-5}$$

$$\sum_{\forall j \in \Gamma(i)} y_{ij}^t - x_i^t \le 0, \sum_{\forall j \in \Gamma(i)} y_{ij}^t \le Q_i^t, \forall i \in C_D, \forall t \in T \tag{3-6}$$

$$y_{ij}^t - x_i^t \le 0, y_{ij}^t \le Q_i^t, \forall (i,j) \in E_M, \forall t \in T \tag{3-7}$$

$$\sum_{\forall i \in \Gamma^{-1}(j)} y_{ij}^t \le Q_j^t, \sum_{\forall \in \Gamma^{-1}(j)} y_{ij}^t + \delta_j^t x_j^t \le \delta_j^t N_j^t, \forall j \in C_M, \forall t \in T \tag{3-8}$$

$$x_i^t - x_i^{t-1} + y_{ij}^{t-1} = d_i^{t-1}, j \in \Gamma(i), \forall i \in C_R, \forall t \in T, x_i^0 = \varsigma_i, \forall i \in C, \tag{3-9}$$

$$y_{ij}^0 = 0, \forall (i,j) \in E \tag{3-10}$$

$$x_i^S = 0, \forall i \in C / C_S, \tag{3-11}$$

$$x_i^t \ge 0, \forall i \in C, \forall t \in T \tag{3-12}$$

$$y_{ij}^t \ge 0, \forall (i,j) \in E, \forall t \in T \tag{3-13}$$

The objective function represents an opportunistic routing for each vehicle captured by the $M_t$ vector as described in Ukkusuri (2002). Constraint (3-2) governs the cell mass conservation relationship for all cells excluding the source and sink cells. Constraint (3-3) is the relaxation of Equation 2. Constraint (3-4) is the equivalent of constraint (3-3) for sink cells. Constraints (3-5) and (3-6) regulate the amount of flow transmitted out of diverging cells *i* for time *t*. Similarly, constraints (3-7) and (3-8) regulate the flow transmitted into a merging cell *j*; Constraint (3-9) expresses the cell mass balance for the source cells and the initial cell volumes; and constraint (3-10) specifies the initial flow conditions; demands $d_i^t$ and initial occupancies $\varsigma_i$ are given and constraint (3-11) is a special case of the mass balance constraint specifying a required final state where all traffic has left the network. Constraints (3-12) and (3-13) state the non-negativity conditions. Next, the extensions of the DUO formulation in (3-1) to (3-13) are presented with simple numerical results.

## 3 Extensions of the DUO formulation

In this section, we extend the formulation in DUO DTA formulation in (3-1) to (3-13) to a stochastic network design problem (NDP) based on the formulation in Ukkusuri and Waller (2004). The investment decisions are assumed to be continuous variables in each cell of the network. A formulation based on stochastic programming is developed where the long term demand is uncertain. A second extension of the formulation is a deterministic bi-level NDP formulation. Brief results of both the problems are presented towards the end of this section.

**Two-Stage Stochastic Linear Programs with Recourse (SLP2) of System Optimal (SO) and User Optimal (UO) Stochastic DTA-based Network Design Problem (NDP)**

The SLP2 SO and UO DTA-based NDP models (denoted as SLP2-SONDP and SLP2-UONDP) are limited to fixed-departure-time single-destination O-D demands. We assume that continuous investment variables change practical capacity but not free flow travel time, and the improvement cost is a linear function of the improvement levels. It is noted that the convex piecewise linear improvement cost functions can easily be incorporated. The formulations of the SO and UO NDP is given below.

**Formulations of SLP2-SONDP and SLP2-UONDP**

$$\min_{b} \quad Eh(b,\tilde{\xi}) \tag{4-1}$$

subject to

$$\sum_{i\in C\setminus C_S} b_i \le TAB \tag{4-2}$$

$$b_i \ge 0 \quad \forall i \in C \setminus C_S \tag{4-3}$$

where $h(b,\tilde{\xi}) = \min_{x,y} \sum_{(i,j)\in E_S} \sum_{t\in T} t y_{ij}^t$ for SLP2 SO DTA-based NDP (4-4a)

$$h(b,\tilde{\xi}) = \min_{x,y} \sum_{(i,j)\in E_S} \sum_{t\in T} M_t y_{ij}^t \text{ for SLP2 UO DTA-based NDP} \tag{4-4b}$$

subject to

$$x_i^t - x_i^{t-1} + \sum_{(i,j)\in FS(i)} y_{ij}^{t-1} - \sum_{(j,i)\in RS(i)} y_{ji}^{t-1} = \tilde{d}_i^t \qquad \forall i \in C \setminus C_s, t \in T \tag{4-5}$$

$$\sum_{(i,j)\in FS(i)} y_{ij}^t - x_i^t \le 0 \quad \forall i \in C, t \in T \tag{4-6}$$

$$\sum_{(j,i)\in RS(i)} y_{ji}^t \le \delta_i^t (\tilde{N}_i^t + \tilde{\chi} b_i - x_i^t) \quad \forall i \in C \setminus (C_R \cup C_S), t \in T \tag{4-7}$$

$$\sum_{(j,i)\in RS(i)} y_{ji}^t \le \tilde{Q}_i^t + \tilde{\phi} b_i \qquad \forall i \in C \setminus (C_R \cup C_S), t \in T \tag{4-8}$$

$$\sum_{(i,j)\in FS(i)} y_{ij}^t \le \tilde{Q}_i^t + \tilde{\phi} b_i \qquad \forall i \in C \setminus C_s, t \in T \tag{4-9}$$

$$x_i^0 = \zeta_i \quad \forall i \in C \tag{4-10}$$

$$y_{ij}^0 = 0 \quad \forall (i,j) \in E \tag{4-11}$$

$$x_i^{|T|} = 0 \quad \forall i \in C \setminus C_S \tag{4-12}$$

$$x_i^t \ge 0 \quad \forall i \in C, t \in T \tag{4-13}$$

$$y_{ij}^t \ge 0 \quad \forall (i,j) \in E, t \in T \tag{4-14}$$

The two formulations employ the same problem parameters, decision variables, and constraints; they are different only in the objective function. The objective function of the SO model minimizes the expected value of TSTT subject to a budget constraint and flow conservations; whereas the objective function of the UO model minimizes a function that makes all vehicles behave in the UO condition subject to a budget constraint and flow conservations. Thus, in the UO model, the objective minimizes the expected value of each individual's travel time rather than the expected value of TSTT.

The timing of what we know and when we know it is of fundamental importance. Here, we articulate the order of the events for clarity. The first-stage decisions are the amount of budget allocated to each cell *i* ($b_i$), and are made with the knowledge of the distribution of stochastic parameters. Thereafter, a realization of stochastic parameters ($\xi^\omega$ of $\tilde{\xi}$) unfolds; there are five interested stochastic parameter types: the long term O-D travel demand originated at cell *i* at time *t* ($\tilde{d}_i^t$), the capacity of cell *i* at time *t* ($\tilde{N}_i^t$), the maximum

flow into or out of cell *i* at time *t* ($\tilde{Q}_i^t$), the increase in $\tilde{N}_i^t$ per a budget unit ($\tilde{\chi}$), and the increase in $\tilde{Q}_i^t$ per a budget unit ($\tilde{\phi}$). The stochastic parameters of the first three types are independent across time intervals and cells, and those of the last two types are time- and space-invariant. After that, the scenario-dependent second-stage decisions are determined from corresponding DTA; i.e., the number of vehicles assigned to cell *i* at time *t* for scenario $\omega$ ($x_i^{t,\omega}$) and the number of vehicles traversing on cell connectors between cell *i* and cell *j* at time *t* for scenario $\omega$ ($y_{ij}^{t,\omega}$). Note that the first-stage decisions cannot depend on the scenario $\omega$ because this would correspond to knowing the future.

Two Monte Carlo bounding techniques, common random numbers (CRN) and independent random numbers (IRN) strategy, are employed to solve the stochastic models. The details are referred to Karoonsoontawong and Waller (2005). The results show that the CRN strategy outperforms IRN on their simple test network resembling a freeway corridor. The network size is sacrificed to gain higher confidence probabilistic behavior and to intuitively understand the effects of different network improvement policies. Although these findings may not necessarily be generalized, they provide interesting and insightful information. First, the SO and UO models allocate investment differently for certain budgets. Subsequently, the results of three comparison cases are summarized: i) for the SO models, it should be more valuable to solve the stochastic than the deterministic models, but it is not always the case for the UO models; ii) the SO models appear more desirable than the UO models for single level analysis; iii) it should be more valuable to solve the stochastic model accounting for more randomness.

**Bi-level NDP Model**

The network design problem is bi-level by nature, and can be seen as a static version of the non-cooperative, two-person game introduced by Von Stackelberg (1952) in the context of unbalanced economic markets. With the assumption of perfect information, the static game means that each player has only one move. The leader (transport planner) goes first, attempts to minimize the total costs, and anticipates all possible responses of his opponent, the follower (road users). The follower observes the leader's decision and reacts such that the follower achieves his/her optimal benefit (individual's travel time) regardless of external effects (system-wide costs). The bi-level NDP model (BLPNDP) is described below.

**Formulation of Bi-level DTA-based Network Design Problem**

$$\min_{b} \sum_{(i,j)\in E_S} \sum_{t\in T} \left(t \cdot y_{ij}^t\right) \tag{5-1}$$

subject to

$$\sum_{i\in C\setminus C_S} b_i \le TAB \tag{5-2}$$

$$b_i \ge 0 \quad \forall i \in C \setminus C_S \tag{5-3}$$

$$\min_{x,y} \sum_{(i,j)\in E_S} \sum_{t\in T} \left(M_t \cdot y_{ij}^t\right) \tag{5-4}$$

subject to

$$x_i^t - x_i^{t-1} + \sum_{(i,j)\in FS(i)} y_{ij}^{t-1} - \sum_{(j,i)\in RS(i)} y_{ji}^{t-1} = d_i^t \qquad \forall i \in C \setminus C_s, t \in T \tag{5-5}$$

$$\sum_{(i,j)\in FS(i)} y_{ij}^t - x_i^t \le 0 \quad \forall i \in C \setminus C_S, t \in T \tag{5-6}$$

$$\sum_{(j,i)\in RS(i)} y_{ji}^t \le \delta_i^t (N_i^t + \chi \cdot b_i - x_i^t) \quad \forall i \in C \setminus (C_R \cup C_S), t \in T \tag{5-7}$$

$$\sum_{(j,i)\in RS(i)} y_{ji}^t \le Q_i^t + \phi \cdot b_i \qquad \forall i \in C \setminus (C_R \cup C_S), t \in T \tag{5-8}$$

$$\sum_{(i,j)\in FS(i)} y_{ij}^t \le Q_i^t + \phi \cdot b_i \qquad \forall i \in C \setminus C_S, t \in T \tag{5-9}$$

$$x_i^0 = \zeta_i \quad \forall i \in C \setminus C_S \tag{5-10}$$

$$y_{ij}^0 = 0 \quad \forall (i,j) \in E \tag{5-11}$$

$$x_i^{|T|} = 0 \quad \forall i \in C \setminus C_S \tag{5-12}$$

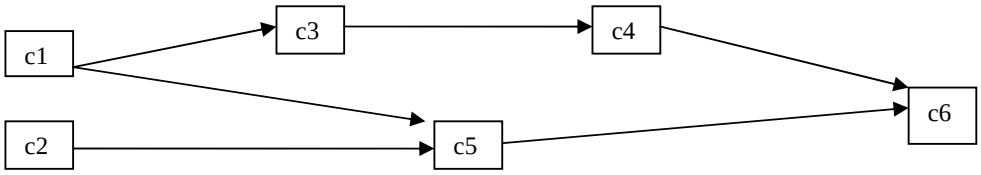
$$x_i^t \geq 0 \quad \forall i \in C \setminus C_S, t \in T \tag{5-13}$$

$$y_{ij}^t \geq 0 \quad \forall (i,j) \in E, t \in T \tag{5-14}$$

The leader's objective function (Eq.5-1) minimizes TSTT subject to a budget constraint (Eq.5-2 to Eq.5-3) and a UO DTA (Eq.5-4 to Eq.5-14). Bard (1998) indicated that there had been nearly two dozen algorithms proposed for solving the linear bi-level programming problem (BLPP), but only some of them appear more successful. He classified the viable algorithms into three categories: vertex enumeration, Karush-Kuhn-Tucker (KKT) condition and penalty approach. This study employs an algorithm that belong to the first category. The vertex-enumeration-type algorithm is the $K^{th}$-best algorithm originally invented by Murty (1968) and coined the name by Bialas and Karwan (1982). This algorithm is adopted because it solves a series of known subproblem UO DTA, and this potentially leads to a heuristic. Moreover, the advantage of this over the others is that if storage or computational limits are reached before convergence, the algorithm can return the best solution found with an optimality gap. The other methods yield a feasible and optimal solution only at the end of the algorithm. This algorithm involves the extreme point ranking procedure in the context of the simplex method. Further details are referred to Karoonsoontawong and Waller (2006).

From the computational experience, the degeneracy leads to extremely long computational time, so a modified $K^{th}$-best algorithm by ignoring the existence of degeneracy was used. This leads to a heuristic with a deterministic error bound. This algorithm is tested on a network shown in Figure 1a. The BLPNDP results are compared with the SONDP and UONDP, shown in Figure 1b-1c. Due to the modification to the original algorithm, the exact solutions cannot be guaranteed. The deterministic error bound of BLPNDP solution is determined from the difference of upper and lower bounds. They employed a known lower bound, the objective value of SONDP. Recall that SONDP is less constrained than BLPNDP. The deterministic error bounds of BLPNDP can be determined from Figure 1c. In Karoonsoontawong and Waller (2006), they employed another solution method to verify that the BLPNDP solutions in Figure 1 are indeed globally optimal.

a.1) Cell Transmission Network of Network 1




a.2) Time-Varying Jam Density ($N_i^t$) and Saturation Flow ($Q_i^t$)

| Cell | $N_i^t$ | $Q_i^t$ |
|---|---|---|
| c1 | +inf | 0.5 |
| c2 | +inf | 0.5 |
| c3 | 2 | 1 |
| c4 | 2 | 1 |
| c5 | 1 | 1 |
| c6 | +inf | +inf |

a.3) Value of $M_t$

| Time Interval | $M_t$ Value |
|---|---|
| t1 | 1 |
| t2 | 1 |
| t3 | 86 |
| t4 | 98 |
| t5 | 99.8 |
| t6 | 99.96 |
| t7 | 99.995 |
| t8 | 100 |

a.4) Time-Dependent Demand* ($d_i^t$)

| Cell | t1 | t2 |
|---|---|---|
| c1 | 2 | 1 |
| c2 | 2 | 1 |

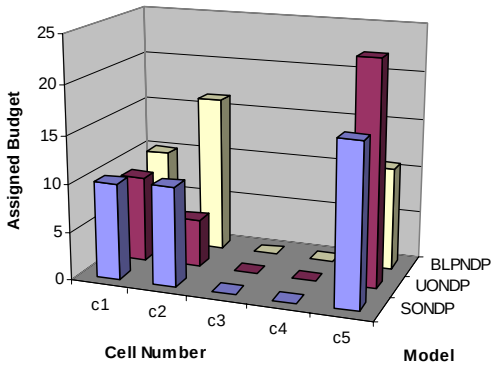
*All other cells and time intervals have zero

a.5) $\phi_i$ and $\chi_i$ Value

| |
|---|
| $\phi_i = 0.05; \ \forall i \in C \setminus C_S$ |
| $\chi_i = 0.03; \ \forall i \in C \setminus C_S$ |

demand.

b) Budget Allocations (Total Budget = 37)




c) Relationships between TSTT and Total Budget

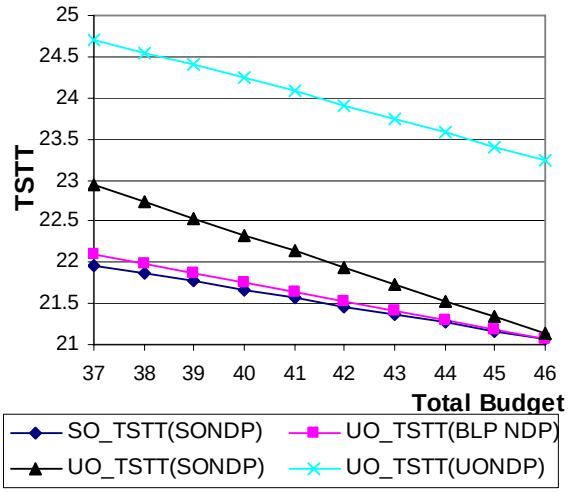



Figure 1. Test Problem and Results for Bi-level DTA-based Network Design Model

## 4 Summary

In this paper, we summarize the work done so far in the development of linear programming based DTA models which incorporate the cell transmission model as the traffic flow model. Other than developing a user optimal DTA formulation with an embedded CTM model, the work of Waller and Ukkusuri (2004) provides an alternate uniqueness and existence condition for the DUO in capacitated networks. Further, the DUO model is extended to account for stochaticity in the demand and capacity in a single network design model. Specifically, Monte Carlo bounding techniques, common random numbers (CRN) and independent random numbers (IRN) strategy, are employed to solve the stochastic models. In addition, the DUO formulation is extended to a bi-level NDP formulation and initial global results are presented for this problem. The advantages of these models lend themselves to numerous extensions. Current work with these models includes a combined hybrid simulation-analytical models that account for dynamic conditions for the network design problem. All these research problems make these models tremendously attractive from the practical and theoretical view points. The authors are currently working on some of these problems.

## References

1. Bard, J. F. *Practical Bilevel Optimization Algorithms and Applications*. Kluwer Academic Publishers, 1998.
2. Bialas, W. F. and M. H. Karwan. On-Two-Level Optimization. *IEEE Transactions Automatic Control*, Vol. AC-27, No.1, pp.211-214, 1982.
3. Boyce, D. E., Lee, D-H. and Ran, B. "Analytical models for the dynamic traffic assignment problem." *Networks and Spatial Economics*, 1(3), 377-390. 2002.
4. Carey, M. "Dynamic traffic assignment with more flexible modelling within links". *Networks and Spatial Economics 1(4)*, 349-375. Sept. 2001.
5. Friesz, T.L., Bernstien, D.H., Smith, T.E., Tobin, R.L. and Wie, B.W. A variational inequality formulation of the dynamic network equilibrium problem. *Operations Research*, 41, 179-191. 1993.

6. Heydecker, B.G. and Addison, J.D. Traffic models for dynamic traffic assignment. In M.G.H. Bell (ed.) Transport Networks: Recent Methodological Advances, Pergamon-Elsevier, Oxford, 35-49, 1998.
7. Karoonsoontawong, A. and S. T. Waller. A comparison of system- and user-optimal stochastic dynamic network design models using Monte Carlo bounding techniques. *Journal of the Transportation Research Board*, No.1923, pp.91-102, 2005.
8. Karoonsoontawong, A. and S. T. Waller. Linear bi-level programming and metaheuristics for dynamic continuous network design problem. Forthcoming in Journal of the *Transportation Research Board*, 2006.
9. Kuwahara, M. and Akamatsu, T. Dynamic user optimal assignment with physical queues for a many-to-many OD pattern. Transportation Research Part B, 35, 461-479, 2001.
10. Kyunghwi, J., Ukkusuri, S. and Waller, S.T. Improving computational efficiency for discrete network design problems accounting for dynamic traffic assignment conditions. Working Paper. 2006.
11. Lo, H.K. and Szeto, W.Y. A cell-based variational inequality formulation of the dynamic user optimal assignment problem. *Transportation Research Part B*, 36, 421-443, 2002.
12. Lo, H.K. and Szeto, W.Y. Modeling advanced traveller information services: static versus dynamic paradigms. *Transportation Research Part B*, 38, 495-515, 2004.
13. Murty, K. G. Solving the Fixed Charge Problem by Ranking the Extreme Points. *Operations Research*, XVI, pp.268-279, 1968.
14. Perakis, G. and Kachani, S. Modeling Travel Times in Dynamic Transportation Networks; A Fluid Dynamics Approach. Currently under review in *Operations Research*. 2004.
15. Tong, C.O. and Wong, S.C. A predictive dynamic traffic assignment model in congested capacity-constrained road networks. *Transportation Research Part B*, 34, 625-544, 2000.
16. Ukkusuri, S. V., and Waller, S. T. Linear programming models for the user and system optimal dynamic network design problem: Formulations, implementations and comparisons. *ASCE Journal of Transportation Engineering.* (2nd review. Submitted November 2005).
17. Ukkusuri, S. V., Karoonsoontawong, A., and Waller, S. T. A stochastic dynamic user optimal network design model accounting for demand uncertainty. In *International Conference on Transportation Systems Planning and Operations (TRANSPO2004) at IIT Madras* (Chennai, India, February 2004).
18. Von Stackelberg, H. *The Theory of the Market Economy*, Oxford University Press, Oxford, 1952.
19. Waller, S. T., and Ukkusuri, S. V. A linear programming formulation for the user optimal dynamic traffic assignment problem. *Transportation Science.* (In review. Submitted February 2003).

# VALIDATION OF A DYNAMIC TRAFFIC ASSIGNMENT MODEL

*Henk Taale: Delft University of Technology, The Netherlands h.taale@tudelft.nl*
*Henk J. van Zuylen: Delft University of Technology, The Netherlands h.j.vanzuylen@tudelft.nl*

## Abstract
In this paper the dynamic traffic assignment model MARPLE was validated, using data from an assessment of traffic management measures during roadworks on the A10-West near Amsterdam. First, an assignment was done using the existing network and OD matrix. Comparing the simulated flows and travel times with the measured ones showed large deviations. A good OD matrix is a very important factor for obtaining good results. Therefore, the OD matrix was calibrated using an OD estimation method and the flow measurements. The results after calibration are very good. The model predicts flows and travel times in good agreement with the measured values. For the validation the situation with roadworks was used. The network and OD matrix were adjusted for this situation and the assignment was run. The results for the flows are less good than before, but, taking into account that the OD matrix used was scaled with a global factor and not per relation, the results are satisfactory. The results for the travel times are very good. The conclusion if this paper is that the dynamic traffic assignment model is capable of simulating large networks with good results.

## 1 Introduction
Dynamic traffic assignment (DTA) models can be used to predict traffic flows in a transport network and to evaluate the effectiveness of traffic management strategies. This can be realised in a number of ways. A common classification of models is based on the way the models deal with traffic flows. This can be done macroscopically (traffic is a flow, propagated through the network using macroscopic relations), microscopically (traffic consists of individual vehicles, which interact with each other and each vehicle has its own behaviour) and mesoscopically (traffic consists of individual vehicles, but is propagated through the network using macroscopic relations).
To produce reliable predictions, a model should resemble reality as close as possible. One could say that microscopic models have an advantage on this aspect, but a good representation of the traffic flows in a network depends on more factors than just the modelling of traffic itself. A good network representation and a decent (dynamic) OD matrix are just as important. For regional networks it is even the question if microscopic models are the best way to predict traffic flows, due to the long calculation times and the large number of parameters one can adjust. For this type of networks a macroscopic approach can be better. To see if a macroscopic DTA model can adequately can represent traffic in a regional network, a validation study was done with MARPLE (Model for Assignment and Regional Policy Evaluation) (Taale *et al.*, 2004).

## 2 Context
In the summer of 2001 large roadworks were planned for a part of the western stretch of the ring road around Amsterdam in The Netherlands, called the A10-West. After years of intensive use, a thorough renovation was needed. To be able to carry out these roadworks, Rijkswaterstaat consulted the municipality of Amsterdam and other stakeholders and together they decided to close one carriageway during the summer holidays. From earlier experiences, it was known that this method would cause the least disruption for the road users. Therefore, in the period from May 26$^{th}$ until August 26$^{th}$, 2001 one carriageway with a length of 5 kilometres was closed and the carriageway in the other direction was used to handle the traffic in both directions on narrow lanes. Furthermore, all on-ramps and off-ramps on the stretch were closed, with some exception for trucks and emergency services. This was done for both carriageways.
Beforehand, it was judged that these roadworks would have a large impact on the traffic, not only on the A10-West itself, but also on the connecting motorways and the Amsterdam road network. To prevent chaos from happening, Rijkswaterstaat, in cooperation with the municipality of Amsterdam and other parties involved, prepared all kinds of measures. The measures consisted of mobility management and traffic management measures, together with a large public relations campaign to inform the public.
Due to the large impact of the loss of capacity on the A10-West motorway, research was done in an early stage to see whether or not traffic management measures would be able to mitigate the expected problems. Also, in an early stage, the AVV Transport Research Centre of Rijkswaterstaat planned three studies: a dynamic modelling study to determine the effects of the traffic management measures and to adjust these measures; an enquiry among road users about their choice behaviour before, during and after the roadworks

and a large evaluation program to monitor the traffic situation and to analyse the effects of the roadworks and measures. Taale *et al.* (2002) formulated the main conclusions of these three studies as follows:

- On the motorway itself no large problems occurred. In peak periods the traffic situation was reasonably good, due to the speed regime and the closure of on and off-ramps, among other things. This was different for the Amsterdam network. Especially on the north-south routes, traffic was severely congested. The volume on the A10-West decreased with 38%, but a large part of this presumably local traffic choose another route to arrive at their destination, due to the closure of the on-ramps and off-ramps. So, more traffic used the neighbouring motorways and the urban network. Therefore, a part of the congestion problem moved from the motorway to the urban network.
- The public relations campaign to inform the road users, before and during the roadworks, was effective. The campaign was partially responsible for the fact that large problems on the motorways did not occur. Of course, also the holiday period contributed to this. A part of the road users was on holiday or a couple of days on leave. Also, the shift to other routes, times of departure or travel modes caused less traffic on the roads.
- The measures related to mobility management (improve existing public transport and extend it with some temporary bus lines, improve the use of the bicycle, P+R facilities, teleworking, etc.) were not very effective. During the roadworks about 10% of the road users using the A10-West, chose alternatives such as public transport and the bicycle. The other mobility management measures were not used. Also the companies in the neighbourhood of the A0-West were not very interested in stimulating mobility management measures.
- The expected traffic problems did not occur. A part of the traffic disappeared or took other routes and departure times, which is in agreement with the findings of Cairns *et al.* (1998) and Goodwin *et al.* (1998). They studied nearly 100 cases covering incidents of road closure and capacity reduction throughout the world. The report looks at the changes in travel choice and behaviour that affect traffic conditions when road capacity is reduced and shows how reducing road capacity can lead to traffic reduction.

# 3 Measurements

Data were gathered from the A10 (speeds and flows per minute), on-ramps and off-ramps, flows on urban roads and travel times for two periods. The first period was in the summer of 2000. The roadworks were realised between May 26$^{th}$ and August 26$^{th}$, 2001, which was the second period. For the months May, June, July and August of 2000 and 2001, all data from the motorways around Amsterdam were gathered and processed. This concerned data per minute on flow and speed for the period between 06:00 and 20:00 hrs. On 25 locations shown in figure 1, flows were measured for both directions on the on-ramps and off-ramps and the urban network. These flows were measured on a 15-minute base for 2 weeks in the summer of 2000 (June 26-30 and August 7-11) and two weeks in the summer of 2001 (June 11-15 and August 6-10).

The travel times on the 8 routes shown in figure 2 were measured for 5 days in the first period and seven days in the after period with roadworks. This was done with probe cars. They drove through the network in the morning and evening peak and recorded the travel time and all abnormalities that occurred, e.g. open bridges, empty fuel tanks, etc.

All these data have been used to analyse the traffic before and during the roadworks. For the validation it is important to note that the amount of traffic decreased with 11% in the situation with roadworks. Unfortunately, it was not possible to distinguish between long distance and local trips, so a global decrease of the OD matrix was used.

# 4 Calibration and validation

For the validation of MARPLE, use has been made of the data from the simulation study and the measurements. The data from the first period (without roadworks) have been used to calibrate the model and the data from the second period with roadworks have been used to validate the model. The process and results of the calibration and validation are described in the next paragraphs. Part of this research is based on the work of Li (2005). She found that, apart from some model parameters related to path set generation, the OD matrix had the largest impact on the results. Therefore, in the calibration process much attention is paid to the OD matrix. For the model parameters, the values suggested in Li (2005) are used.

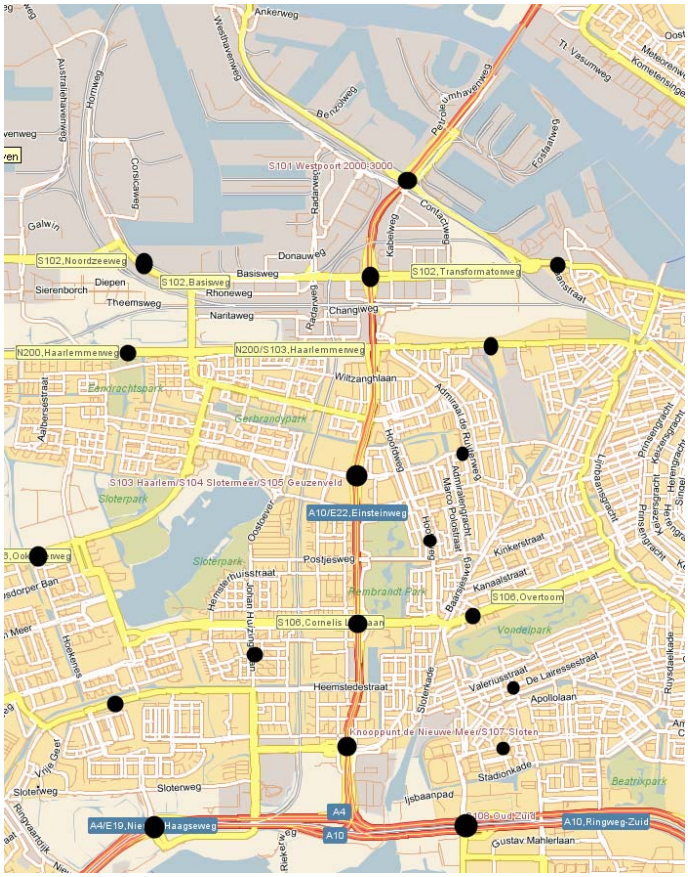



*Figure 1: Measurement locations for flows*

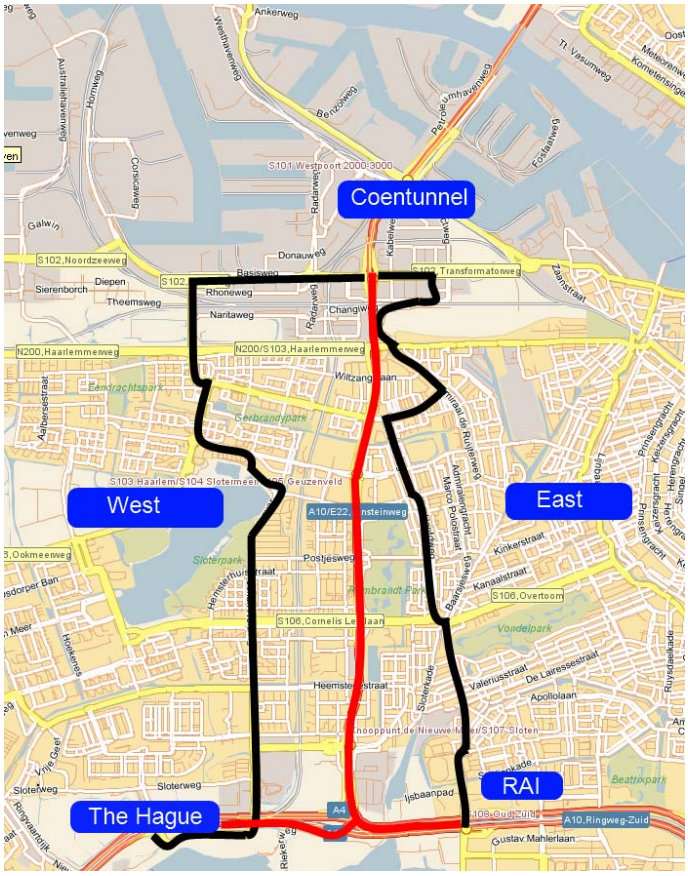



*Figure 2: Routes for travel time measurements*

# 5 Results before calibration of the OD matrix

From the simulation study, mentioned in section 2, the network was used, which is shown in figure 3. Some modifications were necessary for the use with a macroscopic model instead of a microscopic model. For example, for the lower level of detail in the macroscopic model it is not necessary to model all the turning lanes on an intersection. Other modifications dealt with some coding errors and missing links in the network. These errors were corrected and some new links were added.

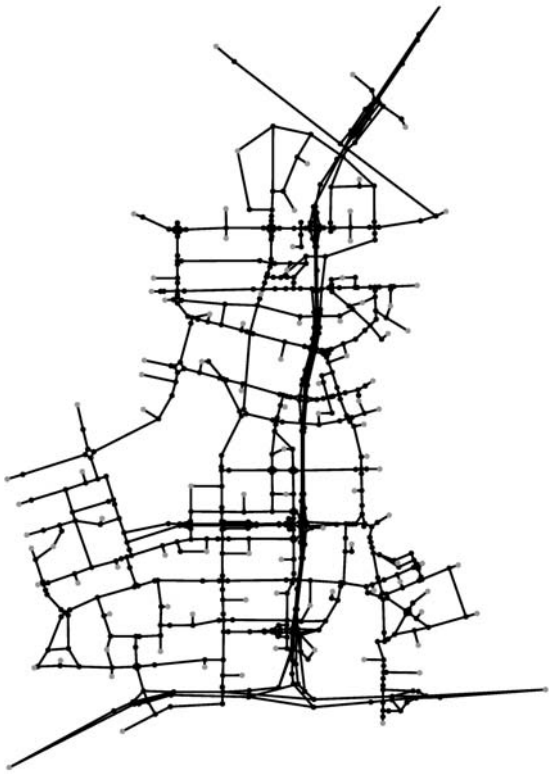


*Figure 3: Model network for the A10-West*

With this adjusted network and the existing OD matrix, an equilibrium assignment was done and flows and travel times from the simulation were compared with the measured ones for the before situation. For 64 measurement locations the flows are compared. For these locations the flow was measured for two and an half hour (15:30 - 18:00 hrs) divided into quarter of an hour periods. For the travel times the average value for the simulation period is used, because the number of measurements is too small to distinguish between the periods. The results for the flows are shown in figure 4 and for the travel times in figure 5. For the flows it can be seen that for the motorway flows (high values), the simulated flows are too high and the variation in smaller flows is large. The simulated travel times show a good resemblance with the measured ones.
To be able to quantify the quality of the model, goodness-of-fit measures are used. One of these measures is the root mean squared percentage error. The RMSPE quantifies the overall error of the model, penalising large errors at a higher rate than smaller ones. It is defined as

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - y_i}{y_i}\right)^2}, \tag{1}$$

where $x_i$ is the simulated value and $y_i$ the measured value. Another popular goodness-of-fit measure is the correlation coefficient (also called Pearson's correlation), which is defined as

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - x')(y_i - y')}{\sigma_x \sigma_y}, \tag{2}$$

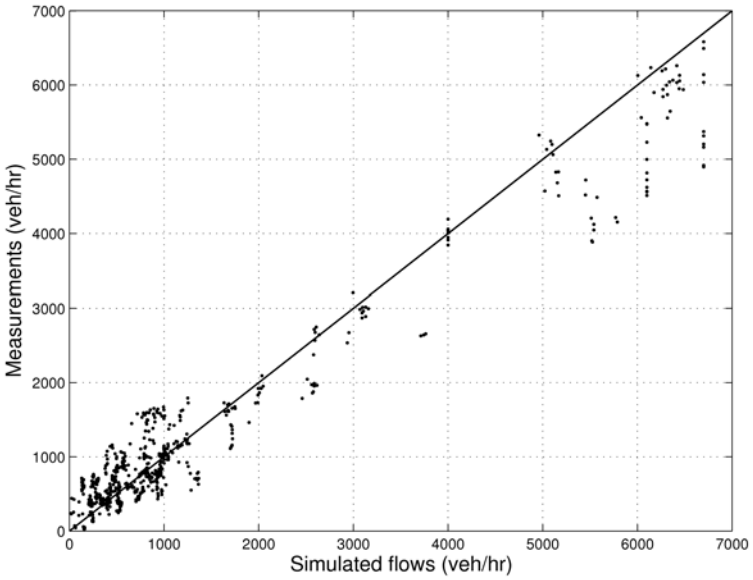



*Figure 4: Flows before calibration of the OD matrix*

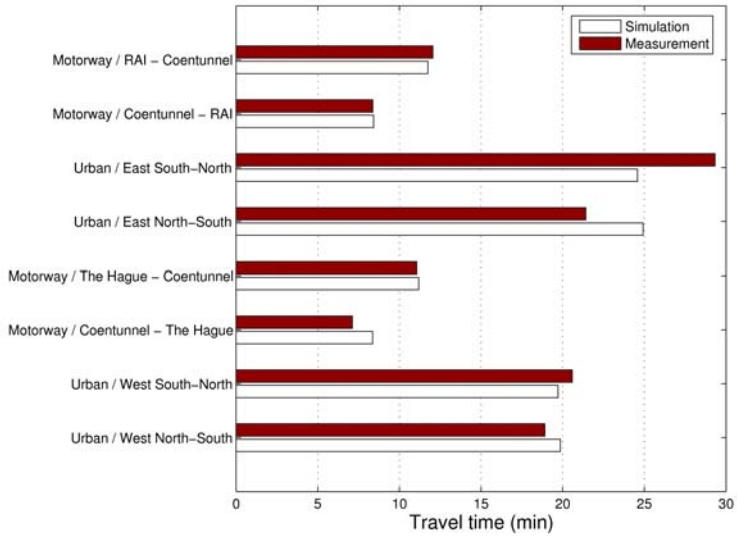



*Figure 5: Travel times before calibration of the OD matrix*

where *x'* and *y'* are the average simulated and measured values and $\sigma_x$ and $\sigma_y$ are the standard deviations of the simulated and measured values. The square of *r* is the well-known $R^2$ measure. According to Hourdakis *et al*. (2003) the RMSPE has an inherent deficiency in considering the disproportional weight of large errors while *r*, although being a good measure, does not provide any additional information to the modeller as to the nature of the error (difference) between real measurements and simulation. Theil (1961), in his work on economic forecasting, developed a goodness-of-fit measure called "Theil's Inequality Coefficient", which is more sensitive and accurate than the RMSPE or *r* and it can also be decomposed into three other metrics that provide specific information about the nature of the error. Theil's Inequality Coefficient is defined as

$$U = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}}, \qquad (3)$$

where $U$ is bounded between 0 and 1: 0 means perfect fit and 1 means no correlation at all. This measure can be decomposed into three proportions of inequality: bias ($U_M$), variance ($U_S$) and covariance ($U_C$). These three components are defined as

$$U_M = \frac{n(y'-x')^2}{\sum_{i=1}^{n}(y_i - x_i)^2}, \tag{4}$$

$$U_S = \frac{n(\sigma_y - \sigma_x)^2}{\sum_{i=1}^{n}(y_i - x_i)^2}, \tag{5}$$

$$U_C = \frac{2n(1-r)\sigma_y\sigma_x}{\sum_{i=1}^{n}(y_i - x_i)^2}. \tag{6}$$

By definition $U_M+U_S+U_C=1$. The bias reflects the systematic error, while the variance proportion indicates how well the simulation model replicates the variability in the observed data. Both should be as close to zero as possible, and in the opposite direction, covariance should be close to 1.
For the flows, separate goodness-of-fit measures are calculated for the motorways, on-ramps and off-ramps, the urban network and the total network. For the travel times only 8 values are available and therefore only the RMSPE and $R^2$ measures are given. All measures are shown in table 1. From this table it can be concluded that for the motorway flows the correlation is high, but there is a strong bias, which can also be seen in figure 4 (the large flows). The explanation of the variance is better, which is also true for the on-ramp and off-ramp flows. For the urban flows the bias is small, but the variance is not explained that well. The prediction of the travel times is very good, which can be seen in figure 5 and the small values for the RMSPE and $R^2$.

*Table 1: Goodness-of-fit measures for flows and travel times before calibration of the OD matrix*

| | | RMSPE | $R^2$ | $U$ | $U_M$ | $U_S$ | $U_C$ |
|---|---|---|---|---|---|---|---|
| Flows | Motorways | 0.2057 | 0.8586 | 0.0823 | 0.4883 | 0.0239 | 0.4930 |
| | On and off-ramps | 0.6742 | 0.4440 | 0.1973 | 0.0906 | 0.0215 | 0.8906 |
| | Urban network | 0.4006 | 0.7662 | 0.1456 | 0.0815 | 0.1822 | 0.7406 |
| | Total network | 0.5419 | 0.9489 | 0.1023 | 0.0007 | 0.2966 | 0.7043 |
| Travel time | | 0.1051 | 0.8914 | | | | |

# 6 Calibration method

To improve the estimation of the flows the OD matrix was re-estimated using the method described by Van Zuylen and Willumsen (1980) and adjusted by Van Zuylen (1981). This method uses a base OD matrix and adjusts OD flows to match the measured flows as close as possible. Originally, the method was meant for static matrices and one time period. In that case the equation to solve is

$$q^{od} = q_0^{od} X_o \prod_a X_a^{p_a^{od}}, \tag{7}$$

where $q^{od}$ is the estimation for OD pair $(o,d)$, $q_0^{od}$ the original demand for OD pair $(o,d)$, $X_o$ is defined by

$$X_o = \frac{\sum_{o,d} q^{od}}{\sum_{o,d} q_0^{od}}, \tag{8}$$

$X_a$ are factors to adjust the matrix and $p_a^{od}$ is the fraction of trips for relation ($o,d$) using link $a$. $X_o$ and $X_a$ are to be solved with

$$V_a = \sum_{o,d} p_a^{od} q_0^{od} X_o \prod_a X_a^{p_a^{od}}, \tag{9}$$

and

$$\sum_{o,d} q_0^{od} = \sum_{o,d} q_0^{od} \prod_a X_a^{p_a^{od}}. \tag{10}$$

The algorithm to solve this starts with initialising the iteration number $n$=0 and setting the factors $X_a^0$=1 for all links $a$ and

$$X_o^0 = \frac{\sum_a V_a}{\sum_a \sum_{o,d} p_a^{od} q_0^{od}}. \tag{11}$$

For each iteration $n$ and for each link $a$, calculate $X_a^{n+1}$ by solving

$$V_a = \sum_{o,d} p_a^{od} X_o^n q_0^{od} \left( \prod_{a'} \left( X_{a'}^n \right)^{p_{a'}^{od}} \right) Y_a, \tag{12}$$

for $Y_a$ and setting

$$X_a^{n+1} = X_a^n Y_a. \tag{13}$$

Calculate the new OD adjustment factors with

$$\widehat{X}_o^{n+1} = \frac{X_o^n \sum_{o,d} q_0^{od} \prod_a \left( X_a^{n+1} \right)^{p_a^{od}}}{\sum_{o,d} q_0^{od}}, \tag{14}$$

and smooth them with

$$X_o^{n+1} = \frac{X_o^n + \widehat{X}_o^{n+1}}{2}. \tag{15}$$

Now the estimated OD matrix for iteration $n+1$, $q_{n+1}^{od}$ is given by

$$q_{n+1}^{od} = q_0^{od} X_o^{n+1} \prod_a \left( X_a^{n+1} \right)^{p_a^{od}}. \tag{16}$$

If the flows, resulting from the assignment with $q_{n+1}^{od}$, are sufficiently close to the measurements, then stop, otherwise increase the iteration number $n$ with 1 and return to equation 12.
The algorithm was used for adjusting the dynamic matrix by applying it for every time period separately. The dynamics are dealt with by taking travel times into account for the calculation of the OD link proportions. The original algorithm was adjusted a little bit and incorporated into a dynamic environment and an iteration loop with the dynamic stochastic assignment in MARPLE (Taale and Van Zuylen, 2003). The convergence criterion used, is the total RMSPE, as defined by equation (1).

# 7 Results after calibration of the OD matrix

The calibration method described in the previous section was used to construct an adjusted OD matrix. For the first runs it turned out that the results for the motorway were not that good, caused by some measurement locations on road sections with congestion. The measurement locations were removed and the process was repeated to obtain a better OD matrix. This matrix was used in a new equilibrium assignment and the results for flows and travel times were again compared with the measured values. The results are shown in the figures 6 and 7. From these figures it is clear that the simulated flows show a much better correspondence with the measured flows. The correspondence for the travel times is a little bit worse, which is caused by the fact that in the simulation there is less congestion on the A10-West (less demand in the new OD matrix). This gives shorter travel times for the motorway routes The Hauge-Coentunnel and RAI-Coentunnel. But still the travel times in the assignment represent the measured ones reasonably well, which is also shown in table 2, that gives the goodness-of-fit measures. It is clear that the calibration improved the results a lot.

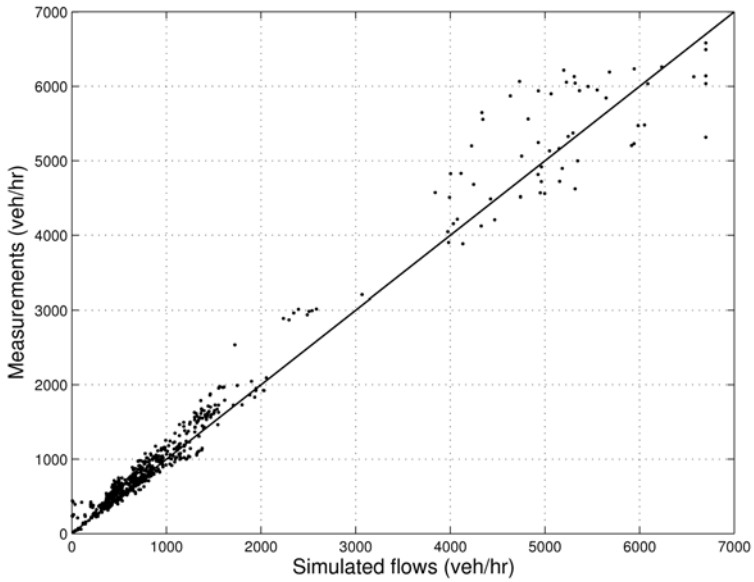



*Figure 6: Flows after calibration of the OD matrix*

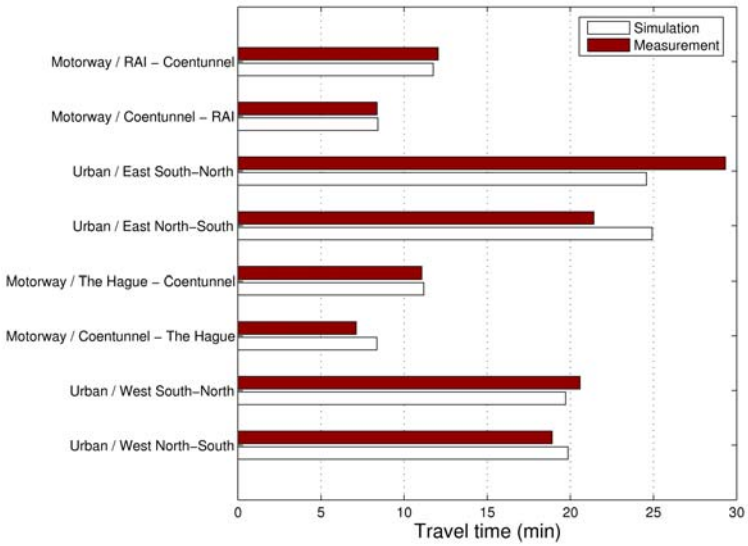



*Figure 7: Travel times after calibration of the OD matrix*

*Table 2: Goodness-of-fit measures for flows and travel times after calibration of the OD matrix*

| | | RMSPE | $R^2$ | $U$ | $U_M$ | $U_S$ | $U_C$ |
|---|---|---|---|---|---|---|---|
| Flows | Motorways | 0.1361 | 0.8802 | 0.0616 | 0.1469 | 0.0547 | 0.8092 |
| | On and off-ramps | 0.1426 | 0.9584 | 0.0715 | 0.4109 | 0.1756 | 0.4153 |
| | Urban network | 0.2451 | 0.9317 | 0.0749 | 0.0656 | 0.2356 | 0.7033 |
| | Total network | 0.1832 | 0.9751 | 0.0638 | 0.1193 | 0.0184 | 0.8637 |
| Travel time | | 0.1403 | 0.9034 | | | | |

# 8 Validation

For the validation, the situation with roadworks was used. For this situation the network was adjusted, because of the closure of on and off-ramps and the reduced capacity and speed limit for the motorway section where the roadworks were done. From the measurements during the roadworks it could be concluded that the demand was decreased with 11%, although the decrease was not measured for every OD relation separately. Therefore, for the situation with roadworks a global scale factor of 0.89 was used for all OD pairs. Of course, also new routes had to be generated because of the differences in the network. An equilibrium assignment with the new network and the new OD matrix gives the results for the flows and travel times as shown in the figures 8 and 9. The results for the goodness-of-fit measures are shown in table 3.

From figure 8 it is clear that the predicted flows for the motorway are a little bit lower than the measured ones, which also results in a larger bias $U_M$, but still the error is reasonably small. For the on and off-ramp and urban flows, the variations are larger, but the error remains within acceptable limits, taking into account that the OD matrix used, was scaled with a global factor and not per relation. Also the simulated travel times reproduce the measured travel times very well.

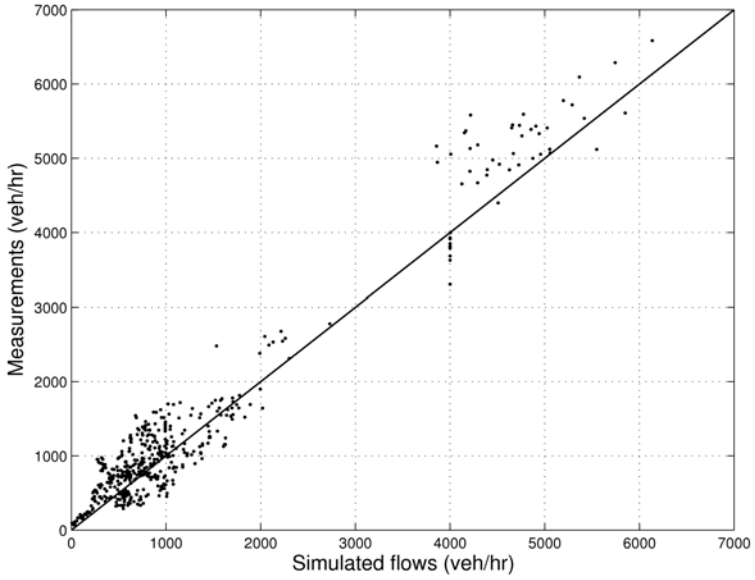



*Figure 8: Flows validation*

*Table 3: Goodness-of-fit measures for flows and travel times for validation*

| | | RMSPE | $R^2$ | $U$ | $U_M$ | $U_S$ | $U_C$ |
|---|---|---|---|---|---|---|---|
| Flows | Motorways | 0.1634 | 0.9350 | 0.0667 | 0.3353 | 0.0483 | 0.6248 |
| | On and off-ramps | 0.3933 | 0.4990 | 0.1872 | 0.2655 | 0.0028 | 0.7357 |
| | Urban network | 0.3844 | 0.5756 | 0.1533 | 0.0352 | 0.0000 | 0.9688 |
| | Total network | 0.3624 | 0.9438 | 0.0978 | 0.1603 | 0.0670 | 0.7744 |
| Travel time | | 0.1425 | 0.9807 | | | | |

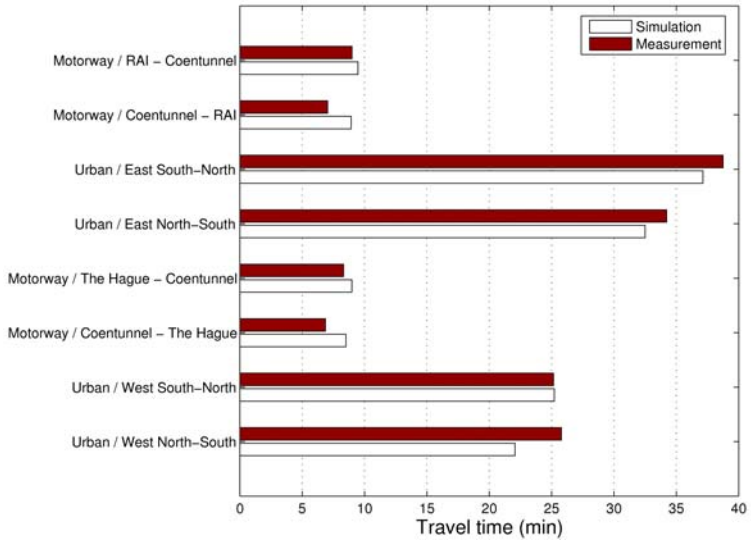



*Figure 9: Travel times validation*

## 9 Conclusions

In this paper the validation of the dynamic traffic assignment model MARPLE is described. Flow and travel time measurements from an assessment study, about the effects of traffic management during road works, were used to calibrate the OD matrix and validate the model. The results showed that the OD calibration method developed is very effective in improving the results of the assignment related to the measurements. The model is able to accurately predict the flows and the travel times in a network.
The results for the validation are less accurate, but this can be attributed to the different situation and demand which were used for the validation.

## References


- Cairns, S., Hass-Klau, C. and Goodwin, P. (1998) Traffic Impact of Highway Capacity Reductions: Assessment of the Evidence, Landor Publishing, London, United Kingdom.
- Goodwin, P., Hass-Klau, C. and Cairns, S. (1998) Evidence on the Effects of Road Capacity Reduction on Traffic Levels, *Traffic Engineering + Control*, Vol. 39, No. 6, pp. 348-354.
- Hourdakis, J., Michalopoulos. P.G. and Kottommannil, J. (2003) Practical Procedure for Calibrating Microscopic Traffic Simulation Models, *Transportation Research Record*, No. 1852, pp. 130-139.
- Li, H. (2005) *Calibration and Validation of MARPLE*, Technical Report, AVV Transport Research Centre.
- Taale, H., Schuurman, H. and Bootsma, G. (2002) *Evaluation Large Roadworks on the A10-West*, Report, AVV Transport Research Centre (in Dutch).
- Taale, H. and Van Zuylen, H.J. (2003) The Effects of Anticipatory Control for Several Small Networks. In: *Proceedings of the 82nd Annual Meeting of the Transportation Research Board*, January 12-16, 2003, Washington D.C., U.S.A.
- Taale, H., Westerman, M., Stoelhorst, H. and Van Amelsfort, D. (2004) Regional and Sustainable Traffic Management in The Netherlands: Methodology and Applications, In: *Proceedings of the European Transport Conference 2004*, October 4-6, 2004, Strasbourg, France, Association for European Transport.
- Theil, H. (1961) *Economic Forecasts and Policy*. North-Holland Publishing Company, Amsterdam, The Netherlands.
- Van Zuylen, H.J. (1981) Some Improvements in the Estimation of an OD Matrix from Traffic Counts, In: *Proceedings of the 8 International Symposium on Transportation and Traffic Theory*, Toronto, Canada.
- Van Zuylen, H.J. and Willumsen, L.G. (1981) The Most Likely Trip Matrix Estimated from Traffic Counts. *Transportation Research Part B*, Vol. 14, No. 3, pp. 281-293.

# MODELING OF COMMUTERS' DAY-TO-DAY LEARNING BEHAVIOR

*Kaan Ozbay, Dept. of Civil and Environmental Engineering, Rutgers University, USA Associate Professor, kaan@rci.rutgers.edu*
*Ozlem Yanmaz-Tuzel, Dept. of Civil and Environmental Engineering, Rutgers University, USA, Graduate Research Assistant, yanmaz@rci.rutgers.edu* (Corresponding author)

## Abstract
This study uses Stochastic Learning Automata (SLA) theory to model commuters' day-to-day learning behavior within the context of departure-time choices. Unlike other learning methods proposed in the literature, SLA is a powerful modeling tool that does not require extensive data because it attempts to find a solution without a priori information on the optimal action. The proposed SLA model addresses the learning/adaptation of travelers on the basis of experienced travel choices and user-specific characteristics. The stochastic nature of the system and heterogeneity among different drivers in day-to-day learning behavior are considered by combining Bayesian Learning approach with the SLA theory, and by developing probability distributions for each model parameter. To train and test the parameters of the model individual commuter data obtained from New Jersey Turnpike a 148 mile-toll road with 27 entry-exit points are used. The proposed model aims to capture the commuters' departure-time choice learning behavior both under undisturbed (before toll change) and disturbed network conditions (after toll change) and to investigate commuters' responses to toll, travel-time, departure/arrival time restrictions while selecting their departure-times. The results have demonstrated the possibility of developing a psychological framework (i.e., learning models) as a viable approach to represent traveler behavior.

## 1. Introduction
Drivers' day-to-day learning behavior and behavioral processes involved in travel decisions in uncertain environments, and how such learning affects their adaptation to the system have become a vital component of travel behavior research. Transportation systems are highly dynamic, non-stationary and uncertain. Moreover, drivers' information is limited and perhaps biased, thus their choices among available alternatives reflect their perceptions of the utilities associated with each alternative (Avineri and Prashker, 2005). Consequently, drivers tend to make decisions based on their day-to-day experiences with the system which result from repeated previous choices. These experiences are the input for drivers' learning behavior and their adaptation to the changes in the dynamic and uncertain system.

In travel-choice decision-making process, travel utilities occurred in the past affect commuters' current decisions. Travel attributes like travel time are not constant and not likely to be known before the current trip occurs, thus Horowitz (1984) suggested an equilibrium model in which the travel choices on each day are based on averages of travel times on previous days. However, recent research (Kahneman and Tversky, 1979; Mahmassani and Cheng, 1985; Avineri and Prashker, 2003; Senbil and Kitamura, 2004; and others) has shown that travelers do not necessarily minimize travel time when making a travel choice, or they are not necessarily utility maximizers. Rather travelers may adopt some simple rules; such that good outcomes, associated with selecting a particular strategy, increase the probability that this strategy would be chosen again. This kind of learning process in travel-choice has been studied by applying Reinforcement Learning, Bayesian Learning and Stochastic Learning Automata.

Roth and Erev (1995) suggested that drivers' learning process can be described as reinforcement models. Reinforcement learning is a form of machine learning used to solve problems of interaction and to learn the optimal action through a learning process (Sutton and Barto 1998). It is a form of trial-and-error learning, where an agent starts interacting with the environment with a random action, and receives rewards when this action leads to successful performance. As the agent explores the environment and finds actions to high reward its behavior changes. To make the decision as to what action to take in a particular state, an agent can draw on previous experience and take the action which, on average, led to the better reward. In transportation area Erev and Roth (1998), Schreckenberg and Selten (2004), Miyagi (2005) and Selten et al. (2006) proposed reinforcement learning models considering route travel times, and conducted experiments to understand individual traveler behavior.

Another form of learning model, Bayesian Learning (BL), describes the traveler-choice behavior as an iterative process, in which at each step, the traveler uses historical frequencies of different travel times and form a belief about the travel choices' expected travel times. Then the traveler makes a choice to minimize the expected (or random) utility, given his/her beliefs. March (1996), Jha et al. (1998), and Chen and Mahmassani (2004) used Bayesian updating of travel time to model how travelers learn from experience.

Stochastic Learning Automata (SLA), a form of Reinforcement Learning model, mimics drivers' day-to-day learning by updating the drivers' choice probabilities based on information received by the drivers and the experience of drivers. The appropriate selection of the proper learning algorithm as well as the parameters it contains is crucial for its success in modeling choice behavior. In simple terms, the SLA approach is an inductive inference mechanism that updates the probabilities of its actions occurring in a stochastic environment in order to improve a certain performance index, i.e. travel cost of users. The learning schemes, in which the action probabilities of an automaton are updated, are based of the responses of the environment. This process is naturally closely related to BL, in which the distribution function of a parameter is updated at each instant on the basis of new information. However, in BL the updating takes place according to Bayes' rule, while it is more general in SLA. In fact, in SLA, no unique procedure exists and the specific scheme used depends on a number of factors such as accuracy, stability, and speed of convergence (Narendra and Thathachar, 1989). Ozbay *et al.* (2001) and Ozbay et al. (2002) proposed to use SLA to model the day-to-day learning behavior of drivers in the context of route and departure time choice behavior.

In this paper, Stochastic Learning Automata (SLA) theory is used to model drivers' day-to-day learning/adaptation behavior within the context of travel-choices. In Section 2, the theory of SLA is discussed in detail. Next, in Section 3, the proposed SLA model developed to understand New Jersey Turnpike users' departure-time choices under time-of-day pricing applications is provided. Then the proposed model trained and tested using NJ Turnpike traffic and traveler survey data. The results are presented in Section 4. Finally, Section 5 is devoted to conclusions and discussions.

# 2. Stochastic Learning Automata

The stochastic automaton attempts a solution of the problem without any information on the optimal action (initially, equal probabilities are attached to all the actions). One action is selected at random, the response from the environment is observed, action probabilities are updated based on that response, and the procedure is repeated. A stochastic automaton acting as described to improve its performance is called a *learning automaton* (Narendra and Thathachar, 1989). The objective in the design of the learning automaton is to determine how the choice of the action at any stage should be guided by past actions and responses.

Theory of learning automata is concerned with the analysis and synthesis of automata which operate in random environments. In this section we describe the random environments, the structure and characteristics of the automata, and the mathematical tools that are applicable to the analysis of such systems.

## 2.1. Environment

In the context of learning automata it is not easy to specify the environment. The definition encompasses a large class of unknown random media in which an automaton can operate. Mathematically, an environment is represented by a triple $\{\alpha, c, \beta\}$ where $\underline{\alpha}$ represents a finite action/input set, $\underline{\beta}$ represents an output set, and $c$ is a set of penalty probabilities, where each element $c_i$ corresponds to one action $\alpha_i$ of the set $\alpha$. The action $\alpha(n)$ of the automaton belongs to the set $\underline{\alpha}$, and is applied to the environment at time $t = n$. The output $\beta(n)$ from the environment is an element of the set $\underline{\beta}$. In the simplest case, the values $\beta_i$ are chosen to be 0 and 1, where 1 is associated with failure/penalty response. Consequently, $c_i$ represents the probability that the action $\alpha_i$ will result in a penalty output. The elements of $\underline{c}$ are defined as:

$$Pr(\beta(n) = 1 \mid \alpha(n) = \alpha_i) = c_i \quad (i = 1,2...,r) \tag{1}$$

There are several models defined by the response set of the environment. Models in which the output can take only one of two values, 0 or 1, are referred to as P-models. In this case, response value of 1 corresponds to an "unfavorable" (failure, penalty) response, while output of 0 means the action is "favorable." A further generalization of the environment allows finite response sets with more than two elements that may take finite number of values in an interval [a, b]. Such models are called Q-models. When the output is a continuous random variable with possible values in an interval [a, b], the model is named S-model. The focus of this paper is P-models.

## 2.2. The Stochastic Automaton

The automaton takes in a sequence of inputs and puts out a sequence of actions. Mathematically, the automaton can be represented by a quintuple $\{\underline{\Phi}, \underline{\alpha}, \underline{\beta}, F(\cdot,\cdot), H(\cdot,\cdot)\}$ where;

- $\underline{\Phi}$ is asset of internal states. At any instant $n$, the state $\phi(n)$ is an element of the finite set $\underline{\Phi} = \{\phi_1, \phi_2, ..., \phi_s\}$

- $\underline{\alpha}$ is a set of actions (or outputs of the automaton). The output/action of an automaton at instant *n*, denoted by $\alpha(n)$, is an element of the finite set $\underline{\alpha} = \{\alpha_1, \alpha_2, ..., \alpha_r\}$
- $\underline{\beta}$ is a set of responses (or inputs from the environment). The input from the environment $\beta(n)$ is an element of the set $\underline{\beta}$ which could be either a finite set or an infinite set, such as an interval on the real line: $\underline{\beta} = \{\beta_1, \beta_2, ..., \beta_m\}$ or $\underline{\beta} = \{(a, b)\}$.
- $F(\cdot,\cdot): \underline{\Phi} \times \underline{\beta} \to \underline{\Phi}$ is a function that maps the current state and input into the next state. *F* can be deterministic or stochastic:

$$\phi(n+1) = F[\phi(n), \beta(n)] \tag{2}$$

- $H(\cdot,\cdot): \underline{\Phi} \times \underline{\beta} \to \underline{\alpha}$ is a function that maps the current state and input into the current output. If the current output depends on only the current state, the automaton is referred to as state-output automaton. In this case, the function $H(\cdot,\cdot)$ is replaced by an output function $G(\cdot): \underline{\Phi} \to \underline{\alpha}$, which can be either deterministic or stochastic:

$$\alpha(n) = G[\phi(n)] \tag{3}$$

Figure 1 shows the relation between the automaton and the environment.

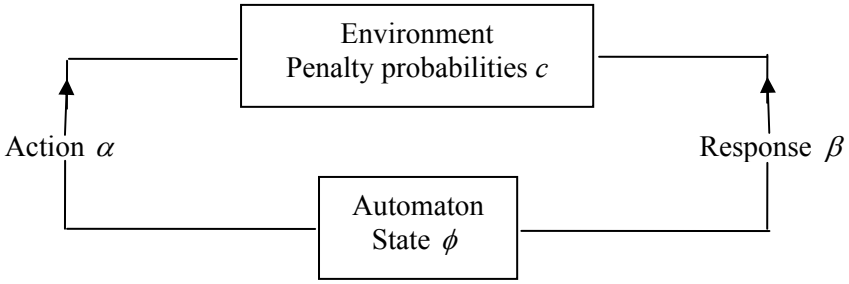



**Figure 1 The automaton and the environment**

In the stochastic automaton one of the two mappings *F* and *G* is stochastic. If the transition function *F* is stochastic, the elements $f_{ij}^{\beta}$ of *F* represent the probability that the automaton moves from $\phi_i$ to $\phi_j$ following an input $\beta$:

$$f_{ij}^{\beta} = Pr\{\phi(n+1) = \phi_j \mid \phi(n) = \phi_i, \beta(n) = \beta\} \quad i, j = 1,2,...s \quad \beta = \beta_1, \beta_2, ...\beta_m \tag{4}$$

For the stochastic mapping *G*, the definition is similar:

$$g_{ij} = Pr\{\alpha(n+1) = \alpha_j \mid \phi(n) = \phi_i\} \quad i, j = 1,2,...r \tag{5}$$

### 2.3. Reinforcement Schemes

A learning automaton generates a sequence of actions on the basis of its interaction with the environment. If the automaton is "learning" in the process, its performance must be superior to an automaton for which the action probabilities are equal. In general stochastic systems the action probabilities are updated at every state using a reinforcement scheme. Reinforcement schemes are categorized based on the behavior type they provide and the linearity of the reinforcement algorithm. In general terms a reinforcement scheme can be represented as

$$p(n+1) = T[p(n), \alpha(n), \beta(n)] \tag{6}$$

where *T* is mapping. If $p(n+1)$ is a linear function of $p(n)$, the reinforcement scheme is said to be linear, otherwise it is termed nonlinear. Since this study utilizes a linear reinforcement scheme with multi-actions, in Section 3.3 the update process of action probabilities in a linear environment is discussed in detail. This kind of learning scheme is called linear reward-penalty learning scheme and denoted by $L_{R-\epsilon P}$. For other reinforcement schemes please refer to Narendra and Thathachar (1989).

### 2.4. Convergence Properties

The concepts associated with the convergence of learning automata require sophistical mathematical tools, and the nature of convergence depends on the kind of reinforcement scheme employed.

The following discussion based on Narendra and Thathachar (1989) about the expediency and optimality of the learning provides some insights about these convergence issues. Average penalty received by the automaton is one of the useful quantities in understanding the behavior of the SLA. At any stage, $n$, if action $\alpha_i$ (departure-time choice $i$) is selected with probability $p_i$, the average penalty (reward) conditioned on $p(n)$ is given as follows:

$$\begin{aligned} M(n) &= E[\beta(n) \mid p(n)] = Pr[\beta(n) = 1 \mid p(n)] \\ &= \sum_{i=1}^{r} Pr[\beta(n) = 1 \mid \alpha(n) = \alpha_i] Pr[\alpha(n) = \alpha_i] \\ &= \sum_{i=1}^{r} p_i(n) c_i \end{aligned} \tag{7}$$

For the linear reward-penalty scheme $L_{R-\epsilon P}$, the pure chance automaton, $M_o$, is calculated as follows:

$$M_o = \frac{\sum_{k=1}^{r} c_k}{r} \tag{8}$$

The learning automaton must at least do better than pure chance automaton at least asymptotically as $n \to \infty$. This asymptotic behavior, defined in Definition 1 is called expediency.

*Definition 1*: A learning automaton is called expedient if there exists a $n_o$ such that for all $n > n_o$,

$$E[M(n) - M_o(n)] < 0 \tag{9}$$

Furthermore, if the average penalty is minimized by the proper selection of actions, the learning automaton is called optimal. This condition is defined in Definition 2.

*Definition 2:* A learning automaton is called optimal if

$$\lim_{n \to \infty} E[M(n)] = \min_i \{c_i\} \tag{10}$$

The multi-action automaton using $L_{R-\epsilon P}$ scheme is expedient for all initial action probabilities and in all stationary random environments.

# 3. SLA Travel Choice Model

Our approach to the problem of drivers' learning behavior makes use of SLA techniques described in the previous sections. We model drivers' day-to-day travel choices in a non-stationary environment. The aim is to design an automata system which can learn the best possible action based on the data collected from toll booths located at the exit locations of NJ Turnpike. The learning automata method we define will be useful in understanding the drivers' behavior in congested and constrained systems and how they adapt their choices as a response to a change in a system i.e., toll changes. The following sections describe the data sources used in this study, the proposed model, and experiment and test results.

## 3.1. NJ Turnpike

NJTPK is a 148 mile-toll road with 27 entry-exit points. The toll-road extends from Delaware Memorial Bridge in the South New Jersey (NJ) to George Washington Bridge in New York City (NYC). Since September, 2000, time-of-day pricing has been applied to encourage peak-period commuters to shift to off-peaks to reduce peak-hour congestion. Peak hour tolls are effective on weekdays from 7:00 to 9:00am and from 4:30 to 6:30pm. Only passenger cars with E-ZPass tag are eligible for time-of-day pricing, and cash users' tolls are higher than E-ZPass users irrespective of time-of-day. During peak hours and shoulders, more than 90% of the vehicles are passenger cars with E-ZPass (Ozbay et al. 2005). On January, 2003, toll levels for each period and vehicle type were increased.

## 3.2. Data Sources

The proposed SLA model is applied to the road section from Exit no 14 (Newark, NJ) and Exit no 18E (George Washington Bridge, NYC). The main reason to select this specific road section is that, NJ Turnpike road sections from an exit location to another exit location include both the demand between these two exit locations and the demand from that particular exit to other exit points which are located further away. Thus, any change in the latter demand will affect the traffic conditions in the selected road section. In order to minimize these outside effects we selected the road section from 14 to 18E, a section isolated from the other portions of NJ Turnpike i.e., the more than 90% of traffic volume observed on this section is due to the

demand between exits 14 and 18E. The traffic data used in training and testing of the SLA model include E-ZPass traffic volume and travel time information. The traffic volume data contain the traffic counts for each 15 min time interval between 6:30 am and 7:30 am from October-2002 to March-2003, three-months before and three-months after the toll increase at NJ Turnpike. The travel time data, on the other hand, not only include mean travel times for the corresponding time period, but also the individual travel times for 20 users at each 15 min interval throughout the time period. From this data, the mean and standard deviation of the travel times at each interval are obtained. Unfortunately, the traffic data do not include any information regarding users' desired/actual arrival times, early/late arrival amount to their locations. This crucial information about drivers' departure-time selection is obtained from descriptive analysis of traveler surveys conducted for NJ Turnpike users.

### 3.3. Model

For the proposed model, an input set $\boldsymbol{X}$ composed of the explanatory variables typically used in the utility functions is considered. Thus $\boldsymbol{X} = \{x_{n1}, x_{n2}, \ldots, x_{ni}\}$ where $n$ is the day, $i$ is the individual user. The output set $\boldsymbol{D} = \{d_1, d_2, d_3, d_4\}$ includes multiple actions composed of four choices (1: interval 6:30-6:45 am, 2: interval 6:45-7:00 am, 3: interval 7:00-7:15 am, 4: interval 7:15-7:30 am). Using the explanatory variables obtained from traffic and traveler survey data a user-specific travel cost function is derived. In terms of departure-time choice decisions, if at day $n$ the travel cost on a time-period is less than the travel costs on other periods, the algorithm increases the probability of choosing that time-period and decreases the probability of choosing other periods using reinforcement learning scheme. Due to non-stationary nature of the system at each day a different time-period can be the minimum travel cost choice. The estimated travel cost function is a function of value of travel time, travel time, value of reliability, travel time variance, value of early/late arrival, early/late arrival amount, and toll value from exit 14 to exit 18E. In this paper only departure-time choice decisions of NJ Turnpike users are addressed. In general, drivers choose among both different departure-times and different routes. However, for the NJ Turnpike case descriptive analysis of traveler surveys indicate statistically insignificant shift to other modes/routes from NJTPK (Ozbay et al. 2005), stating that NJTPK users make only departure-time choices depending on their valuation of travel costs.

$$\tau_{n,i,k} = VOTT_i tt_{n,i,k} + VOR_i \eta_{(n-1),i,k} + VOEA_i \gamma_{n,i,k} \delta_{n,i,k} + VOLA_i \varpi_{n,i,k} \xi_{n,i,k} + toll_k \qquad (11)$$

where;

$k$ = Departure-time interval index (1: pre-peak, 2: peak, 3: post-peak)

$\tau_{n,i}$ = Total travel cost for driver $i$ on day $n$ ($)

$VOTT_i$ = Value of travel time for driver $i$ ($/min)

$tt_{n,i}$ = travel time for driver $i$ on day $n$ (min)

$VOR_i$ = Value of reliability for driver $i$ ($/min)

$\eta_{(n-1),i}$ = Travel time variance experienced by driver $i$ up to day $(n-1)$ (min)

$VOEA_i$ = Value of early arrival for driver $i$ ($/min)

$\delta_{n,i,k}$ = Amount of early arrival experienced by driver $i$ on day $n$ when departure-time $k$ is selected (min)

$VOLA_i$ = Value of late arrival for driver $i$ ($/min)

$\xi_{n,i,k}$ = Amount of late arrival experienced by driver $i$ on day $n$ when departure-time $k$ is selected (min)

$Toll_k$ = Toll amount at departure-time interval $k$ ($)

$$\gamma_{n,i,k} = \begin{cases} 1 & \text{if the driver arrives early to the destination when he/she has chosen departure - time k} \\ 0 & \text{otherwise} \end{cases}$$

$$\varpi_{n,i,k} = \begin{cases} 1 & \text{if the driver arrives late to the destination when he/she has chosen departure - time k} \\ 0 & \text{otherwise} \end{cases}$$

In this paper, unlike the previous studies using constant values, the Bayesian Learning approach is combined with the SLA theory and for each individual driver, the parameters of the total cost function are sampled from probability distributions (Table 1). Each probability distribution is estimated using either traffic or traveler survey data. With the use of probability distributions, the stochastic nature of the system and heterogeneity among different drivers can be successfully included in the day-to-day learning and adaptation process of users.

Based on the proposed travel cost function, traffic volume data collected between October 2002 and November 2002 for each 15 min interval are used to calibrate the reward ($a$) and penalty ($b$) parameters of the SLA model. Unlike other SLA models in the literature, we assume that reward and penalty parameters are specific to each individual driver and follow a probability distribution.

**Table 1 Parameter information**

| Parameter | Distribution | Data Source |
|---|---|---|
| VOTT ($/hour) | Normal (18,4) | Traveler Survey (Ozbay et al., 2006) |
| VOR ($/hour) | Normal (10,3) | Assumed |
| VOEA ($/hour) | Normal (20,5) | Assumed based on Traveler Survey results |
| VOLA ($/hour) | Normal (20,5) | Assumed based on Traveler Survey results |
| Travel Time (min) | Normal | Traffic data (for each departure-time interval) |
| Total Trip Duration (min) | Normal (65, 15) | Traveler Survey Data |
| Desired Arrival time (time) | Normal (7:30, 75 min) | Traveler Survey Data |
| Toll ($) | constant | NJ Turnpike Authority website |
| Allowable time | (Desired arrival time) – (Departure time) | |
| Early/late arrival amount | (Allowable time) – (Total trip Duration) | |

In this study linear reinforcement scheme $(L_{R-\epsilon P})$ is considered for the proposed multi-action SLA model. For an *r*-action learning automaton, the general definition of $(L_{R-\epsilon P})$ can be obtained as follows:

$$\begin{aligned} & g_k(p(n)) = a \cdot p_k(n) \\ & h_k(p(n)) = \frac{b}{r-1} - b \cdot p_k(n) \\ & 0 < a, b < 1 \end{aligned} \tag{12}$$

where $g_k$ and $h_k$ ($k=1,2,...r$) are continuous, nonnegative functions with the following assumptions:

$$\begin{aligned} & 0 < g_k(p(n)) < p_k(n) \quad \forall p_k(n) \in (0,1) \\ & 0 < \sum_{k \neq i}^{r} [p_k(n) + h_k(p_k(n))] < 1 \quad i = 1,2...r \end{aligned} \tag{13}$$

Thus, the scheme corresponding to general linear schemes is given as:

If $\alpha(n) = \alpha_i$,

$$\begin{aligned} \beta(n) = 0 \quad \Rightarrow \quad & \begin{cases} p_j(n+1) = (1-a) \cdot p_j(n) & \forall \quad j \neq i \\ p_i(n+1) = p_i(n) + a \cdot [1 - p_i(n)] \end{cases} \\ \beta(n) = 1 \quad \Rightarrow \quad & \begin{cases} p_j(n+1) = \frac{b}{r-1} + (1-b) \cdot p_j(n) & \forall \quad j \neq i \\ p_i(n+1) = (1-b) \cdot p_i(n) \end{cases} \end{aligned} \tag{14}$$

where 0<a<1 is the reward parameter, and 0<b<1 is the penalty parameter of the reinforcement scheme.

**Example**: Suppose we have only 2 commuters (m= 1, 2) who make the following choices for 3 consecutive days, where during these days for commuters 1 and 2 minimum cost periods are period-3 and period-2, respectively. Moreover, assume after day 4 commuter 1 chooses departure-time periods 1, 2, 3 and 4 with probability 0.15, 0.15, 0.55, and 0.15, respectively, while commuter 2 chooses departure-time periods 1, 2, 3 and 4 with probability 0.15, 0.55, 0.15, and 0.15, respectively. (Table 2)

**Table 2 Sample choice set for commuter 1**

| Commuters | Commuter 1 | | | Commuter 2 | | |
|---|---|---|---|---|---|---|
| **Choice** | **Day 1** | **Day 2** | **Day 3** | **Day 1** | **Day 2** | **Day 3** |
| **Observed choice** | 2 | 1 | 3 | 3 | 4 | 2 |
| **Optimum choice** | 3 | 3 | 3 | 2 | 2 | 2 |

Based on this information, commuter *i*'s probability of selecting a period can be calculated for each day as shown in Table 3.

**Table 3 Update of Choice probabilities at each day**

| Commuter | Choice | Day 1 | Day 2 | Day 3 | Day 4 |
|---|---|---|---|---|---|
| **1** | $j \neq 3$ | 0.25 | $0.25 \cdot (1-b)$ | $0.25 \cdot (1-b)^2$ | $(1-a) \cdot (0.25 \cdot (1-b)^2)$ |
| | $j = 3$ | 0.25 | $\frac{b}{3} + 0.25 \cdot (1-b)$ | $\frac{b}{3} + (1-b) \cdot \left[\frac{b}{3} + 0.25 \cdot (1-b)\right]$ | $(1-a) \cdot \left(\frac{b}{3} + (1-b) \cdot \left[\frac{b}{3} + 0.25 \cdot (1-b)\right]\right) + a$ |
| **2** | $j \neq 2$ | 0.25 | $0.25 \cdot (1-b)$ | $0.25 \cdot (1-b)^2$ | $(1-a) \cdot (0.25 \cdot (1-b)^2)$ |
| | $j = 2$ | 0.25 | $\frac{b}{3} + 0.25 \cdot (1-b)$ | $\frac{b}{3} + (1-b) \cdot \left[\frac{b}{3} + 0.25 \cdot (1-b)\right]$ | $(1-a) \cdot \left(\frac{b}{3} + (1-b) \cdot \left[\frac{b}{3} + 0.25 \cdot (1-b)\right]\right) + a$ |

Since we know the true choice probabilities after day 4, we can find the parameters of reward (a) and penalty (b) distributions by maximizing the likelihood function given by:

$$max \qquad L(\theta; P) = \prod_{j=1}^{4} \prod_{i=1}^{2} g_{\theta}(p_{ij}) \tag{15}$$

In this expression, $\theta$ represents the unknown parameters that govern the distribution of choice probabilities as a function of the reward and penalty parameters which follow a particular probability distribution, $p_{ij}$ is the probability of selecting departure-time $j$ for commuter $i$ during day 4 (as shown in Table 3).

Unfortunately, this reinforcement learning scheme requires us to keep track of each commuter choice at each day of the analysis period, in order to update the probability values. However, our database does not provide any information regarding the choice of a particular commuter throughout the analysis period; rather it provides the resulting traffic volumes at each departure-time period for any day. Thus, it is not possible for the research team to calculate individual choice probabilities at each day $n$, and update the user probability values based on the proposed update mechanism. To overcome this drawback, instead of calculating individual choice probabilities, for any day, $n$, we calculate the probability of selecting a departure-time choice $j$ using the following equation:

$$p_j(n) = \frac{\sum_{i=1}^{T_n} choice_{ni}}{T_n} \tag{16}$$

where;

$p_j(n)$ = Probability of choosing departure-time $j$ during day $n$

$$choice_{ni} = \begin{cases} 1 & \text{if individual } i \text{ has selected choice } j \text{ during day } n \\ 0 & \text{otherwise} \end{cases}$$

$T_n$ = Total number of commuters at during day $n$

Then, training procedure determines the optimum probability distributions (and the parameters of the distributions) for reward and penalty parameters, and updates the action probabilities $p_j(n+1)$ at end of each day $n$ based on $L_{R-\epsilon P}$ scheme, such that at the end of the calibration process, the difference between the observed equilibrium departure-time choice probabilities and the calculated $p_j(N)$ values at day $N$ (last day of the calibration period) is minimized.

After calibrating the model parameters, the proposed SLA model is tested using traffic data collected on December 2002 when the system is at equilibrium, and data collected between January 2003 and March 2003 when the system is disturbed (i.e., toll level change on January 2003). Sensitivity analysis of the model both in equilibrium and disturbed conditions helps us to determine the performance of the proposed model under system changes, how drivers respond to these changes, and how long does it take for the model to converge.

### 3.4. Training Results

This section focuses on the training of the proposed SLA model, and estimation of reward and penalty distributions using October-November 2002 database. The calibration process resulted in the following distributions for reward ($a$) and penalty parameters ($b$):

$$\begin{aligned} a &\sim Normal(0.032, 0.005) \\ b &\sim Normal(0.0025, 0.001) \end{aligned} \tag{17}$$

On the basis of $L_{R-\epsilon P}$ scheme, values around 0.02 and 0.002 are suggested for $a$ and $b$ parameters, respectively for learning automata in various disciplines (Ozbay et al., 2001). However, as obtained from training results, for NJ Turnpike users reward and penalty parameters are higher and come from a distribution rather than constant values. These relatively high values for $a$ and $b$ parameters state that NJ Turnpike users are more risk prone users, and can adapt themselves to the changes in the system quickly. This is in fact an expected behavior for NJ Turnpike users since most of the E-ZPass users are frequent users of NJ Turnpike and are familiar with the daily traffic conditions. To this extent, proposed SLA model with these estimated parameters seem to accurately mimic the day-to-day behavior of NJ Turnpike users.

Figure 2 shows the mean standard deviations (MSD) for each day, calculated from the difference between observed traffic values and the assigned traffic volumes using the proposed SLA model. The mean standard deviations show that small portion of the days have high MSD (between 0.18 and 0.22), while on other days the error is around 0.10. To investigate the reason for these days with high MSD values, travel time distributions for the whole time period are analyzed, as shown in Figure 3. The analysis results indicate that on the days with high MSD, the previous day travel time values were exceptionally high compared with other days. Since the proposed learning model is mainly built on travel time, travel time values directly affect users' arrival time and consequently their departure-time choices. Thus, any unexceptional travel time value causes very high travel costs, and disturbs user-optimum departure-time choice behavior. Even though, the learning level is relatively low for these relatively high travel-time days, when the system returns back to the equilibrium state, the proposed model quickly adapts itself and in four to five days the MSD values drop down to values around 0.10, i.e. the model converges.

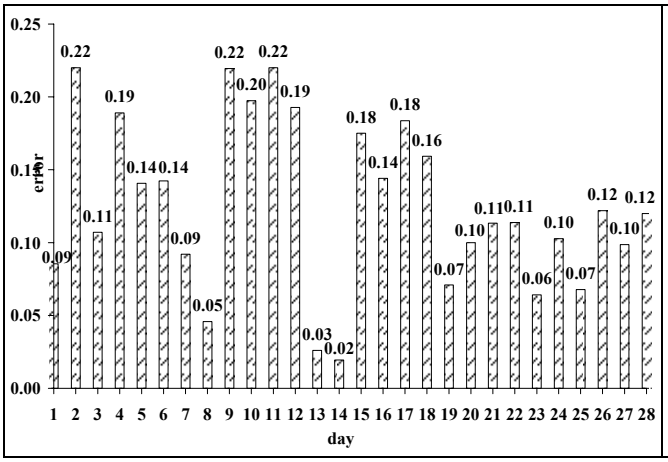



**Figure 2 Mean standard deviation**

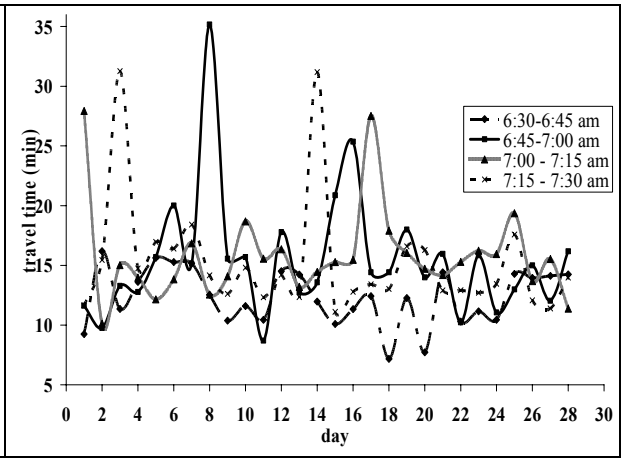



**Figure 3 Travel time distribution**

The overall calibration results show that proposed SLA model can successfully mimic the NJ Turnpike users' learning/adaptation behavior and can quickly adapt itself to the system disturbances. Next section focuses on the testing of the proposed SLA model using December 2002-March 2003 database.

### 3.5. Test Results

This section tests the estimated SLA model both on undisturbed (before the toll change) and disturbed network conditions (after toll change). Test cases focus on the NJ Turnpike users' responses to toll, travel-time, departure/arrival time restrictions while selecting their departure-times. Figure 4 shows the change in MSD from December 2002 to March 2003. The values indicate that during December 2002 the MSD values are around 0.1. On the other hand, between January 2003 and February 2003 we observe relatively high MSD values indicating lower learning levels for the SLA model. This low level of learning can be explained by the fact that the time period from January 2003 to February 2003 is right after the toll increase. Thus, the system is still fluctuating as a response to the new toll levels, and has not reached to equilibrium yet, which results in higher MSD values for the proposed SLA level. On March 2003 the system reaches a new equilibrium, and MSD values reduce up to 0.1, indicating that the SLA model successfully captures this new equilibrium and correctly mimics drivers' day-to-day behavior.

Table 4 presents the observed and estimated percent shares of traffic at each departure-time choice. The results indicate that on January 2003 and February 2003 converged probability values for departure-time choice are slightly different than the observed probability values. On the other hand, on December 2002, and March 2003, where the system is at equilibrium, the proposed SLA model successfully converged to true probability values.

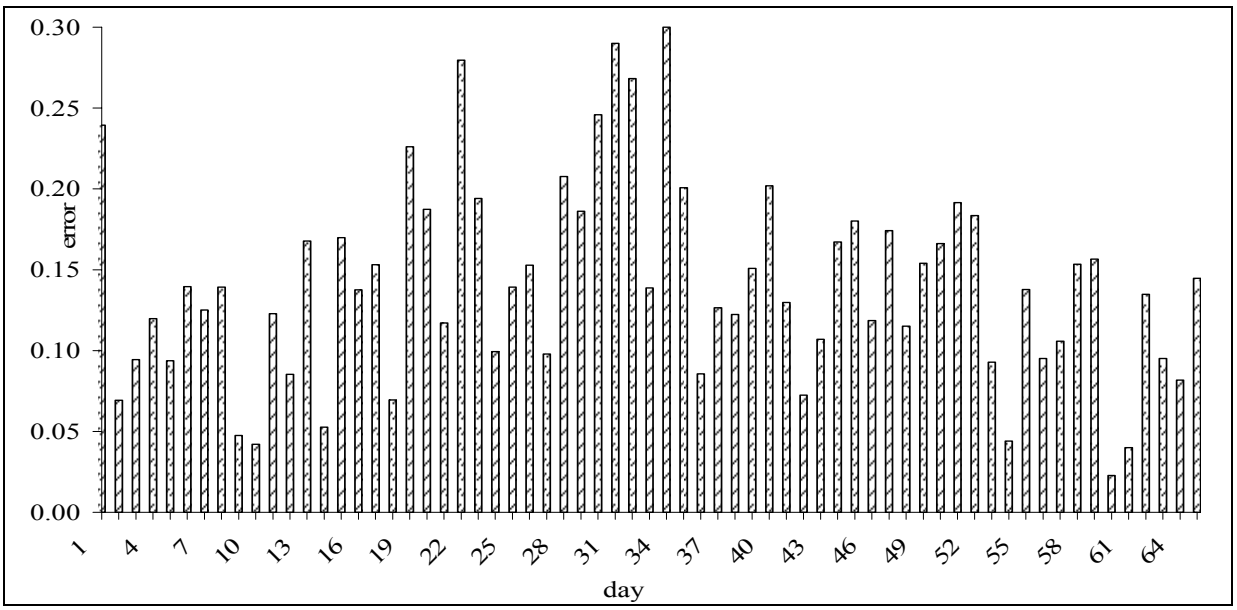



**Figure 4 Mean standard deviation, test data**

**Table 4 Percent share of traffic at different periods**

| | 6:30-6:45 am | | 6:45-7:00 am | | 7:00 - 7:15 am | | 7:15-7:30 am | |
|---|---|---|---|---|---|---|---|---|
| | Observed | Estimated | Observed | Estimated | Observed | Estimated | Observed | Estimated |
| **Dec-02** | 0.198 | 0.193 | 0.269 | 0.270 | 0.288 | 0.295 | 0.245 | 0.242 |
| **Jan-03** | 0.213 | 0.199 | 0.236 | 0.270 | 0.287 | 0.272 | 0.265 | 0.259 |
| **Feb-03** | 0.211 | 0.195 | 0.248 | 0.254 | 0.265 | 0.284 | 0.276 | 0.267 |
| **Mar-03** | 0.201 | 0.199 | 0.260 | 0.260 | 0.284 | 0.280 | 0.255 | 0.258 |

# 4. Conclusions and Discussions

In this paper, SLA theory is used to model commuters' day-to-day learning behavior, and to evaluate the effect of a feedback mechanism on decision-making under uncertainty. A linear reward-penalty reinforcement scheme $L_{R-\epsilon P}$ is proposed to represent day-to-day learning behavior of NJ Turnpike users both in undisturbed (before toll change) and disturbed network conditions (after toll change), and to investigate commuters' responses to toll, travel-time, departure/arrival time restrictions while selecting their departure-times.

The proposed learning model attempts to model drivers' day-to-day travel choices in a stochastic environment where multiple choices are available to the users. The following improvements are achieved compared with the past studies: (a) unlike previous studies in the literature, in this study reward and penalty parameters are assumed to be different for each commuter, and follow a probability distribution, rather than just constant values, (b) while determining the drivers' travel choices a wide variety of explanatory variables are considered other than the travel time, including value of travel time, travel time, value of reliability, travel time variance, value of early/late arrival, early/late arrival amount, and toll value (c) in order to consider the stochastic nature of the system and heterogeneity among different drivers in the context of day-to-day learning behavior, Bayesian Learning approach is combined with the SLA theory and for each individual driver, the parameters of the total cost function are sampled from probability distributions, (d) the proposed model is trained and tested using real traffic and traveler survey data.

The proposed SLA model is calibrated using NJ Turnpike E-ZPass traffic database (road section from Exit no 14 (Newark, NJ) and Exit no 18E (George Washington Bridge, NYC)) collected at each 15 minute time interval between 6:30 am to 7:30 am from October-2002 to November-2002. Using the explanatory variables obtained from traffic and traveler survey data, a user-specific travel cost function is derived. The user-specific travel cost function is found to be a function of travel time, early/late arrival amount, travel-time variance, value of travel time, value of early/late arrival amount, and value of travel time reliability. The calibration results show that proposed SLA model can successfully mimic the NJ Turnpike users'

learning/adaptation behavior and can quickly adapt itself to the system disturbances. Next, the calibrated model is tested using December 2002-March 2003 database in order to determine the performance of the proposed SLA model both for undisturbed (before the toll change) and disturbed network conditions (after toll change), and to investigate NJ Turnpike users' responses to toll, travel-time, departure/arrival time restrictions while selecting their departure-times. The test results show that the proposed SLA model can successfully converge to observed probability values in four to five days.

This study has attempted to gain insights into the learning/adaptation behavior of commuters in uncertain and dynamic environments. The experiment results have demonstrated the possibility of developing a psychological framework (i.e., learning models) as an alternative to represent traveler behavior.

## 5. References


1. Avineri E. and Prashker J.N. (2005) Sensitivity to Travel Time Variability: Travelers' Learning Perspective. *Transportation Research Part C*, **13,** 157-183.
2. Horowitz J.L. (1984) The Stability of Stochastic Equilibrium in a Two-link Transportation Network. *Transportation Research B*, **18**, 13-28.
3. Kahneman D. and Tversky A. (1979) Prospect theory: An analysis of Decisions under Risk. *Econometrica*, **47 (2),** 263-291.
4. Mahmassani H.S. and Chang G.L. (1985) Dynamic Aspects of Departure Time Choice Behavior in Commuting System: Theoretical Framework and Experimental Analysis. *Transportation Research Record*, **1037**, 88-101.
5. Senbil M. and Kitamura R. (2004) Reference Points in Commuter Departure Time Choice: A Prospect Theoretic Test of Alternative Decision Frames. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, **8**, 19-31.
6. Avineri, E. and J.N. Prashker (2003) Sensitivity to Uncertainty: The Need for a Paradigm Shift. *Transportation Research Record*, **1854**, 90-98.
7. Roth A.E. and Erev I. (1995) Learning in Extensive-form Games: Experimental Data and Simple Dynamic Models in Intermediate Term, *Games and Economic Behavior, Special Issue: Nobel Symposium*, **8**, 164-212.
8. Erev I. and Roth A.E. (1998) Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *American Economic Review*, **88(4)**, 848-881.
9. Schreckenberg M. And Selten R. (2004) Experimental Investigation of Day-to-Day Route-Choice Behaviour and Network Simulations of Autobahn Traffic in North Rhine-Westphalia. *Human Behaviour and Traffic Networks*, 1-22.
10. Miyagi T. (2005) A Reinforcement Learning Model for Simulating Route Choice Behaviors in Transport Network. *Proceedings of the 16th Mini - EURO Conference and 10th Meeting of EWGT.*
11. Selten R. Schreckenberg M., Pitz T., Chmura T. and Kube S. (2006) Experiments on Route Choice-Behavior. *Games and Economic Behavior*, accepted for publication.
12. March, J.G. (1996) Learning to be risk averse, *Psychological Review*, **10**, 309-319.
13. Jha, M., S. Madanat S. ad Peeta S. (1998). Perception Updating and Day-to-day Travel Choice Dynamics in Traffic Networks with Information Provision. *Transportation Research Part C: Emerging Technologies*, **6(3)**, 189-212.
14. Chen, R. B. and Mahmassani H.S. (2004) Travel Time Perception and Learning Mechanisms in Traffic Networks. [CD-ROM] *Transportation Research Board*, TRB, National Research Council, Washington, D.C.
15. Sutton R. S., and Barto S. (1998). *Reinforcement learning.* Cambridge, MA: MIT Press.
16. Narendra, K.S. & Thathachar, M.A.L. (1989). *Learning Automata*. Englewood Cliffs, N.J.: Prentice-Hall, Inc.
17. Ozbay K., Datta, A., and Kuchroo, P. (2001) Modeling Route Choice Behavior using Stochastic Learning Automata. *Transportation Research Record: Journal of the Transportation Research Board*, **1752**, 38-46
18. Ozbay K., Datta, A., and Kuchroo, P. (2001) Application of Stochastic Learning Automata for Modeling Departure Time and Route Choice Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, **1807**, 154-162.
19. Ozbay, K., J. Holguín-Veras, O. Yanmaz-Tuzel, S. Mudigonda, A. Lichtenstein, M. Robins, B. Bartin, M. Cetin, N. Xu, J.C. Zorrilla, S. Xia, S. Wang, and M. Silas. (2005) Evaluation Study of New Jersey Turnpike Authority's Time-of-day Pricing Initiative. *Publication, FHWA-NJ-2005-012.FHWA, U.S. Department of Transportation*.

# DOUBLY DYNAMIC SIMULATION MODEL FOR TRAFFIC ASSIGNMENT

*N.C.Balijepalli: Institute for Transport Studies, University of Leeds, England,* tra9bnc@leeds.ac.uk
*D.P.Watling: Institute for Transport Studies, University of Leeds, England,* tra6dwat@leeds.ac.uk
*R..Liu: Institute for Transport Studies, University of Leeds, England,* trarl@leeds.ac.uk

## Abstract

This research forms an important element in the context of developing unified traffic assignment models combining deterministic and stochastic approaches. Developing a *doubly dynamic* simulation assignment model involves specifying a *day-to-day* route choice model as a stochastic process combined with a driver learning and adjusting model, and a *within day* dynamic network loading model for moving the vehicles along the links of the network while capturing the interactions amongst vehicles that departed in the same/successive departure periods. Numerical tests aim to show the stationarity of the stochastic process, while illustrating the consistency of the link flow model results with properties such as First-In, First-Out (FIFO), in case of a network serving multiple origins-destinations for which the O-D demand varies over multiple departure periods.

## 1 Introduction

Traditionally, dynamic traffic assignment in the literature refers to the modelling of traffic flows on street networks due to the variations in the demand within a day, and capturing the spatio-temporal congestion effects through suitable dynamic link travel time functions. Usually such models are aimed at solving for either dynamic system optimal or dynamic user equilibrium, and as they consider the traffic flow as a deterministic variable the solutions naturally tend to be deterministic representing an average situation at each moment. Clearly, the within day deterministic models cannot explain the random variations in traffic flow, besides being unable to represent the transient states in the evolution towards equilibrium (Cascetta, 1989). In fact, the purview of dynamic traffic assignment is much wider and includes day-to-day variations in the demand. There are a few examples of pure day-to-day dynamic models such as Cascetta (1989), Watling (1996), Watling and Hazelton (2003), which considered the evolution of the traffic flow as a stochastic process, but are limited by the use of static within day cost-flow functions. However, a more generalised framework of assignment should include day-to-day and within day variations in traffic flows in order to be able to represent a realistic scenario, and such models are called *doubly dynamic traffic assignment* models; these models are the main subject of the present paper.

Cascetta and Cantarella (1991) developed such a doubly dynamic simulation model in which they defined the route flows on any day as a stochastic process, and included a queuing model to capture the delays on the links. Friesz (1996) also developed a doubly dynamic assignment model, but assuming deterministic flow variables, which carries with it some of the limitations described in the previous paragraph. In this research, we aim to develop a stochastic process model capable of modelling the variations in day-to-day and within day traffic. This paper is a continuation of Balijepalli and Watling (2005) in which the overall context of the research was explained, but for the benefit of unfamiliar readers a brief summary is included here.

The main aim of Balijepalli and Watling's (2005) research is to bring the deterministic and stochastic modelling approaches into a common modelling framework. This should be plausible because the deterministic model can be seen as a special case of a stochastic model where the variability element is assumed to be zero. Their proposition was based on earlier research results, notably those by Davis and Nihan (1993), which proves in their asymptotic results that a fairly broad class of stochastic process traffic assignment models converge to a stationary multivariate Normal distribution, approximately, and that the mean of this distribution is equal to the stochastic user equilibrium (SUE) flows. In order to describe a Normal distribution completely, one would require its mean and variance (i.e. variance-covariance matrix). Since the mean, represented by the SUE flows, can be estimated by a method such as the method of successive averages, we only need to estimate the variance to complete the approximation. Hazelton and Watling (2004) followed this lead and estimated the variance in the case of day-to-day dynamic but within

day static conditions. In order to include the within day variability, we need to extend their modelling framework to replace the static cost-flow functions with dynamic cost-flow functions. However, in such a case, working out some of the parameters required to proceed with the approximation, viz., jacobians of the travel time functions with respect to the path inflows gets increasingly complicated, and Balijepalli and Watling (2005) includes a detailed specification for analytically computing the required jacobians.

Given the time varying demand profile and the network specification, the expected output of the stochastic process model includes, during each departure period within a day, the mean traffic flow and the variance of route flows when the process is stationary. Thus this research provides an initial step in advancing our understanding of the variability of traffic flows on street networks and more importantly combines the deterministic and stochastic models - the two seemingly parallel approaches to carrying out the traffic assignment. The specific aims of this paper are to specify a stochastic process model for the route choice process incorporating the drivers' day-to-day learning and adjustment process through a simple weighted averaging method, combined with a continuous time dynamic network loading method to obtain the drivers' experienced travel costs. It is aimed that the doubly dynamic traffic assignment is solved using a simulation framework. A simple grid network with multiple origins and destinations will be used to illustrate the principles described.

# 2 Methodology

Consider a network of directed links serving O-D demand represented by $\mathbf{Q} = \{...., q_k, ...\}$ where $q_k$ is the O-D demand for a particular *commodity* k, each commodity defining a combination of origin, destination and (discrete) departure period, it being assumed that the total period of analysis is divided into L departure periods. Each commodity k is served by a set of routes $R_k$ with $|R_k|$ elements; the full route set across all commodities thus has dimension $\rho = \sum_{k=1}^{K} |R_k|$. Let **f** be the ρ- vector of commodity route flows and **c**(**f**) be the vector of commodity route costs.

It is assumed that all the trip makers of commodity k are rational in their behaviour when choosing their route, in an attempt to minimise their perceived cost of travel. For each commodity k and route $r \in R_k$, the perceived travel cost $\hat{C}_r^{(n)k}$ at the start of day k is given by

$$\hat{C}_r^{(n)k} = C_r^{(n-1)k} + \eta_r^{(n)k} \qquad (1)$$

where $C_r^{(n-1)k}$ is the population-mean perceived cost for commodity k and route r at the end of day n-1, and $\eta_r^{(n)k}$ is a random variable describing unobserved attributes contributing to the population-dispersion of the perceived attractiveness of route r by commodity k. The $\rho$-vector $\mathbf{C}^{(n-1)}$ represents the collection of population-mean perceived costs across all routes and commodities. The probability of choosing route r on day n is then given by:

$$\mathrm{p}_r^k(\mathbf{C}^{(n-1)}) = \mathrm{Prob}\left(C_r^{(n-1)k} + \eta_r^{(n)k} < C_i^{(n-1)k} + \eta_i^{(n)k}\right) \qquad \forall i \neq r \qquad (2)$$

$\mathbf{p}^k(.)$ then represents the vector (of dimension $|R_k|$) of route choice probabilities for the commodity k, and $\mathbf{p}(.)$ denotes the collection of these choice probability vectors over all the commodities (i.e. p(.) is a vector of dimension ρ). The functional form of the path choice probabilities depends on the joint probability density function assumed for the residuals $\{\eta_r^{(n)k} : r \in R_k\}$ for each commodity k, resulting (for example) in a logit

model, if independent Gumbel distributions are assumed, and a probit model for a multivariate normal distribution.

While the behavioural choice-side of the model is quite conventional, a simple linear learning filter is used to replicate drivers building up their experience of travel costs on a day-by-day basis following the completion of each day's trip. In this research, we assume a simple weighted average approach akin to many other simulation experiments, for example, Horowitz (1984), Cascetta (1989) and Nakayama et al (1999). Thus following the completion of trips on any day (n-1), the population-mean perceived costs are updated based on a weighted average of costs actually incurred in a finite number of previous days m, using the form:

$$\mathbf{C}^{(n)} = s(\lambda)^{-1}\left\{ \mathbf{c}(\mathbf{F}^{n-1}) + \lambda\mathbf{c}(\mathbf{F}^{n-2}) + .... + \lambda^{m-1}\mathbf{c}(\mathbf{F}^{n-m}) \right\} \qquad \forall\ 0 < \lambda < 1 \tag{3}$$

$$s(\lambda) = \sum_{j=1}^{m} \lambda^{j-1} = (1-\lambda^{m})/(1-\lambda) \tag{4}$$

where s(λ) is simply a scaling factor to make the weights sum to unity and **c**(.) is the commodity route cost-flow function as defined above, and where $\mathbf{F}^{n}$ is a vector random variable of dimension ρ denoting the network path flows by commodity on day n. Assuming that for any day n and for each commodity k, all $q^k$ drivers wishing to travel make their travel choices independently, conditional on their experiences in past days, then the number of drivers taking each possible route on day n by each commodity k, conditional on the costs (3) experienced in the past, is obtained as:

$$\mathbf{F}^{(n)k} \,\big|\, \mathbf{C}^{(n-1)} \sim Multinomial(q^k, \mathbf{p}^k(\mathbf{C}^{(n-1)})) \quad \text{independently for k = 1,2,,….K} \tag{5}$$

where $\mathbf{F}^{(n)k}$ is the vector of route flows on day n by the commodity k.

In order to be able to capture the interactions amongst the vehicles departing in the same/successive departure periods, we need to subdivide each departure period into a number of smaller time steps. Let δ be the time increment of this discretisation, and denote the complete analysis period by (0, Nδ] for some positive integer N. The time increments are thus the intervals (t-δ, t] for t = δ, 2δ, …, Nδ, which are referred to as minor time steps. Below, when we refer to a time step (or interval) t, it is to be understood that we are referring to the period (t-δ, t]. We assume that δ is chosen so as to be smaller than the free flow time to traverse any link. This is an assumption we shall use implicitly on a number of occasions – implying that a vehicle could not enter and exit a link in the same increment of time. The OD demand rates are assumed (for notational convenience) to be specified over a common discretisation of the whole analysis period (0,Nδ], divided it into L major time periods, also referred to as departure periods ($w_{j-1}$, $w_j$] (for j = 1,2,,..L) such that $(w_0, w_1] \cup (w_1, w_2] \cup ..... \cup (w_{L-1}, w_L] = (0, N\delta]$. These match exactly the departure periods defined in the previous section, and for convenience are assumed to be of the same duration, i.e. $w_j - w_{j-1} = \kappa$ for all $j = 1,2,...,L$ and some given $\kappa$.

Let A be the set of links on the network, such that A = {…., $a_i$, …} for i = 1,2,… Assuming that whole link travel time models of linear form (Friesz 1993) are defined on each of the links on route r, the travel time function and exit time functions for any link $a_i$ may be expressed as a nested path cost operators. Then the expressions for travel time and the exit time are as given below:

$$\tau_{a_i}(g_{a_{i-1}}(t)) = \alpha_{a_i} + \beta_{a_i}(x_{a_i}(g_{a_{i-1}}(t))) \tag{6}$$

$$g_{a_i}(t) = g_{a_{i-1}}(t) + \tau_{a_i}(g_{a_{i-1}}(t)) \tag{7}$$

where, $\tau_{a_i}(.)$ is the travel time on the link $a_i$ , $\alpha_{a_i}$ is the free flow time on the link, $\beta_{a_i}$ is the inverse of the exit capacity of the link $a_i$, $x_{a_i}(.)$ is the number of vehicles on the link $a_i$ , and $g_{a_i}(.)$ is the exit time from the link $a_i$ .

As the model discretises time into a finite number of minor time steps, we have the knowledge of travel times computed only at the discrete time steps. But this will be insufficient to compute the path travel time on any path with multiple links, especially from the second link onwards where the travel time needs to be computed at some real time and not just integers. We propose to counter this by computing the travel time in equation (6) using linear interpolation, which is given below:

$$\tau_{a_i}(t) \approx \hat{\tau}_{a_i}(<t/\delta>.\delta) + \frac{t - <t/\delta>.\delta}{\delta}\left[\ \hat{\tau}_{a_i}((<t/\delta>+1).\delta) - \hat{\tau}_{a_i}(<t/\delta>.\delta)\ \right] \quad (8)$$

for, $(t \geq 0;\ i = 1,2,...,n)$

where, $\hat{\tau}_{a_i}(.)$ = travel time on link $a_i$ at integer time, and

<t/δ> = integer part of time t.

Then for example, the path travel time for vehicles entering the link $a_1$ at time t on route r (with $a_n$ being the last link on route r before discharging the vehicles to the their destination) is simply given as the difference between the exit time from link n and the entry time at the origin, expressed as $[g_{a_n}(t) - t]$.

$$c(t) = [g_{a_n}(t) - t] \quad (9)$$

Finally, the departure time dependent mean travel time for route r with uniform inflow rate in any departure time period T bounded by $(w_{j-1}, w_j]$, may be expressed as,

$$c_r^T = \left\{\ \frac{1}{(w_j - w_{j-1})}\ \right\} \sum_{t=(j-1)n\delta+1}^{jn\delta} c(t) \quad (10)$$

where, n is the number of minor time steps in major time period T. Departure time dependent mean travel time obtained from (10) is used for updating the drivers' memorised travel cost in (3) which then is used to work out the route choice for the following day.

## 3 Numerical Example

In order to illustrate the principles described in the previous section, a simple grid network of 12 links serving two origins and three destinations is used (Figure 1). Note that all the links are one-way, and there are 14 routes in all and the link-path incidence is shown in Table 1. It is assumed that dynamic linear travel time functions with parameters shown in Table 2 are defined on all the links of the network. The demand for each of the six possible O-D pairs is assumed to be known in each departure period, and is as shown in Figure 2. The route choice is assumed to follow the logit principle with the dispersion parameter θ = 0.1, unless otherwise mentioned. In this example, we included four departure periods of 15 minutes each, and we assumed a minor step length of one minute each. Drivers were assumed to remember up to a couple of days, and the memory weight was taken to be 0.5.

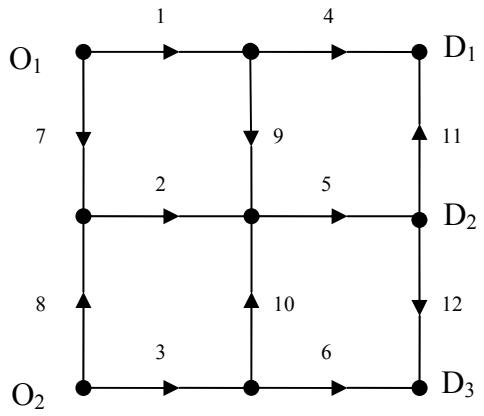



**Figure 1 Grid Network**

**Table 1 Link – Path Incidence**

| OD Pair | Path | Links |
|---|---|---|
| $O_1$-$D_1$ | 1 | 1-4 |
| | 2 | 1-9-5-11 |
| | 3 | 7-2-5-11 |
| $O_1$-$D_2$ | 4 | 1-9-5 |
| | 5 | 7-2-5 |
| $O_1$-$D_3$ | 6 | 1-9-5-12 |
| | 7 | 7-2-5-12 |
| $O_2$-$D_1$ | 8 | 8-2-5-11 |
| | 9 | 3-10-5-11 |
| $O_2$-$D_2$ | 10 | 8-2-5 |
| | 11 | 3-10-5 |
| $O_2$-$D_3$ | 12 | 3-6 |
| | 13 | 8-2-5-12 |
| | 14 | 3-10-5-12 |

**Table 2 Network Link Parameters**

| Link | Free flow time, $\alpha_a$ minutes | Service Rate, $\beta_a$ minutes/vehicle | Exit Capacity, Vehicles/hour |
|---|---|---|---|
| 1 | 6 | 0.025 | 2400 |
| 2 | 4 | 0.040 | 1500 |
| 3 | 5 | 0.029 | 2069 |
| 4 | 4 | 0.021 | 2857 |
| 5 | 5 | 0.015 | 4000 |
| 6 | 5 | 0.030 | 2000 |
| 7 | 3 | 0.018 | 3333 |
| 8 | 2 | 0.024 | 2500 |
| 9 | 4 | 0.019 | 3158 |
| 10 | 3 | 0.022 | 2727 |
| 11 | 6 | 0.01 | 6000 |
| 12 | 5 | 0.01 | 6000 |

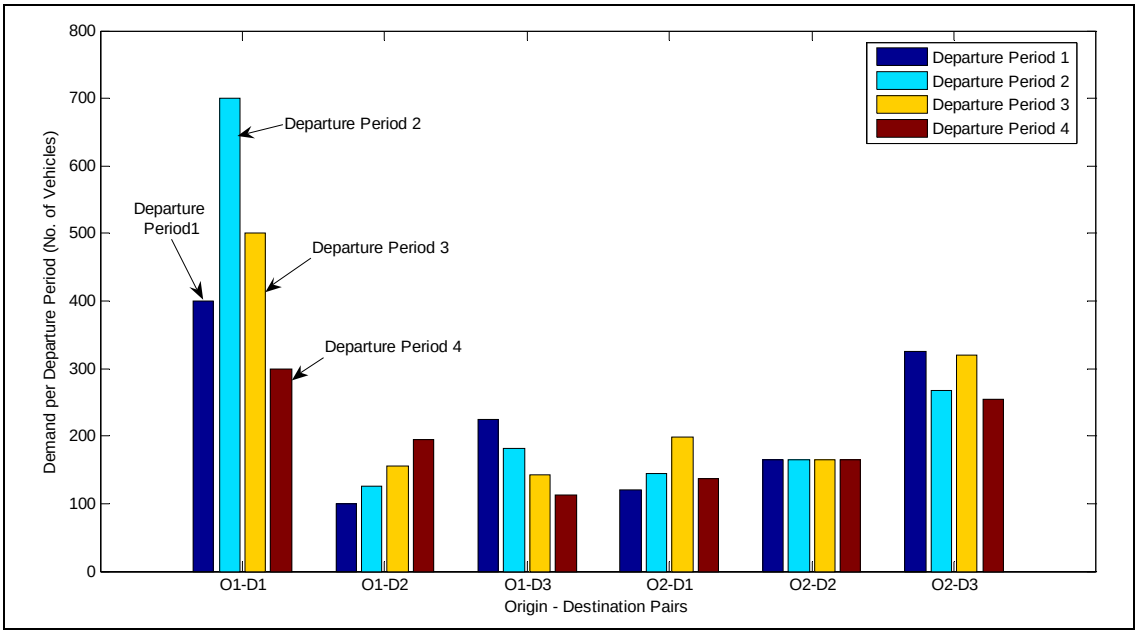



Figure 2 Origin–Destination Demand

## 3.1 Comments on day-to-day results

Total travel time measured by the vehicle-hours on the network indicates the intensity of travel over the network, and if monitored over the period of simulation, will indicate the day-to-day evolution of the intensity of travel. Figure 3 shows the plot of total travel on the network. It indicates that in a realisation of 500 days, total travel time on the network settles down to about its mean value (= 3079 veh-hrs), with a standard deviation of 18.4 veh-hrs. While monitoring the total travel on the network, an initial burn in period equivalent to 10% of the simulated days has been discounted. Figure 4 shows the day-to-day evolution of travel over 1000 days and provides some visual reassurance that the process is stable. In order to further ensure that the stochastic process is stable, we have analysed the autocorrelations of route flows based on 1000 simulated days. Autocorrelations are expected to die down (and approach their equilibrium levels) with larger lags for a stationary series and indicate that the random variable under consideration is stable about its mean value. Figure 5 shows the autocorrelations in path flows on routes 1,2 and 3 for up to 15 days of lag over a realisation of 1000 days. As the correlation of the flows with themselves is unity, the first bar (with '0' lag) reflects the same. From then on, the autocorrelations can be observed to reduce with increasing lags. Insignificant autocorrelations compared to standard errors at some lag k>0, indicate that the flows on any route do not depend on the flows on the same route beyond k days, during the same departure period. This condition also implies that the process is stationary. Figure 5 includes error bars (based on Bartlett's formula for large lag standard error) for each of the routes 1,2 and 3, for some lag k>0 beyond which the theoretical autocorrelation function deemed to have died out. However, a more affirmative test of stationarity of the time series would be that the determinant of the autocorrelation matrix and all the minors should be greater than zero, thus requiring a large number of conditions to be satisfied. It can also be commented that the route flows in departure period 1 settle down relatively quickly, compared to the flows over the rest of the departure periods, this is intuitively supportive to the notion that the delays in later time periods are affected by the flows from the earlier time periods, and hence the flows in later time periods settle down much slower compared to the flows in earlier departure periods.

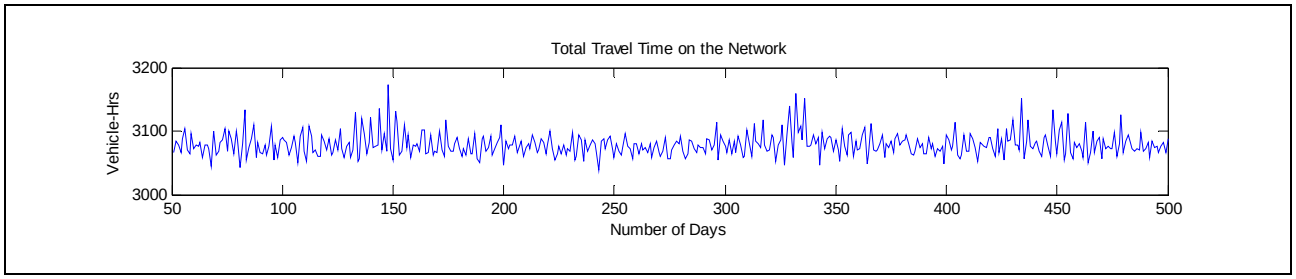



Figure 3 Total Travel time on the Network over 500 Days

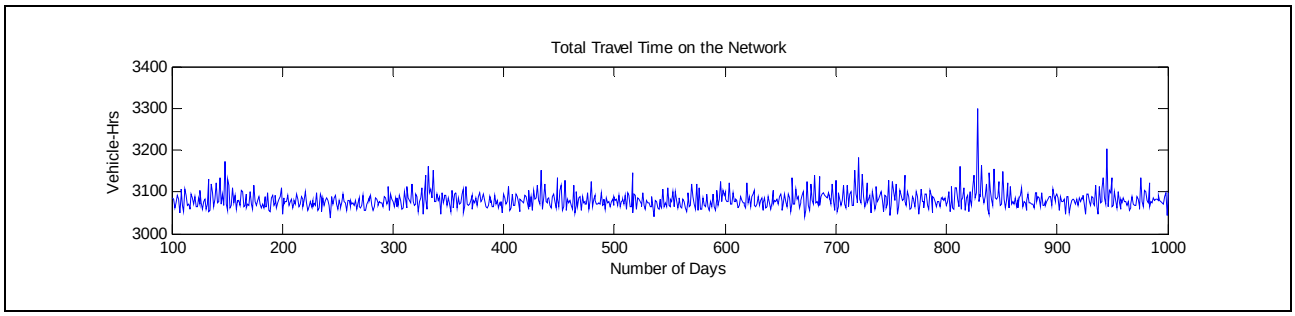



Figure 4 Total Travel time on the Network over 1000 Days

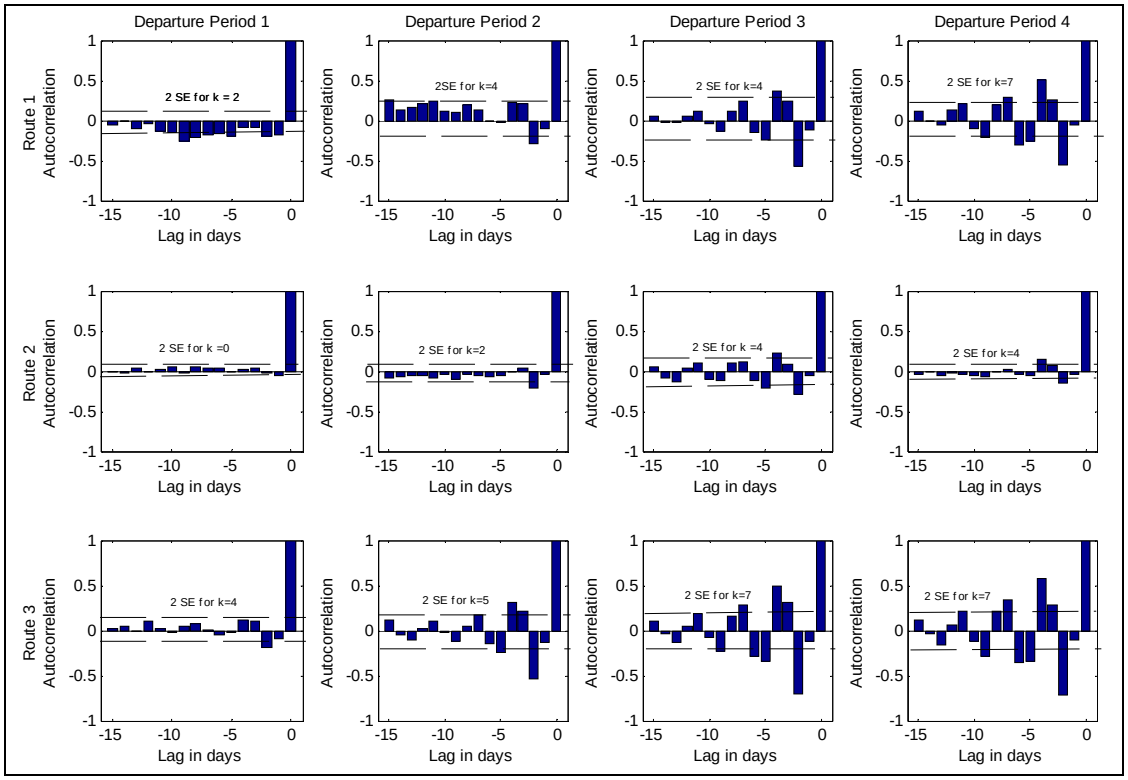



Figure 5 Correlogram for Flows on Routes 1, 2 and 3

## 3.2 Comments on within day results

Figure 6 shows the link-time plot for routes 1,2 and 3 in each of the departure periods. They indicate the travel times are fanning out in general, meaning that the congestion builds up as we progress with the dynamic loading of vehicles over the network. Especially on links 1 and 2, this phenomenon is very clear. On the other hand, parallel travel time lines indicate that the links are uncongested and operate below the capacity, as is the case with most of the links on routes 1,2 and 3. Figure 6 also indicates that the model results are consistent with FIFO property as we do not have any intersecting link travel time lines. The figure is also indicative of satisfying the FIFO property at the path level. As link 2 is used by several paths (see Table 1), in order to illustrate the dispersion of outflows over larger periods than the inflow periods the link inflow and outflow profiles have been drawn (Figure 7). Figure 7 illustrates that the vehicles on link 2 no longer operate under free flow speeds and started experiencing higher travel times due to the increase in the level of congestion.

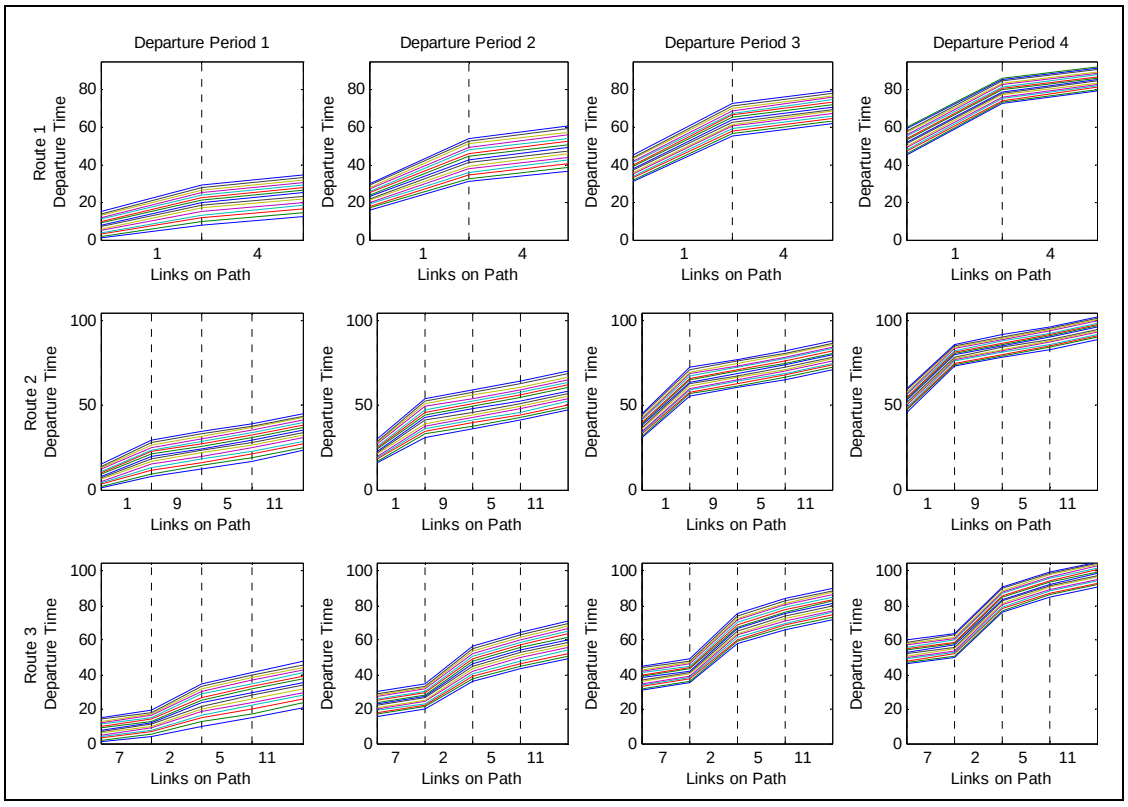



Figure 6 Link-Time Plots for Routes 1,2 and 3

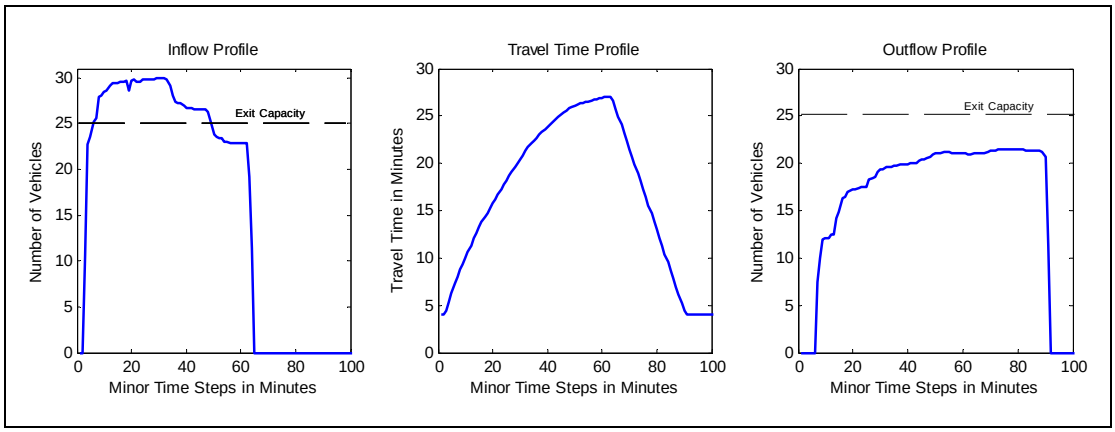



Figure 7 Inflow and Outflow Profiles for Link 2

# 4 Conclusions

This research is an important element in the context of developing a unifying modelling framework combining deterministic and stochastic assignment approaches. In specific, simulation modelling offers a benchmark against which the results of the new paradigm can be compared. In its own right, simulation modelling provides solutions to complex traffic assignment problems, such as the doubly dynamic traffic assignment described in this paper, through a fairly simple and transparent process. However, the main difficulty lies in interpreting the volume of results, and also requires a high level of expertise in identifying the main features of the results. For example, identifying the stability of the stochastic process could be tricky. Simulation models require very long and intense computational processing to solve the real life practical problems. For example, for the 12-link grid network described in section 3, when a higher degree of stability was sought by increasing the length of the realisation, the computer program ran for over three hours on a machine with Pentium D processor with 3.0 GHz having 1GB RAM. Precisely, this is the

motivation behind the attempts of developing variance approximation method for the stochastic process (Hazelton and Watling 2004, Balijepalli and Watling 2005).

## Acknowledgement

The first author gratefully acknowledges the University of Leeds Centenary Chair Research Project for funding this part of the research work.

## References

Balijepalli, N.C. and Watling, D.P. (2005) 'Doubly Dynamic Equilibrium Distribution Approximation Model for Dynamic Traffic Assignment', in Transportation and Traffic Theory: Flow, Dynamics and Human Interaction, H.Mahmassani (Ed), Elsevier, Oxford, UK. pp 741-760.

Cascetta, E. (1989) A stochastic process approach to the analysis of temporal dynamics in transportation networks, Transportation Research B, 23(1), 1-17.

Cascetta, E. and Cantarella, G.E. (1991) A Day-to-day and Within-day Dynamic Stochastic Assignment Model, Transportation Research A 25(5), 277-291

Davis, G.A. and Nihan, N.L. (1993) Large Population Approximations of a General Stochastic Traffic Assignment Model, Operations Research 41(1), 169-178.

Friesz, T.L., Bernstein, D., Smith, T.E., Tobin, R.L. and Wie, B.W. (1993) A Variational Inequality Formulation of the Dynamic Network User Equilibrium Problem, Operations Research 41(1), 179-191.

Friesz, T.L., Bernstein, D., Stough, R. (1996) Dynamic Systems, Variational Inequalities and Control Theoretic Models for Predicting Time-Varying Urban Network Flows, Transportation Science, 30(1), 14-31.

Hazelton, M. and Watling, D. (2004) Computation of Equilibrium Distributions of Markov Traffic Assignment Models, Transportation Science 38(3), 331-342.

Horowitz, J.L. (1984) The stability of stochastic equilibrium in a two-link transportation network, Transportation Research B 18(1), 13-28.

Nakayama, S., Kitamura, R., and Fujii, S. (1999) Driver's learning and network behaviour: dynamic analysis of the driver-network system as a complex system, Transportation Research Record 1676, 30-36.

Watling, D.P. (1996). Asymmetric problems and stochastic process models of traffic assignment, Transportation Research B 30(5), 339-357.

Watling, D.P. and Hazelton, M. (2003) The Dynamics and Equilibria of Day-to-day Assignment Models, Networks and Spatial Economics, 3(3), 349-370

# THE DYNAMIC ASSIGNMENT OF TOURS IN CONGESTED NETWORKS WITH PRICING

*JW Polak: Imperial College London, England* *j.polak@imperial.ac.uk*
*BG Heydecker: University College London, England* *ben@transport.ucl.ac.uk*

## ABSTRACT

Models of route and departure time choice for individual trips through congested networks can provide insight into the dynamics of peak periods and sensitivity of travellers' behaviour in response to a range of transport policy measures, including time-varying pricing. The consequence of travellers' scheduling process is an inter-linked spatial and temporal pattern of activity participation and travel episodes. The degree to which travellers will adjust the timing of a particular trip will in general depend on the overall structure of the activity pattern in which it is embedded. Moreover, the impact of a policy measure, such as time-dependent road pricing, affecting one part of the activity pattern may have ramifications elsewhere in the pattern. In the present work, we explore a model that considers the choice of the schedule of a tour that visits several locations, linked by a congestable network. Traveller's scheduling decisions are made on the basis of a trade off between the utility associated with participating in activities at successive stops on the tour, taking into account and the disutility of travel in the network between the stops. Under mild hypotheses, the equilibrium conditions for this model are identified and a number of properties explored.

## 1 Introduction

The decision entailed in undertaking a trip from one location to another is influenced by the benefits that would be gained by remaining at the origin, the conditions that are encountered in the making the trip, and the benefits gained through arriving at the destination. A traveller who undertakes a trip will generally do so because the benefits of being at the destination outweigh the losses associated with absence from the origin and the costs of making the trip. In making a trip, travellers who use public access transport systems of a kind that can become congested at peak times will incur additional travel costs because of that. Thus the collective behaviour of travellers influences the conditions experienced by each of them in travelling.

From the point of view of the individual travellers, the requirement to travel arises from the range of locations at which different activities can be undertaken. This view of travel leads to a microscopic analysis of trip-making behaviour by individuals. On the other hand, the collective effect of travellers can be to cause congestion that will impact on the travellers themselves through increased journey times and decreased convenience of travel. This view leads to a macroscopic analysis of travel in congested networks.

The present analysis integrates the two distinct approaches of utility-based analysis for activity participation and scheduling (building on the work of Polak and Jones, 1994; Ashiru *et al*, 2004 and Ettema *et al*, 2004) and equilibrium-based departure-time choice for travel (building on the tradition of Vickrey, 1969; Arnott *et al*, 1993 and others). The benefit of undertaking the activities forming a tour is represented in the form of utility that depends on the timing and duration of participation at specific locations, whilst the cost of travel between locations includes congestion delays that are incurred as a consequence of the need to travel during peak periods. The analysis of tours is based upon consideration of the individuals' needs for travel, and hence is microscopic in nature. The analysis of departure time choice is based upon consideration of the effects of congestion in the network, and hence is macroscopic in nature. Here, we show how these two distinct approaches to the analysis of travel behaviour can be combined into a self-consistent model. This brings a utility-based analysis of tours that describes the requirement to travel together with an equilibrium-based analysis of the choices that are available to travellers, and in particular those of departure-time and route.

According to this model, benefits (and also some kinds of costs, which can be accommodated directly in the present analysis) derive from time spent at the different locations, whilst costs are incurred through travel between them. This represents the role of travel as a means to the end of gaining access to facilities at a range of locations and hence emphasises its nature as a derived demand. In this model, we represent the benefit of attendance in the form of utility that depends on timing and duration of attendance. The resulting analysis is based upon consideration of the individuals' requirements for travel, and is framed in terms of the benefits that they gain through having travelled. We represent the cost of travel between locations, including congestion delays that are incurred as a consequence of travel during peak periods, in the form of travel time. The present approach combines these elements in a single framework of travel equilibrium, for which we present analytical results. This model and analysis can be used to investigate the effects of transport policy interventions, such as road pricing, on travel patterns. Here, we apply them to a simple example and show how in equilibrium, different travellers can achieve identical net utility through different combinations of utility and travel cost by scheduling their travel at different times. We then examine the effect on individual travel behaviour of introducing a toll charge that eliminates congestion, and explore how the societal benefits of this are represented in the present model.

# 2 A model of trip making

## 2.1 Introduction

The problem of unifying models of travel with those of activity patterns has been recognised in the literature. Kitamura (1988) noted that although the two kinds of model are complementary, establishing an integrated formulation would be challenging. Oppenheim (1995, 300-30) discussed combined formulations of activity at different locations and travel between them using discrete choice models for the locations together with static assignment to routes between the locations. Recker (1995) developed a discrete choice formulation of

this that includes departure time choice and congested travel with activities that have representative utilities that are constant over time. Lam and Yin (2001) extended this to time-dependent utilities at the different locations, and developed a discrete choice framework in discrete-time to select the sequence of activities. They adopted a variational inequality–based formulation of dynamic equilibrium assignment of traffic between locations and iterated to achieve mutual consistency between activity choices and consequent travel times. Zhang, Yang, Huang and Zhang (2005) developed a discrete-choice modelling formulation to calculate stationary distributions of departure times for journeys through a congested network.

Here, we consider a population of travellers that is homogeneous in respect of their travel needs and their trip making decisions. We suppose that each of these travellers undertakes a tour that starts and ends at the same location (home) and visits a series of locations that we take in the first instance to be predetermined. The timing of these trips depends on the timeliness of attendance at each of the origin and the destination locations in respect of the benefits that accrue to the individuals. The duration of each trip depends on the traffic conditions that are encountered, and can be estimated by use of a traffic model. We suppose that travel between locations is undertaken when the benefit of attendance at the destination surpasses that of remaining at the origin when allowance is made for the time and cost of travel. The timing of the trip is then a resolution of the tension between the benefits for the individual of being at the origin and at the destination. This is balanced against the cost of travel through the network, which varies according to the congestion caused collectively by travellers. Hence the departure rates and consequent levels of congestion in the network are endogenous to the present analysis.

### 2.2 Analysis of trip-making

Consider a single trip $j$ $(1 \leq j \leq J)$ made by a traveller that forms part of the day-long tour that include $J$ trips and up to that many locations. Suppose that the traveller departs from location $j - 1$ on trip $j$ at time $s_j$ and consequently arrives at location $j$ at time $\tau_j(s_j)$ so that the duration of this trip is $\tau_j(s_j) - s_j$. The arrival time for the trip is determined from the departure time by use of a traffic model according to the conditions that are encountered. The total travel time during a tour of this kind is then $\Sigma_j [\tau_j(s_j) - s_j]$. If a time-dependent toll $c_j(s_j)$ (expressed in equivalent travel time) is charged for starting trip $j$ at time $s_j$, then the total cost incurred during a tour of this kind is $\Sigma_j [\tau_j(s_j) - s_j + c_j(s_j)]$.

Following Polak and Jones (1994) we suppose that the time from $t$ to $s$ spent at location $j$ confers a benefit to an individual, which we represent as $f_j(t, s)$. For convenience, we express this in terms equivalent to savings in travel time. This benefit can depend separately and jointly on each of the start time $t$, the duration of attendance $s - t$, and the end time $s$. In a tour, the start time $t$ at location $j$ is given by the arrival time $\tau_j(s_j)$ of journey $j$, and the end time $s$ is given by the departure time $s_{j+1}$ of journey $j+1$. Thus the benefit derived from attendance at location $j$ is $f_j[\tau_j(s_j), s_{j+1}]$ and the benefit accumulated during a tour of this kind is $\Sigma_j f_j[\tau_j(s_j), s_{j+1}]$.

The net benefit to an individual of undertaking a tour of this kind with departure times **s** is then

$$V(\mathbf{s}) = f_0[0, s_1] + \sum_{j=1}^{J} f_j[\tau_j(s_j), s_{j+1}] - [\tau_j(s_j) - s_j] - c_j(s_j) \tag{1}$$

where by convention we set $s_{J+1}$ = 24 h so that the final part of the day is spent at location $J$.

Variations in departure time $s_j$ will affect the benefit that is obtained at the origin location $j-1$, the toll charged, the duration of the journey and hence its cost, and the arrival time at the destination location and hence the benefit obtained there.

### 2.3 Analysis of network equilibrium

Suppose that the departure rate of individuals on trip $j$ ($1 \le j \le J$) at time $s$ is $e_j(s)$. In equilibrium, the value $V(\mathbf{s})$ achieved by each individual is identical – otherwise, some would have an incentive to change their departure times. Following Heydecker and Addison (2005), we note that while the departure rate is non-zero, the value of net utility $V$ is invariant with respect to time so that

$$e_j(s_j) > 0 \quad \Rightarrow \quad \frac{\partial V}{\partial s_j} = 0 \quad (1 \le j \le J) \tag{2}$$

Let the partial derivatives of the functions **f** be $\partial f_j(t, s)/\partial t = f_j^1$ and $\partial f_j(t, s)/\partial s = f_j^2$. Then we can express the equilibrium condition (2) as

$$e_j(s_j) > 0 \quad \Rightarrow \quad f_{j-1}^2(s_j) + f_j^1[\tau_j(s_j)]\,\dot{\tau}_j(s_j) - \dot{\tau}_j(s_j) + 1 - \dot{c}_j(s_j) = 0 \quad (1 \le j \le J). \tag{3}$$

Rearranging this gives

$$e_j(s_j) > 0 \quad \Rightarrow \quad \dot{\tau}_j(s_j) = \frac{1 + f_{j-1}^2(s_j) - \dot{c}_j(s_j)}{1 - f_j^1[\tau_j(s_j)]} \quad (1 \le j \le J). \tag{4}$$

Now flow propagation (see, for example, Heydecker and Addison, 1996) on trip $j$ ($1 \le j \le J$) means that the arrival rate $g_j(t)$ at location $j$ satisfies

$$e_j(s_j) = g_j[\tau_j(s_j)]\,\dot{\tau}_j(s_j)\,. \tag{5}$$

Thus the equilibrium departure profile $e_j(s)$ from location $j$–1 ($1 \le j \le J$) is generated by the arrival profile $g_j(t)$ at location $j$ using the flow propagation relationship (5) together with the invariance relationship (4) as:

$$e_j(s_j) = \left( \frac{1 + f_{j-1}^2(s_j) - \dot{c}_j(s_j)}{1 - f_j^1[\tau_j(s_j)]} \right) g_j[\tau_j(s_j)] \quad (1 \le j \le J)\,. \tag{6}$$

The arrival rate profile $g_j(t)$ at location $j$ can be found from a suitable traffic model together with knowledge of departure profiles $e_j(s)$ from location $j$ - 1 at times for which $\tau(s) \le t$: Mun (2001) has investigated the suitability of various models for this purpose.

We note that if, as is often the case, marginal increases in attendance at location $j$ confer benefits, $f_j^1 \le 0$ (for greater benefit with earlier arrival) and $f_j^2 \ge 0$ (for greater benefit with later departure).

In expression (1) for the net benefit $V$ to an individual of making a tour, the effect of changes in travel time and in time spent at one or other of the locations have opposite signs. This highlights the role of travel as being a costly necessity in attending the various locations of the tour. We also note that travellers cannot change their allocation of time to a single component of the tour without affecting that to other ones: rather they can change their travel pattern in a way that transfers time between components. The UK Department for Transport (2004) provides typical perceived monetary values for transfer of time from commuting to the same amount of time spent at work (£22.11/h) and at leisure (£5.04/h). When converted to equivalent travel time, these values correspond to the quantities $1+f^1$ and $1-f^2$.

From expression (6), we can calculate a toll that will eliminate congestion on trip $j$. Thus when

$$\dot{c}_j(s_j) = f_{j-1}^2(s_j) + f_j^1[\tau_j(s_j)], \tag{7}$$

the equilibrium departure rate is equal to the consequent arrival rate so that the travel time is constant.

In order for travel on journey $j$ to be confined to a bounded interval of time, either the marginal value of attendance at the origin should decrease over time, or the rate of change of toll should increase over time, or the marginal value of attendance at the destination should increase over time, or some combination of these. Equilibrium is achieved through variations in travel time between origin and destination that are complementary to these variations in utility.

### 2.4 Volume of travel

The volume of travel that takes place on each trip *j* is determined as the time integral of the departure rate. Travel starts when the net utility $V$ of a trip made in uncongested conditions rises to the equilibrium value. Travel then continues until the net utility of further trips falls below the equilibrium value, even when the network is uncongested. Thus the volume $E_j$ of travel of journey $j$ is given by $E_j = \int_s e_j(s)\,ds \ (1 \le j \le J)$,

where the integrand has non-zero value only within a certain departure time interval within which travel incurs the equilibrium cost. Thus the start time of the interval during which travel takes place determines the end time, and together with the inflow profile they determine the volume. For conservation of flow along the tour, we require that the volume $E_j$ be equal for all trips $j$ $(1 \le j \le J)$.

The present model thus provides a relationship between the volume of trips made and the net utility $V$ achieved through making them, as given by (1). We expect that as the volume of travel increases, so congestion will increase travel costs and also reduce the time available to gain benefit from attendance at the

locations, so that utility decreases on each of these grounds. This relationship can be used in conjunction with a demand function $D(V)$ to establish a demand-performance equilibrium in which

$$E_j = D(V) \quad (1 \le j \le J) \tag{8}$$

where $E_j = \int_s e_j(s)\, ds$ $(1 \le j \le J)$, with $e_j(s_j)$ given by (6), and

$V$ is given by (1), evaluated at times $\mathbf{s}$ when $\mathbf{e}(\mathbf{s}) > \mathbf{0}$

so that the equilibrium assignment of volume $E$ induces the net benefit $V$.

### 2.5 Evaluation

To undertake a macroscopic evaluation of a certain travel profile $\mathbf{e}(s)$ for a tour, we combine the individual utilities (1) according to the rate at which they are achieved. Thus for a trip $j$, the contribution to the total cost incurred in travel is

$$W_j^t = \int_s \left[\tau_j(s) - s + c_j(s)\right] e_j(s)\, ds \ . \tag{9}$$

Because the tolls $c_j(s)$ charged for a trip are transferred within the system, the total revenue $C$ appears as a benefit in a way that cancels this element of the total cost incurred by individuals. We can calculate the total revenue generated and credited in this way as

$$\begin{aligned} C &= \sum_j C_j \\ &= \sum_j \int_s c_j(s)\, e_j(s)\, ds \ . \end{aligned} \tag{10}$$

Similarly, for activity $j$, the contribution to the total benefit of the tour is

$$\begin{aligned} W_j^a &= \frac{1}{E_{j+1}} \int_s \int_t f_j(t, s)\, g_j(t)\, dt\, e_{j+1}(s)\, ds \\ &= \frac{1}{E_{j+1}} \int_{s_j} \int_{s_{j+1}} f_j\left(\tau_j(s_j), s_{j+1}\right) e_j(s_j)\, ds_j\ e_{j+1}(s_{j+1})\, ds_{j+1} \ . \end{aligned} \tag{11}$$

The total net benefit $W$ of the tours undertaken in equilibrium can then be expressed using (9), (10) and (11) as

$$W = \sum_j W_j^a - W_j^t + C_j \ . \tag{12}$$

Using expression (1) for the net benefit $V$ that is achieved in equilibrium, we have $\sum_j W_j^a - W_j^t = \int_s V\, e_j(s)\, ds$. This, together with expression (8) for the throughput in equilibrium and the definition (12) of $W$ then gives

$$W = V\, D(V) + C \,. \tag{13}$$

## 3 Example calculations

By way of example, we consider a two-trip tour from home to work and back. We suppose that the marginal value of time at home is high early on during the day, and then falls to a constant value near the start of the morning peak period, representing a reluctance to depart too early: this is illustrated in Figure 1. Similarly, we suppose that there is a premium on attendance at work during core hours, and that some reward is given for flexible working beyond those: a marginal value of time at work that achieves this is illustrated in Figure 2.

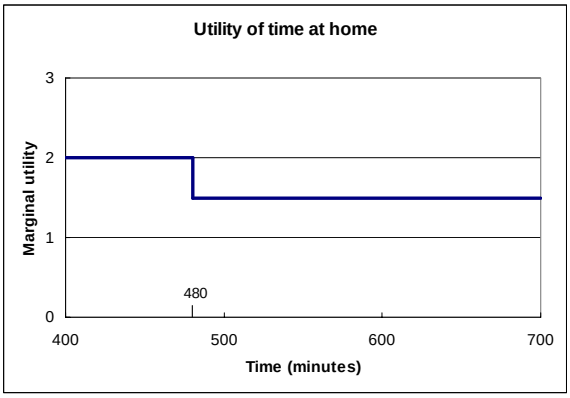



Figure 1: Marginal utility of time at home

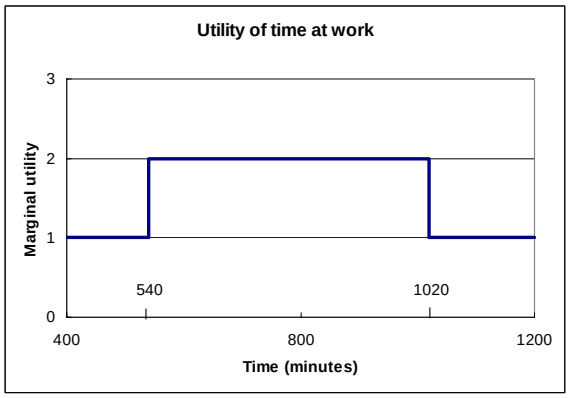



Figure 2: Marginal wage rate at work

Let the marginal value of time at home, expressed in units of travel time saved, be 2.0 before 08:00 (480 minutes) and be 1.5 after that. Thus the utility $f_0(s)$ of remaining home until time $s$ is $f_0(s) = 1.5\, s + 0.5\, \mathrm{Min}(s,\ s^0)$, where $s^0 = 8$ h (480 minutes). Suppose that the value of time at work is determined by three elements: the time of arrival, the duration of stay, and the time of departure. Let the wage rate, expressed in units of equivalent travel time saved, be 1.0 outside the core hours of 09:00 (540 minutes) to 17:00 (1020 minutes), and 2.0 within these core hours. Thus the benefit $f_1(t, s)$ of working from time $t$ to time $s$ is given by $f_1(t, s) = s - t + \mathrm{Min}(s, s^1) - \mathrm{Max}(t, t^1)$, where $s^1 = 9$ h (540 minutes) and $t^1 = 17$ h (1020 minutes). Because time at home is valued at 1.5 throughout the day after 08:00, the return journey from work to home is influenced by that.

According to this specification, we have the following relationships for the marginal utilities that affect the two journeys:

$$f_0^2(s) = \begin{cases} 2.0 & s < s^0 \\ 1.5 & s > s^0 \end{cases},$$

$$f_1^1(t) = \begin{cases} -1.0 & t < t^1 \\ -2.0 & t > t^1 \end{cases} \text{ and } f_1^2(s) = \begin{cases} 2.0 & s < s^1 \\ 1.0 & s > s^1 \end{cases}, \text{ and}$$

$$f_2^1(t) = -1.5\,. \tag{13}$$

We suppose that the free-flow travel time for each journey is 30 minutes, and that the capacity of the network in each direction is 1800 vehicles/h . We use a deterministic queueing model (see, for example, Vickrey, 1969; Arnot de Palma and Lindsey, 1993; Mun, 2001) with these parameters to describe the travel times and their variations due to congestion.

When the start time of departures from home to work is 07:40 (460 minutes), the end time of departures is 09:40 (580 minutes) and the total volume of travel is 3,600 trips. In order to achieve the same volume for the return journey, departures from work to home start at time 16:00 (960 minutes), and end at time 18:00 (1,080 minutes). The departure profiles and consequent travel times that achieve equilibrium are shown in Figures 3 and 4: the initial peak in the profile of departures from home is a consequence of the higher marginal value of time at home before 08:00 (480 minutes).

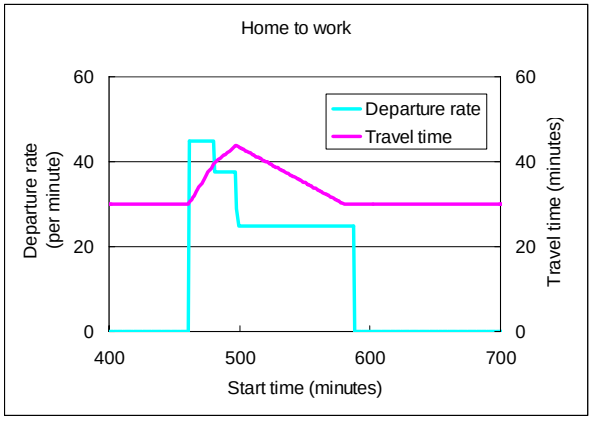



Figure 3: Journey from home to work

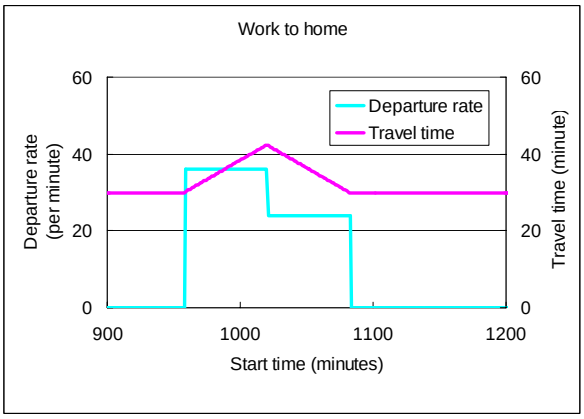



Figure 4: Journey from work to home

The travel time at the time of the first departure from home at 07:40 (460 minutes) is 30 minutes, and the utility of having remained home until that time is equivalent to having saved 920 minutes of travel time. An individual who departs at this time will then arrive at work at 08:10 (490 minutes): if they remain only until the time of the first departure from work to home, at 16:00 (960 minutes), they will gain utility equivalent to a saving of 890 minutes of travel time. Departing for home at that time gives a travel time of 30 minutes, resulting in arrival at home at time 16:30 (990 minutes) where they will gain a further 675 minutes of utility during the remainder of the day. The net utility for this individual is then the sum of the two home utilities plus the work utility minus the two travel costs, giving 2,425 minutes, which will be identical for all travellers in this homogeneous group. These results are summarised in Table 1.

Table 1: Utility obtained by an individual departing at the start of the morning peak, 07:40 .

| | Home | | Travel | Work | | | Travel | Home | |
|---|---|---|---|---|---|---|---|---|---|
| | Duration (minutes) | Depart | Duration (minutes) | Arrive | Duration (minutes) | Depart | Duration (minutes) | Arrive | Duration (minutes) |
| Time | 460 | 07:40 | 30 | 08:10 | 470 | 16:00 | 30 | 16:30 | 450 |
| Utility | 920 | | -30 | | 890 | | -30 | | 675 |

Consider now another individual who departs from home at the most congested time of 08:16 (496 minutes) and will arrive at work at 09.00 (540 minutes) after a travel time of 44 minutes. Because in this example each of the marginal utilities (13) of attendance does not depend on other start and end times, the journey from home to work can be equilibrated separately from that from work to home. Compared with the earliest departing individual, this one will gain an equivalent of 64 minutes of additional utility at home, but will lose 14 minutes through increased travel time and a further 50 equivalent minutes of wages through later arrival at work. The results of this variation in departure time from home are shown in Table 2. This illustrates that in equilibrium, different individuals achieve identical net utilities through different interrelated combinations of costs and benefits by timing their journeys and activities differently.

Table 2: Utility obtained by an individual departing at the height of the morning peak, 08:16 .

| | Home | | Travel | Work | | | Travel | Home | |
|---|---|---|---|---|---|---|---|---|---|
| | Duration (minutes) | Depart | Duration (minutes) | Arrive | Duration (minutes) | Depart | Duration (minutes) | Arrive | Duration (minutes) |
| Time | 496 | 08:16 | 44 | 09:00 | 420 | 16:00 | 30 | 16:30 | 450 |
| Utility | 984 | | -44 | | 840 | | -30 | | 675 |

If the individual whose tour is described in Table 1 had remained at home instead of going to work, they would have had an additional 530 minutes there, each valued as equivalent to 1.5 minutes of travel time so leading to a benefit equivalent to having saved 795 minutes of travel time. This benefit is less that the net 830 minutes (890 minutes from work minus $2 \times 30$ minutes spent travelling) gained by attending at work after allowing for the costs of travel. Similarly, the individual whose tour is described in Table 2 could have gained benefit equivalent to having saved 741 minutes of travel time by remaining at home, which again is less than the net 766 minutes (840 minutes minus 44 minutes minus 30 minutes) gained by working. According to this analysis an individual could even plan rationally to depart from home at 09:40 (580 minutes), the end of the morning peak, arriving at work at 10:10 (610 minutes), and then depart from work just 350 minutes later at 16:00 (960 minutes) when the evening peak begins. This is because they would gain net benefit equivalent to 640 minutes of travel time saving ($2 \times 350 - 2 \times 30$ minutes), which is greater than the benefit equivalent to 615 minutes of travel time saving ($1.5 \times 410$ minutes) that they would gain by remaining home for the whole of that time.

Now suppose that a time-varying charge is levied for travel during the morning peak, and that this is calculated according to (7) to achieve the same arrival profile but without any congestion. In this case, the travel time will be identical for all travellers and their utility maintained at the same constant value through variations in the charge. For a traveller who departs home at 08:16 (496 minutes), the travel time of 30 minutes will lead to arrival at 08:46 (526 minutes). The timings and utilities for this trip are shown in Table 3, where comparison is made between this trip and the untolled trips summarised in Tables 1 and 2. Compared with the case shown in Table 2, this traveller will save 14 minutes travel cost and consequently gain additional payment for employment equivalent to saving 14 minutes travel time: the charge required to render this in equilibrium is then equivalent to 28 minutes of travel time rather than the 14 minutes that are saved. In this case, the individual has a shorter journey and also earns more at work, but transfers both of these benefits through the toll charged.

Table 3: Utility obtained by an individual who departs from home at 08.16 (496 minutes) under congestion-eliminating charging, paying 28 minutes toll.

| | Home | Travel | Work |
|---|---|---|---|
| | Duration (minutes) | Duration (minutes) | Duration (minutes) |
| Time | 496 | 30 | 434 |
| Cf Table 1 | 36 | 0 | -36 |
| Cf Table 2 | 0 | -14 | 14 |
| Utility | 984 | -30 -28(toll) | 854 |
| Cf Table 1 | 64 | -28 | -36 |
| Cf Table 2 | 0 | -14 | 14 |

For a traveller who arrives at work at 09:00 (540 minutes), the tolled travel time of 30 minutes will require departure at 08:30 (510 minutes). The timings and utilities for this trip are shown in Table 4, where comparison is made between this trip and the untolled trips summarised in Tables 1 and 2. Compared with the case shown in Table 2 where the traveller arrives at the same time but only after experiencing congestion, this traveller will save 14 minutes of travel time, but in this case will have spent that time remaining at home for longer. The benefit to this traveller is then equivalent to a saving of 21 minutes of travel time through the increased time at home plus the saving of 14 minutes of travel time on the journey. This is balanced exactly by the toll equivalent to 35 minutes of travel time. In this case then, the individual has a shorter journey and spends the time saved at home before travelling, but transfers both of these benefits through the toll charged.

Table 4: Utility obtained by an individual who departs from home at 08.30 (510 minutes) under congestion-eliminating charging, paying 35 minutes toll.

| | Home | Travel | Work |
|---|---|---|---|
| | Duration (minutes) | Duration (minutes) | Duration (minutes) |
| Time | 510 | 30 | 420 |
| Cf Table 1 | 50 | 0 | -50 |
| Cf Table 2 | 14 | -14 | 0 |
| Utility | 1005 | -30 -35(toll) | 840 |
| Cf Table 1 | 85 | -35 | -50 |
| Cf Table 2 | 21 | -21 | 0 |

We now consider the way in which the present analysis represents the macroscopic effects of this congestion-eliminating charging. The toll is calculated so that the arrival profile at the destination is unaffected by it. All of the changes are therefore caused at the origin and on the journey. Under this toll, the departure profile is similar to the arrival profile except that it occurs earlier by a time that corresponds to the free-flow travel time. This is achieved by deferring some departures so that extra time is spent at the origin. The total time spent at the origin is then $T_0^a = \int_s s\, e_1(s)\, ds$, which increases by 26,400 person-minutes when the toll is introduced. The benefit to travellers of this can be calculated according to (11) as $W_0^a = \int_s f_0(s)\, e_1(s)\, ds$, which increases by 67,500 person minutes of equivalent travel time saved when the toll is introduced. Because the total volume of travel is 3,600 persons, these equate respectively to a mean of 7 1/3 minutes of additional time at home per person, with a mean benefit equivalent to saving 18 ¾ minutes of travel time per person. The toll revenue $C$ generated by this is equivalent to 67,500 person-minutes of travel time saved, with the same mean charge equivalent to saving 18 ¾ minutes of travel time per person.

From this, we see that the congestion-eliminating toll has the effect of converting congestion to toll revenue and the time saved through elimination of delay to additional time at the origin. For each individual traveller, the net effect of this is neutral. However, the benefits of this are then two-fold: first, the cash revenue that is generated can be allocated beneficially, which contrasts with the dead loss of congestion delays. Second, the travellers spend more time at their origin, with the possibility of conferring external benefits to those at that location: in the case of time at home, this corresponds to a benefit to the household, whilst in the case of time at work, it corresponds to a benefit to the employer.

## 4 Summary

The present model provides a representation of the way in which travellers organise their tours according to an equilibrium principle. We have seen how this can be used to represent travellers who choose their departure times in tours according to the benefits that they gain through attendance at different locations and

the travel conditions that they encounter between them. Choices made by individuals can, in equilibrium, include a range of possible balanced combinations of costs of travel and benefits obtained in attending the locations. We have shown how this can be used to investigate the effects of changes that might be made to travel provision (for example, to the free-flow travel time, capacity or monetary tolls made for use of the network) and to the marginal utilities of attendance at the locations. The influence of changes of these kinds on trip making has been illustrated using the example of a congestion-eliminating toll, and ways in which this confers benefits to society have been explored. From this analysis, we see that the marginal value of time spent at each location is of central importance in estimating travel activity.

Several issues remain to be investigated in analysis of this kind. These include the effect of various kinds of heterogeneity, including that in marginal values of time at each of the locations, in values of cash paid as tolls, and in network behaviour. Furthermore, in multi-stop tours, equilibrium behaviour could entail balance between different sequences of locations in tours. The issues also arise of equity in charging of tolls, especially in heterogeneous populations of travellers, and in distribution of toll revenues. In due course, the present model will be extended to consider these various aspects.

## Acknowledgements

An earlier version of this paper was presented at the Behaviour in Networks conference, Seoul, 2005.

## References

Arnott, R, de Palma, A and Lindsey, R (1993) A structural model of peak period congestion: A traffic bottleneck wit elastic demand. *American Economic Review*, **83**(1), 161-79.

Ashiru, O, Polak, JW and Noland, RB (2004) The utility of schedules: A model of departure time choice and activity time allocation with application to individual activity schedules, *Transportation Research Record* (in press).

Department for Transport (2004) Values of time and operating costs. *Transport Analysis Guidance* **Unit** 3.5.6. London: Department for Transport.

Ettema, D, Ashiru, O and Polak, JW (2004) Modelling timing and duration of activities and trips in response to road pricing policies, *Transportation Research Record* (in press).

Heydecker, BG and Addison, JD (1998) Analysis of traffic models for dynamic equilibrium traffic assignment. **In**: *Transportation Networks: Recent Methodological Advances* (ed MGH Bell). Oxford, Pergamon, ISBN: 0-08-043052-X, 35-49.

Heydecker, BG and Addison, JD (2005) Analysis of dynamic traffic equilibrium with departure time choice. *Transportation Science*, **39**(1), 39-57.

Kitamura, R (1988) An evaluation of activity-based travel analysis. *Transportation*, **15**, 9-34.

Lam, WHK and Yin, Y (2001) An activity-based time-dependent traffic assignment model. *Transportation Research*, **35B**(6), 549-74.

Mun, J-s (2001) A divided linear travel time model for dynamic traffic assignment. *Proceedings* of the 9th World Conference on Transport Research, Seoul, **D1-05**, 4189.

Oppenheim, N (1995) Urban travel demand modelling: from individual choices to general equilibrium. Chichester: Wiley.

Polak, J and P Jones (1994) Travellers' choice of time of travel under road pricing, paper presented at the 73rd Annual Meeting of the Transportation Research Board, Washington DC.

Recker, WW (1995) The household activity pattern problem: general formulation and solution. Transportation *Research*, **29B**(1), 61-77.

Vickrey, WS (1969) Congestion theory and transport investment. *American Economic Review (Papers and Proceedings)* **59**, 251-61.

Zhang, X, Yang, H, Huang, H-J and Zhang, HM (2005) Integrated scheduling of daily work activities and morning-evening commutes with bottleneck congestion. *Transportation Research*, **39A**(1), 41-60.

# Day-to-day Congestion Pricing Policies towards System Optimal

*Fan Yang: ESRI Inc., 380 New York St, Redlands, CA, 92373, fyang@esri.com*
*W. Y. Szeto: Department of Civil, Structural, and Environmental Engineering, Trinity College Dublin, Dublin 2, Ireland, szetow@tcd.ie*

## Abstract
The marginal social cost pricing policy has been proposed for many years to ensure that the resulting equilibrium flows are system optimal ones. However, how to implement a socially efficient congestion pricing scheme, taking into account drivers' day-to-day route choice disequilibrium behavior, is not well studied. In this paper, day-to-day static and dynamic congestion pricing schemes have been proposed to guide the day-to-day dynamic flow evolving towards system optimal instead of user equilibrium, considering a general drivers' behavior adjustment process. The strong dynamic optimal toll is defined to be such that the dynamic total system cost is monotonically decreasing along the day-to-day dynamic flow trajectory until day-to-day dynamic flows converge to system optimal flows. A simple solution to the strong dynamic optimal toll problem has been developed. The convergence of the tolled day-to-day dynamics, the equivalency between stationary link flow states and system optimal states in the tolled day-to-day dynamics, and the conditions where system optimal can be locally or globally asymptotically stable are discussed. The results are illustrated using the well-known Braess network.

## 1 Introduction
Congestion pricing has become one of the priorities on the transport policy agendas given the increasing congestion level throughout the world. The goal of traffic management agencies is to minimize the (traffic) system's total travel cost, i.e., to reach the system optimal (SO) states. However, drivers only concern their own travel cost. The end result is that the system reaches the user equilibrium (UE) state or other unstable states at the end rather than the SO one. Traditionally, we believe that we can employ the marginal cost toll to shift UE to SO (Beckmann et al., 1956). This marginal cost pricing, nevertheless, only considers possibly a final UE state of the system without considering the realistic driver's day-to-day disequilibrium learning and route switching behaviour. Some important questions, including how the congestion pricing can impact the drivers' day-to-day learning process and whether and how we can achieve SO taking into account the drivers' day-to-day disequilibrium behavior, have not been studied in the literature as far as we know.

Assuming that the marginal toll can guide the system towards SO. How can a traffic manager realize SO to be reached from the day to day perspective? Intuitively, we can first investigate the equivalence between path flow stationary and SO flow pattern since drivers consider paths rather than links in their switching decisions and consequently day-to-day dynamics usually deal with path flows. Nonetheless, measuring and monitoring path flow stationary is very difficult. Therefore, does link flow stationary imply SO flow pattern?

Other than the realization problem, is there any other static tolling strategy that can lead to SO as well? It is well known that marginal cost pricing assuming that all roads can be tolled but this is not the case in reality. More importantly, can we have a better tolling strategy comparing with the static toll, in terms of reaching SO faster or the system cost to be decreasing towards SO? Obviously, reaching SO faster is desirable as the goal of the traffic management agencies is to minimize the total system travel cost. The property that the total system cost is monotonically decreasing along the day-to-day dynamic flow trajectory until day-to-day dynamic flows converge to system optimal flows is desirable too as this can tell the traffic management agencies that the tolling strategy is working well, the congestion level keeps being improved along the drivers' day-to-day behavior adjustment process, and the system goes with the right direction and reaches SO at the end. Intuitively, a (day-to-day) dynamic toll can help in these two aspects.

Suppose that we can have positive answers to the above questions. Is the final state globally (or locally) stable? In other words, will any initial flow pattern be driven to system optimal by the drivers' behavior adjustment process? Or, will any flow pattern near system optimal always stay close given some small

perturbation? These questions are crucial to transportation planners and analysts, as they prefer to have a tolling strategy that results in a stable state so that a small temporal change in supply can result in the same desirable state quickly.

This paper gives answers to all the questions above by considering Yang and Zhang's (2006) rational behavior adjustment process (RBAP). This adjustment process agrees with Zhang's (2001) one, and is more general than the existing adjustment process including Smith (1984), Friesz et al. (1994), Nagurney and Zhang (1996), and Yang (2005) according to Yang and Zhang (2006). This paper also differs from Friesz et al. (2004), the only one we can find in the literature about day-to-day congestion pricing up to date, that they deal with elastic demand but we deal with fixed demand. We thus fill the gap in the literature. In addition, this study introduces a novel and simple day-to-day dynamic congestion toll (called the strong dynamic optimal toll) such that the total system cost is monotonically decreasing along the day-to-day dynamic flow trajectory until day-to-day dynamic flows converge to system optimal flows, under the mild assumption that the link travel cost function is differentiable. This dynamic toll can be viewed as the dynamic version of the classic marginal social cost one, by replacing the link flow and link travel cost at SO with the ones on current "day". This dynamic toll is also shown to drive the day-to-day flow pattern to SO faster than the day-to-day static toll. Moreover, this paper proves the convergence of the day-to-day static and dynamic tolls, the equivalence between the stationary link flow pattern and system optimal flow pattern under RBAP, and the system optimal state to be globally or locally asymptotically stable under mild assumptions. For completeness, next section will give a brief review on UE, SO, marginal cost pricing, the existing day-to-day path flow dynamics, and the tolled day-to-day dynamics. The rest of the paper is organized as follows: Section 3 discusses the day-to-day static tolls, including marginal tolls in the context of SO. Section 4 extends the discussion to the day-to-day dynamic tolls. Section 5 examines the stability issue. Section 6 is the numerical studies. Lastly, section 7 is the concluding remarks.

## 2 Review of UE, SO and Day-to-day Dynamics with and without Tolls

Typically, a transportation network can be considered as a fully-connected directed graph denoted as $G(\mathcal{N},\mathcal{A})$, consisting of a set of nodes $\mathcal{N}$ and a set of links $\mathcal{A}$. Let the set of O-D pairs be denoted by $\mathcal{W}$, the fixed travel demand for O-D pair $w \in \mathcal{W}$ by $d^w$, the set of paths connecting the O-D pair $w \in \mathcal{W}$ by $\mathcal{P}^w$, the flow on path $p \in \mathcal{P}^w$ by $f_p^w$, the path travel cost on path $p$ by $C_p^w$, the flow on link $a \in \mathcal{A}$ by $x_a$, the travel cost on link $a$ by $\tau_a$, the minimum travel cost vector by $\pi$, the link-path incidence matrix by $A$, the feasible path flow set by $K$ and the Cartesian product of each $K^w = \{f \in R^{n^r} : f \geq 0, \text{ and } \sum_p f_p^w = d^w, w \in \mathcal{W}\}$, and the link toll vector by $\beta$ with $\beta \geq 0$.

The well-known Wardrop first principle states that at the user equilibrium state, all the paths actually used have the equal and minimal path travel cost. A feasible path flow vector $f^*$ is the user equilibrium solution if and only if $f^{*'}(C^* - \pi^*) = 0$, $f^* \in K$, where the superscript $'$ denotes the transpose of a vector (matrix). Equivalently, $f^*$ is the solution of the variational equality problem $C^{*'}(f - f^*) \geq 0$, $\forall f \in K$ (Nagurney, 1993). The total system cost is given by the dot product $C'f$. The system optimal flow $\overline{f}$ is the solution of the optimization problem $\min C'f$, $f \in K$.

Let $C(\beta) = C + A'\beta$ denote the tolled path travel cost vector under the (static or dynamic) link toll vector $\beta$. Then the tolled user equilibrium $f^*$ is the solution of the variational equality problem $C(\beta)^{*'}(f - f^*) \geq 0$, $\forall f \in K$. Let $\overline{x}$ and $\overline{\tau}$ be the link flow and link travel cost at system optimal states, respectively. The classic marginal cost pricing scheme is to add the marginal cost $\overline{\beta} = D\overline{\tau}'\overline{x}$ to each link, where $D\overline{\tau}$ is the Jacobian matrix of the link travel cost function $\tau$ with respect to the link flow $x$ at $\overline{x}$. A toll scheme $\beta$ is called an optimal toll if adding this toll results in the situation where the tolled user equilibrium becomes (untolled) system optimal. Obviously the toll $\overline{\beta} = D\overline{\tau}'\overline{x}$ is an optimal toll scheme.

The arguments above only consider in the long-term day-to-day static case in toll determination and ignore the realistic and important driver's day-to-day dynamic adjustment behavior. To account for this realistic adjustment process in determining optimal tolls, we must consider day-to-day dynamics. Day-to-day dynamics describe the aggregated path flow evolution over time. They play an important role in transportation network analysis, both for gaining a deeper understanding of the properties of the standard traffic equilibrium model, and for practical applications related to the monitoring and controlling of traffic flow evolution.

In the literature, day-to day dynamics can be modelled in the continuous time setting, and take the form of ordinary differential equations. They assume that drivers have information on path flows and path costs on current "day" (including toll) before making their route choice decisions. The continuous time day-to day dynamics also have the property that there exists a unique solution trajectory of each dynamic and the fixed point of each dynamic is the corresponding user equilibrium. They include the proportional-switch adjustment process (e.g., Smith, 1984; Smith and Wisten, 1995), the projected dynamical system (e.g., Nagurney and Zhang, 1996), the network tatonnement dynamical system (e.g., Friesz, et al., 1994), and the Brown-von Neumann-Nash (BNN) dynamic (Sandholm, 2001; Yang, 2005).

The proportional-switch adjustment process (PAP) originates from Smith (1984), although it is not named by Smith (1984). It assumes that travelers on a higher cost path will switch to other lower cost paths in next "day", while the switching rate depends on the cost difference between this path and others. The PAP can be written as

$$\dot{f}_p^w = \sum_{q \in \mathscr{P}^w} f_q^w [C_q^w - C_p^w]_+ - f_p^w \sum_{q \in \mathscr{P}^w} [C_p^w - C_q^w]_+ , \tag{2.1}$$

where $[x]_+ = \max\{0, x\}$ ; $\dot{f}$ denotes the derivative of day-to-day dynamic path flow w.r.t. "day" $t$ ; $f_p^w$ and $C_p^w$ are respectively the flow and cost on path $p$ between OD pair $w$ on "day" $t$ , in which the subscript $t$ is dropped from the dynamic variables hereafter to simplify notations if there is no confusion. By replacing the untolled path travel cost $C$ by the tolled one $C(\beta)$ , we have the tolled dynamical system under the toll $\beta$

$$\dot{f}_p^w = \sum_{q \in \mathscr{P}^w} f_q^w [C_q^w(\beta) - C_p^w(\beta)]_+ - f_p^w \sum_{q \in \mathscr{P}^w} [C_p^w(\beta) - C_q^w(\beta)]_+ . \tag{2.2}$$

The projected dynamical system (PDS) describes disequilibrium trajectories of traffic dynamics until reaching user equilibrium. It has been widely used to solve the variational inequality problem (Nagurney, 1993). The individual driver's behavior explanation for PDS is that travelers will switch to paths with positive probabilities if the path travel costs are lower than the special form of the weighted average travel cost (Sandholm and Lahkar, 2005; Yang and Liu, 2006). PDS can be expressed by

$$\dot{f} = \prod_{\mathrm{K}}(f, -C), \tag{2.3}$$

where the operator $\prod_{\mathrm{K}}$ is defined as $\prod_{\mathrm{K}}(x, y) = \lim_{\varepsilon \to 0} \frac{P_{\mathrm{K}}(x + \varepsilon y) - x}{\varepsilon}$ and $P_{\mathrm{K}}(x) = \arg\min_{z \in \mathrm{K}} \|x - z\|$ is the projection of vector $x$ into the feasible set $K$ . Similarly, the tolled projected dynamical system under the toll $\beta$ can be stated as

$$\dot{f} = \prod_{\mathrm{K}}(f, -C(\beta)) . \tag{2.4}$$

The network tatonnement dynamical system (NTDS) describes drivers' dynamic behavior adjustment under incomplete information. Recall the definition of NTDS

$$\dot{f} = \delta[P_K(f - \alpha(C - \pi)) - f] , \tag{2.5}$$

where $\delta$, $\alpha$ are small positive constants. Then, the network tatonnement dynamical system under the toll $\beta$ can be expressed as

$$\dot{f} = \delta[P_K(f - \alpha(C(\beta) - \pi(\beta)) - f] . \tag{2.6}$$

The BNN dynamic is a canonical dynamic in microeconomics to model players' dynamical evolving behavior. It was first introduced by Brown and von Neumann (1950) for symmetric zero-sum games and recently studied extensively in Swinkels (1993), Hofbauer (2000), Sandholm (2001) and Hofbauer and Sigmund (2003). Yang (2005) was one of the first attempts to model the BNN traffic path flow dynamic. The BNN path flow dynamic describes a certain driver's learning process where the frequency to choose paths with travel cost above average decreases, while the frequency with travel cost below average increases, as long as the path flow is changing. Furthermore, drivers choose the paths with positive probabilities whose travel costs are less than the weighted average travel cost (Yang and Liu, 2006).

Let the average path cost be $\bar{C}^w = \frac{1}{d^w} \sum_{p \in \mathcal{P}^w} C_p^w f_p^w$ for O-D pair $w$, and the excess travel cost of path $p$ relative to the average travel cost for O-D pair $w$ be $[\hat{C}_p^w]_+ = \max\{0, -C_p^w + \bar{C}^w\}$. Then the BNN dynamic can be formulated as

$$\dot{f}_p^w = d^w [\hat{C}_p^w]_+ - f_p^w \left( \sum_{q \in \mathcal{P}^w} [\hat{C}_q^w]_+ \right). \tag{2.7}$$

After levying the static link toll $\beta$, we have the average tolled path cost $\bar{C}^w(\beta) = \frac{1}{d^w} \sum_{p \in \mathcal{P}^w} C_p^w(\beta) f_p^w$ and the excess tolled path cost on path $p$ is $[\hat{C}_p^w(\beta)]_+ = \max\{0, -C_p^w(\beta) + \bar{C}^w(\beta)\}$. The tolled BNN path flow dynamic under the toll $\beta$ becomes

$$\dot{f}_p^w = d^w [\hat{C}_p^w(\beta)]_+ - f_p^w \left( \sum_{q \in \mathcal{P}^w} [\hat{C}_q^w(\beta)]_+ \right). \tag{2.8}$$

It is well known that for the (untolled) PAP, PDS, NTDS and BNN dynamics, their fixed points $f^*$'s are equivalent to user equilibrium, i.e., $f^{*'}(C^* - \pi^*) = 0$. Are the fixed points of dynamical systems under the day-to-day static toll $\beta$ levied on every link on every day equivalent to SO flows? The next section investigates this.

## 3 Day-to-day Static Optimal Toll and System Optimal Flow

**Proposition 1** Given any day-to-day static toll $\beta$, then for all PAP, PDS, NTDS and BNN dynamics, their fixed points $f^*$'s are the tolled user equilibrium flows. In addition, for the day-to-day static scheme $\bar{\beta} = D\bar{\tau}'\bar{x}$, the fixed points $f^*$'s of the tolled dynamical systems above are system optimal flows.

**Proof**: Since the PAP, PDS, NTDS and BNN dynamics do not require any condition on the (untolled) path travel cost, we can replace the untolled path travel cost $C^*$ by the tolled one $C^*(\beta)$, then $f^{*'}(C^*(\beta) - \pi^*(\beta)) = 0$. This concludes the first part. The second part follows because $\bar{\beta} = D\bar{\tau}'\bar{x}$ is an optimal toll. □

The above statement only concerns the marginal static toll. In fact, there is more than one optimal toll to achieve the same purpose: the fixed point is SO. These optimal tolls form a polyhedral set if the link cost function is differentiable (Bergendorff et al., 1997). Let the set of system optimal flows $\bar{S} = \arg\min\{C'f \mid f \in K\}$, the set of tolled user equilibrium flows $U_\beta^* = \{f^* \mid C(\beta)^{*'}(f - f^*) \geq 0,\ \forall f \in K\}$. Then Bergendorff et al. (1997) proves that the optimal toll set is $\Gamma = \bigcup_{\bar{f} \in \bar{S}} \{\beta \in U_\beta^* \mid U_\beta^* \subseteq \bar{S}\}$. Our question is whether under any toll scheme in this polyhedral set, the stationary (tolled) path flows will be the system optimal flows. The next proposition gives an answer to this question.

**Proposition 2** For any $\beta \in \Gamma$, the fixed points of the tolled dynamical systems above are system optimal states.
The proof follows the definition of the optimal toll set $\Gamma$ and the proof of Proposition 1. □

In words, for all PAP, PDS, NTDS and BNN dynamics, when the day-to-day dynamic path flows are stationary, the stationary path flows will be system optimal flows, provided that the toll scheme is given from the optimal toll polyhedral set. However, to measure day-to-day dynamic path flow pattern is usually difficult, while it becomes feasible to monitor dynamic link flow pattern with the fast development of intelligent transportation systems (ITS). Therefore, we investigate the equivalence between stationary link flow states under static tolls and SO. We can actually prove that this equivalence holds in the rational behavior adjustment process (RBAP). RBAP includes the PAP, PDS, NTDS and BNN dynamics (Yang and Zhang, 2006), and is actually the most general one of the proposed behaviour adjustment mechanisms in Zhang et al. (2001), although this mechanism is not formally named by Zhang et al. (2001). This RBAP requires that all the feasible directions of motion reduce the aggregate travel cost. In other words, if path flows are changing, they will move in the direction to reduce the aggregate travel cost. Furthermore, the fixed point will be the user equilibrium state. Mathematically, it can be written as:

$$\dot{f}\begin{cases}\in \mathrm{T} & \text{if } \mathrm{T} \neq \phi \\ = 0 & \text{if } \mathrm{T} = \phi\end{cases} \text{ where } T = \{z(t) \mid \sum_{p \in \mathcal{P}^w} z_p^w(t) = 0,\ C(t)' z(t) < 0\}. \tag{2.9}$$

RBAP also has a property that the stationary link flow pattern implies the stationary path flow pattern and user equilibrium as well (Zhang et al., 2001). Therefore, we have the following Lemma to state the equivalence of stationary link flow states and system optimal ones.

**Lemma 3** For all PAP, PDS, NTDS and BNN dynamics, given any $\beta \in \Gamma$, the stationary link flow states are system optimal states.

**Proof**: Yang and Zhang (2006) prove that all these dynamics satisfy RBAP, which has the property that the stationary link flow pattern implies the stationary path flow pattern (Zhang et al., 2001). Moreover, the stationary path flows implies system optimal flows (Proposition 2). Therefore, the stationary link flows implies the system optimal ones. □

# 4 Day-to-day Dynamic Optimal Toll and System Optimal Flow

The statements above show that the static optimal toll schemes $\beta \in \Gamma$ will function as guiding day-to-day dynamic traffic flow towards system optimal. More importantly, for any static optimal toll, if the link flows are steady, then they are the system optimal flows. However, the total system travel cost can be non-decreasing along the dynamic flow trajectory as shown in the following example.

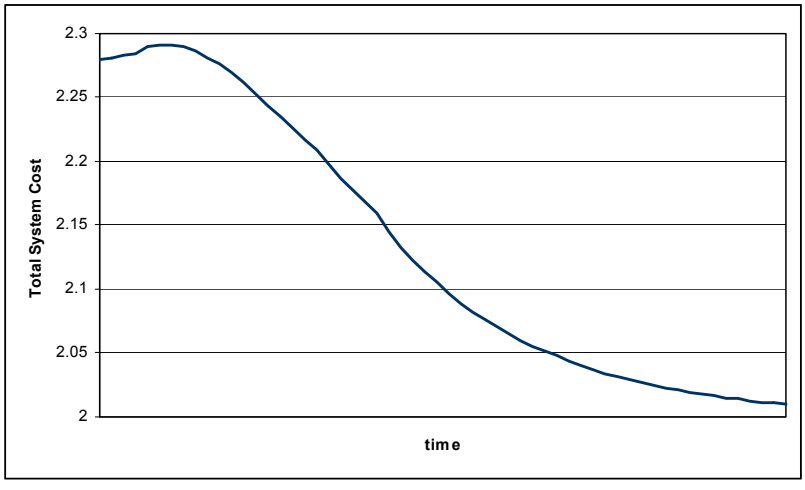



Figure 1 Total System Cost over Time under the Tolled BNN Dynamic

Consider a simple transportation network with one O-D pair and three parallel links. Suppose the demand is 1, and the link travel cost function is $\tau_1 = 2x_1 + x_2 + 4x_3$, $\tau_2 = 4x_1 + 2x_2 + x_3$ and $\tau_3 = x_1 + 4x_2 + 2x_3$. The initial link flow condition is randomly chosen as $x_0 = (0.2\ 0.6\ 0.2)'$ and the static optimal toll $\bar{\beta} = (4\ 2\ 1)'$ has been levied all the time along the dynamic flow trajectory. Matlab

ODE15 solver is used to solve the tolled BNN dynamic. The result is shown in Figure 1, which indicates that the dynamic total system cost actually increases in the early time.

This is not good from the viewpoint of traffic managers, as this might indicate that the tolls are charged wrongly or the tolling scheme is not effective. However, we can get around it through implementing a strong dynamic optimal toll. To explain what a strong dynamic optimal toll is, let us introduce the strong dynamic optimal toll problem (SDOTP). The (SDOTP) is defined as:

To find a strong dynamic optimal toll scheme $\bar{\beta}(t) \geq 0$ such that

(i) $\dot{f}(t) = 0 \Rightarrow f(t) \in \bar{S}$ and (SDOTP)

(ii) $\frac{d}{dt}(C(t)' f(t)) \leq 0$, and the equality holds only if $\dot{f}(t) = 0$,

where $f(t)$ and $C(t)$ denote respectively path flow and path travel cost vectors for day $t$; $\frac{d}{dt}(C(t)' f(t))$ is the derivative of the total system cost with respect to time.

A day-to-day dynamic toll is then called the strong dynamic optimal toll if the toll solves (SDOTP). That is, the toll results in the following:
(a) the fixed points of the tolled dynamical systems are the system optimal states, and
(b) the total system cost is monotonically decreasing along the dynamic flow trajectory until the dynamic flows become the system optimal flows.

Three points are worthwhile to mention. First, although the day-to-day static optimal congestion toll scheme satisfies (i) of (SDOTP), it might not necessarily be a strong dynamic optimal toll as (ii) might not be satisfied. One counter example is actually shown in Figure 2. Second, from the chain rule, we have $\frac{d}{dt}(C(t)' f(t)) = grad(C(t)' f(t))' \dot{f} = (C(t) + DC(t)' f(t))' \dot{f}$. Hence $\frac{d}{dt}(C(t)' f(t)) = C(\beta(t))' \dot{f}$ if the dynamic toll is given as $\beta(t) = D\tau(t)' x(t)$. Therefore, the dynamic toll $\beta(t) = D\tau(t)' x(t)$ is the strong dynamic optimal toll if $C(\beta(t))' \dot{f} \leq 0$ and the equality holds only if $\dot{f} = 0$. Third, the strong dynamic optimal toll turns out to be simple and its mathematical form is close to the marginal toll, as shown later.

Before showing this result, we introduce $\beta$-RBAP, the RBAP under the dynamic toll $\beta(t)$, which is obtained by substituting the tolled path travel cost $C(\beta(t)) = C(t) + A'\beta(t)$ into RBAP (2.9):

$$\dot{f}(t) \begin{cases} \in \mathrm{T}_\beta & \text{if } \mathrm{T}_\beta \neq \phi \\ = 0 & \text{if } \mathrm{T}_\beta = \phi \end{cases} \text{ where } T_\beta = \{z(t) \mid \sum_{p \in \mathcal{P}^w} z_p^w(t) = 0,\ C(\beta(t))' z(t) < 0\}. \tag{2.10}$$

We can do this substitution as RBAP does not impose any condition on link travel cost functions. Now we state the result.

**Theorem 4** Let the day-to-day dynamic toll scheme be $\bar{\beta}(t) = D\tau(t)' x(t)$. Then for any tolled dynamical system satisfying $\beta$-RBAP, its fixed point $f^*(t)$ is system optimal flow. In addition, $\bar{\beta}(t) = D\tau(t)' x(t)$ is a strong dynamic optimal toll and solves (SDOTP).

**Proof**: Since the fixed point of RBAP (2.9) is user equilibrium, the fixed point of any $\beta$-RBAP (2.10) $f^*(t)$ will be the tolled user equilibrium. Define $x^*(t) = Af^*(t)$, $\tau^*(t) = \tau(x)|_{x=x^*(t)}$ and $D\tau^*(t) = D\tau(x)' x|_{x=x^*(t)}$. Therefore, $f^*(t)'(C^*(t) + A'D\tau^*(t)' x^*(t) - \pi^*(t)) = 0$, which is the Karesh-Kuhn-Tucker (KKT) condition of solving the system optimal problem. Hence $\beta^*(t) = D\tau^*(t)' x^*(t)$ is an optimal toll and $f^*(t) \in \bar{S}$. This proves that $\bar{\beta}(t) = D\tau(t)' x(t)$ satisfies (i) of (SDOTP). The rest of the proof is based on the fact that the definition of $\beta$-RBAP ensures $C(\beta(t))' \dot{f} \leq 0$ and the equality holds only if $\dot{f} = 0$. □

**Remark:** All PAP, PDS, NTDS and BNN dynamics satisfy RBAP (see Yang and Zhang, 2006). Therefore, $\bar{\beta}(t) = D\tau(t)^{'} x(t)$ is the strong dynamic optimal toll for these dynamics. In the next section, again, $t$ is dropped for simplicity if this causes no confusion.

# 5 Stability Results of System Optimal Flow

The followings are some definitions on stability and Lyapunov functions.

Let $B(f, r)$ denote the open ball with center $f$ and radius $r$. A fixed point $\bar{f}$ of the dynamical system $\dot{f} = v(f)$ is called *stable* if for any $\varepsilon > 0$, there exists a $\delta > 0$ such that, for every $f \in B(\bar{f}, \delta)$, the solution $f(t)$ with $f(0) = f$ is defined, and $f(t) \in B(\bar{f}, \varepsilon)$ for all $t > 0$.

A fixed point $\bar{f}$ of the dynamical system $\dot{f} = v(f)$ is called *locally asymptotically stable* if it is stable and there exists a $\delta > 0$ such that for every solution $f(t)$ with $f(0) \in B(\bar{f}, \delta)$, we have $\lim_{t \to \infty} f(t) = \bar{f}$. $\bar{f}$ is called *globally asymptotically stable* if for any $f(0) \in K$, we have $\lim_{t \to \infty} f(t) = \bar{f}$.

Let $\bar{f}$ be the fixed point of the dynamical system $\dot{f} = v(f)$. If $L : U \to \mathbb{R}$ is a continuous function defined on an open neighbourhood $U$ of $\bar{f}$, differentiable on $U \setminus \{\bar{f}\}$, such that

1. $L(\bar{f}) = 0$ and $L(f) > 0$ if $f \neq \bar{f}$, and;

2. for all $f \in U$, if $f(t)$ is the solution to $\dot{f} = v(f)$ with $f(0) = f$, then $\dot{L}(f) = \frac{d}{dt}(L(f(t)))|_{t=0} \leq 0$, and the equality holds only if $f = \bar{f}$. Then $L$ is called *a locally strict Lyapunov function* for $\bar{f}$. If such conditions hold for any $U \subseteq K$, then $L$ is called *a globally strict Lyapunov function* for $\bar{f}$.

**Remark** If there exists a locally strict Lyapunov function $L$ for the fixed point $\bar{f}$, then $\bar{f}$ is locally asymptotically stable. If there exists a globally strict Lyapunov function $L$ for the fixed point $\bar{f}$, then $\bar{f}$ is globally asymptotically stable (Robinson, 1995).

**Theorem 5** Let the day-to-day dynamic toll scheme $\bar{\beta} = D\tau^{'} x$. For all tolled PAP, PDS, NTDS and BNN dynamics, if the system optimal $\bar{f}$ is isolated, then $\bar{f}$ is locally asymptotically stable.

**Proof:** $\bar{f}$ is locally asymptotically stable if there exists a locally strict Lyapunov function for $\bar{f}$. Let $L(f) = C^{'} f - C^{'} f|_{f=\bar{f}}$, i.e., the difference of the total system cost between $f$ and $\bar{f}$. It suffices to prove that $L(f)$ is actually a locally strict Lyapunov function for $\bar{f}$. First, since the system optimal $\bar{f}$ is isolated, there exists an open neighbourhood $B(\bar{f}, \delta)$ such that $C^{'} f|_{f=\bar{f}} < C^{'} f$ for any $f \in B(\bar{f}, \delta) \setminus \bar{f}$. Second, since $\bar{\beta} = D\tau^{'} x$ solves (SDOTP) for the tolled PAP, PDS, NTDS and BNN dynamics, by Theorem 4, $\dot{L}(f) = \frac{d}{dt}(C^{'} f - C^{'} f|_{f=\bar{f}}) = \frac{d}{dt}(C^{'} f) = C(\beta)^{'} \dot{f} \leq 0$, and the equality holds only if $f = \bar{f}$. □

**Theorem 6** Suppose the total system cost $C^{'} f$ is strictly convex. Then under the day-to-day dynamic toll scheme $\bar{\beta} = D\tau^{'} x$, for all tolled PAP, PDS, NTDS and BNN dynamics, the system optimal $\bar{f}$ is globally asymptotically stable.

**Proof:** since the total system cost $C^{'} f$ is strictly convex, the system optimal $\bar{f}$ is unique. Then $C^{'} f|_{f=\bar{f}} < C^{'} f$ for any $f \in K \setminus \bar{f}$. Let $L(f) = C^{'} f - C^{'} f|_{f=\bar{f}}$. Therefore $L(\bar{f}) = 0$ and $L(f) > 0$ if $f \neq \bar{f}$.

Since $\bar{\beta} = D\tau' x$ solves (SDOTP) for the tolled PAP, PDS, NTDS and BNN dynamics, by Theorem 4, $\dot{L}(f) = \frac{d}{dt}(C'f - C'f|_{f=\bar{f}}) = \frac{d}{dt}(C'f) = C(\beta)'\dot{f} \le 0$, and the equality holds only if $f = \bar{f}$. Hence $L(f)$ is a globally strict Lyapunov function for $\bar{f}$. □

Therefore, under our day-to-day dynamic toll scheme $\bar{\beta} = D\tau' x$, for any day-to-day behavior adjustment satisfying $\beta$-RBAP (e.g., the tolled PAP, PDS, NTDS and BNN dynamics), when the link flow pattern is stationary, it can be concluded that this state is the system optimal one. Furthermore, this system optimal state will be locally asymptotically stable if it is isolated and globally asymptotically stable if the system cost function is strictly convex.

# 6 Experimental Studies

In the section, we will compare the different toll schemes including no congestion pricing, static marginal congestion cost pricing, another pricing from the static optimal toll set $\Gamma$, and the day-to-day dynamic optimal toll pricing. For illustrative purposes, the Smith's PAP is used as the example of the four dynamics above to describe drivers' rational day-to-day route choice behavior. The test network is the well-known Braess network as shown in Figure 2. There is only one O-D pair, (1,2), and its demand is 6 units. The unique system optimal flows are 3 units on paths (1,3,2) and (1,4,2). The unique user equilibrium flows are 2 units on paths (1,3,2), (1,4,2) and (1,3,4,2). The initial condition is given as $f = [5\ 1\ 0]'$, i.e., 5 units on path (1,3,2), 1 unit on path (1,4,2) and no flows on path (1,3,4,2), when where link 5 (3→4) is just added into the network at time $t = 0$. ODE15 solver in Matlab is used to solve for the dynamic solutions.

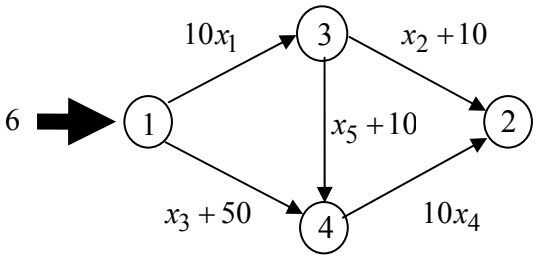



Figure 2 Braess Network

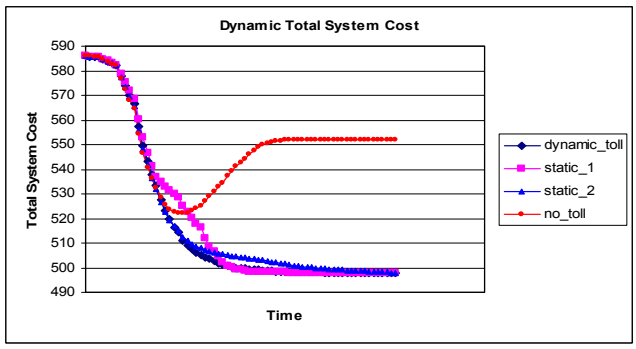



Figure 3 Dynamic Total System Cost under Different Tolls

Since the total system cost is strictly convex for this network, the unique system optimal is globally asymptotically stable under our dynamic optimal toll. In Figure 3, "static_1" stands for the static marginal congestion cost pricing, and "static_2" denotes another static optimal toll (Bergendorff et al., 1997), $\beta = (0,0,0,0,13)'$. Under all these three toll schemes, the total system cost keeps decreasing monotonically until reaching its minimal. Figure 3 shows that the dynamic optimal toll makes the total system cost converges slightly faster than other toll schemes.

It is of our interest to study why the total system cost could increase without toll while it is monotonically decreasing with the dynamic optimal toll, as shown in Figure 3. For the case without toll, the dynamic evolution of total system cost does behave as it is expected – keep decreasing – in the early stage, while later it behaves oppositely. Therefore, it seems that there exists a "transition" in the dynamic evolution of total system cost. This is our motivation to understand the dynamic behavior of total system cost.

Since drivers make route choices in terms of (tolled) path costs, it is important to study the (tolled) path cost evolution. Figure 4 compares the dynamic path flow and path cost between no toll and the dynamic optimal toll cases. For the case without toll, when $t < t_1$, we have $C_2 < C_3 < C_1$. Therefore, according to the individual behavior behind the proportional-switching adjustment process, flows on path 1 switch to both paths 2 and 3, and flows on path 3 to path 2 as well. This causes the rapid flow increase on path 2, so does its path travel cost. This leads to the situation where path 3 becomes the most "attractive" path from $t > t_1$

because its path cost is minimal. It is expected that flows from both paths 1 and 2 switch to this path, which causes congestion on link 5 and deterioration of total system cost.

After levying tolls on links, drivers aim at finding paths with minimal tolled path cost $C(\beta)$. For the case with the dynamic optimal toll, if $t < t_2$, $C(\beta)_2 < C(\beta)_3 < C(\beta)_1$. Similarly, flows switch from path 1 to paths 2 and 3, and from path 2 to path 3. Since $C(\beta)_2 << C(\beta)_3$ at $t < t_2$, most of flows on path 1 will move to path 2. Because of the dynamic toll, the tolled cost of path 3 is greater than that of path 2 all the time, which removes the situation where flows move to path 3 from other paths as without toll case. Instead, at $t > t_2$, path 3 becomes the most expensive path, therefore, few flows on path 3 keep switching to the other two paths until the flow dynamics converge to system optimal.

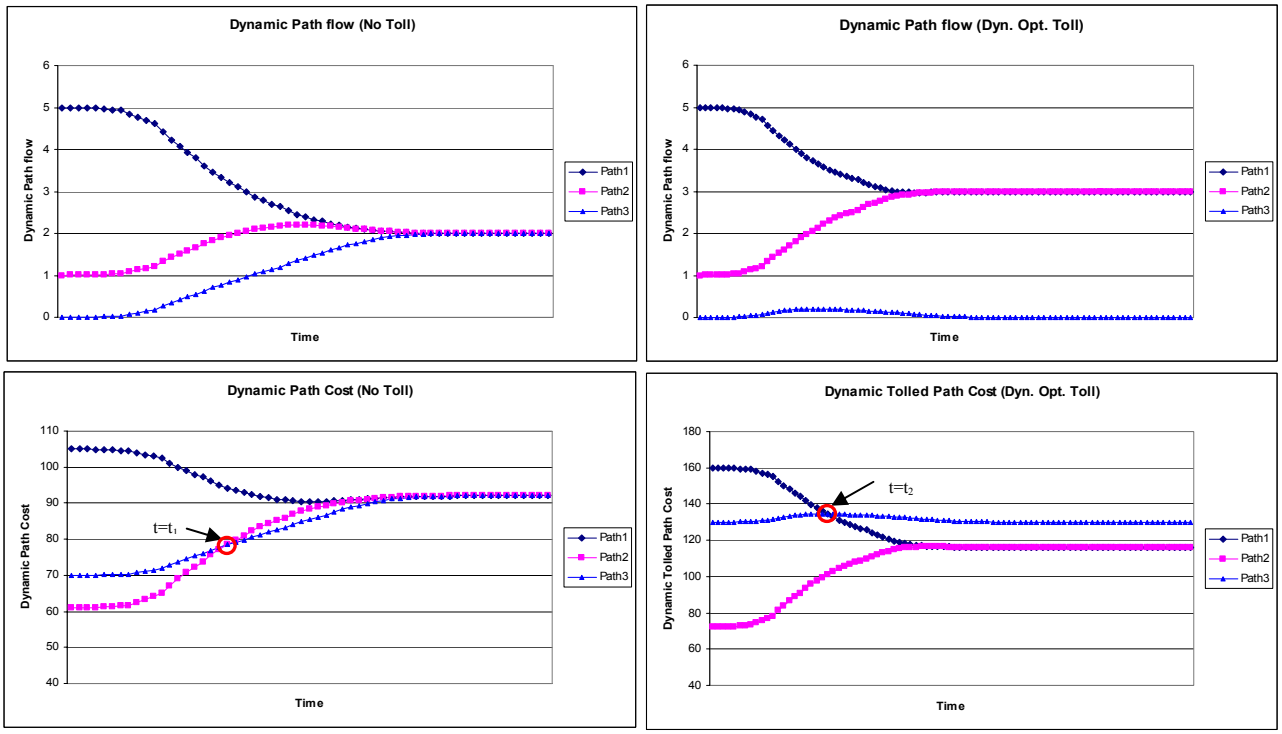



Figure 4 Dynamic Evolution of Congestion

# 7 Concluding Remarks

In this paper, we study the convergence of both day-to-day static and dynamic tolled dynamical systems, the equivalency between stationary link flow states and SO states in the tolled dynamical system, and the conditions where system optimal can be locally or globally asymptotically stable. A notion of the strong dynamic optimal toll is also introduced and investigated. To our best knowledge, this is the one of the first efforts to find the strong dynamic optimal toll solution such that the dynamic total system cost is monotonically decreasing along the day-to-day dynamic flow trajectory until day-to-day dynamic flows converge to system optimal flows. More realistically, the traffic flows in the real world are more likely evolving at disequilibrium states due to the possible changes of the network supply or the travel demand. Therefore, our dynamic toll scheme addresses how to levy an optimal toll over the more likely disequilibrium flows than equilibrium flows. An interesting observation one can make is that our strong dynamic optimal toll scheme seems to be the dynamic version of the classic marginal social cost one, by replacing the system optimal link flow and link travel cost with the current ones. Therefore, similar to Bergendorff et al. (1997), there might exist multiple solutions to the strong dynamic optimal toll problem. However, we did not prove that the strong dynamic toll can always drive the flow pattern towards SO quicker than the static tolls and other feasible dynamic tolls. Furthermore, we only assume homogeneous driver's behaviour. How to accommodate the heterogeneous behaviour in our framework is an interesting direction. We leave them to future studies.

## References

Beckmann, M. J., McGuire, C. B., and Winsten, C. B., 1956. Studies in the Economics of Transportation. Yale University Press, New Haven, CT.

Bergendorff P., Hearn D., and Ramana M., 1997. Congestion Toll Pricing of Traffic Networks. P. M. Pardalos, D.W. Hearn and W.W. Hager (Eds.), Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Vol. 450, 51-71.

Brown, G.W. and Neumann, J. von, 1950. Solutions of Games by Differential Equations, Ann. Math. Studies 24, 73-79.

Friesz, T.L., Berstein, D.H., Mehta, N.J., Tobin, R.L., and Ganjalizadeh, S., 1994. Day-to-day Dynamic Network Disequilibrium and Idealized Traveler Information Systems. Operations Research. 42, 1120-1136.

Friesz, T.L., Bernstein, D. and Kydes, N., 2004. Dynamic Congestion Pricing in Disequilibrium. Networks and Spatial Economics, 4, 181-202.

Hofbauer, J., 2000. From Nash and Brown to Maynard Smith: Equilibria, Dynamics and ESS, Selection 1, 81-88.

Hofbauer, J. and Sigmund, K., 2003. Evolutionary Game Dynamics. BULLETIN (New Series) of the American Mathematical Society 40(4), 479-519.

Nagurney, A., 1993. Network Economics: A Variational Inequality Approach. Kluwer Academic Publishers, Norwell, MA, USA.

Nagurney, A., and Zhang, D., 1996. Projected Dynamical Systems and Variational Inequalities with Applications. Kluwer, Boston.

Robinson, C., 1995. Dynamical Systems: Stability, Symbolic Dynamics, and Chaos. CRC Press, Florida.

Sandholm, W., 2001. Potential Games with Continuous Player Sets. Journal of Economic Theory 97, 81-108.

Sandholm, W. and Lahkar R., 2005. The Payoff Projected Dynamic. Working paper, University of Wisconsin at Madison.

Smith, M.J., 1984. The Stability of a Dynamic Model of Traffic Assignment – an Application of a Method of Lyapunov. Transportation Science, 18, 259-304.

Smith, M. J., and Wisten, M. B., 1995. A continuous Day-to-day Traffic Assignment Model and the Existence of a Continuous Dynamic User Equilibrium. Annals of Operations Research, 60, 59-79.

Swinkels, J., 1993. Adjustment Dynamics and Rational Play in Games, Games Econ. Behav. 5, 455-84.

Yang, F., 2005. An Evolutionary Game Theory Approach to the Day-to-day Traffic Dynamics. Dissertation, University of Wisconsin-Madison.

Yang, F., and Liu H., 2006. A Unifying Framework for Travelers' Day-to-Day Route Adjustment Processes. Partially accepted by the 17th International Symposium on Transportation and Traffic Theory, London, United Kingdom, 2006.

Yang, F. and Zhang, D., 2006. Day-to-day Stationary Link Flow Pattern. Working paper.

Zhang, D., Nagurney A. and Wu J., 2001. On the Equivalence between Stationary Link Flow Patterns and Traffic Network Equilibria, Transportation Research 35B(8),731- 748.

# A Computable Theory of Dynamic Congestion Pricing

*Terry L. Friesz: Penn State University, USA, tfriesz@psu.edu*
*Changhyun Kwon: Penn State University, USA, chkwon@psu.edu*
*Reetabrata Mookherjee: Penn State University, USA, reeto@psu.edu*

**Abstract**

In this paper we present a theory of dynamic congestion pricing for the day-to-day as well as the within-day time scales. The equilibrium design problem emphasized herein takes the form of an MPEC, which we call the Dynamic Optimal Toll Problem with Equilibrium Constraints, or DOTPEC. The DOPTEC formulation we employ recalls an important earlier result that allows the equilibrium design problem to be stated as a single level problem, a result which is surprisingly little known. The DOPTEC maintains the usual design objective of minimizing the system travel cost by appropriate toll pricing. In addition, we present a direct dynamic generalization of the static efficient toll problem and show that such a generalization involves certain ambiguities that do not arise when the DOTPEC is formed as we suggest herein. A numerical example for the DOTPEC is provided.

*Keywords:* Dynamic congestion pricing; Dynamic user equilibrium; Differential Variational Inequality; Optimal Control

## 1 Introduction

The advent of new commitments by municipal, state and federal governments to construct and operate roadways whose tolls may be set dynamically has brought into sharp focus the need for a computable theory of dynamic tolls. Moreover, it is clear from the policy debates that surround the issue of dynamic tolls that pure economic efficiency is not the sole or even the most prominent objective of any dynamic toll mechanism that will be implemented. Rather, equity considerations as well as preferential treatment for certain categories of commuters must be addressed by such a mechanism. Accordingly, we introduce in this paper the dynamic user equilibrium optimal toll problem and discuss several plausible algorithms for its solution; we also provide detailed numerical results that document the performance of one of those algorithms.

The dynamic user equilibrium optimal toll problem should not be confused with the traditional congestion pricing paradigm associated with static user equilibrium and usually accredited to Beckmann, McGuire and Winsten (1956). Rather, the dynamic user equilibrium optimal toll problem is closely related to the equilibrium network design problem which is now widely recognized to be a specific instance of a mathematical program with equilibrium constraints (MPEC). In fact it will be convenient to refer to the dynamic user equilibrium optimal toll problem as the dynamic optimal toll problem with equilibrium constraints or DOTPEC, where it is understood that the equilibrium of interest is a dynamic user equilibrium.

The relevant background literature for the DOTPEC includes a paper by Friesz, Bernstein and Kydes (2002) who discuss a version of the DOTPEC but for the day-to-day time scale rather than the dual (within-day as well as day-to-day) time scale formulation emphasized in this paper. Also pertinent is the paper by Friesz, Bernstein and Stough (1996) which discusses dynamic disequilibrium network design and the review by Liu (2004) which considers multi-period efficient tolls.

As noted earlier, the DOTPEC is not the same as the problem of determining efficient tolls including the latter's multiperiod generalization. Yet the exact nature of the differences and similarities is not known and has never been studied. To study the dynamic efficient toll problem it is necessary to employ some form of dynamic user equilibrium model. We elect the formulation due to Friesz, Bernstein, Suo and Tobin (2001) and Friesz and Mookherjee (2006) and its varieties analyzed by Ban, Liu, Ferris and Ran (2005) and others. The dynamic efficient toll formulation will be constructed by direct analogy to the static efficient toll problem formulation of Hearn and Yildirim (2002).

Clearly the main focus of this paper is the formulation and solution of the DOTPEC. To this end, again using the DUE formulation reported in Friesz et al. (2001) and Friesz and Mookherjee (2006), we will form a Stackelberg game that envisions a central authority minimizing social costs through its control of link tolls subject to DUE constraints with additional side constraints for equity and other policy considerations. Also, as we will allow multiple target arrival times of the users, the within-day scale model can be easily extended to include day-to-day time scale properly. Of course there are several ways such a model may be formulated. The formulations we shall emphasize are based on our work on differential variational inequalities and equilibrium network design.

In particular, Tan, Gershwin and Athans (1979) show that a system of inequalities is equivalent to a static user equilibrium, and Friesz and Shah (2001) state the equilibrium network design problem as a single level mathematical program. An extension of this result to the dynamic setting allows us in this paper to have an equivalent optimal control problem statement of the DOTPEC. We consider two principal methods for solving this optimal control problem: (1) descent in Hilbert space without time discretization, and (2) a finite dimensional approximation solved as a nonlinear program. In both approaches we employ a numerical scheme like that in Friesz and Mookherjee (2006) for dealing with time lags that arise in the flow propagation constraints. Friesz and Mookherjee (2006) establish through numerical examples that such an approach for dealing with time lags works and is computationally practical.

In an example provided in the last part of this paper, we study a small network numerically and determine its optimal dynamic tolls.

## 2 Notation and Model Formulation

In this section we purposely repeat key portions of the time-lagged DUE formulation given in Friesz et al. (2001). The network of interest will form a directed graph $G(\mathcal{N},\mathcal{A})$, where $\mathcal{N}$ denotes the set of nodes and $\mathcal{A}$ denotes the set of arcs; the respective cardianlities of these sets are $|\mathcal{N}|$ and $|\mathcal{A}|$. An arbitrary path $p \in \mathcal{P}$ of the network is

$$p \equiv \left\{a_1, a_2, ..., a_i, ..., a_{m(p)}\right\}$$

where $\mathcal{P}$ is the set of all paths and $m(p)$ is the number of arcs of $p$. We also let $t_e$ denote the time at which flow exists an arc, while $t_d$ is the time of departure from the origin of the same flow. The exit time function $\tau^p_{a_i}$ therefore obeys

$$t_e = \tau^p_{a_i}(t_d)$$

The relevant arc dynamics are

$$\begin{aligned} \frac{dx^p_{a_i}(t)}{dt} &= g^p_{a_{i-1}}(t) - g^p_{a_i}(t) \qquad \forall p \in \mathcal{P}, \quad i \in \{1,2,...,m(p)\} \\ x^p_{a_i}(t) &= x^p_{a_{i,0}} \qquad \forall p \in \mathcal{P}, \quad i \in \{1,2,...,m(p)\} \end{aligned}$$

where $x^p_{a_i}$ is the traffic volume of arc $a_i$ contributed by path $p$, $g^p_{a_i}$ is flow exiting arc $a_i$ and $g^p_{a_{i-1}}$ is flow entering arc $a_i$ of path $p \in \mathcal{P}$. Also, $g^p_{a_0}$ is the flow exiting the origin of path $p$; by convention we call this the flow of path $p$ and use the symbolic name

$$h_p = g^p_{a_0}$$

Furthermore

$$\delta_{a_i p} = \begin{cases} 1 & \text{if } a_i \in p \\ 0 & \text{if } a_i \notin p \end{cases}$$

so that

$$x_a(t) = \sum_{p \in \mathcal{P}} \delta_{ap} x^p_a(t) \qquad \forall a \in \mathcal{A}$$

is the total arc volume.

Arc unit delay is $D_a(x_a)$ for each arc $a \in \mathcal{A}$. That is, arc delay depends on the number of vehicles in front of an auto as that auto enters an arc. Of course total path traversal time is

$$D_p(t) = \sum_{i=1}^{m(p)} \left[\tau^p_{a_i}(t) - \tau^p_{a_{i-1}}(t)\right] = \tau^p_{a_{m(p)}}(t) - t \qquad \forall p \in \mathcal{P}$$

It is expedient to introduce the following recursive relationships that must hold in light of the above development:

$$\begin{aligned}\tau^p_{a_1}(t) &= t + D_{a_1}\left[x_{a_1}(t)\right] \quad \forall p \in \mathcal{P} \\ \tau^p_{a_i}(t) &= \tau^p_{a_{i-1}}(t) + D_{a_i}\left[x_{a_i}(\tau^p_{a_{i-1}}(t))\right] \quad \forall p \in \mathcal{P}, \quad i \in \{2,3,...,m\,(p)\}\end{aligned}$$

from which we have the nested path delay operators first proposed by Friesz et al. (1993):

$$D_p(t,x) \equiv \sum_{i=1}^{m(p)} \delta_{a_i p} \Phi_{a_i}(t,x) \quad \forall p \in \mathcal{P},$$

where

$$x = (x^p_{a_i} : p \in \mathcal{P}, i \in \{1,2,...,m\,(p)\}$$

and

$$\begin{aligned}\Phi_{a_1}(t,x) &= D_{a_1}(x_{a_1}(t)) \\ \Phi_{a_2}(t,x) &= D_{a_2}(x_{a_2}(t+\Phi_{a_1})) \\ \Phi_{a_3}(t,x) &= D_{a_3}(x_{a_3}(t+\Phi_{a_1}+\Phi_{a_2})) \\ &\vdots \\ \Phi_{a_i}(t,x) &= D_{a_i}(x_{a_i}(t+\Phi_{a_1}+\cdots+\Phi_{a_{i-1}})) \\ &= D_{a_i}(x_{a_i}(t+\sum_{j=1}^{i-1}\Phi_{a_j})).\end{aligned}$$

To ensure realistic behavior, we employ asymmetric early/late arrival penalties

$$F\left[t + D_p\,(t,x) - t_A\right]$$

where $t_A$ is the desired arrival time and

$$\begin{aligned}t + D_p(t,x) > t_A \Longrightarrow &\quad F(t + D_p(t,x) - t_A) = \chi^L(x,t) > 0 \\ t + D_p(t,x) < t_A \Longrightarrow &\quad F(t + D_p(t,x) - t_A) = \chi^E(x,t) > 0 \\ t + D_p(t,x) = t_A \Longrightarrow &\quad F(t + D_p(t,x) - t_A) = 0 \\ \chi^L(t,x) > \chi^E(t,x)\end{aligned}$$

Let us define the arc tolls $y_a$ for each arc $a \in \mathcal{A}$. It is assumed that users pay the toll at the entrance of the arc. Then the path tolls $y_p$ for each path $p \in \mathcal{P}$ are

$$y_p\,(t) = \sum_{i=1}^{m(p)} \delta_{a_i p} y_{a_i}\left(t + \Phi_{a_{i-1}}(t,x)\right) \quad \forall p \in \mathcal{P}$$

where $\Phi_{a_0}(t,x) = 0$. If the tolls are paid when users exit arcs, then the path toll becomes

$$y_p\,(t) = \sum_{i=1}^{m(p)} \delta_{a_i p} y_{a_i}\,(t + \Phi_{a_i}(t,x)) \quad \forall p \in \mathcal{P}$$

We now combine the actual path delays and arrival penalties to obtain the *effective delay operators*

$$\Psi_p(t,x) = D_p(t,x) + F\left\{t + D_p(x,t) - T_A\right\} \qquad \forall p \in \mathcal{P} \tag{1}$$

Since the volume which enters and exits an arc should satisfy the conservation law, we must have

$$\int_0^t g^p_{a_{i-1}}\,(t)\,dt = \int_{D_{a_i}(x_{a_i}(0))}^{t+D_{a_i}(x_{a_i}(t))} g^p_{a_i}\,(t)\,dt \quad \forall p \in \mathcal{P}, i \in [1, m(p)] \tag{2}$$

where $g^p_{a_0}(t) = h_p(t)$. Differentiating the both sides of (2) with respect to time $t$ and using the chain rule, we have

$$\begin{aligned} h_p(t) &= g^p_{a_1}(t + D_{a_1}(x_{a_1}(t)))(1 + D'_{a_1}(x_{a_1}(t))\dot{x}_{a_1}) \qquad \forall p \in \mathcal{P} \\ g^p_{a_{i-1}}(t) &= g^p_{a_i}(t + D_{a_i}(x_{a_i}(t)))(1 + D'_{a_i}(x_{a_i}(t))\dot{x}_{a_i}) \qquad \forall p \in \mathcal{P}, \quad i \in [2, m(p)] \end{aligned}$$

These are *proper flow progression constraints* derived in a fashion that make them completely *consistent with the chosen dynamics and point queue model of arc delay.* These constraints involve a state dependent time lag $D_{a_i}(x_{a_i}(t))$ but make no explicit reference to the exit time functions. These flow propagation constraints describe the expansion and contraction of vehicle platoons; they were first presented by Friesz, Tobin, Bernstein and Suo (1995), Astarita (1995) and Astarita (1996) independently proposed flow propagation constraints that may be readily placed in the above form.

### 2.1 Dynamic User Equilibrium

Given the traveling cost $\Theta_p$ for path $p$, the infinite dimensional variational inequality formulation for dynamic network user equilibrium itself is: find $(g^*, h^*) \in \Omega$ such that

$$\langle \Theta(t, x(h^*)), (h - h^*) \rangle = \sum_{p \in \mathcal{P}} \int_{t_0}^{t_f} \Theta_p[t, x(h^*)] \left[ h_p(t) - h^*_p(t) \right] dt \geq 0 \tag{3}$$

for all $(g, h) \in \Omega$, all of whose solutions Friesz *et al* Friesz et al. (2001) show are dynamic user equilibria[1]. In particular the solutions of (3) obey

$$\Theta_p(t, x^*) > \mu_{ij} \Longrightarrow h^*_p(t) = 0 \tag{4}$$

$$h^*_p(t) > 0 \Longrightarrow \Theta_p(t, x^*) = \mu_{ij} \tag{5}$$

for $p \in \mathcal{P}_{ij}$ where $\mu_{ij}$ is the lower bound on achievable costs for any $ij$-traveler, given by

$$\mu_p = \operatorname{ess\,inf} \{\Theta_p(t, x) : t \in [t_0, t_f]\} \geq 0$$

and

$$\mu_{ij} = \min \left\{ \mu_p : p \in \mathcal{P}_{ij} \right\} \geq 0$$

We call a flow pattern satisfying (4) and (5) a *dynamic user equilibrium.* The behavior described by (4) and (5) is readily recognized to be a type of Cournot-Nash non-cooperative equilibrium. It is important to note that these conditions do not describe a stationary state, but rather a time varying flow pattern that is a Cournot-Nash equilibrium (or user equilibrium) at each instant of time.

## 3 The Efficient Toll

Hearn and Yildirim (2002) studied the efficient toll in the static setting with the traveling cost which is linear in the traffic flow. The objective of the efficient toll is to make the user equilibrium traffic flow equivalent to the system optimum by appropriate congestion pricing. To study the dynamic efficient toll, in addition to the effective delay operator defined in (1), let us have the *tolled effective delay operators* as

$$\Theta_p(t, x, y_p) = D_p(t, x) + F\{t + D_p(x, t) - T_A\} + y_p(t) \qquad \forall p \in \mathcal{P}$$

where $y_p$ denotes the toll for path $p$. Of course we have the relationship

$$\Theta_p(t, x, y_p) = \Psi_p(t, x) + y_p(t) \tag{6}$$

The system optimum is achieved by solving

$$\min \; J = \int_{t_0}^{t_f} \sum_{p \in \mathcal{P}} e^{-rt} \Psi_p(t, x) \, h_p(t) \, dt$$

[1] Although we have purposely supressed the functional analysis subtleties of the formulation, it should be noted that (3) involves an inner product in a Hilbert space, namely $\left(L^2[0, T]\right)^{|\mathcal{P}|}$.

subject to

$$\frac{dx_{a_i}^p(t)}{dt} = g_{a_{i-1}}^p(t) - g_{a_i}^p(t) \qquad \forall p \in \mathcal{P}, \quad i \in \{1,2,...,m(p)\} \tag{7}$$

$$x_{a_i}^p(t) = x_{a_{i,0}}^p \qquad \forall p \in \mathcal{P}, \quad i \in \{1,2,...,m(p)\}$$

$$h_p(t) = g_{a_1}^p(t + D_{a_1}(x_{a_1}(t)))(1 + D'_{a_1}(x_{a_1}(t))\dot{x}_{a_1}) \qquad \forall p \in \mathcal{P} \tag{8}$$

$$g_{a_{i-1}}^p(t) = g_{a_i}^p(t + D_{a_i}(x_{a_i}(t)))(1 + D'_{a_i}(x_{a_i}(t))\dot{x}_{a_i}) \qquad \forall p \in \mathcal{P}, \quad i \in [2, m(p)] \tag{9}$$

$$\sum_{p \in \mathcal{P}_{ij}} \int_{t_0}^{t_f} h_p(t)\, dt = Q_{ij} \quad \forall (i,j) \in \mathcal{W} \tag{10}$$

$$x \geq 0 \qquad g \geq 0 \qquad h \geq 0 \tag{11}$$

Let us define the set of mixed (state and control) constraints

$$\Omega \equiv \{(h,x) : (8), (9), (10) \text{ and } (11) \text{ are satisfied}\} \tag{12}$$

and the Hamiltonian function

$$H_1(h_p, \lambda, t) \equiv \sum_{p \in \mathcal{P}} e^{-rt} \Psi_p(t,x)\, h_p(t) + \sum_{p \in \mathcal{P}} \sum_{i=1}^{m(p)} \lambda_{a_i}^p \left( g_{a_{i-1}}^p(t) - g_{a_i}^p(t) \right)$$

The first-order necessary conditions for the system optimum are

$$0 \leq \sum_{p \in \mathcal{P}} \frac{\partial H_1\left(h_p^S, \lambda^S, t\right)}{\partial h_p} \left(h_p - h_p^S\right) = \sum_{p \in \mathcal{P}} \left[ e^{-rt} \left\{ \Psi_p^S(t,x) + \frac{\partial \Psi_p^S(t,x)}{\partial h_p} h_p^S \right\} + \lambda_{a_1}^{p,S} \right] \left(h_p - h_p^S\right) \quad \forall h \in \Omega \tag{13}$$

for each time instant $t \in [t_0, t_f]$, the state dynamics (7) and adjoint dynamics,

$$\begin{aligned} -\frac{d\lambda_{a_i}^{p,S}}{dt} &= \frac{\partial H_1^S}{\partial x_{a_i}^p} = e^{-rt} \frac{\partial \Psi_p\left(t, x^S\right)}{\partial x_{a_i}^p} \qquad \forall p \in \mathcal{P}, \quad i \in [1, m(p)] \\ \lambda_{a_i}^{p,S}(t_f) &= 0 \qquad \forall p \in \mathcal{P}, \quad i \in [1, m(p)] \end{aligned} \tag{14}$$

where the superscript $S$ denotes the corresponding values at the system optimum.

However, the dynamic tolled user equilibrium condition is

$$\sum_{p \in \mathcal{P}} \int_{t_0}^{t_f} e^{-rt} \left\{ \Theta_p\left[t, x\left(h^U\right), y_p^U\right] \right\} \left[ h_p(t) - h_p^U(t) \right] dt \geq 0 \quad \text{for all } h \in \Omega \tag{15}$$

subject to

$$\begin{aligned} \frac{dx_{a_i}^p(t)}{dt} &= g_{a_{i-1}}^p(t) - g_{a_i}^p(t) \qquad \forall p \in \mathcal{P}, \quad i \in \{1,2,...,m(p)\} \\ x_{a_i}^p(t) &= x_{a_{i,0}}^p \qquad \forall p \in \mathcal{P}, \quad i \in \{1,2,...,m(p)\} \\ (h,x) &\in \Omega \end{aligned}$$

where the superscript $U$ denotes the corresponding values at the tolled user equilibrium. We formulate the fictitious optimal control problem for the user equilibrium variation inequality problem, whose objective is:

$$\min \sum_{p \in \mathcal{P}} \int_{t_0}^{t_f} e^{-rt} \Theta_p\left[t, x\left(h^U\right), y_p^U\right] h_p(t)\, dt$$

with the same constraint set. The Hamiltonian function for this problem is

$$H_2(h_p, \mu, t) \equiv \sum_{p \in \mathcal{P}} e^{-rt} \Theta_p\left[t, x\left(h^U\right), y_p^U\right] h_p(t) + \sum_{p \in \mathcal{P}} \sum_{i=1 a_i}^{m(p)} \lambda_{a_i}^p \left( g_{a_{i-1}}^p(t) - g_{a_i}^p(t) \right)$$

The first-order necessary conditions are

$$0 \leq \sum_{p\in\mathcal{P}} \frac{\partial H_2\left(h_p^U, \mu^U, t\right)}{\partial h_p}\left(h_p - h_p^U\right) = \sum_{p\in\mathcal{P}}\left[e^{-rt}\left\{\Theta_p\left[t, x\left(h^U\right), y_p^U\right]\right\} + \lambda_{a_1}^{p,U}\right]\left(h_p - h_p^U\right) \tag{16}$$

for each time instant $t \in [t_0, t_f]$, the state dynamics (7) and adjoint dynamics,

$$\begin{aligned} -\frac{d\lambda_{a_i}^{p,U}}{dt} &= \frac{\partial H_2^U}{\partial x_{a_i}^p} = e^{-rt}\frac{\partial \Theta_p\left[t, x\left(h^U\right), y_p^U\right]}{\partial x_{a_i}^p} \qquad \forall p \in \mathcal{P}, \quad i \in [1, m(p)] \\ \lambda_{a_i}^{p,U}\left(t_f\right) &= 0 \qquad \forall p \in \mathcal{P}, \quad i \in [1, m(p)] \end{aligned} \tag{17}$$

Assume that the the traffic flow at the tolled user equilibrium achieves the system optimum, i.e., $h^U(t) = h^S(t)$, which is the result we want to obtain by the efficient toll. Comparing (13) and (16), we have

$$\begin{aligned} e^{-rt}\left\{\Psi_p^S(t,x) + \frac{\partial \Psi_p^S(t,x)}{\partial h_p} h_p^S\right\} + \lambda_{a_1}^{p,S} &= e^{-rt}\left\{\Theta_p\left[t, x\left(h^U\right), y_p^U\right]\right\} + \lambda_{a_1}^{p,U} \\ &= e^{-rt}\left\{\Psi_p\left(t, x^U\right) + y_p^U(t)\right\} + \lambda_{a_1}^{p,U} \end{aligned}$$

or, immediately

$$y_p^U(t) = \frac{\partial \Psi_p\left(t, x^S\right)}{\partial h_p} h_p^S + e^{rt}\left\{\lambda_{a_1}^{p,S} - \lambda_{a_1}^{p,U}\right\} \qquad \forall t \in [t_0, t_f] \tag{18}$$

The dynamic efficient toll (18) we obtained is hard to evaluate by its nature that involves two-point-boundary-value problems for two different problems, the system optimum and the user equilibrium.

# 4 Dynamic Optimal Toll Problem with Equilibrium Constraints (DOTPEC)

The DOTPEC is a design problem to achieve the maximum of social benefit constrained by the user equilibrium conditions. That is

$$\min \ J = \int_{t_0}^{t_f} \sum_{p\in\mathcal{P}} e_p^{-rt} \Psi_p(t,x)\, h_p(t)\, dt \tag{19}$$

subject to

$$\sum_{p\in\mathcal{P}} \int_{t_0}^{t_f} e^{-rt}\Theta_p\left[t, x(t), y_p(t)\right]\left[\bar{h}_p(t) - h_p(t)\right] dt \geq 0 \ \ \forall\ \bar{h} \in \Omega \tag{20}$$

$$\frac{dx_{a_i}^p(t)}{dt} = g_{a_{i-1}}^p(t) - g_{a_i}^p(t) \qquad \forall p \in \mathcal{P}, \quad i \in \{1, 2, ..., m(p)\} \tag{21}$$

$$x_{a_i}^p(t) = x_{a_{i,0}}^p \qquad \forall p \in \mathcal{P}, \quad i \in \{1, 2, ..., m(p)\} \tag{22}$$

$$x \geq 0 \tag{23}$$

$$y_a(t) \geq 0 \qquad \forall a \in \mathcal{A} \tag{24}$$

$$h \in \Omega \tag{25}$$

where the set $\Omega$ is defined as (12). The infinite dimensional optimization problem (19)-(25) is in the class of mathematical programs with equilibrium constraints, or MPECs, which are usually hard to find the solutions. The main difficulty arise in the presence of the equilibrium constraint. However, let us consider the following alternative statement of the DUE conditions:

**Theorem 1** *Given that the traveling cost for path $p$ is $\Theta_p$, a nonnegative path flow $h$ is a user equilibrium if and only if*

$$\Theta_p \geq \frac{\sum_{p\in\mathcal{P}_{ij}} \int_{t_0}^{t_f} \Theta_p\left[t, x(t), y_p(t)\right] h_p(t)\, dt}{\sum_{p\in P_{ij}} \int_{t_0}^{t_f} h_p(t)\, dt} = \mu_{ij} \qquad \forall p \in \mathcal{P}_{ij}, \quad (i,j) \in \mathcal{W}$$

*are satisfied*

**Proof.** The dynamic user equilibrium condition stated in (4) and (5) can be modeled as an equivalent complementarity problem, that is

$$\left[\Theta_p\left(t,x^*\right)-\mu_{ij}\right]h_p^*\left(t\right)=0,\qquad \Theta_p\left(t,x^*\right)-\mu_{ij}\geq 0,\qquad h_p^*\left(t\right)\geq 0 \tag{26}$$

for all $t\in\left[t_0,t_f\right],p\in\mathcal{P}_{ij},(i,j)\in\mathcal{W}$. Integrating the complementarity equation in (26) over time horizon and summing for all paths, we obtain

$$\sum_{p\in\mathcal{P}_{ij}}\int_{t_0}^{t_f}\left[\Theta_p\left(t,x^*\right)-\mu_{ij}\right]h_p^*\left(t\right)dt=0\qquad\forall\left(i,j\right)\in\mathcal{W}$$

or

$$\sum_{p\in\mathcal{P}_{ij}}\int_{t_0}^{t_f}\Theta_p\left(t,x^*\right)h_p^*\left(t\right)dt=\mu_{ij}\sum_{p\in\mathcal{P}_{ij}}\int_{t_0}^{t_f}h_p^*\left(t\right)dt\qquad\forall\left(i,j\right)\in\mathcal{W} \tag{27}$$

In addition, due to the nonnegativity conditions in (26), the result (27) is also equivalent to the dynamic user equilibrium conditions. ■

Now we may replace the DUE constraint (20) by the following equality and inequality constraints:

$$\begin{aligned}\mu_{ij} &= \frac{\sum_{p\in\mathcal{P}_{ij}}\int_{t_0}^{t_f}\Theta_p\left[t,x\left(t\right),y_p\left(t\right)\right]h_p\left(t\right)dt}{\sum_{p\in P_{ij}}\int_{t_0}^{t_f}h_p\left(t\right)dt}\qquad\forall\left(i,j\right)\in\mathcal{W}\\ \Theta_p &\geq \mu_{ij}\qquad\forall p\in\mathcal{P}_{ij},\quad(i,j)\in\mathcal{W}\end{aligned}$$

With the above replacements, we now obtain an infinite dimensional mathematical program with inequality and equality constraints. To solve this problem, one may consider a descent in Hilbert spaces used in Friesz and Mookherjee (2006), or a nonlinear program approach with appropriate time discretizations. When we discretize the planning horizon, we may use off-the-shelf solvers like GAMS/MINOS.

# 5 Multiple Time Scales

Let $\tau\in\Upsilon\equiv\{1,2,...,L\}$ be one typical day within the planning horizon, and take the length of each day to be $\Delta$, while the clock time within each day $\tau$ is presented by $t\in\left[\left(\tau-1\right)\Delta,\tau\Delta\right]$ for all $\tau\in\{1,2,...,L\}$. The planning horizon consists of $L$ consecutive days. We assume the travelling demand for each day changes based on the moving average of congestion experienced over previous days. We postulate that the travelling demand $Q_{ij}^{\tau}$ for day $\tau$ between a given O-D pair $(i,j)\in\mathcal{W}$ determined by the following system of difference equations:

$$\begin{aligned}Q_{ij}^{\tau+1} &= \left[Q_{ij}^{\tau}-\eta_{ij}^{\tau}\left\{\frac{\sum_{p\in\mathcal{P}_{ij}}\sum_{j=0}^{\tau-1}\int_{j\cdot\Delta}^{(j+1)\cdot\Delta}\Psi_p\left[t,x\left(h^*,g^*\right)\right]dt}{\left|\mathcal{P}_{ij}\right|\cdot\tau\cdot\Delta}-\chi_{ij}\right\}\right]^{+}\qquad\forall\tau\in\{1,2,...L-1\}\\ Q_{ij}^{1} &= \tilde{Q}_{ij}\end{aligned} \tag{28}$$

where $\tilde{Q}_{ij}\in\mathbb{R}_+$ is the fixed traveling demand for the O-D pair $(i,j)\in\mathcal{W}$ for the first day. The operator $[x]^+$ is equivalent to $\max\left[0,x\right]$.

# 6 Algorithms for Solving the DOTPEC

In this secion, we provide two different algorithms for solivng DOTPEC: (1) descent in Hilbert spaces without time discretization, and (2) a finite dimensional approximation solved as a nonlinear program. In both approaches, the state-depende time shifts, which are path delays in DOTPEC, are hard to treat. In the implicit fixed point theorem for the dynamic user equilibrium Friesz and Mookherjee (2006) used control and state information from the previous iteration to approximate the current time delay, which enables to have constant time delays instead of state-dependent delays at each iteration. Adopting the idea, we propose a heuristic numerical scheme to solve the DOTPEC as following:

1. Pick a feasible set of control variables.
2. Determine the state-dependent time delays according to the current controls.
3. Solve a DOTPEC with the fixed time delays.
4. Update the control and repeat Step 2 and Step 3 until the control variables converge to the solution.

In solving a DOTPEC with the fixed time delays at each iteration, we have two above-mentioned numerical algorithms.

## 6.1 Descent in Hilbert Spaces

To study the descent in Hilbert spaces, let us consider an abstract optimal control problem with state dependent time shifts articulated as follows:

$$\min \ J = \int_{t_0}^{t_f} F(x, u, u_D, t)dt \tag{29}$$

subject to

$$x(u, u_D, t) \in \Omega = \left\{ x : \frac{dx}{dt} = f(x, u, u_D, t), x(0) = 0, G(x, u, u_D, t) = 0, x \geq 0 \right\} \in (\mathcal{H}^1[t_0, t_f])^n$$

and

$$\begin{array}{rcl} u & \in & U \subseteq (L^2[t_0, t_f])^m \\ u_D & \equiv & u(t + D(x)) : (\mathcal{H}^1[t_0, t_f])^n \times \mathbb{R}^1_+ \longrightarrow (L^2[t_0, t_f])^m \\ f & : & (\mathcal{H}^1[t_0, t_f])^n \times (L^2[t_0, t_f])^{2m} \times \mathbb{R}^1_+ \longrightarrow (L^2[t_0, t_f])^m \\ F & : & (\mathcal{H}^1[t_0, t_f])^n \times (L^2[t_0, t_f])^{2m} \times \mathbb{R}^1_+ \longrightarrow (L^2[t_0, t_f])^m \\ G & : & (\mathcal{H}^1[t_0, t_f])^n \times (L^2[t_0, t_f])^{2m} \times \mathbb{R}^1_+ \longrightarrow (L^2[t_0, t_f])^m \end{array}$$

where $(L^2[t_0, t_f])^m$ is the $m-$fold product of the space of square integrable function space $L^2[t_0, t_f]$ and $(\mathcal{H}^1[t_0, t_f])^n$ is the $n-$fold product of the Sobolev space $\mathcal{H}^1[t_0, t_f]$ for the real interval $[t_0, t_f] \subset \mathbb{R}^1_+$.

Applying descent in Hilbert space method to the problem, we penaltize the equality constraints and the non-negativity constraints as follows:

$$\min \ \tilde{J} = \int_{t_0}^{t_f} F(x, u, u_D, t)dt + \frac{1}{2}\int_{t_0}^{t_f} \sum_i \eta_i (G_i(x, u, u_D, t))^2 dt + \frac{1}{2}\int_{t_0}^{t_f} \sum_i \rho_i \min(0, x_i)^2 dt$$

subject to

$$x(u, u_D, t) \in \tilde{\Omega} = \left\{ x : \frac{dx}{dt} = f(x, u, u_D, t), x(0) = x_0 \right\} \in (\mathcal{H}^1[t_0, t_f])^n,$$

where $\xi_i, \rho_i$ and $\eta_i$ are increasing sequences.

This problem can be solved using a continuous time gradient projection method or a discrete time gradient projection method supplemented by spline approximations. For the penalized problem, the algorithm can be stated as following:

- **Step 0**. **Initialization**. Pick $u^k(t) \in U$ and set $k = 0$.
- **Step 1**. **Finding state variables**. Solve the state dynamics

$$\begin{array}{rcll} \dfrac{dx}{dt} & = & f(x, u^k, u^k_D, t) & (30) \\ x(0) & = & x_0 & (31) \end{array}$$

and call the solution $x^k(t)$.

- **Step 2**. **Finding adjoint variables**. Solve the adjoint dynamics

$$-\frac{d\lambda}{dt} = \nabla_x H^k|_{x=x^k} \quad (32)$$
$$\lambda(t_f) = 0 \quad (33)$$

where $H$ is the Hamiltonian for the penalized optimal control problem based on current information:

$$H^k = F(x^k, u^k, u_D^k, t) + \frac{1}{2}\sum_i \rho_i^k \min(0, x_i^k)^2 + \frac{1}{2}\sum_i \eta_i^k (G_i(x^k, u^k, u_D^k, t))^2 + \lambda^T f(x^k, u^k, u_D^k, t)$$

Call the solution $\lambda^k(t)$.

- **Step 3**. **Finding the gradient**. Determine

$$\nabla_u J^k(t) \equiv \nabla_u H^k$$

- **Step 4**. **Updating the current control**. For a suitably small step size

$$\theta_k \in \mathbb{R}^1_{++}$$

determine

$$u^k(t) = P_U[u^k(t) - \theta_k \nabla_u J^k]$$

- **Step 5**. **Stopping Test**. For $\epsilon \in \mathbb{R}^1_{++}$, a preset tolerance, stop if

$$||u^{k+1} - u^k|| < \epsilon$$

and declare

$$u^* \approx u^{k+1}$$

Otherwise set $k = k + 1$ and go to Step1.

## 6.2 Discrete-time Approximation of DOTPEC

The optimal control problem (19)-(25) may be given as the following discrete time approximation:

$$\min\ J = \sum_{k=0}^{N}\sum_{p\in\mathcal{P}} \phi(k)\, e_p^{-rt_k}\, \Psi_p[t_k, x(t_k)]\, h_p(t_k)\, \Delta$$

subject to

$$\mu_{ij} = \frac{\sum_{p\in\mathcal{P}_{ij}}\sum_{k=0}^{N}\phi(k)\,\Theta_p[t_k, x(t_k), y_p(t_k)]\,h_p(t_k)\,\Delta}{\sum_{p\in P_{ij}}\sum_{k=0}^{N}\phi(k)\,h_p(t_k)\,\Delta} \qquad \forall(i,j)\in\mathcal{W}$$
$$\Theta_p(t_k) \geq \mu_{ij} \qquad \forall k\in[0,N],\quad p\in\mathcal{P}_{ij},\quad (i,j)\in\mathcal{W}$$
$$x_{a_i}^p(t_{k+1}) = x_{a_i}^p(t_k) + \Delta\left[g_{a_{i-1}}^p(t_k) - g_{a_i}^p(t_k)\right]$$
$$\forall k\in[0,N-1],\quad p\in\mathcal{P},\quad i\in[1,m(p)]$$
$$x_{a_i}^p(t_0) = x_{a_{i,0}}^p \qquad \forall p\in\mathcal{P},\quad i\in\{1,2,...,m(p)\}$$
$$x(t_k)\geq 0 \qquad \forall k\in[0,N]$$
$$h_p(t_k) = g_{a_1}^p(t_k + D_{a_1}(x_{a_1}(t_k)))(1 + D'_{a_1}(x_{a_1}(t_k))\dot{x}_{a_1}) \quad \forall k\in[0,N], p\in\mathcal{P}$$
$$g_{a_{i-1}}^p(t_k) = g_{a_i}^p(t_k + D_{a_i}(x_{a_i}(t_k)))(1 + D'_{a_i}(x_{a_i}(t_k))\dot{x}_{a_i})$$
$$\forall k\in[0,N], p\in\mathcal{P}, i\in[2,m(p)]$$
$$\sum_{p\in\mathcal{P}_{ij}}\sum_{k=0}^{N}\phi(k)\,h_p(t_k)\,\Delta = Q_{ij} \quad \forall(i,j)\in\mathcal{W}$$
$$y_a(t_k)\geq 0 \qquad \forall a\in\mathcal{A},\quad k\in[0,N]$$
$$x(t_k)\geq 0 \qquad g(t_k)\geq 0 \qquad h(t_k)\geq 0 \qquad \forall k\in[0,N]$$

where $k$ takes non-negative integer values, $\Delta$ is the discrete time step that divides the time interval $[t_0, t_f]$ into $N$ equal segments, $\phi(k)$ is the coefficient which arises from a trapezoidal approximation of integrals, that is

$$\phi(k) = \begin{cases} 0.5 & \text{if } k = 0 \text{ and } N \\ 1 & \text{otherwise} \end{cases}$$

and

$$t_k = k\Delta$$

One advantage of time discretization is that we can now completely eliminate state variables (arc volumes) from the problem by noting that

$$x^p_{a_i}(t_{k+1}) = x^p_{a_{i,0}} + \sum_{r=0}^{k} \Delta \left[ g^p_{a_{i-1}}(t_r) - g^p_{a_i}(t_r) \right]$$

$$\forall k \in [0, N-1], \quad p \in \mathcal{P}, \quad i \in \{1, 2, ..., m(p)\}$$

As a consequence one obtains a finite dimensional mathematical program, which may be solved by conventional algorithms developed for such problems.

# 7 Numerical Example

In what follow, we consider a 3 arc, 3 node network shown in Fig 1. The arc labeling and arc delay functions for this network are summarized in the following table:

| Arc name | From node | To node | Arc delay, $D_a(x_a(t))$ |
|---|---|---|---|
| $a_1$ | 1 | 2 | $2 + (x_{a_1}/200)$ |
| $a_2$ | 2 | 3 | $1 + (x_{a_2}/150)$ |
| $a_3$ | 2 | 3 | $3 + (x_{a_3}/100)$ |

There are 2 paths connecting the single OD pair formed by nodes 1 and 3, namely:

$$\mathcal{P}_{13} = \{p_1, p_2\}, \qquad p_1 = \{a_1, a_2\}, \qquad p_2 = \{a_1, a_3\}$$

The conrols (path flows and arc exit flows) and states (path-specific arc traffic volumes) associated with the network are:

| Path | Path Flow | Arc Exit Flow | Traffic Volume of Arc |
|---|---|---|---|
| $p_1$ | $h_{p_1}$ | $g^{p_1}_{a_1}, g^{p_1}_{a_2}$ | $x^{p_1}_{a_1}, x^{p_1}_{a_2}$ |
| $p_2$ | $h_{p_2}$ | $g^{p_2}_{a_1}, g^{p_2}_{a_3}$ | $x^{p_2}_{a_1}, x^{p_2}_{a_3}$ |

We consider three-day toll planning in which each day is 24 hours, hence, $\Delta = 24$ and $L = 3$. We assume there is the initial travel demand $\tilde{Q} = 150$ units from node 1 (origin) to node 3 (destination). The threshhold for traveling cost is $\chi = 8300$ and the speed of changes in traveling demand is $\eta = 0.01$. The desired arrival time for each day is $t_A = 12$, and we employ the symmetric early/late arrival penalty

$$F[t + D_p(x,t) - t_A] = 5[t + D_p(x,t) - t_A]^2$$

Further, without any loss of generality, we take

$$x^p_{a_i}(0) = 0 \quad \forall i \in [1, m(p)], p \in \mathcal{P}$$

We forgo the detailed symbolic statement of this example, and, instead, provide numerical results in graphical form for the solution. Path flows and arc exit flows for paths $p_1$ and $p_2$ are presented in Figures 2 and 3, while path flows and toll at each arc are in Figures 4, 5 and 6. In this case, we have tolls tend to be proportional to the path flows. When, for path $p_1$, we compare the effective path delay operator with toll (6) with path flow (which is the departure rate) by plotting both for the same time scale, Figure 7 is obtained. This figure shows that departure rate peaks when the associated effective path delay achieves a local minimum, thereby demonstrating that an user equilibrium has been found. Similar comparisons are made for paths $p_2$ in Figure 8.

The daily changes of traveling demand from the origin to destination according to the difference equation (28) are given in Figure 9.

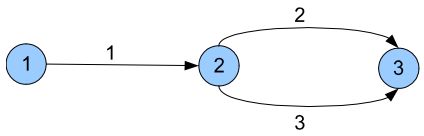



Figure 1: 3-Arc 3-Node Traffic Network

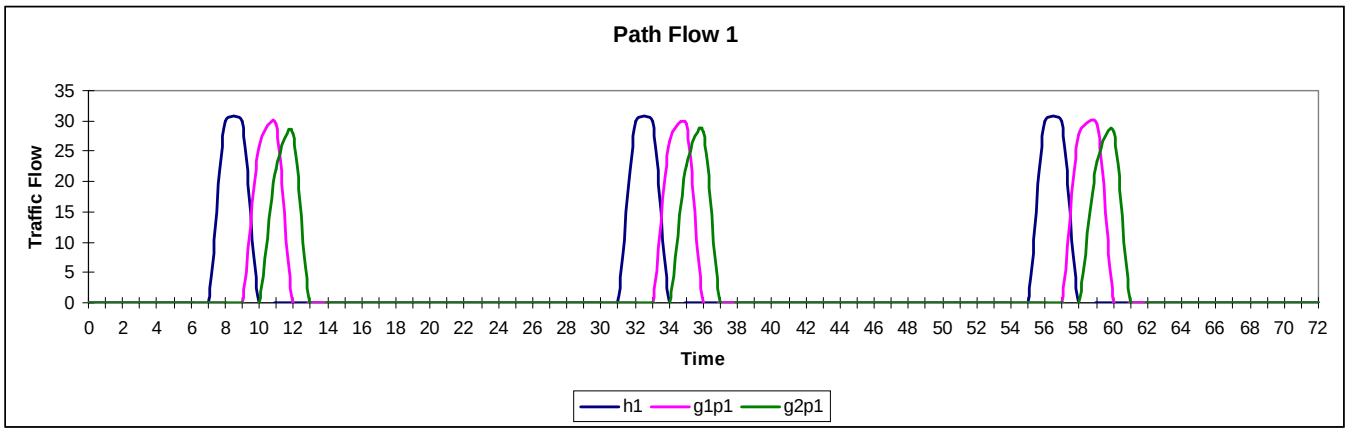



Figure 2: Path and arc exit flows for path $p_1$.

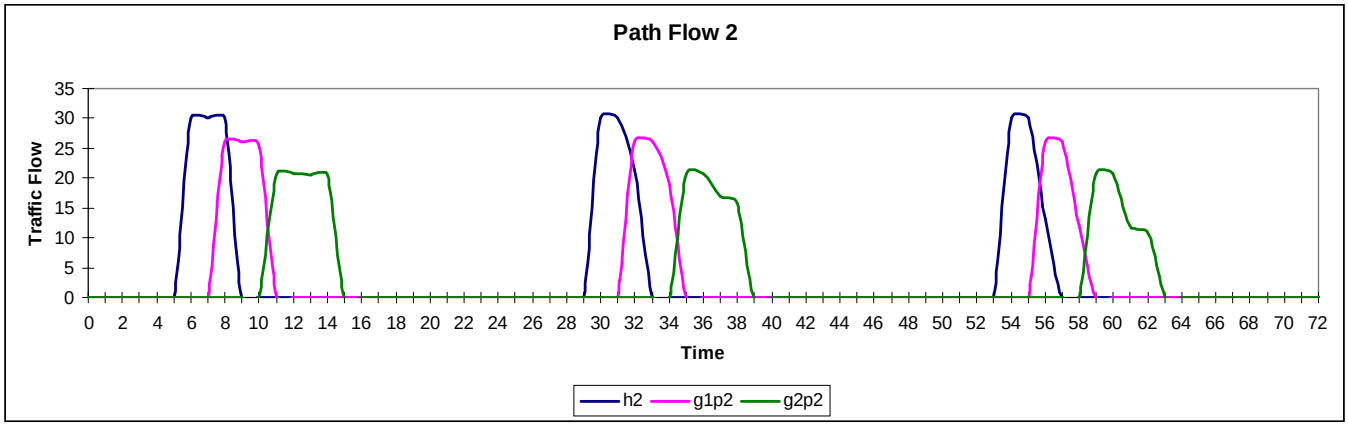



Figure 3: Path and arc exit flows for path $p_2$.

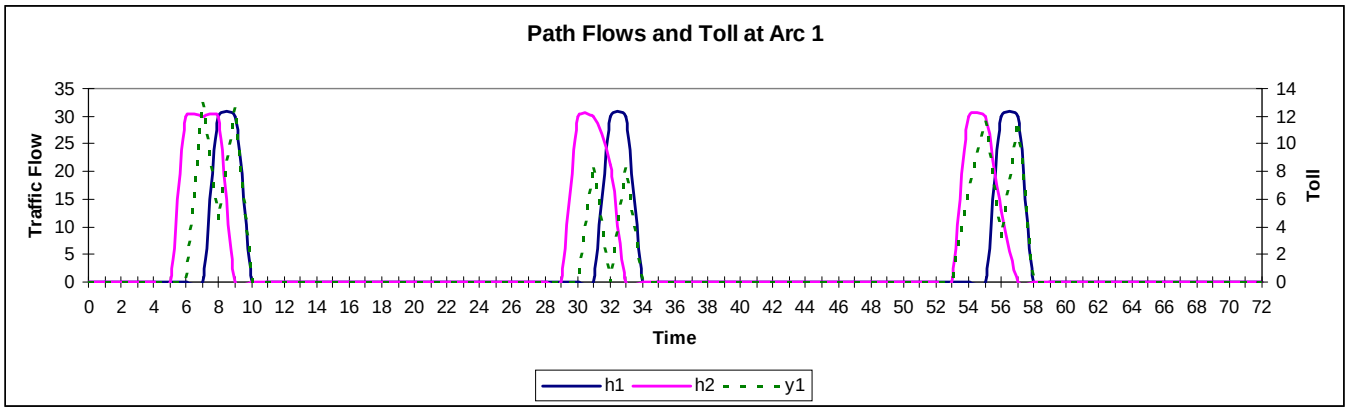



Figure 4: Path flows and toll at arc $a_1$.

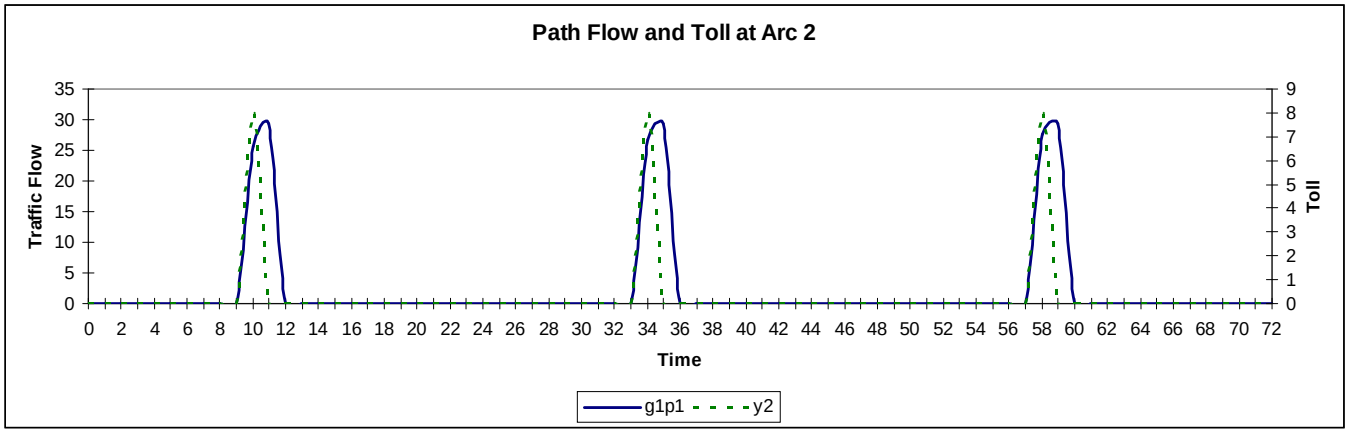



Figure 5: Path flow and toll at arc $a_2$.

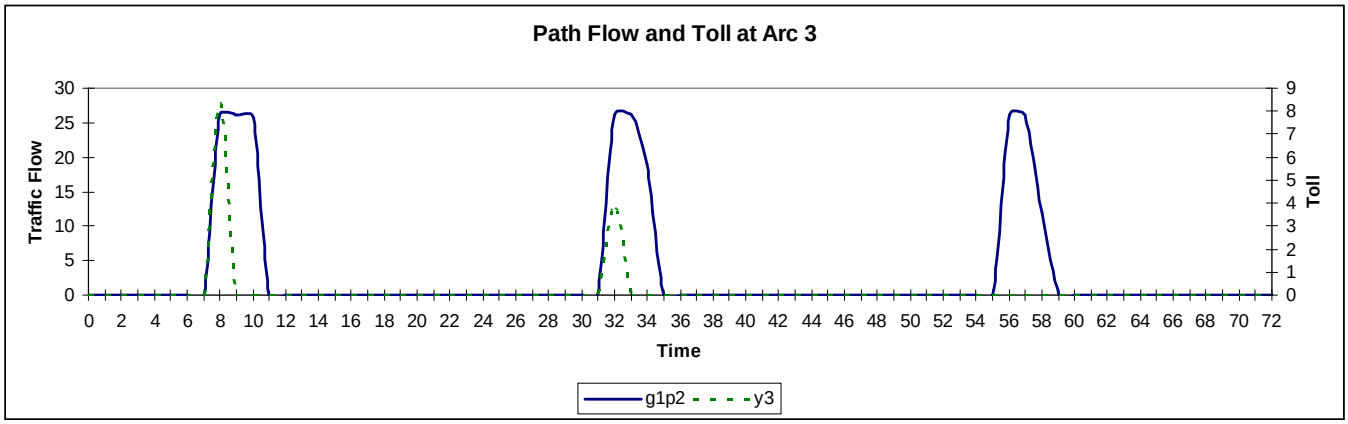



Figure 6: Path flow and toll at arc $a_3$.

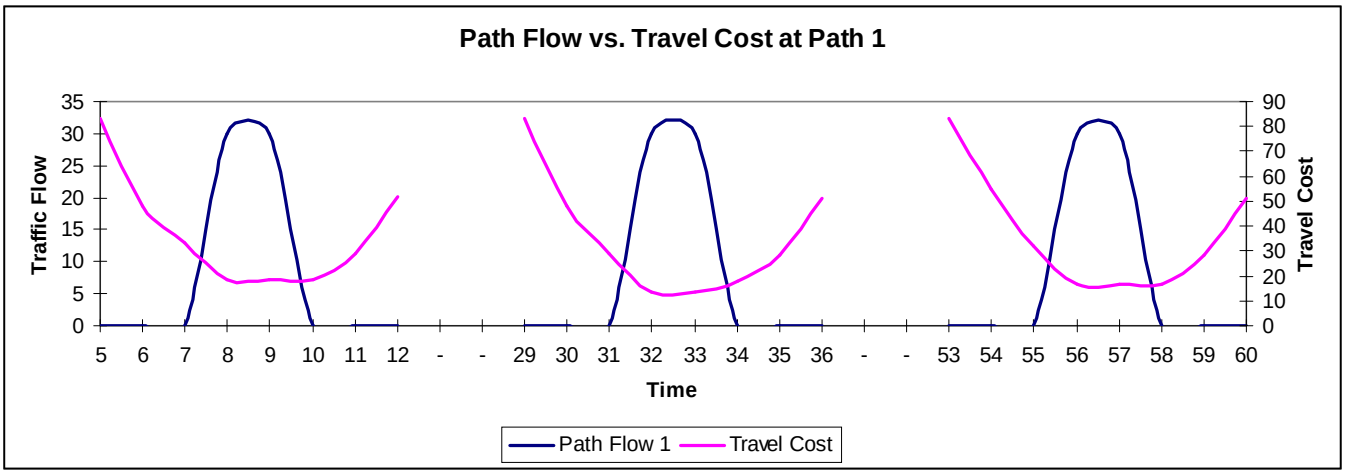



Figure 7: Comparison of path flow and associated unit travel costs for path $p_1$.

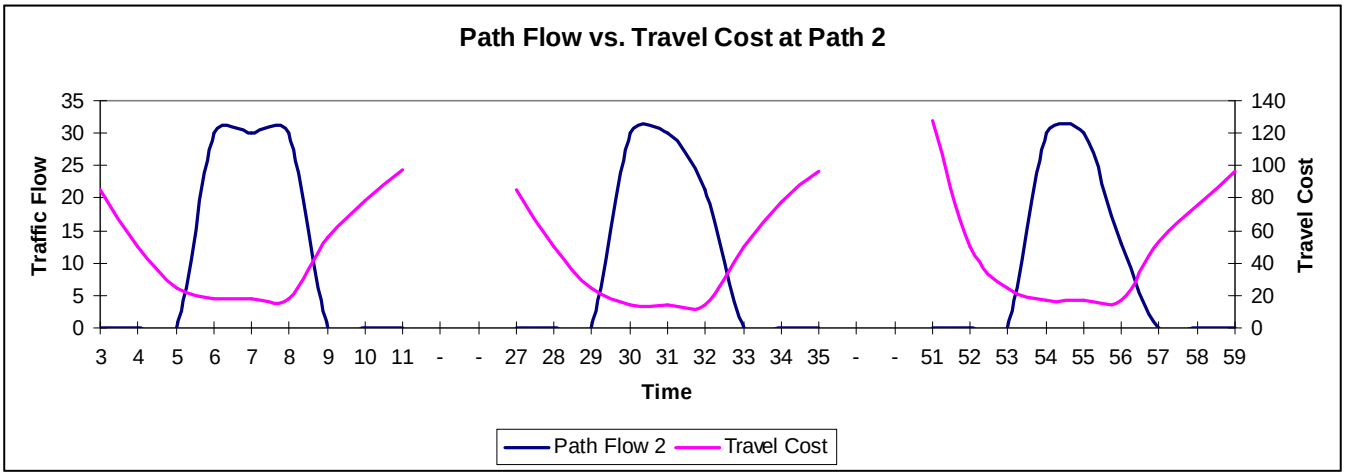



Figure 8: Comparison of path flow and associated unit travel costs for path $p_2$.

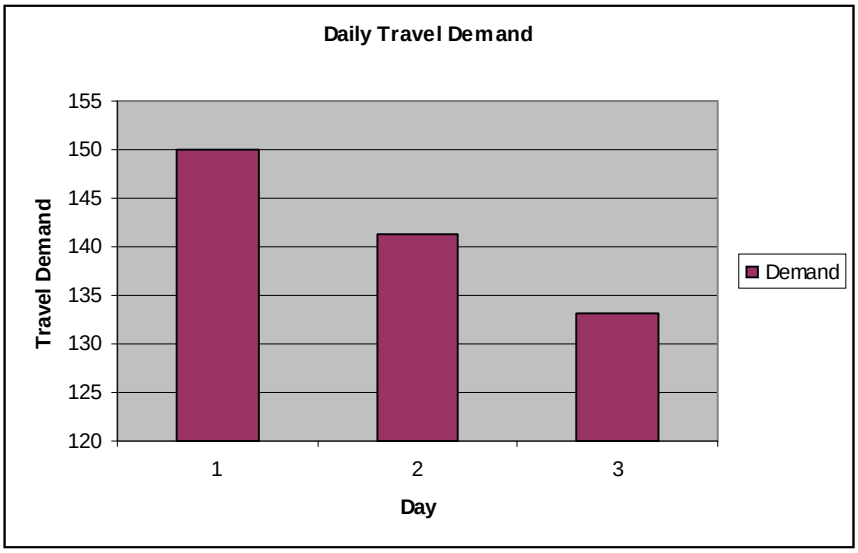



Figure 9: Daily changes of travel demand from the origin (node 1) to the destination (node 3)

# 8 Concluding Remarks

We have presented a study of the dynamic efficient toll and a version of the DOTPEC that are central to the field of dynamic congestion pricing. We have shown that the dynamic efficient toll is hard to solve so that the DOTPEC formulation of toll pricing becomes essential. For the DOTPEC, we recast the equilibrium constraints to the systems of inequality and equality constraints so that we can convert the DOTPEC to an infinite dimensional mathematical program, and it may be solved by using a descent method in Hilbert spaces or by any available solvers for nonlinear programs with time discretizations. Adopting the numerical scheme in Friesz and Mookherjee (2006) we were able to deal the theoretically rigorous flow propagation constraints within a non-simulation based model.As shown in the numerical example provided, the dynamic optimal toll tends to be proportional to the arc flow.

Future reserach should foucs on further analyzing the dynamic efficient toll and providing numerically rigorous approaches for the problem. In addition, we intend to include multiple target arrival times so that the model reflects the actual commuting networks and more.

# References

Astarita, V.: 1995, Flow propagation description in dynamic network loading models. YJ Stephanedes, F. Filippi, eds, *Proc. IV Internat. Conf. Appl. Adv. Tech. Transportation Engrg.(AATT)* pp. 599–603.

Astarita, V.: 1996, A Continuous Time Link Based Model for Dynamic Network Loading Based on Travel Time Function., *Proceedings of the 13th International Symposium on Transportation and Traffic Theory* .

Ban, J., Liu, H., Ferris, M. and Ran, B.: 2005, A link based quasi-vi formulation and solution algorithm for dynamic user equilibria, *INFORMS 2005, San Francisco, CA USA*.

Beckmann, M., McGuire, C. B. and Winsten, C. B.: 1956, *Studies in the Economics of Transportation*, Yale University Press.

Friesz, T., Bernstein, D. and Kydes, N.: 2002, Congestion pricing in disequilibrium, *Networks and Spatial Economics* **4**, 181–202.

Friesz, T., Bernstein, D., Suo, Z. and Tobin, R.: 2001, Dynamic network user equilibrium with state-dependent time lags, *Networks and Spatial Economics* **1**, 319–347.

Friesz, T. L., Bernstein, D. and Stough, R.: 1996, Dynamic systems, variational inequalities and control theoretic models for predicting urban network flows, *Transportation Science* **30**(1), 14–31.

Friesz, T. L. and Mookherjee, R.: 2006, Solving the dynamic network user equilibrium with state-dependent time shifts, *Transportation Research Part B* **40**, 207–229.

Friesz, T. and Shah, S.: 2001, An overview of nontraditional formulations of static and dynamic equilibrium network design, *Transportation Research Part B* **35**, 5–21.

Friesz, T., Tobin, R., Bernstein, D. and Suo, Z.: 1995, Proper flow propagation constraints which obviate exit functions in dynamic traffic assignment, *INFORMS Spring National Meeting, Los Angeles, April 23* **26**.

Hearn, D. W. and Yildirim, M. B.: 2002, *A Toll Pricing Framework for Traffic Assignment Problems with Elastic Demand*, Kluwer Academic Publishers, pp. 135–145.

Liu, L. N.: 2004, Multi-period congestion pricing models and efficient tolls in urban road, *Review of Network Economics* **3**, 381–391.

Tan, H.-N., Gershwin, S. and Athans, M.: 1979, Hybrid optimization in urban traffic networks, *Technical report*, LIDS Technical Report, MIT, Cambridge.

# TEMPORAL EXTERNALITY IN DYNAMIC USER EQUILIBRIUM WITH HETEROGENEOUS TRAVELLERS – WHO MAKES CONGESTION WORSE?

*Takamasa Iryo: Kobe University, Japan iryo@kobe-u.ac.jp*

## Abstract

Temporal externality in dynamic user equilibrium is analyzed in consideration of departure time choice behaviour. It is known that the increase of demand at earlier time of the congestion make the congestion at later time worse. Knowing such temporal externality is important to find out travellers to be the targets of congestion alleviating scheme. It is also known that considering departure time choice is necessary when travellers have their schedule constraint at their destinations. Such situation can occur in morning commute, for example. This study aims to find out travellers making congestion worse in consideration of departure time choice with schedule constraint at destinations. An optimization problem which is equivalent to equilibrium is stated first, and analysis of external cost is made by utilizing this optimization problem. It is revealed that travellers who need greater schedule cost to move out of the congestion have stronger external effect.

## 1 Introduction

It is known that change of demand in certain time period can affect not only the travellers travelling in the same time period but the travellers travelling at the different time in the bottleneck model. Increase of a demand in an earlier period of congestion increases travel time at later time. This can be interpreted that "new travellers joining the queue at the beginning of congestion makes congestion worse than travellers joining the queue in a later time period" (Kuwahara, 2001). Thus, travellers travelling in earlier time can be considered as "travellers who make congestion worse". Policies affecting their behaviour are more effective than policies on other people.

Problems finding "high social cost travellers" can be more complicated when travellers can choose their departure travel times. Considering departure time choice is necessary when travellers have their schedule constraint at their destinations because the capacity constraint at the bottleneck does not allow all travellers to arrive at destinations at their desired arrival time in peak hours. The model with the bottleneck model and departure time choice was introduced by Vickrey (1969) and many expanded studies are achieved. Among these previous studies, for example, Arnott et al.(1989)(1990) has calculated total travel cost in consideration of specific type of schedule cost function. Recently, an analysis with general schedule cost function is made by Lindsey (2004), which also discusses the change of travel cost of each traveller when demand increases.

This study aims to find out the "travellers who make congestion worse" in the model where travellers choose their departure time in consideration of the schedule constraint at the destinations. The model adopted by this study is similar to the model used by Lindsey, though a discrete time scheme is used instead of the continuous time scheme. Unlike many previous studies other than the Lindsey's one, the model of this study can handle completely heterogeneous travellers without imposing any special restriction on the schedule cost function. An optimisation problem which is equivalent to equilibrium is stated first and is utilized to calculate the change of travel cost caused by the increase of travellers having certain characteristic increases. Then, it is discussed which type of characteristics of travellers can make congestion worse than other types.

## 2 Definition of the system

This study considers a network with only one link and bottleneck. All travellers must pass through this bottleneck regardless of their origins or destinations. No traveller has route choice option. They can only choose their departure time from their origin. Each traveller has his/her own desired arrival time at his/her destination. Travel time is constant anywhere in the network other than the bottleneck. Only the bottleneck can have delay. The bottleneck capacity is given as a constant value $\mu$. A first-in first-out (FIFO) service is assumed at this bottleneck. An example of a study network is shown in figure 1.

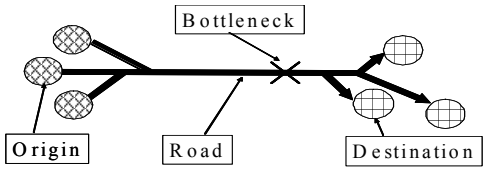



Figure 1 : Study network

Assuming that all travellers know actual travel time before they depart, choosing departure time from the origin is equivalent to choosing departure time from the bottleneck. This assumption can be used if equilibrium is established, where all travellers have perfect information on travel time. This study assumes that travellers choose departure time from the bottleneck instead of departure time from origins.

Travellers are heterogeneous. They can have different origins, destinations, and preferences in their arrival time at destinations. A concept of "class" is introduced to express the heterogeneousness. All travellers are classified into one of the classes. The total number of the class is $m$. The travellers belonging to the same class is homogeneous.

Each traveller choose his/her departure time from the bottleneck so as to minimize his/her travel cost. Travel cost is determined as

$$\pi(\tau, j) = w(\tau) + p(\tau, j), \tag{2.1}$$

where $w(\tau) \geq 0$ is the bottleneck delay of travellers departing from the bottleneck at time $\tau$ and $p(\tau, j)$ is the schedule cost of the travellers in the $j$ th class departing from the bottleneck at time $\tau$. Note that the schedule cost $p(\tau, j)$ and travel cost $\pi(\tau, j)$ is measured by the unit of delay time. Because travel time from the bottleneck to the destination is constant, this schedule cost describes traveller's preference in arrival time at his/her destination.

This study adopts the discrete time scheme. Time $\tau$ is replaced by $\tau = (t-1)\Delta t + T_0$, where $T_0$ is beginning time of study hours, $\Delta t$ is a duration of each time section, and $t$ is an integer. The maximum number of $t$ is defined as $n$. Time $\tau$ in the equation (2.1) is replaced by the discrete time $t$. The capacity of the bottleneck per $\Delta t$ is defined as $A = \mu \Delta t$.

Traveller's behaviour is expressed by the "choice variable" $X(t, j) \geq 0$, which describes the number of travellers who belong to the $j$ th class and choose time $t$ as their bottleneck departure travel time.

$X(t, j)$ has two constraints. One is the "demand constraint", which is described as

$$\sum_{t'=1}^{n} X(t', j) = D(j) \quad \text{for } \forall j, \tag{2.2}$$

where $D(j)$ is a number of travellers belonging to the $j$ th class. The other is the "capacity constraint", which is described as

$$\sum_{j'=1}^{m} X(t, j') \leq A \quad \text{for } \forall t. \tag{2.3}$$

Equilibrium is defined as "no traveller can have the option which decreases his/her travel cost", which is expressed by

$$w(t)+p(t,j)\geq w(t_e)+p(t_e,j) \quad \text{for } \forall t,j,t_e \text{ if } X(t_e,j)>0, \tag{2.4}$$

where $t_e$ is an integer satisfying $1\leq t_e\leq n$.

Also it is assumed that the capacity of the bottleneck is fully utilized by travellers if the delay at the bottleneck is greater than zero. It is natural due to the property of the bottleneck model. This assumption is described as

$$\sum_{j'=1}^{m} X(t,j')=A \quad \text{for } \forall t \quad \text{if } w(t)>0. \tag{2.5}$$

This equation (2.5) must be also established in equilibrium. Note that all other constraints on the system, such as non-negative constraint on $X(t,j)$ and $w(t)$ and the demand and capacity constraint shown by (2.2) and (2.3), must be also satisfied in equilibrium.

Due to the FIFO service at the bottleneck, $w(t)$ must satisfy another condition such as

$$w(t+1)-w(t)\leq \Delta t. \tag{2.6}$$

This means that there is a maximum limit on the increasing speed of the delay at the bottleneck service. A proof is shown in appendix 1.

# 3 An optimisation problem equivalent to the equilibrium

This section shows an optimisation problem which is equivalent to the equilibrium. Consider an optimisation problem determined as

$$\begin{aligned} &\text{minimise } S=\sum_{t'=1}^{n}\sum_{j'=1}^{m} X(t',j')\,p(t',j') \\ &\text{subject to } X(t,j)\geq 0,\ \sum_{t'=1}^{n} X(t',j)=D(j) \text{ and } \sum_{j'=1}^{m} X(t,j')\leq A \text{ for } \forall t,j \end{aligned} \tag{3.1}$$

This optimisation problem solves travellers' choices of departure time which minimises total schedule cost imposed on all travellers. This is a linear programming problem and has a corresponding dual problem such as

$$\begin{aligned} &\text{maximise } S'=\sum_{j'=1}^{m} D(j')\theta(j')-A\sum_{t'=1}^{n} w_e(t') \\ &\text{subject to } w_e(t)\geq 0 \text{ and } \theta(j)\leq w_e(t)+p(i,j) \text{ for } \forall t,j, \end{aligned} \tag{3.2}$$

where $w_e(t),\theta(j)$ are new variables. A complementarity condition is:

$$\left(w_e(t)+p(t,j)-\theta(j)\right)X(t,j)=0 \quad \text{for } \forall t,j \tag{3.3a}$$

$$w_e(t)\left(A-\sum_{j'=1}^{m} X(t,j')\right)=0 \quad \text{for } \forall t \tag{3.3b}$$

$$X(t,j)\geq 0,\ \sum_{t'=1}^{n} X(t',j)=D(j) \text{ and } \sum_{j'=1}^{m} X(t,j')\leq A \quad \text{for } \forall t,j \tag{3.3c}$$

$$w_e(t)\geq 0 \text{ and } \theta(j)\leq w_e(t)+p(i,j) \quad \text{for } \forall t,j \tag{3.3d}$$

Note that (3.3c) and (3.3d) are the same as the constraints contained by the primal problem (3.1) and dual problem (3.2) respectively. A process to derive out (3.2) and (3.3) is described in the appendix 2. Due to the complementarity slackness theorem, the condition (3.3) is equivalent to the primal problem (3.1) and the dual problem (3.2).

The variable $w_e(t)$ in the complementarity condition (3.3) satisfies the definition of equilibrium if they are recognised as the bottleneck delay of travellers departing from the bottleneck at time $t$ in equilibrium. The equation (3.3a) can be replaced by

$$\left(w_e(t)+p(t,j)\right)X(t,j)=\theta(j)X(t,j) \qquad \text{for } \forall t,j \qquad (3.4)$$

and therefore

$$\theta(j)=w_e(t)+p(t,j) \quad \text{if } X(t,j)>0. \qquad \text{for } \forall t,j \qquad (3.5)$$

is obtained. Combining (3.5) and the second inequality of (3.3d), It can be confirmed that $w_e(t)$ satisfies (2.4). $w_e(t)$ also satisfies (2.5) due to the equation (3.3b). Conversely, the equations (3.3) can be derived from the definition of the equilibrium described by (2.4) and (2.5) and the demand constraint and capacity constraint shown by (2.2) and (2.3). Now assume that $w_e(t)$ indicates the bottleneck delay in equilibrium and define $\theta(j)$ as

$$\theta(j)=\min_t\left\{w_e(t)+p(t,j)\right\}. \qquad (3.6)$$

This definition means that $\theta(j)$ represents the travel cost for travellers belonging to the $j$ th class. $\theta(j)$ defined by (3.6) satisfies the second inequality of (3.3d). Considering the definition of the equilibrium shown by (2.4) and the second inequality of (3.3d), (3.5) can be derived. A combination of (3.5) and the second inequality of (3.3d) can derive (3.4). Therefore (3.3a) and (3.3d) is established. (3.3b) is directly derived from (2.5). (2.2) and (2.3) is identical to (3.3c) along with non-negative constraint on $X(t,j)$. Thus, it can be concluded that the complementarity condition (3.3) is equivalent to the definition of equilibrium. It means that the definition of equilibrium is equivalent to the optimisation problems defined by (3.1) and (3.2). $X(t,j)$ in equilibrium can be calculated by the primal problem (3.1) and $w_e(t)$ can be calculated by the dual problem (3.2). Also, travel cost of travellers belonging to $j$ th class can be calculated as $\theta(j)$ by the dual problem (3.2).

Note that the calculation shown above neglects the condition of FIFO shown by (2.6). This condition must be examined after calculating $w_e(t)$ by the dual problem (3.2). If $w_e(t)$ satisfies (2.6), it can be concluded that the delay $w_e(t)$ can be made by the bottleneck where FIFO service is provided.

The equivalence shown in this chapter has three important meaning. One is that "equilibrium of the departure time choice problem can be solved by linear programming". This is useful to find equilibrium where all of the constants (capacity, demand and schedule costs) are given. Another one is that "disregarding FIFO constraint, at least one equilibrium state exists". This can be said because the feasible set of the primal problem (3.1) is finite. Last one is that "the sum of schedule cost measured by the unit of delay time is minimised in equilibrium".

The last characteristic is important to consider congestion alleviation policies. Assume that an administrator can control each traveller's departure time at his/her origin. Under this assumption, the bottleneck delay can be erased with delaying his/her departure time for the delay which they have experienced in equilibrium.

This means that all of the travellers wait at their origins instead of joining the queue at the bottleneck and the sum of delay at the bottleneck is minimised by this policy. Note that departure time from the bottleneck does not change. Because the sum of the schedule costs has been minimised in equilibrium already, this policy achieves the minimum total schedule cost of all travellers which is available under the given capacity constraint and the amount of the demand. Thus, it can be said that the policy shown above achieves minimum total travel cost of all travellers if the cost is measured by the unit of delay time.

# 4 Increase of travel cost made by additive demand

This section shows how to calculate the increase of travel cost made by additive demand. It is shown in the section 3 that changing all travellers' departure times can minimise the total travel cost of all travellers. Though this policy is the best solution to mitigate congestion, however, controlling all travellers' departure times is not realistic. Controlling behaviour of a part of travellers is more realistic. For example, policies letting travellers belonging to certain class escape from the congestion (i.e. choosing off-peak time or alternative transport) may be less difficult as the targets of the policy may not be so many. Knowing who makes congestion worse is important to determine the target of the polity.

Calculating the increase of the total travel cost caused by additive travellers is necessary to find out travellers making congestion worse. This problem can be considered as a sensitive analysis of the optimisation problem (3.2). Now let the demand of the $j_D$ th class increase by $\Delta D > 0$. All variables before adding travellers are described with an superscript of 0, such as $X^0(i,j)$. All variables after adding travellers are described with an asterisk, for example $X^*(i,j)$, and a change of a variable by additive demand is described with a delta followed by a variable, for example $w_e^0(t) + \Delta w_e(t) = w_e^*(t)$. According to the knowledge of the sensitive analysis in linear programming, there is an upper bound of $\Delta D$ for conserving the value of variables in the dual problem (3.2), that is,

$$w_e^0(t) = w_e^*(t) \quad \text{for all } t, \tag{4.1}$$

$$\theta^0(j) = \theta^*(j) \quad \text{for all } j. \tag{4.2}$$

The upper bound of $\Delta D$ is defined as $\Delta D_m \geq 0$. Because of (4.2), all travellers who have stayed in the congestion before adding demand does not experience the change of their travel cost. This means that the change of total travel cost can be explained with the travel cost which the additive travellers have. No externality is observed here.

The value of $\Delta D_m$ is determined with considering the change of the optimal function of the primal problem (3.1). Applying (4.2), this can be described as

$$\begin{aligned} \Delta S &= \sum_{t'=1}^{n}\sum_{j'=1}^{m} \Delta X(t',j')\, p(t',j') = \sum_{t'=1}^{n}\sum_{j'=1}^{m} \Delta X(t',j')\left(\theta^0(j') - w_e^0(t') + \gamma^0(t',j')\right) \\ &= \Delta D\theta^0(j_D) - \sum_{t'=1}^{n} w_e^0(t') \sum_{j'=1}^{m} \Delta X(t',j') + \sum_{t'=1}^{n}\sum_{j'=1}^{m} \Delta X(t',j')\gamma^0(t',j') \end{aligned} \tag{4.3}$$

where $\gamma^0(t,j)$ is a slack for the second constraint in the dual problem (3.2), which is defined as

$$\gamma^0(t,j) = w_e^0(t) + p(t,j) - \theta^0(j). \tag{4.4}$$

Due to duality of the linear programming, $\Delta S$ in the primal problem shown in (4.3) must be the same as the $\Delta S'$ in the dual problem (3.2). Also, considering (3.3b) and $w_e^0(t) = w_e^*(t)$, the second term of the right-hand side of (4.3) is zero (see appendix 3). Thus

$$\sum_{t'=1}^{n}\sum_{j'=1}^{m}\Delta X\left(t',j'\right)\gamma^{0}\left(t',j'\right)=0. \tag{4.5}$$

is obtained. Note that $\Delta S'=\Delta D\theta^{0}\left(j\right)$ due to (4.1) and (4.2). Therefore

$$\Delta X\left(t,j\right)=0 \quad \text{if } \gamma^{0}\left(t,j\right)>0 \quad \text{for } \forall t,j. \tag{4.6}$$

must be satisfied in order to satisfy $\Delta S=\Delta S'$. Note that $\Delta X\left(t,j\right)$ cannot be negative if $\gamma^{0}\left(t,j\right)>0$ because $X^{0}\left(t,j\right)=0$ in this case due to (3.3a). Conversely, if (4.6) is satisfied, $\Delta S=\Delta S'$ is established because the second and third term of the right-hand of the equation (4.3) become zero. Other constraints are derived from the demand and capacity constraint. Due to the demand constraint, $\Delta X\left(t,j\right)$ also have to satisfy

$$\sum_{t'=1}^{n}\Delta X\left(t',j\right)=\begin{cases}\Delta D & \text{if } j=j_D\\ 0 & \text{otherwise}\end{cases} \tag{4.7}$$

and the capacity constraint requests that

$$\sum_{j'=1}^{m}\Delta X\left(t,j'\right)\le A-\sum_{j'=1}^{m}X^{0}\left(t,j'\right) \tag{4.8}$$

is satisfied. The non-negative constraint on $X\left(t,j\right)$ also derives a condition of

$$\Delta X\left(t,j\right)\ge -X^{0}\left(t,j\right). \tag{4.9}$$

To show the set of $\left(t,j\right)$ where $\Delta X\left(t,j\right)$ can have a value other than zero, a "selection graph" is introduced. The horizontal axis of the selection graph indicates time $t$ and the vertical axis indicates traveller's class $j$. Nodes are plotted at $\left(t,j\right)$ where $\gamma^{0}\left(t,j\right)=0$. Note that only $\left(t,j\right)$ where nodes are plotted can be chosen by travellers. An example of the selection graph is shown in the figure 2.

The selection graph shows conservative laws of $\Delta X\left(t,j\right)$ shown by (4.7), (4.8), and (4.9). The condition (4.7) shows the conservative law of $\Delta X\left(t,j\right)$ along the horizontal direction of the selection graph. Connecting all nodes having the same $j$ with "horizontal links", it can be said that the sum of the $\Delta X\left(t,j\right)$ belonging to the same link must be equal to $\Delta D_m$ where $j=j_D$, or must be zero where $j\ne j_D$. If a node $\left(t,j\right)$ is not connected to any horizontal link, $\Delta X\left(t,j\right)$ must be $\Delta D_m$ where $j=j_D$, or must be zero where $j\ne j_D$. Also, the condition (4.8) shows the conservative law of $\Delta X\left(t,j\right)$ along the vertical direction of the selection graph. Connecting all nodes having the same $t$ with "vertical links", it can be said that the sum of the $\Delta X\left(t,j\right)$ belonging to the same vertical link must not be greater than $A-\sum_{j'=1}^{m}X^{0}\left(t,j'\right)$. This means that only the nodes where the capacity is not fully used can absorb the additive demand. Such nodes are named as "anchors". The anchors must be placed the time $t$ where $w_e^{0}\left(t\right)=0$. If a node $\left(t,j\right)$ is not connected to any vertical link and is not an anchor, $\Delta X\left(t,j\right)$ must be zero. These conservative laws conclude that $\Delta X\left(t,j\right)$ must be zero unless the node $\left(t,j\right)$ is connected by both horizontal and vertical

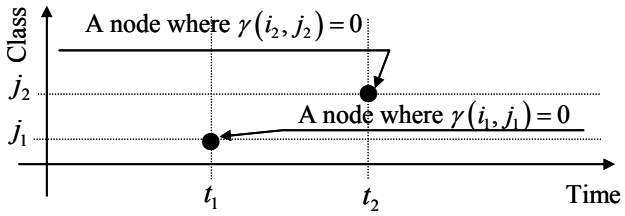



Figure 2 : An example of nodes on a selection graph

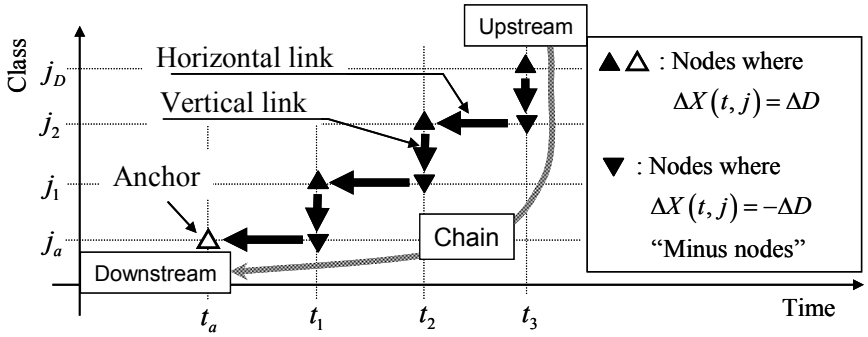



Figure 3 : An example of a chain

link, connected by vertical link and $j = j_D$, or connected by horizontal link and is an anchor.

The increase of demand must be transmitted to the anchor(s) via the network made by the horizontal and vertical links due to the conservative laws expressed by vertical and horizontal links. An example is shown in figure 3. To treat the additive demand $\Delta D$, $\Delta X(t, j_D)$ at some $t$ must be greater than zero. In figure 3, only one node whose $\Delta X(t, j_D)$ can be positive is a node $(t_3, j_D)$ where a vertical link is connected. So, let $\Delta X(t_3, j_D) = \Delta D$. This $\Delta X(t_3, j_D)$ must be cancelled by $\Delta X(t_3, j_2)$, which is connected to $\Delta X(t_3, j_D)$ by a vertical link. Thus, $\Delta X(t_3, j_2) = -\Delta D$. The decrease of $\Delta X(t_3, j_2)$ must be cancelled by $\Delta X(t_2, j_2)$, which is connected to $\Delta X(t_3, j_2)$ by a horizontal link. Thus, $\Delta X(t_2, j_2) = \Delta D$. Such procedure must be repeated until it reaches an anchor where $\Delta X(t_a, j_a) = \Delta D$.

A route consisting of nodes and links which is traced by this process is named as "chain". Chains have directions. The direction toward an anchor is defined as the downstream and the direction toward nodes at $j = j_D$ is defined as the upstream. Adding the demand $\Delta D$, $\Delta X(t, j)$ at nodes on the chain will be as follow:

$$\Delta X(t,j) = \begin{cases} \Delta D & \text{if } (t,j) \text{ is } \begin{cases} \text{- a node on upstream end of a vertical link} \\ \quad \text{and on downstream end of a horizontal link, or} \\ \text{- an anchor, or} \\ \text{- } j = j_D \text{ and a node on upstream end of a vertical link} \end{cases} \\ -\Delta D & \text{if } (t,j) \text{ is a node on downstream end of a vertical link} \\ & \quad \text{and on upstream end of a horizontal link} \\ 0 & \text{Otherwise} \end{cases} \tag{4.10}$$

The nodes where $\Delta X(t, j)$ decreases are named as "minus node". Two or more chains connecting nodes at $j = j_D$ to the anchor(s) can exist and each route can treat the additive demand separately.

The constraint of (4.9) restricts the amount of additive demand which can be treated by each route. (4.9) gives a lower limit on $\Delta X(t,j)$ at each minus node and therefore limits the amount of transmitted demand $\Delta D$ via a chain such as

$$\Delta D \le \min\left\{ X^0(t,j) \middle| (t,j) \in \text{ set of minus nodes on the chain} \right\}. \tag{4.11}$$

In the example in fig 3, $\Delta D \le \min\left\{ X^0(t_1,j_a), X^0(t_2,j_1), X^0(t_3,j_2) \right\}$, for instance. If two or more chains exist, they can be used to transmit excess demand to anchor(s). If no other chain is found, no more demand can be added. Also the amount of the demand which can be absorbed by an anchor is limited. This can be derived from the constraint (4.8), that is, $\Delta D \le A - \sum_{j'=1}^{m} X^0(t_a,j')$ must be satisfied at the time $t_a$ where the anchor exists. If $\Delta D$ exceeds this limit, another anchor at a different time which can be connected by another chain must be used to absorb excess demand. If no other anchor is found, no more demand can be added. These restrictions determine $\Delta D_m$. For example, $\Delta D_m$ of the example in fig 3 is calculated as

$$\Delta D_m = \min\left\{ X^0(t_1,j_a), X^0(t_2,j_1), X^0(t_3,j_2), A - \sum_{j'=1}^{m} X^0(t_a,j') \right\} \tag{4.12}$$

The nodes where $\Delta X(t,j) = -X^0(t,j)$ is named as "disconnected node". The anchor where $\sum_{j'=1}^{m} X^*(t_a,j') = A$ is named as "disconnected anchor".

If $\Delta D > \Delta D_m$, the disconnected node(s) must be abandoned and a new node must be placed on the selection graph in order to create a new chain which transmits an excess demand to the existing anchor, or a new anchor must be added to absorb an excess demand. Though there are many possibilities to set a new node and form a new chain, this must meet the optimisation problem defined by (3.1). This implies that the new node will be placed at $(t,j)$ where $\gamma^0(t,j)$ can be as small as possible. It can be derived from the equation (4.3) and a proof is shown in appendix 4. An example of this procedure is shown in the figure 4. A place where the new node is set is referred as $(t_p, j_p)$.

The chain can also be used to calculate the travel cost of each traveller $\theta^0(j)$. Combining (4.4) at two nodes belonging the same vertical link,

$$\theta^0(j_2) - \theta^0(j_1) = p(t,j_2) - p(t,j_1) \tag{4.13}$$

can be obtained, where $(t,j_1)$ and $(t,j_2)$ are nodes on the some vertical link. The equation (4.13) describes the difference of the travel cost between $j_1$ and $j_2$. Accumulating this difference along a chain, $\theta^0(j)$ can be calculated. An example is shown in the figure 5. Note that this accumulation can continue beyond $j_D$ if the vertical and / or horizontal links can be aligned continuously via nodes.

The recombination of a chain caused by the excess demand can increase the travel cost of travellers who does not belong to the $j_D$ th class, meaning that the externality is made. Due to (4.13),

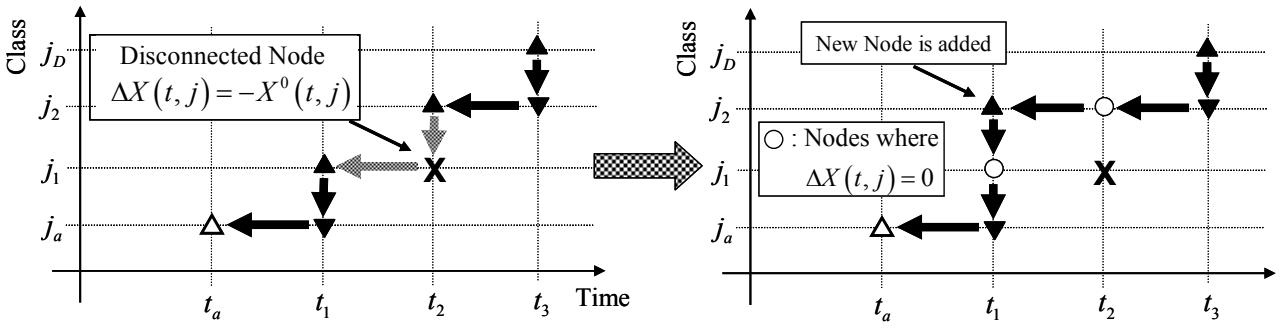



Figure 4 : A disconnected node and the reformation of the chain

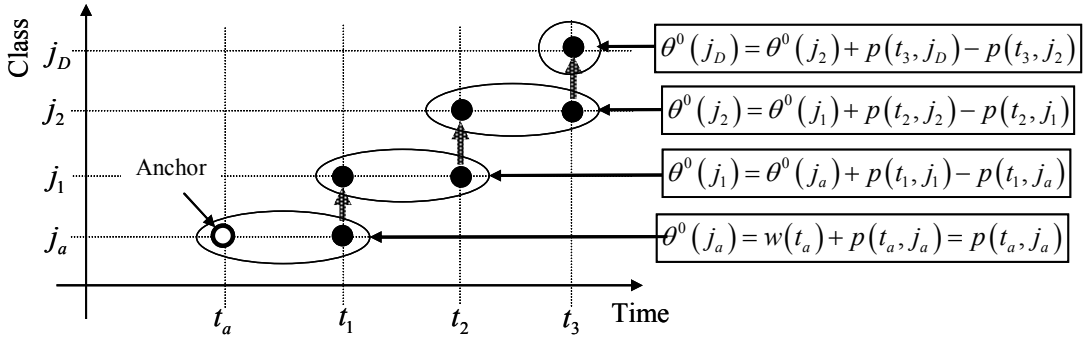



Figure 5 : Calculation of travel cost

$$\Delta\theta(j) = \gamma^0(t_p, j_p) \quad \text{if } j \text{ belongs to a node } (t_p, j_p) \text{ or a node on the upstream side than } (t_p, j_p) \tag{4.14}$$

See appendix 5 for a proof of (4.14). Because $\gamma^0(t_p, j_p) > 0$, total travel cost increases. An example is shown in figure 6. It indicates that the external cost made by additive demand is transmitted along the horizontal and vertical links toward the upstream direction. Also it implies that travellers who belong to nodes on upstream have more chance to incur external cost than travellers on downstream nodes.

## 5 Characteristic of travellers who makes congestion worse

The results shown in the previous section indicates that the reformation of the chain increases travel costs. Therefore, the travellers who induce the reformation of the chain more often make congestion worse than others.

Increasing the number of travellers belonging to the nodes on upstream side, congestion becomes worse with smaller additive demand. The equation (4.12) indicates that $\Delta D_m$ become smaller if the chain has more "minus nodes", which are candidates for the disconnected node. Smaller $\Delta D_m$ means that smaller amount of the additive demand can change the chain, meaning that congestion becomes worse with fewer additive travellers. Such case occurs when the node on the upstream side is chosen as the class where travellers are added. An example is shown in figure 7. In this figure, adding travellers belonging to $j_2$ th class can have three candidates for disconnected nodes, whereas $j_1$ th class can only have two candidates.

Travellers who incur greater schedule cost to move out of the congestion, which is named as "high schedule cost travellers" belong to upstream nodes. Considering the optimisation problem (3.1), high schedule cost travellers must choose the time in the middle of the congestion, whereas other travellers choose times nearer to the off-peak time. This means the nodes of other travellers are included by the chain from the nodes of the high schedule cost travellers to an anchor, meaning that high schedule cost travellers are placed on the upstream side. This suggests that high schedule cost travellers makes congestion worse.

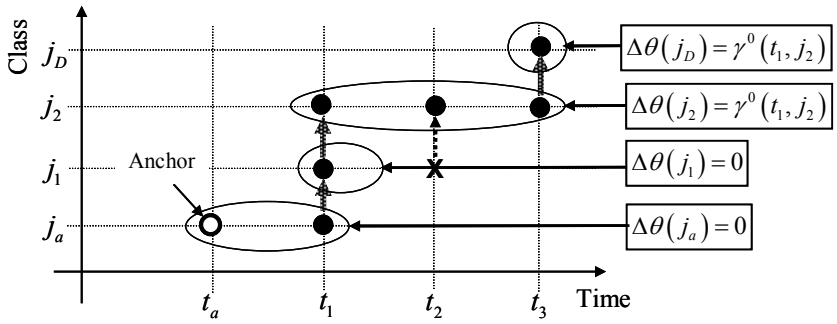



Figure 6 : The change of travel cost

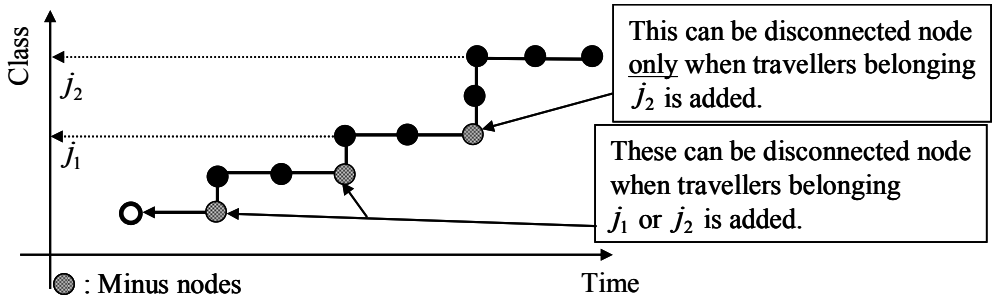



Figure 7 : Candidates of disconnected nodes

## 6 Discussions

This study shows a method to calculate travel cost and its change due to additive demand in equilibrium on the assumption of the bottleneck model and schedule constraint at destinations. An optimisation problem is stated to calculate equilibrium. Then, a concept of chain is introduced to calculate the change of travel cost when additive demand is induced. The analysis in this study indicates that travellers incurring higher schedule cost make congestion worse.

The concept of the chain proposed in this study can be used to examine the change of travel cost and travellers behaviour when the number of traveller changes. It can also shows the externality of additive demand, that is, whose travel cost increases when new travellers join into the congestion. An analysis carried out in this study is limited and further analysis in many other cases should be done with the concept of the chain to obtain more information of the externality of the bottleneck congestion with departure time choice.

## Acknowledgement

This research is partly supported by the Grant-in-Aid for Scientific Research of Japan Society for the Promotion of Science #16760427.

## References

Arnott, R., de Palma, A. and Lindsey, R. (1989) Schedule Delay and Departure Time Decisions with Heterogeneous Commuters. *Transportation Research Record*, 1197, pp.56-67.

Arnott, R., de Palma, A. and Lindsey, R. (1990) Departure Time and Route Choice for the Morning Commute. *Transportation Research*, Vol.24B, No.3, pp.209-228.

Kuwahara, M. (2001) A Theoretical Analysis on Dynamic Marginal Cost Pricing. *Proceedings of the Sixth Conference of Hong Kong Society for Transportation Studies*, pp.28-39.

Lindsey, R. (2004) Existence, Uniqueness, and Trip Cost Function Properties of User Equilibrium in the Bottleneck Model with Multiple User Classes. *Transportation Science*, Vol. 38, No.3, pp.293–314.

Vickrey, W. S. (1969) Congestion Theory and Transportation Investment. *American Economic Review*, Vol.59, pp.251-260.

# 7 Appendix

## Appendix 1

The equation (2.6), which is the condition on FIFO, is proved here. FIFO service is defined as "a vehicle arriving at the bottleneck earlier departs from the bottleneck earlier than the vehicle arriving later". This proposition can be formulated as

$$t_2 - w(t_2) < t_1 - w(t_1) \quad \Rightarrow \quad t_2 < t_1, \tag{7.1}$$

if the times $t_1$ and $t_1$ is continuous. The proposition (7.1) is equivalent to

$$t_2 - t_1 \geq 0 \quad \Rightarrow \quad t_2 - t_1 \geq w(t_2) - w(t_1). \tag{7.2}$$

In the discrete time scheme, (7.2) is reformulated as

$$t_2 - t_1 \geq 0 \quad \Rightarrow \quad \Delta t(t_2 - t_1) \geq w(t_2) - w(t_1). \tag{7.3}$$

Substituting an equation $t_2 = t_1 + 1$ into (7.3),

$$w(t_1 + 1) - w(t_1) \leq \Delta t \tag{7.4}$$

is obtained. It is a necessary condition of FIFO. This condition (7.4) can be also a sufficient condition of (7.3). Assuming the left-hand side of inequality of (7.3),

$$\sum_{t=t_1}^{t_2-1} \{w(t+1) - w(t)\} = w(t_2) - w(t_1) \tag{7.5}$$

is obtained. Taking a summation of inequality (7.4) from $t_1$ to $t_2 - 1$ and substituting the equation (7.5) into this summation,

$$w(t_2) - w(t_1) \leq \Delta t(t_2 - t_1) \tag{7.6}$$

is obtained. This is the same as the right-hand side of the proposition (7.3). Thus, it is proved that the condition (7.4) is equivalent to the FIFO condition. (QED)

## Appendix 2

Here the dual problem (3.2) and the complementarity slackness condition (3.3) are derived from the primal problem (3.1). The optimisation problem (3.1) can be rewritten as

$$\begin{aligned} &\text{minimise } S = \sum_{t'=1}^{n} \sum_{j'=1}^{m} X(t', j')\, p(t', j') \\ &\text{subject to } X(t, j) \geq 0,\ X(t, 0) \geq 0,\ \sum_{t'=1}^{n} X(t', j) = D(j),\ X(t, 0) + \sum_{j'=1}^{m} X(t, j') = A \ \text{ for } \forall t, j \end{aligned} \tag{7.7}$$

where $X(t, 0)$ is a slack. Apart from non-negative conditions, (7.7) has two equations as constrains. They can be reformulated as

$$\sum_{t'=1}^{n}\sum_{j'=0}^{m}\delta(t',t)X(t',j')=A \quad \text{for } 1\le\forall t\le n \tag{7.8}$$

and

$$\sum_{t'=1}^{n}\sum_{j'=0}^{m}\delta(j',j)X(t',j')=D(j) \quad \text{for } 1\le\forall j\le m\,, \tag{7.9}$$

where the Kronecker's delta is defined as

$$\delta(i,j)=\begin{cases}1 & \text{if } i=j\\ 0 & \text{otherwise}\end{cases}. \tag{7.10}$$

Combining the equations (7.8) and (7.9),

$$\sum_{t'=1}^{n}\sum_{j'=0}^{m}\varepsilon(t',j',l)X(t',j')=B(l) \quad \text{for all } 1\le l\le n+m\,, \tag{7.11}$$

where $B(l)$ is defined as

$$B(l)=\begin{cases}A & \text{if } l\le n\\ D(l-n) & \text{if } l>n\end{cases} \tag{7.12}$$

and $\varepsilon(t',j',l)$ is defined as

$$\varepsilon(t',j',l)=\begin{cases}\delta(t',l) & \text{if } l\le n\\ \delta(j',l-n) & \text{if } l>n\end{cases}. \tag{7.13}$$

Adopting (7.11) as a constraint of (7.7) and letting $p(t,0)=0$ for all $t$,

$$\begin{aligned}&\text{minimise } S=\sum_{t'=1}^{n}\sum_{j'=0}^{m}X(t',j')\,p(t',j')\\ &\text{subject to } \sum_{t'=1}^{n}\sum_{j'=0}^{m}\varepsilon(t',j',l)X(t',j')=B(l) \quad \text{for } 1\le\forall l\le n+m\\ &\qquad\qquad X(t,j)\ge 0 \quad \text{for } 1\le\forall t\le n, 0\le\forall j\le m\end{aligned} \tag{7.14}$$

is obtained. (7.14) is a slack form of linear programming and its dual problem can be derived as

$$\begin{aligned}&\text{maximise } S'=\sum_{l=1}^{n+m}v(l)B(l)\\ &\text{subject to } \sum_{l=1}^{n+m}\varepsilon(t,j,l)v(l)\le p(t,j) \text{ for } 1\le\forall t\le n, 0\le\forall j\le m\end{aligned} \tag{7.15}$$

A complementarity condition can be written as

$$\left( p(t,j) - \sum_{l=1}^{n+m} \varepsilon(t,j,l) v(l) \right) X(t,j) = 0 \quad \text{for } 1 \le \forall t \le n, 0 \le \forall j \le m \tag{7.16}$$

$$\sum_{t'=1}^{n} \sum_{j'=0}^{m} \varepsilon(t',j',l) X(t',j') = B(l) \quad \text{for } 1 \le \forall l \le n+m \tag{7.17}$$

$$X(t,j) \ge 0 \quad \text{for for } 1 \le \forall t \le n, 0 \le \forall j \le m \tag{7.18}$$

$$\sum_{l=1}^{n+m} \varepsilon(t,j,l) v(l) \le p(t,j) \text{ for } 1 \le \forall t \le n, 0 \le \forall j \le m \tag{7.19}$$

(7.17), (7.18), and (7.19) are the same as the constraints included in the primal problem (7.7) and the dual problem (7.15). (7.17) and (7.18) correspond to the constraints in the primal problem, which is identical to (3.3c). On the other hand, (7.19) corresponds to the dual problem. Let

$$v(t) = -w_e(t) \quad \text{for } 1 \le t \le n \tag{7.20}$$

and

$$v(j-n) = \theta(j) \quad \text{for } 1 \le j \le m\,. \tag{7.21}$$

When $j = 0$, (7.19) becomes

$$w_e(t) \ge 0 \quad \text{for } 1 \le \forall t \le n\,, \tag{7.22}$$

otherwise it is

$$\theta(j) - w_e(t) \le p(t,j) \quad \text{for } 1 \le \forall t \le n,\ 1 \le \forall j \le m\,, \tag{7.23}$$

which is identical to (3.3d). (7.16) can be modified into

$$\left( p(t,j) - \theta(j) + w_e(t) \right) X(t,j) = 0 \quad \text{for } 1 \le \forall t \le n, 1 \le \forall j \le m \tag{7.24}$$

and

$$w_e(t) X(t,0) = 0 \quad \text{for } 1 \le \forall t \le n\,. \tag{7.25}$$

They are identical to (3.3a) and (3.3b) respectively. Finally, the dual problem (7.15) can be rewritten as

$$\begin{aligned} &\text{maximise } S' = \sum_{j'=1}^{m} D(j') + A \sum_{t'=1}^{n} w_e(t') \\ &\text{subject to } w_e(t) \ge 0,\ \theta(j) - w_e(t) \le p(t,j) \quad \text{for } 1 \le \forall t \le n,\ 1 \le \forall j \le m \end{aligned} \tag{7.26}$$

which is the same as (3.2). (QED).

# Appendix 3

Here it is proved that the second term of the right-hand side of (4.3) is zero. Considering (3.3b) and $w_e^0(t) = w_e^*(t)$,

$$w_e^0(t) \left( A - \sum_{j'=1}^{m} X^0(t,j') \right) = w_e^0(t) \left( A - \sum_{j'=1}^{m} X^*(t,j') \right) \quad \text{for } \forall t\,. \tag{7.27}$$

Therefore,

$$\sum_{j'=1}^{m} \Delta X\left(t, j'\right) = 0 \quad \text{if } w_e^0\left(t\right) > 0 \tag{7.28}$$

can be derived (QED).

## Appendix 4

The change of optimal function shown by (4.3) must be reconsidered here. First, the second term of (4.3) must be zero unless $w_e^0\left(t\right) > 0$ and $\sum_{j'=1}^{m} \Delta X\left(t, j'\right) < 0$ are established at certain time $t$. This condition means that capacity is not fully utilized at the time where congestion exists before adding demand. This cannot be achieved in an optimal solution because the residue of the capacity must be used to minimise the sum of the schedule costs. Thus, only the third time must be considered to get the minimum $\Delta S$ in (4.3). Adding a new node corresponds to let $\Delta X\left(t, j\right) > 0$ where $\gamma^0\left(t, j\right) > 0$. This means that choosing a new node $\left(t, j\right)$ having smaller $\gamma^0\left(t, j\right)$ makes $\Delta S$ smaller. Therefore, a new node must be chosen by the following procedure:

1. Select all nodes which can form a new chain transmitting excess additive demand $\Delta D - \Delta D_m$ to the anchor(s).
2. Find the node having the smallest $\gamma^0\left(t, j\right)$ from the nodes selected at the step 1.

## Appendix 5

The equation (4.14) is proved here. Assume that the situation shown in the figure 6, where a new node is added on $\left(t_p, j_p\right) = \left(t_1, j_2\right)$. Due to (4.4),

$$\begin{aligned} \gamma^0\left(t_1, j_2\right) &= w_e^0\left(t_1\right) + p\left(t_1, j_2\right) - \theta^0\left(j_2\right) \\ 0 &= w_e^0\left(t_1\right) + p\left(t_1, j_1\right) - \theta^0\left(j_1\right) \end{aligned} \tag{7.29}$$

is established . Subtracting the second equation of (7.29) from the first equation,

$$\gamma^0\left(t_1, j_2\right) = p\left(t_1, j_2\right) - p\left(t_1, j_1\right) - \left\{\theta^0\left(j_2\right) - \theta^0\left(j_1\right)\right\} \tag{7.30}$$

is obtained. Applying (4.13) to the situation where the demand has been added,

$$\theta^*\left(j_2\right) - \theta^*\left(j_1\right) = \theta^0\left(j_2\right) + \Delta\theta\left(j_2\right) - \theta^0\left(j_1\right) = p\left(t_1, j_2\right) - p\left(t_1, j_1\right) \tag{7.31}$$

is obtained. Substituting (7.31) into (7.30), (4.14) is obtained. (QED)

# STABILITY DOMAINS OF TRAFFIC EQUILIBRIUM: DIRECTING TRAFFIC SYSTEM EVOLUTION TO EQUILIBRIUM

*Hong K LO: The Hong Kong University of Science and Technology, Hong Kong cehklo@ust.hk*
*Jing BIE: The Hong Kong University of Science and Technology, Hong Kong jbie@ust.hk*

## Abstract

This study investigates the global and local asymptotic stability of user equilibria in a transportation network. A dynamical system is said to possess global asymptotic stability when the maximal domain of attraction of its equilibrium point contains the entire feasible state space. That is, any feasible traffic state of the system will evolve toward user equilibrium. For dynamical systems without global asymptotic stability, the system evolution starting with traffic states outside the domain of attraction will not arrive at user equilibrium. In this case, the feasible state space can be divided into various local domains of attraction, each corresponding to a different attractor of the dynamical system. By developing a good understanding of the relationships between the various domains of attraction, in this study, we demonstrate the possibility of directing the system evolution toward user equilibrium through temporary network alterations. The instruments for temporary network alteration may include traffic signal control, lane closure or addition, or pricing, etc. We believe that this approach of directing traffic system evolution onto a certain desirable course will open up innovative ways for traffic network management.

## 1. Introduction

Ever since the notion of user equilibrium (UE) was proposed (Wardrop, 1952), it has become a cornerstone for traffic assignment analysis. Past research primarily focused on the issues of existence and uniqueness of equilibrium. The assumption is that if equilibrium exists, then it will also occur. This, of course, is an idealization. In fact, it has been shown that quite the contrary can happen. Horowitz (1984) demonstrated that even for a well-behaved system whose UE solution is known to exist, depending on the dynamic route adjustment process, the system may still fail to converge to UE. Therefore, it is not sufficient to ask whether equilibrium exists or not; it is equally important to ask whether and how the system can achieve the equilibrium state. In other words, the dynamic route adjustment processes of travelers in search of better routings constitute a key part of equilibrium analysis.

To address this problem further, it is imperative to conduct a more elaborate study on the effects and properties of dynamic route adjustment processes. Studies on asymptotic stability (e.g., Smith (1979), Smith (1984) and Watling (1999)) address the attractiveness of a certain equilibrium state. Asymptotic stability is important in the sense that some equilibrium states may turn out to be unstable and thus could not be attained unless one is fortunate enough to start with an initial state that is identical to the equilibrium state. In general, asymptotic stability associated with an equilibrium state can be classified as either local or global. In local asymptotic stability, one can define the domain of attraction associated with the equilibrium state; any initial state within this domain of attraction will eventually converge to the equilibrium by the specified dynamic route adjustment process. On the other hand, in global asymptotic stability, the domain of attraction contains the entire state space; therefore, any initial state will converge to the equilibrium. In other words, when the asymptotic stability associated with an equilibrium is not global, whether the system will converge to the equilibrium or not depends on the initial state; those states that are outside the domain of attraction will not evolve toward the equilibrium. Instead, they may evolve to other attractors, such as periodic cycles or (aperiodic) chaos.

By developing a good understanding of the domains of attraction associated with the equilibrium states, one can find ways to alter the system evolution. In this study, we show with a small example how traffic management techniques can be applied to direct the day-to-day dynamical system to equilibrium when the

system cannot achieve this on its own. In other words, given an initial state that is outside the domain of attraction of the equilibrium, by temporarily altering the network configuration via link cost adjustments (accomplished through pricing or signal control), we can direct the system to a state within the domain of attraction, thereby allowing the system to arrive at the equilibrium on its own. We believe that this approach of directing traffic dynamics onto a certain desirable course will open up innovative ways for traffic network management. On the other hand, such a finding implies that temporary network modifications, such as roadwork repair work or lane closure, could produce unintentional long-term effects on the system evolution, which may last long after the closure is finished.

Although the technique in this study can be extended to within-day (i.e. continuous-time) dynamics and multiple equilibria, for simplicity, in this first study, we restrict our attention to day-to-day (i.e. discrete-time) dynamics with a unique equilibrium. In section 2, we will introduce models of day-to-day dynamics in the spirit of Cantarella and Cascetta (1995). Global asymptotic stability is discussed in section 3, together with other attractors associated with a dynamical system. Section 4 shows how the methodology can be implemented to direct traffic dynamics to equilibrium. Finally, some concluding remarks are provided in Section 5.

## 2. Day-to-Day Dynamics

### 2.1. Traffic Assignment and Wardrop Equilibrium

Consider a network with $N$ origin-destination (OD) pairs. Each OD pair $i$ ($i = 1,2,...,N$) is connected by a set of routes, denoted as $\mathbf{R}_i$, with $m_i = |\mathbf{R}_i|$ and $M = \sum_{i=1}^{N} m_i$. On each day $n$, a demand of $d_i^{(n)}$ users on OD pair $i$ make their travel choices over the route set $\mathbf{R}_i$. The $M$-vector $\mathbf{x}^{(n)} \in \mathbf{D}^{(n)}$ denotes a feasible route flow assignment where $\mathbf{D}^{(n)}$ is the feasible set:

$$\mathbf{D}^{(n)} = \{\mathbf{x}^{(n)} \in \Re_+^M : \sum_{r \in \mathbf{R}_i} x_r^{(n)} = d_i^{(n)}, \forall i = 1,2,...,N\}. \quad (1)$$

The route cost function is defined as $\mathbf{c}^{(n)} = c(\mathbf{x}^{(n)})$ where $\mathbf{c}^{(n)}$ is an $M$-vector; a flow $\mathbf{x}^{(n)}$ is termed Wardrop equilibrium (Smith, 1979) if and only if

$$x_r^{(n)} > 0 \Rightarrow c_r(\mathbf{x}^{(n)}) \le c_s(\mathbf{x}^{(n)}), \forall r,s \in \mathbf{R}_i, r \ne s, i = 1,2,...,N. \quad (2)$$

This Wardrop equilibrium flow $\mathbf{x}^*$ forms a steady state. Every trip-maker on the same OD pair experiences the same cost regardless of which route they choose. Each unused route has an equal or higher cost. As such, the system offers no incentive for any user to switch route. Thus, the system is in equilibrium.

### 2.2. Dynamical System

A first-order discrete-time dynamical system is defined by the following recurrence equation:

$$\mathbf{x}^{(n)} = y(\mathbf{x}^{(n-1)}). \quad (3)$$

A point $\mathbf{x}^*$ is called a fixed-point (or equilibrium point) if $\mathbf{x}^* = y(\mathbf{x}^*)$. The function $y$ maps the state of the system on day $n-1$ to day $n$.

The day-to-day dynamics of network flow can be modeled by the following system:

$$\mathbf{x}^{(n)} = \mathbf{Q}^{(n)}\mathbf{W}^{(n)}\mathbf{x}^{(n-1)} + (\mathbf{I} - \mathbf{W}^{(n)})\mathbf{x}^{(n-1)}, \quad (4)$$

where $\mathbf{W}^{(n)} = W(\mathbf{x}^{(n-1)})$ is a diagonal $M \times M$ matrix whose entry $w_{kk}^{(n)}$ represents the probability that users on route $k$ in the previous day reconsider their route choices. $\mathbf{Q}^{(n)} = Q(\mathbf{x}^{(n-1)})$ is a block diagonal $M \times M$ matrix, containing $m_i \times m_i$ blocks $\mathbf{Q}_{[i]}^{(n)}$, one for each OD pair $i$. An entry within each block, $q_{kj}^{(n)}$ for $k,j \in \mathbf{R}_i$ represents the probability of users on route $j$ in the previous day switch to route $k$. Note that a positive $q_{jj}^{(n)}$ implies that users on route $j$ on day $n$ continue to use the same route as on day $n-1$.

Furthermore, in this study, we consider the travel demand per OD pair as fixed over time, i.e. $d_i^{(n)} = d_i, \forall i$. Therefore, the feasible set $\mathbf{D}^{(n)} = \mathbf{D}$ does not vary day by day. To fulfill this feasibility requirement, it is necessary to specify that:

$$w_{kk}^{(n)} \in [0,1], \forall k \text{ and } \sum_{k \in \mathbf{R}_i} q_{kj}^{(n)} = 1, \forall j \in \mathbf{R}_i, \forall i. \quad (5)$$

Under this condition, $\mathbf{x}^{(n-1)} \in \mathbf{D} \Rightarrow \mathbf{x}^{(n)} \in \mathbf{D}$. On the other hand, the Wardrop equilibrium $\mathbf{x}^*$ should concur as a fixed-point for this dynamical system, meaning that starting from the equilibrium point, the system should remain at that point. This requirement is represented as:

$$\mathbf{x}^* = Q(\mathbf{x}^*)W(\mathbf{x}^*)\mathbf{x}^* + (\mathbf{I} - W(\mathbf{x}^*))\mathbf{x}^*. \quad (6)$$

The details of the dynamical system (4) can be specified and simplified in many ways. For example, on the same OD pair, if the probability of route choice reconsideration does not depend on the route chosen in the previous day, then $w_{kk}^{(n)} = w_i^{(n)}, \forall k \in \mathbf{R}_i$. In another example, the route switching matrix can be assumed to be invariant over time; that is $\mathbf{Q}^{(n)} = \mathbf{Q}$ or that the probability of route switching between each pair of routes remains unchanged. On the other hand, one can also define more elaborate routes for the route adjustment processes, making them functions of both the route flows and route costs in the previous day.

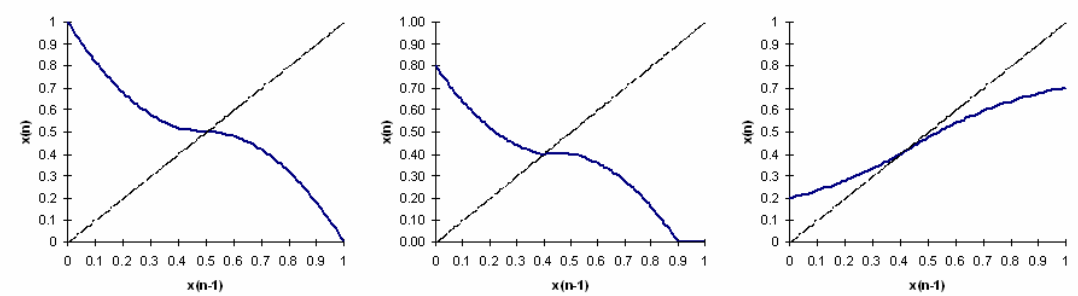
### 2.3. The Two-link Example Network

For simplicity, we restrict our attention to a two-link network with the fixed demand of one unit and use the same example throughout this study. The single OD pair is connected by two different routes, each consisting of one link. The flows on the two routes are denoted as: $\mathbf{x} = [x_1, x_2]^T$, with the feasible set defined as: $D = \{\mathbf{x} : 0 \le x_1 \le 1, x_2 = 1 - x_1\}$. The travel cost functions are linear and separable: $c_1(x_1) = 0.6x_1 + 0.4$ for route 1 and $c_2(x_2) = 0.4x_2 + b$ for route 2, where $b$ is a parameter to be varied between $0 < b < 1$ for illustration.

Under fixed demand, there is only one degree of freedom to express the two route flows. If we denote $x = x_1$ as the flow on route 1, then the flow on route 2 is $x_2 = 1 - x$. Therefore a single variable $x$ is sufficient to specify the network flow dynamics. Let $g(x) = c_1(x) - c_2(1-x) = x - b$ be the travel cost difference between the two routes, then solving the equation $g(x) = 0$ gives the equilibrium flow, i.e., $x^* = b$.

The day-to-day dynamics is specified by setting $\mathbf{W}^{(n)} = \begin{bmatrix} p_1^{(n)} & 0 \\ 0 & p_2^{(n)} \end{bmatrix}$ and $\mathbf{Q}^{(n)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ in (4). We further define $p_1^{(n)} = \min\left\{1, \alpha[g(x^{(n-1)})]_+\right\}$ and $p_2^{(n)} = \min\left\{1, \alpha[-g(x^{(n-1)})]_+\right\}$, where $[z]_+ = \max\{0, z\}$. In this setting, a proportion of $p_1$ ($p_2$) users on route 1 (route 2) reconsider their previous route choices. And that all those who reconsider their previous route choices will switch to the better route on the current day. The probability of route reconsideration is proportional to the expected travel cost reduction scaled by $\alpha$, and bounded from above by 1. The expected travel cost reduction, $[g(x^{(n-1)})]_+$ for route 1 and $[-g(x^{(n-1)})]_+$ for route 2, is equal to the previous day's cost difference between the chosen route and the unused route, bounded from below by 0. This is consistent with the behavior that under higher travel cost differences, users are more willing to switch routes. Finally, one can verify that this route adjustment process from day to day fulfills condition (5). Therefore, its feasibility is maintained over time. Finally, one can verify that condition (6) also holds, i.e., Wardrop equilibrium is a fixed point of this dynamical system.

The flow evolution function $x^{(n)} = y(x^{(n-1)})$ is plotted in Figure 1. Three cases are shown: (a) $b = 0.5, \alpha = 2$, (b) $b = 0.4, \alpha = 2$, and (c) $b = 0.4, \alpha = 0.5$. An additional straight line, $y = x$, is drawn on each graph. The intersection between this straight line and the evolution function indicates the equilibrium point. As can be seen, user behavior in terms of route switching is very active in cases (a) and (b), whereas it is not so active in case (c). Another point worth mentioning is that in case (b), the probability bound of route reconsideration forms a binding constraint for the horizontal region between $[0.9, 1.0]$.




(a) $b = 0.5, \alpha = 2$ (b) $b = 0.4, \alpha = 2$ (c) $b = 0.4, \alpha = 0.5$

Figure 1. Day-to-day dynamics

# 3. Global Asymptotic Stability

## 3.1. Asymptotic Stability

We start by clarifying some concepts. The definitions of some properties for a fixed-point $\mathbf{x}^*$ of the dynamical system $\mathbf{x}^{(n)} = y(\mathbf{x}^{(n-1)})$ are listed below:

**Convergence** A fixed-point $\mathbf{x}^*$ is called *convergent* if there exists $\delta > 0$ such that $\lim_{n \to +\infty} \mathbf{x}^{(n)} = \mathbf{x}^*, \quad \forall \mathbf{x}^{(0)} : | x_r^{(0)} - x_r^* | < \delta, r = 1, 2, ...M$ .

**Stability** A fixed-point $\mathbf{x}^*$ is called *stable* if for any $\varepsilon > 0$ , there exists $\delta > 0$ such that $| x_r^{(0)} - x_r^* | < \delta, r = 1, 2, ...M \Rightarrow | x_r^{(n)} - x_r^* | < \varepsilon, r = 1, 2, ...M$ for all $n \geq 1$.

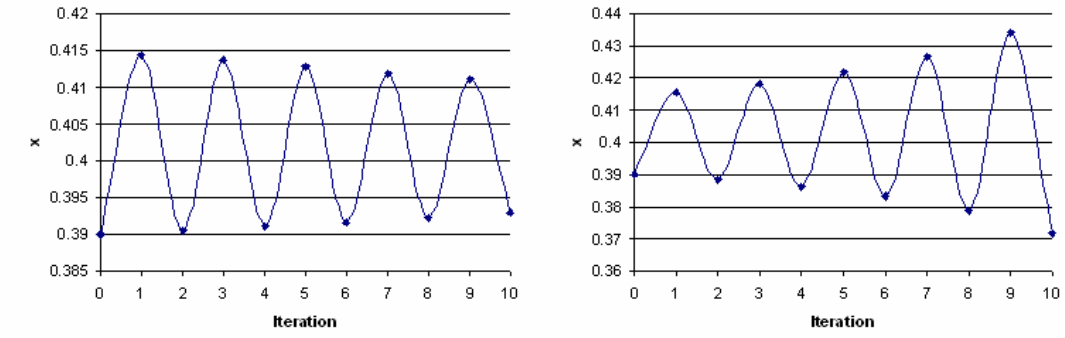
**Asymptotic stability** A fixed-point $\mathbf{x}^*$ is called *asymptotically stable* if it is both stable and convergent.

The difference between convergence and stability can be illustrated by the following two examples. Consider the dynamical system $x^{(n)} = x^{(n-1)}, \forall x \in \Re$. Every point $x$ is a stable fixed-point yet each is not convergent, because if one starts with some $x^{(0)}$ such that it satisfies $| x^{(0)} - x^* | < \delta$ but $x^{(0)} \neq x^*$, the process will never move $x^{(n)}$ closer to $x^*$. In the second example, consider the recurrence equation:

$$x^{(n)} = \begin{cases} \dfrac{1}{x^{(n-1)}} & \text{if } x^{(n-1)} = \dfrac{1}{k} \text{ for some non-zero integer number } k, \text{ or} \\ 0 & \text{otherwise.} \end{cases}$$

The fixed-point $x^* = 0$ is convergent but not stable. If one start with some $x^{(0)}$ that is not $\pm 1$, it will always converge to zero, at most in two cycles, some in one cycle. It is not stable when $n = 1$ for certain $x^{(0)}$; some $x^{(0)}$ is magnified by the process to $x^{(1)}$, violating the stability definition when $n = 1$. Actually, the smaller is the difference between $x^{(0)}$ and $x^* = 0$, the larger is the magnification for $x^{(1)}$. For $n \geq 2$, the process is fine, however. Also, for this recurrence equation, the fixed-points $x^* = \pm 1$ are neither stable nor convergent.

Referring to the two-link example as shown in Figure 1, it can be derived that if $\alpha < 1/\left[ b(1-b) \right]$, the equilibrium point $x^* = b$ is asymptotically stable. According to this condition, for the case of $b = 0.4$, the critical value of $\alpha$ is 4.167 . In Figure 2, two trajectories from the same initial point of $x^{(0)} = 0.39$ are shown for the first ten days, one for $\alpha = 4.0$; the other for $\alpha = 4.2$. It can be seen from these trajectories that case (a) is gradually converging to the equilibrium point $x^* = 0.4$, whereas case (b) is not.




(a) $b = 0.4, \alpha = 4.0$ (b) $b = 0.4, \alpha = 4.2$

Figure 2. System evolution under different values of $\alpha$

### 3.2. Domain of Attraction and Global Asymptotic Stability

The *domain of attraction* (or *attraction basin*) for a fixed-point $\mathbf{x}^*$, $\mathbf{B}(\mathbf{x}^*)$, is the set of all states which will dynamically evolve to the fixed-point: [1]

$$\mathbf{B}(\mathbf{x}^*) = \{\mathbf{x}^{(0)} : \lim_{n \to +\infty} \mathbf{x}^{(n)} = \mathbf{x}^*\}. \quad (7)$$

Obviously, if a state lies within $\mathbf{B}(\mathbf{x}^*)$, then all states on its trajectory [2] also belong to $\mathbf{B}(\mathbf{x}^*)$. If the fixed-point $\mathbf{x}^*$ is asymptotically stable, then there must be a neighborhood around $\mathbf{x}^*$ that forms a subset of $\mathbf{B}(\mathbf{x}^*)$.

A fixed-point $\mathbf{x}^*$ is said to possess global asymptotic stability if

$$\lim_{n \to +\infty} \mathbf{x}^{(n)} = \mathbf{x}^*, \forall \mathbf{x}^{(0)} \in \mathbf{D}. \quad (8)$$

That is, every state in the feasible set converges to the fixed-point. It is equivalent to say that $\mathbf{x}^*$ is stable and $\mathbf{B}(\mathbf{x}^*) = \mathbf{D}$.

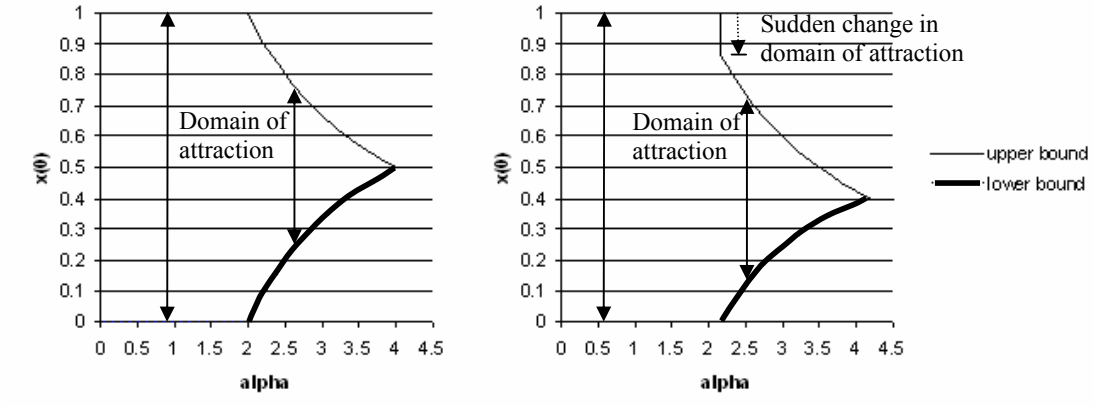
Referring to the two-link example as shown in Figure 1, it can be derived that if $\alpha < \frac{1+\sqrt{1+\frac{4}{0.5-|0.5-b|}}}{2}$, the equilibrium point $x^* = b$ possesses global asymptotic stability, i.e. $\mathbf{B}(x^* = b) = [0,1]$. When $\alpha \geq \frac{1+\sqrt{1+\frac{4}{0.5-|0.5-b|}}}{2}$, $\mathbf{B}(x^* = b)$ is a subset of $[0,1]$. Figure 3 shows the domain of attraction in relation to $\alpha$ and $b$. Case (a) refers to the instance of $b = 0.5$ and case (b) the instance of $b = 0.4$. The two-headed arrows shown in Figure 3 (a) illustrate the domains of attraction for two different $\alpha$ values. For $\alpha < 2$, the entire feasible region overlaps with the domain of attraction, i.e., $\mathbf{B} = [0,1]$ or the equilibrium point possesses global asymptotic stability. For $2 \leq \alpha < 4$, the domain of attraction is bounded by the top and bottom curves, denoted as $\mathbf{B} = (l,u)$ ($l$ refers to the lower bound and $u$ the upper bound). Finally, for $\alpha \geq 4$, the domain of attraction becomes a point, $\mathbf{B} = \{0.5\}$. An important point to note is that the domain of attraction for $2 \leq \alpha < 4$ is an open set rather than a closed set, which is to be discussed further in Section 3. In Figure 3(b), for the instance of $b = 0.4$, $\mathbf{B} = [0,1]$ for $\alpha < 2.158$, $\mathbf{B} = (l,u)$ (i.e., the open set bounded by the top and bottom curves) for $2.158 \leq \alpha < 4.167$, and $\mathbf{B} = \{0.4\}$ for $\alpha \geq 4.167$. For both cases, with an increasing $\alpha$, the domain of attraction gradually shrinks from the entire feasible set to a single point (i.e., the fixed-point). For case (b), there is a sudden change in $\mathbf{B}$ around the point $\alpha = 2.158$, which is due to the binding constraint on the proportion of users reconsidering their route choices (cf. Figure 1(b)).

Figure 4 shows the evolution of the dynamic process for $b = 0.5$ and different values of $\alpha$. Although the equilibrium point possesses global asymptotic stability for $\alpha < 2$, the system evolution processes are different for different values of $\alpha$. When $\alpha$ is small, as in Figure 4(a), the system evolves steadily toward the

[1] In some studies, the term "maximal domain of attraction" is used and any subset of which is called a domain of attraction.

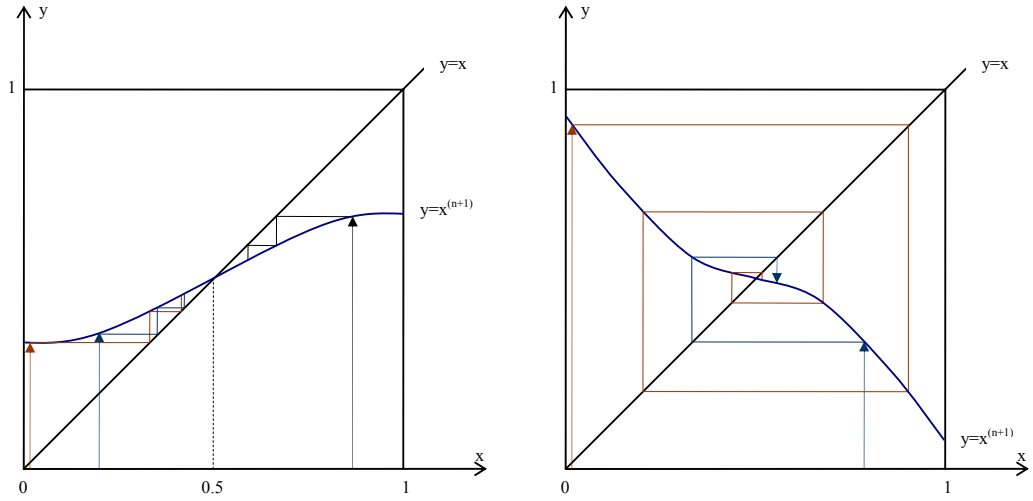
[2] As defined by the recurrence equation (3).

equilibrium. For a larger $\alpha$, as in Figure 4(b), the system swings back and forth around the equilibrium before arriving at it.

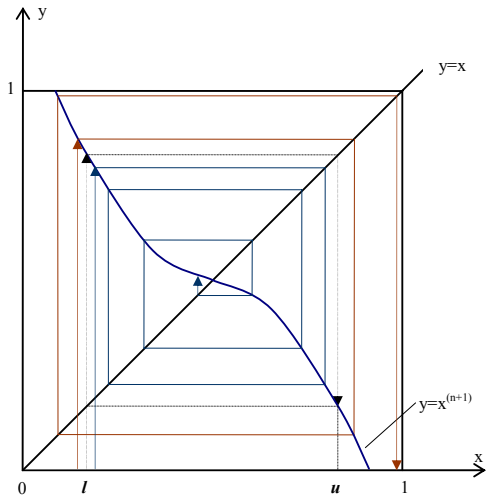



(a) $b = 0.5$ (b) $b = 0.4$

Figure 3. Domain of attraction




(a) Phase portrait of $x^{(n)}$: $b = 0.5, \alpha = 0.5$ (b) Phase portrait of $x^{(n)}$: $b = 0.5, \alpha = 1.9$




(c) Phase portrait of $x^{(n)}$: $b = 0.5, \alpha = 2.4$

Figure 4. Dynamical evolution

When the system does not possess global asymptotic stability, or when the domain of attraction for the fixed-point (and in the case of multiple fixed-points, the union of all attraction basins) does not cover the entire feasible set, there exist initial flow patterns that do not evolve to any fixed-point. Under this situation, the system may converge to other attractors.

Generally there are three types of attractors for a dynamical system: *fixed-points*, *cycles*, and *chaos*. Fixed-points can be considered as cycle attractors with period one. Thus, when referring to cycle attractors, we typically mean cycles with a finite period of at least 2. In a similar way, chaotic attractors can be considered as cycles with an infinite period. Starting from any initial point in the feasible set, the dynamical system always evolves towards one of these three attractors. Earlier in this section, we illustrate the domain of attraction of a fixed-point. In a similar manner, cycle and chaotic attractors also have their domains of attraction. As a result, the union of attraction basins for all types of attractors spans the entire feasible set.

Generally, finding the domains of attraction for all the attractors associated with a dynamical system can be arduous, indicating the complexity of the problem. In the following, we denote an $n$-cycle attractor by an $n$-element set $\mathbf{X}_C^* = \{\mathbf{x}_{C,1}^*, \mathbf{x}_{C,2}^*, ..., \mathbf{x}_{C,n}^*\}$, which consists of the $n$ periodic values on the cycle. The subscript $C = I, II, ...$ is an index to distinguish between different attractors associated with the dynamical system. And the domain of attraction for cycle C is denoted by $\mathbf{B}(\mathbf{X}_C^*)$. In the interest of space, we use the two-link example discussed earlier for this illustration and only state the results without deriving them.

In the two-link example, other than fixed-points, 2-cycle attractors also exist, but there are no attractors with periods longer than 2. For case (a), with $b = 0.5$, when $\alpha = 2$, the 2-cycle attractor is: $\mathbf{X}_I^* = \{0,1\}$ where $x_{I,1}^* = 0$ and $x_{I,2}^* = 1$ are the two values of $x$ on the 2-cycle. The domain of attraction contains only the two points on the cycle itself, i.e., $\mathbf{B}(\mathbf{X}_I^*) = \{0,1\}$. When $2 < \alpha < 4$, two 2-cycles exist, including: $\mathbf{X}_{II}^* = \{0,1\}$ and $\mathbf{X}_{III}^* = \{l,u\}$ with their domains of attraction being $\mathbf{B}(\mathbf{X}_{II}^*) = [0,l) \cup (u,1]$ and $\mathbf{B}(\mathbf{X}_{III}^*) = \{l,u\}$, respectively. When $\alpha \geq 4$, there is only one 2-cycle attractor: $\mathbf{X}_{IV}^* = \{0,1\}$, with its domain of attraction being $\mathbf{B}(\mathbf{X}_{IV}^*) = [0,0.5) \cup (0.5,1]$.

For case (b), with $b = 0.4$, when $\alpha = 2.158$, there is one 2-cycle attractor: $\mathbf{X}_V^* = \{0, \alpha b\}$ whose domain of attraction is $\mathbf{B}(\mathbf{X}_V^*) = \{0\} \cup [\alpha b, 1]$. The region $[\alpha b, 1]$ represents the sudden change as shown in Figure 3(b). When $2.158 < \alpha \leq 2.5$, there are two 2-cycle attractors: $\mathbf{X}_{VI}^* = \{0, \alpha b\}$ and $\mathbf{X}_{VII}^* = \{l,u\}$ whose domains of attraction are, respectively, $\mathbf{B}(\mathbf{X}_{VI}^*) = [0,l) \cup (u,1]$ and $\mathbf{B}(\mathbf{X}_{VII}^*) = \{l,u\}$. When $2.5 < \alpha < 4.167$, there are two 2-cycle attractors: $\mathbf{X}_{VIII}^* = \{0,1\}$ and $\mathbf{X}_{IX}^* = \{l,u\}$, whose domains of attraction are $\mathbf{B}(\mathbf{X}_{VIII}^*) = [0,l) \cup (u,1]$ and $\mathbf{B}(\mathbf{X}_{IX}^*) = \{l,u\}$, respectively. When $\alpha \geq 4.167$, there is only one 2-cycle attractor: $\mathbf{X}_X^* = \{0,1\}$, whose domain of attraction is $\mathbf{B}(\mathbf{X}_X^*) = [0,0.4) \cup (0.4,1]$.

Cycles *I*, *III*, *VII* and *IX* are unstable attractors, whose domains of attraction contain only the periodic points on the cycles themselves. Initial flows between the two periodic points of the unstable 2-cycles are attracted to the equilibrium point, whereas initial flows on one side of these cycles are attracted to the boundary cycle (cycles *II*, *VI* and *VIII*). On the other hand, cycles *II*, *IV*, *VI*, *VIII* and *X* are stable. Neighborhoods around them belong to their domains of attraction.

Figure 4(c) shows the case for $b = 0.5, \alpha = 2.4$ where the equilibrium stability is not global but is restricted to a local neighborhood between $(l,u)$. Only initial states within the $(l,u)$ region are attracted to the equilibrium. The two points $\{l,u\}$ form an unstable 2-cycle (i.e., the same as cycle *III* above). All initial states within $[0,l) \cup (u,1]$ are attracted to the boundary cycle $\{0,1)$.

# 4. Directing Traffic System Evolution to Equilibrium

As discussed earlier, when the equilibrium of a dynamical system does not possess global asymptotic stability, initial points lying outside its domain of attraction will not converge to the equilibrium. However, this does not say that the system cannot be modified so that such initial points will eventually converge to the equilibrium. For this purpose, gaining an understanding of the domains of attraction of the various attractors associated with a dynamical system is instrumental in modifying the system dynamics toward the desired outcome.

One way to modify the system dynamics is to alter travelers' behavior in route switching. In the two-link example, referring to Figure 3, this can be achieved by lowering $\alpha$ or travelers' propensity to route switching. The result is also observed in previous studies (e.g., Szeto and Lo, 2004), which showed that more aggressive route switching behavior generally led to over-reaction and smaller chances of convergence. Route switching behavior is, however, intrinsic to the network users and not something easy to change. From the network management point of view, it is much more feasible to alter the system behavior through traffic control, lane closure or addition, or pricing means. In particular, we illustrate with the following example that temporary alterations in the network configuration are sufficient to bring about fundamental shifts in the traffic system evolution. During the temporary network alteration, the system evolution will be shifted to a point that lies within the domain of attraction of the equilibrium. Subsequently, upon restoration of the original network configuration, the shifted point will evolve toward the equilibrium on its own. In other words, the system evolution can be "directed" by this temporary network modification. The key is to find out where and when to implement the temporary network alteration and for how long.

### 4.1. General Approach

Let the domain of attraction of the equilibrium point $\mathbf{x}^*$ be $\mathbf{B}(\mathbf{x}^*)$, which is a subset of the feasible set, i.e., $\mathbf{B}(\mathbf{x}^*) \subset \mathbf{D}$. And let the initial flow assignment be $\mathbf{x}^{(0)}$, which is outside $\mathbf{B}(\mathbf{x}^*)$, i.e. $\mathbf{x}^{(0)} \in \mathbf{D}$, $\mathbf{x}^{(0)} \notin \mathbf{B}(\mathbf{x}^*)$, and thus will not by itself evolve towards $\mathbf{x}^*$. By altering the network configuration temporarily, we construct a new transitional alternative equilibrium point $\mathbf{x}^*_{alt}$ such that

$$\mathbf{x}^{(0)} \in \mathbf{B}(\mathbf{x}^*_{alt}) \text{ and } \mathbf{x}^*_{alt} \in \mathbf{B}(\mathbf{x}^*). \tag{9}$$

By allowing $\mathbf{x}^{(0)}$ to evolve sufficiently close to $\mathbf{x}^*_{alt}$ during the temporary network alteration, the system will gradually move inside the domain of attraction of $\mathbf{B}(\mathbf{x}^*)$ or $\mathbf{x}^{(n)} \in \mathbf{B}(\mathbf{x}^*)$. Afterward, the temporary alternation can be removed and the network is restored to its original form. This modified initial state $\mathbf{x}^{(n)}$ will then continue to evolve toward $\mathbf{x}^*$. In the interest of space, we leave the details here but illustrate this approach with the two-link example discussed earlier.

### 4.2. The Two-Link Example

Only two variable parameters are presented in the two-link example: $\alpha$ represents travelers' route switching propensity, whereas $b$ is the free-flow cost on the second route. This route cost can be altered quite readily by traffic signal, lane closure, pricing, etc.

Let's consider the case of $b = 0.4$ and $\alpha = 2.5$. As shown in Figure 3(b), the domain of attraction for the equilibrium $x^* = 0.4$ is $\mathbf{B}(x^*) = (0.121, 0.734)$. An initial flow $x^{(0)} \notin \mathbf{B}(x^*)$, say, $x^{(0)} = 0.1$, will not evolve toward $x^* = 0.4$. Indeed the initial flow of $x^{(0)} = 0.1$ is attracted gradually to the boundary cycle, as shown in Figure 5 for days 0 ~ 9. On day 10, we change $b$ from 0.4 to 0.25 for one day. The equilibrium for this modified network is $x^*_{alt} = 0.25$, whose domain of attraction is $\mathbf{B}(x^*_{alt}) = [0,1]$. Since $x^{(10)} \in \mathbf{B}(x^*_{alt})$, the system recourses towards $x^*_{alt} = 0.25$ on day 11, moving to $x^{(11)} = 0.625$, which falls within the domain of attraction

of the original network, i.e. $\mathbf{B}(x^*) = [0.121, 0.734]$ . Upon restoration of the original link cost functions on day 11, $x^{(n)}, n = 12,...$ eventually arrives at the equilibrium point $x^* = 0.4$ in a few more days. In this example, it turns out that one-day of network alternation is sufficient to modify the system evolution to achieve the equilibrium state of the original network.

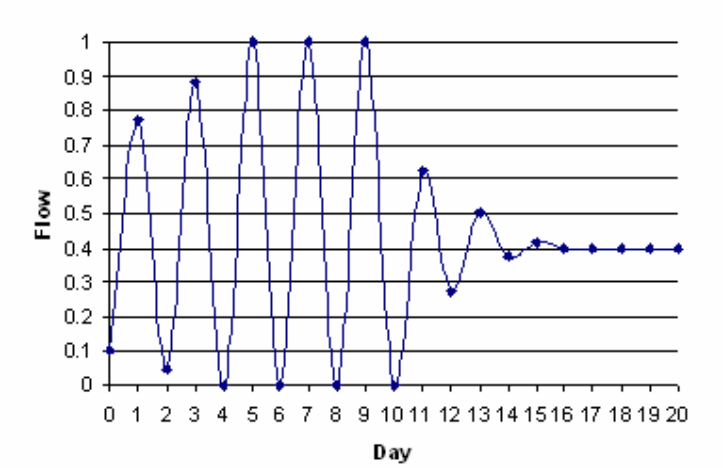



Figure 5. Directing the system dynamics: $b = 0.25$ for day 10 and $b = 0.4$ on all other days

### 4.3. Implications for Network Traffic Management

The results shown here may have implications on network modifications, such as road pricing and the opening of new routes, say for example. One can consider the variable $b$ in the two-link example to be the toll charge on route 2. With an initial flow of $x = 0.1$ and a toll charge of $0.4$ (in travel cost term), the system will result in non-stop cyclic route switching without a fixed point. If the toll charge is lowered to $0.25$ (e.g., an initial discount) for one day, the system will calm down and evolve to the equilibrium flow without further route switching thereafter. Likewise, the opening of a new route to an existing network could introduce non-equilibrium cycles, which can be avoided by an initial partial opening of the route followed by its full opening.

Although illustrated on a specific two-link example, this methodology of directing traffic system evolution through temporary network alteration is applicable to more general cases. Whenever the equilibrium points of a dynamical system do not possess global asymptotic stability, there exist such opportunities of modifying the system evolution through temporary network alteration. In other situations, when the system has multiple equilibrium points, one can also apply this technique to direct the system to the equilibrium that is the most beneficial to the traffic management objective.

Many methods to temporarily alter the network configuration are possible. Generally they can be achieved by modifying the route cost functions. Some examples include changing the traffic signal setting, lane closure or addition, pricing, etc. We note that certain temporary roadway construction or closures may unintentionally introduce the same effect as discussed above, which could lead to long-term changes in the system evolution after the restoration of the network. It is interesting to find out whether this actually occurs in a real transportation network. For the time being, this approach represents a theoretical exploration of what are plausible.

## 5. Concluding Remarks

In this study we provided a methodology through which we can direct the traffic system evolution to the equilibrium state for initial states that lie outside the domain of attraction. This can be achieved by temporary alternations of the network configuration, without modifying the demand constraint, feasible region, or route switching behavior terms. We believe that this technique can be implemented to a wide range of networks and

will open up new ways for contemplating traffic management and network design. Our current research is to extend this approach for a general transportation network.

## Acknowledgement

This study is partially supported by the Competitive Earmarked Research Grant from the Research Grants Council of the Hong Kong Special Administrative Region (HKUST6283/04E).

## References

Cantarella, G.E. and Cascetta, E. (1995) Dynamic processes and equilibrium in transportation networks: towards a unifying theory. *Transportation Science* 29, 305-329.

Horowitz, J.L. (1984) The stability of stochastic equilibrium in a two link transportation network. *Transportation Research B* 18, 13-28.

Smith, M.J. (1979) The existence, uniqueness and stability of traffic equilibria. *Transportation Research B* 13, 295-304.

Smith, M.J. (1984) The stability of a dynamic model of traffic assignment—an application of a method of Lyapunov. *Transportation Science* 18, 245-252.

Szeto, W.Y. and Lo, H. (2005) Non-equilibrium Dynamic Traffic Assignment. *Traffic and Transportation Theory*. Edited by H. Mahmassani. Elsevier Science, 427-446.

Watling, D.P. (1999) Stability of the stochastic equilibrium assignment problem: a dynamical systems approach. *Transportation Research B* 33, 281-612.

Wardrop, J.G. (1952) Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers*, Part II(1), 325-378.

# EQUILIBRIUM DYNAMIC TRAFFIC ASSIGNMENT WITH ADAPTIVE ROUTING CHOICES

*Song Gao, Caliper Corporation, USA, song@caliper.com*

## Abstract

This paper establishes a user-equilibrium traffic assignment model where users make adaptive routing choices, defined as routing policies, in a stochastic time-dependent network. Waldrop's First Principle is generalized to be the equilibrium condition: each user follows a routing policy with minimum perceived disutility at his/her departure time and no user can unilaterally change routing policies to improve his/her perceived disutility. A general framework is provided and the equilibrium problem is formulated as a fixed point problem with three components: the optimal routing policy generation module, the routing policy choice model and the policy-based dynamic network loader. An MSA (method of successive averages) heuristic is designed. Computational tests are carried out in a hypothetical network, where random incidents are the source of stochasticity. The heuristic converges satisfactorily in the test network under the proposed test settings. The adaptiveness in the routing policy based model leads to travel time savings at equilibrium. As a byproduct, travel time reliability is also enhanced. The value of online information is an increasing function of the incident probability. Travel time savings are high when market penetrations are low. However, the function of travel time saving against market penetration is not monotonic. This suggests that in a travelers' information system or route guidance system, the information penetration needs to be chosen carefully to maximize benefits.

## 1 Introduction

Stochasticity in transportation systems is both intuitively prevalent and experimentally shown. Travelers' routing decisions in a stochastic network with online information is conceivably different from those in a deterministic network. It is generally believed that adaptive routing will save travel time and enhance travel time reliability. For example, in a network with random incidents, if one does not adapt to an incident scenario, he/she could be stuck in the incident link for a very long time. However, if adequate online information is available about the incident and the traveler adapts to it by taking an alternative route, he/she can save travel time compared to the non-adaptive case. The adaptiveness also ensures that the travel time is not prohibitively high in incident scenarios, and thus provides a more reliable travel time. The problem of optimal adaptive routing decision making for individual travelers has been studied by various researchers and a complete literature review can be found in Gao and Chabini [1]. A general conclusion from the above studies is that in a flow-independent stochastic time-dependent (STD) network, an individual user's expected travel time from being adaptive (in one way or another, depending on the problems studied specifically) is always no higher than that from being non-adaptive, i.e. following a simple path.

After understanding how an individual traveler makes adaptive routing decisions, another research question would be: what will be the network-level impact if many travelers make adaptive routing decisions? In a congested network, the stochastic nature of traffic variables affects travelers' routing decisions, which in turn affect traffic conditions. The interaction between supply and demand in a stochastic dynamic network needs to be captured to evaluate the network-level impact. This interaction in a deterministic network (with possible perception errors from the demand side) is captured by a conventional dynamic traffic assignment (DTA) model. This paper then establishes a user-equilibrium traffic assignment model where users make adaptive routing decisions, defined as routing policies, in a general stochastic time-dependent network. There is quite limited study of equilibrium dynamic traffic assignment models in the literature, where adaptive routing decisions are an integral part of a user's behavior model. Hamdouch et al. [2] proposed a strategic model for dynamic traffic assignment, as an extension to the static model studied by Marcotte et al. [3]. The strategic model is built around the concept "strategy", commonly used in the transit assigment literature to describe travelers' adaptive behavior. A strategy consists in a rule that assigns to each node of the network a set of arcs in the forward star of that node, sorted

according to some preference order. A traveler follows the first available arc from the sorted set, where the availability of an arc is dictated by its rigid capacity. An equilibrium assignment is reached when expected delays of active strategies are minimal, for every origin-destination pair and every departure time. Some key assumptions of the model are: 1) Travel delays happen only at nodes, when the arc that a traveler wants to access has reached its rigid capacity. Travel times on arcs are fixed. Travelers are in a vertical queue at a node before entering outgoing arcs. 2) Randomness in travel time comes from the fact that the position of any traveler in the arrival flow at a node is random, and thus there is a certain probability that a traveler cannot access his/her preferred arc when the capacity of that arc is reached.

The first assumption is suitable in transit networks(rigid capacity and holding of traffic allowed), yet its applicability to a general traffic network is to be validated. The second assumption implies that external random factors, such as random incidents cannot be modeled, which limits the model's ability to assess an advanced traveler information system (ATIS) which usually plays an important role when random incidents happen in a traffic network. Furthermore, online information is not explicitely modeled, and thus sensitivity analysis with respect to online information is not available.

The paper is organized as follows. In Section 2, a conceptual framework for the policy-based stochastic DTA model is introduced with three components described and a solution algorithm presented. Computational tests are described in Section 3. Throughout the paper, a symbol with a $\sim$ over it is a random variable, while the same symbol without the $\sim$ is one specific value of the random variable. A "support point" is defined as a distinct value that a discrete random variable can take or a distinct vector of values that a discrete random vector can take, depending on the context. Thus a probability mass function (PMF) of a random variable(vector) is a combination of support points and the associated probabilities.

## 2 A Framework for the Policy-Based Stochastic DTA Model

We present a framework for the policy-based stochastic dynamic traffic assignment model to give a big picture on the input, output, model components' interaction, and data flow, as shown in Figure 1. The input to the overall DTA model is the stochastic dynamic demand $\tilde{D}$ and supply $\tilde{S}$ represented by a joint discrete distribution with $R$ support points, each of which has a probability $p_r, r = 1, ..., R$. The demand is assumed to be inelastic, i.e. the demand distribution is fixed. In a discrete time representation, any realization of random demand is given as a matrix of time-dependent numbers of O-D trips during all time intervals. $\tilde{D} = \{D^1, D^2, ..., D^R\}$, where $D^r$ is the demand matrix for the $r^{th}$ support point. $D^r = \{D^r_{jd,t}, t = 0, 1, 2, ..., \forall \text{OD pair}\{j, d\}\}$, where $D^r_{jd,t}$ is the number of trips between origin $j$ and destination $d$ for departure time $t$ for the $r^{th}$ support point. The random supply can be represented through the random occurrence, duration and severity of an incident or any other random supply factors: $\tilde{S} = \{S^1, S^2, ..., S^R\}$. Note that the same probability $p_r$ is associated with the outputs computed from $S^r, D^r$. In the remaining of the paper, whenever a support point has a superscript $r$, its associated probability is $p_r$, otherwise indicated. The output is an equilibrium distribution of link travel times $\tilde{C} = \{C^r_{jk,t}, \forall\{j, k\} \in A, \forall t, r = 1, 2, ..., R\}$, where $A$ is the set of links of the traffic network, and the corresponding routing policy splits $f = \{f^i_{jd,t}\}$, where $\{j, d\}$ is an OD pair, $t$ is the departure time, and $i$ is the index of policies. Note that the distributions of all relevant traffic random variables are discrete, as our definition of a routing policy is based on a discrete distribution of link travel times.

There are three major components of the stochastic DTA model: the users' routing policy choice model, denoted as $U$, the policy-based dynamic network loading model, denoted as $L$, and the optimal routing policy algorithm, denoted as $O$.

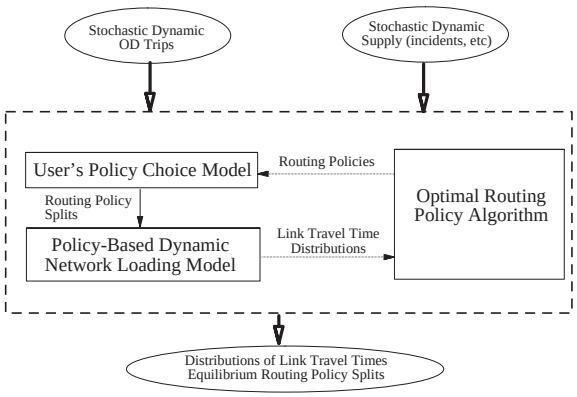



Figure 1: A Conceptual Framework of Stochastic Dynamic Traffic Assignment Model

## 2.1 Users' Routing Policy Choice Model

The users' routing policy choice model takes as input a set of routing policies $G = \{\mu_1, \mu_2, ..., \mu_i, ...\}$ generated by the optimal routing policy algorithm, and a joint distribution of link travel times$\tilde{C} = \{C^r_{jk,t}, r = 1, ..., R\}$ generated by the policy-based dynamic network loading model. The method of generating the choice set will be discussed in Section 2.4. Based on the relevant attributes of candidate routing policies, such as expected OD travel time and travel time standard deviation, the users' policy choice model outputs policy splits $f$ among the routing policies for each OD pair and each departure time. $f = U(G,\tilde{C})$ We keep the "large sample" assumption and assume policy splits are equal to corresponding policy choice probabilities. Note that we use "splits" rather than "flows" here: policy splits are deterministic, while policy flows could be stochastic, if the demand is stochastic. Policy splits will be translated into policy flows in the network loading model. The notion of policy flow can be understood as a generalization of path flow. Since a routing policy will manifest itself as a specific path for a given realization of link travel times, a policy flow will become a path flow for each support point of link travel times. Thus a policy flow can be viewed as a set of path flows, each with some probability.

## 2.2 Policy-Based Dynamic Network Loading Model

The demand is then loaded onto the network according the policy flow splits, by the policy-based dynamic network loading model. The stochastic demand and supply play their roles in the loading process. For each support point of the random demand and/or supply, the network loading model outputs a single realization of the link travel time distribution. Therefore through the loading, we obtain the PMF of link travel times from the PMF of demand/supply. Note that although the input demand/supply support points are distinct from each other, the output link travel time realizations are not necessarily distinct. This is why the word "realization" is used here, rather than support point. Nevertheless, the PMF of link travel times is still expressed through the $R$ realizations with the corresponding probabilities. $\tilde{C} = L(f, \tilde{D}, \tilde{S})$

## 2.3 Optimal Routing Policy Algorithm

The routing policy generation algorithm then takes as input the link travel time distribution and produces an optimal routing policy for each destination, which again will be used to generate the choice set for the users' policy choice model. $\mu_i = O(\tilde{C})$, $G = G \cup \mu_i$. The two equations can be combined as $G = G(\tilde{C})$.

What follows is a summary of optimal routing policy problems described in Gao and Chabini [1]. Let $G = (N, A, T, P)$ be a **stochastic time-dependent network**. $N$ is the set of nodes and $A$ is the set of links. The number of nodes and links are denoted respectively as $|N| = n$ and $|A| = m$. The network has a single destination node $d$. $T$ is the set of time periods $\{0, 1, ..., K-1\}$. Travel time on each link $(j, k)$ during each time

period $t$ is a random variable $\tilde{C}_{jk,t}$ with finite number of discrete, positive and integral support points. Beyond time period $K-1$, travel times are static and deterministic, i.e. the travel time of link $(j,k)$ at any time $t \geq K-1$ is equal to $C_{jk,K-1}$.

$P$ is the probabilistic description of link travel times. Let $P = \{v_1, v_2, ..., v_R\}$ be the set of support points of the link travel time distribution. The $r$th support point has a probability $p_r$, and $\sum_{r=1}^{R} p_r = 1$ . $C^r_{jk,t}$ is the travel time on link $(j,k)$ at time $t$ for the $r$th support point.

We assume the traveler knows *a priori* the probabilistic description $P$ of the network. The traveler can make decisions only at nodes. The decision is what node $k$ to take next, based on the *current state* $x = \{j, t, I\}$, where $j$ is the *current node*, $t$ is the *current time*, and $I$ is the *current information*. Current information $I$ is defined as a set of available realized link travel times at the current time and current node that are useful for making inferences about future link travel times. It represents the traveler's knowledge about the network conditions. This knowledge could be dependent on time, location of the traveler, mode of transportation, etc. Current information $I$ therefore should be regarded as $I(j,t)$, but we usually use $I$ only since $I$ is always associated with a state where $j$ and $t$ are well defined. An ideal case is when travelers have perfect online information, where all link travel time realizations up to the current time are available, but generally the information is local, e.g. one learns the travel time realization of some downstream links when he/she passes a Variable Message Sign (VMS). One can be in many different states traveling in the stochastic time-dependent network, and we have the following definition: A routing policy $\mu(x)$ is a mapping from network states to decisions (next nodes specifically).

This definition indicates that the routing decision in a stochastic time-dependent network is far from being set *a priori*. Rather, it is closely related to the network conditions, and this notion is critical in any ATIS application. The generic optimality condition for optimal routing policy problems and an operational algorithm for the perfect online information variant can be found in Gao and Chabini [1].

## 2.4 Policy-Based Equilibrium

The three components interact with each other, and a fixed-point formulation of the policy-based equilibrium can be derived based on the interaction. $\tilde{C} = L\left(U\left(G\left(\tilde{C},\right),\tilde{C}\right),\tilde{D},\tilde{S}\right)$

The equilibrium can be described by the following generalized Wadrop's First Principle: a traffic network is in policy-based stochastic dynamic equilibrium, if each user follows the routing policy with minimum perceived disutility at his/her departure time, and no user can unilaterally change routing policies to improve his/her perceived disutility.

The idea of the solution algorithm is to find a solution to the fixed point problem by an iterative process on policy splits. At each iteration, the policy splits are updated by combining the results from the current iteration and previous iterations. Since no proof of convergence is available at this moment, the method is heuristic for the DTA problem. The algorithm is presented as follows (**Policy-Based Stochastic DTA Heuristic**):

**Step 0** (Initialization)
0.1: $N$ = maximal number of iterations; MSA counter $i = 1$
0.2: $C^r_{(0)}$ = free flow link travel times, $r = 1, ..., R$; Policy choice set $G_{(0)} = \{paths\}$; Policy splits $f_{(0)} = 0$
**Step 1** (Main Loop)
1.1: Generate an optimal routing policy $\mu_i = O\left(\tilde{C}_{(i-1)}\right)$; Choice set update $G_{(i)} = G_{(i-1)} \cup \{\mu_i\}$
1.2: Users' choice model $f' = U\left(G_{(i)}, \tilde{C}_{(i-1)}\right)$; MSA update $f_{(i)} = (1-\alpha)f_{(i-1)} + \alpha f'$, where $\alpha = 1/i$
1.3: Loader $\tilde{C}_{(i)} = L\left(f_{(i)}, \tilde{D}, \tilde{S}\right)$
**Step 2** (Stopping Criterion) If $i = N$, STOP; Otherwise, $i = i + 1$, and go to Step 1

A reasonable value for the maximum number of iterations will be obtained by running the heuristic for a suf-

ficiently large number of iterations and observing the convergence property. Experimental results on this topic will be presented in the next chapter. In Step 0.4, we initialize the policy choice set to include all paths that would have been included in a choice set for a path-based DTA model. Note that for each OD pair and each departure time, there is a choice set, and the initialization is done for all choice sets. The subscripts for OD pair and departure time are omitted to avoid heavy notation. In Step 0.5, we initialize policy splits to be zeros for all OD pairs and departure times. These are infeasible policy splits, and the initialization is just for the convenience of writing a formula in Step 1.4. $f_{(0)}$ is not taken into account in the MSA update, as when $i = 1$, its coefficient is zero.

# 3 Computational Tests

## 3.1 Comparison of Four Models

The motivation for the policy-based DTA model is to be able to model users' adaptive choices and analyze the effects of online information in a truly stochastic network. We develop four models for comparison purposes as shown in Table 1.

| | Base | Path | Online Path | Policy |
|---|---|---|---|---|
| Distributions of demand/supply | No | Yes | Yes | Yes |
| Online information | No | No | Yes | Yes |
| Optimal online choice | No | No | No | Yes |

Table 1: Four Equilibrium Models

In each column, we have one equilibrium model: base model, path model, online path model, and policy model, respectively. We have three features listed: distributions of demand/supply, online information, and optimal online choice, which specify respectively whether distributions of random demand/supply are considered, whether online information is utilized in routing decision making, and whether online information is utilized optimally by applying the optimal routing policy algorithm developed in Gao and Chabini [1]. We elaborate on the models one by one.

The first model is the base equilibrium model. It is an assignment model in a deterministic network with deterministic demand. It ignores stochastic disturbances in supply, e.g. assumes no incidents at all in a network. On the other hand, the demand is set at its expected value, if any stochasticity in demand exists. This corresponds to the case where users have no idea about the incident at all and just follow their habitual paths in a normal network. After the equilibrium path flows are obtained, they are loaded onto the true network with stochastic demand and supply, and the resulting measures of effectiveness are calculated.

The second model is the path based model with equilibrium in distribution. In this model, the distributions of both demand and supply are known and are used in the assignment. We seek equilibrium in the distribution of link travel times. Users are assumed to take paths with minimum expected travel time. We emphasize that a path is a fixed set of concatenated links. If a user follows a path, then s/he will traverse this set of links one by one, regardless of any online information. Note that online information includes information at the origin node and it should not be restricted to information collected *en route*. Basically it is any information beyond the *a priori* knowledge about the distribution of link travel times.

The third model is the online path based model with equilibrium in distribution. It makes use of online information as compared to the previous model. With the equilibrium link travel time distribution, a user makes routing

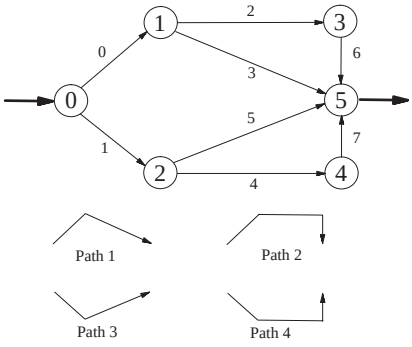



Figure 2: Test Network

| | Link 0(1) | Link 2(4) | Link 3(5) | Link 6(7) |
|---|---|---|---|---|
| Length (mi) | 0.5357 | 0.7576 | 0.8470 | 0.3788 |
| # of Lanes | 2 | 1 | 1 | 1 |
| Free Flow Speed (mph) | 40 | 30 | 20 | 30 |
| Free Flow Time (sec) | 48 | 91 | 152 | 45 |
| Jam Density (veh/link/meter) | 0.30 | 0.15 | 0.15 | 0.15 |
| Output Capacity (veh/link/sec) | 1.1 | 0.5 | 0.5 | 0.5 |

Table 2: Link Data of the Test Network

decisions as follows. For any given state, i.e. node, time and current information, the conditional link travel time distribution is obtained and a path with minimum expected travel time is sought. The user then takes the first link of the path. When s/he arrives at the next node with an arrival time and updated online information, a new conditional distribution is obtained with a new minimum expected time path. The user continues on the first link of this new path. The above steps are repeated until the destination is reached. We remark that the outcome of the process is also a routing policy, in the sense that it is a mapping from any state to a next node. It is just that the routing policy is not generated optimally, as each decision is made assuming that no further information will be available. We term a routing policy generated in the above stated process as "online path". On the other hand, since the information is updated quite often (at the same pace as in an optimal routing policy algorithm), the online path could be a good approximation to an optimal routing policy.

The last model is the optimal routing policy based model with equilibrium in distribution. It is different from the online path model, in the sense that it makes optimal use of online information. We note that the routing decisions in both models are link based, in the sense that only a next link is chosen at each decision node. However, the attractiveness or utility of a link is evaluated differently in these two models. In the online path choice model, the utility of a link is based on only one path; while in the optimal routing policy model, the utility of a link is based on a set of paths that have this link in common. Intuitively the second approach should lead to better decisions.

By comparing results of the base model and the path model, we can study the value of *a priori* information on stochasticity of demand/supply, as the base model ignores the stochasticity of demand/supply while the path model makes use of the *a priori* knowledge on distributions of demand/supply. By comparing results of the path model and the last two models (online path model and policy model), we can study the value of online information. Finally, by comparing results of the online path model and policy model, we can study the value of making optimal use of online information.

## 3.2 Experimental Design

### 3.2.1 The Test Network

We conduct computational tests on the simple hypothetical network shown in Figure 2. The network has 6 nodes and 8 directed links. It is symmetric with respect to the horizontal line passing through nodes 0 and 5. The link data is summarized and shown under the network.

We deal with one OD pair between node 0 and node 5. We assume zero flows between any other OD pair. Four paths exist for OD pair (0,5) as shown in Figure 2, with online diversion possibilities at nodes 0, 1 and 2. The study period is from 6:30am to 8:00am. The time resolution is 1 minute for the optimal routing policy algorithm and users' behavior model. The loader works at a finer resolution (5 sec) for the simulation, but the post-processed link (path) travel times are also by minute. Therefore we have 90 time periods in the tests.

### 3.2.2 Random Incidents

We have random incidents in the network. An incident is defined by the segment ID, start time, duration and capacity reduction factor. A segment is part of a link, and a link can be composed of one or multiple segments. In our network, each link is composed of only one segment. If an incident starts from 8:00am and lasts for 20 minutes with a capacity reduction factor 0.5 on link 0, then the output capacity of link 0 will be $0.5 \times 1.1 = 0.55$ veh/link/sec from 8:00am to 8:20am, and will revert to the original value 1.1 veh/link/sec from 8:20am on. As the capacity reduction is with respect to output capacity, an incident could only happen at the end of a link.

The random incident is defined as follows: 1. There is at most one incident during the study period for any given day; 2. The incident has a positive probability of occurring on link 0, 2, 3 and 6, but zero on links 1, 4, 5 and 7; 3. The probability of incident occurrence on a link is proportional to the link's length (for links 0, 2, 3 and 6); 4. If an incident occurs on a link, the start time can be 6:30am, 6:40am, 6:50am, ..., 7:50am with equal probability; 5. The duration of any incident is fixed at 10min, and the capacity reduction factor is fixed at 0.3; 6. The probability of no incident in the network is $1 - p$.

Based on the above description, the random incident can be described by the joint distribution of link ID $l$ and start time $t_0$. Denote $l_0, l_2, l_3, l_6$ as the length of link 0, 2, 3 and 6 respectively and $L = \sum_{i=0,2,3,6} l_i$.

$$(l, t_0) = \begin{cases} \text{(0, 6:30 or 6:40 ... or 7:50)}, & w.p.\ p \times l_0/L/9 \\ \text{(2, 6:30 or 6:40 ... or 7:50)}, & w.p.\ p \times l_2/L/9 \\ \text{(3, 6:30 or 6:40 ... or 7:50)}, & w.p.\ p \times l_3/L/9 \\ \text{(6, 6:30 or 6:40 ... or 7:50)}, & w.p.\ p \times l_6/L/9 \\ \text{(non-exist, non-exist)}, & w.p.\ 1 - p \end{cases}$$

### 3.2.3 Demand

We assume that the demand for OD pair (0, 5) is always deterministic. The flow rate is 2880 veh/hour between 6:30am and 7:00am, and 4680 veh/hour between 7:00am and 8:00am.

Users are assumed to minimize expected travel time with perception errors. The coefficient of expected travel time is negative with a large enough absolute value (-6.0) to approximate a fastest policy (path) choice situation. All users have perfect online information in the online path model and policy model, i.e. knowledge of travel time realizations on all links up to the current time. Obviously, users have no online information in the base model and the path model.

## 3.3 Results

We discuss the solutions of the four models and compare them when appropriate. For the sake of brevity, not all results are presented. Note that we focus on the statistics collection period 7:00am through 7:30am, although statistics for all time intervals are presented. Special caution should be taken when reading statistics close to 8:00am, as there are unfinished trips during that period and the calculation of travel times could be mistaken.

We present the equilibrium OD travel time distribution of the path model (dashed lines) and the online path model (solid lines) as a function of departure time for all 37 support points in Figure 3. Each plot in the figure is of the 37 discrete supporting points for the distribution of OD travel times. The x-axis represents departure time, while the y-axis represents OD travel time. Recall that in support points 1 through 9, the incident is on link 0 and with 9 different start times from 6:30am to 7:50am. Then in support points 10 through 18, the incident is on link 2; in support points 18 through 27, on link 3; and in support points 28 through 36, on link 6; and finally in support point 37, there is no incident in the network. The incident link ID and incident start time are listed on the top of each graph in the figure. Generally, the online path model gives lower OD travel time, and the savings are

quite outstanding in some cases (e.g. when incidents are on link 4 and start from 7:00 and 7:10). This is largely due to the flexibility gained through adaptive routing. Figure 4 gives time-dependent path flow distributions for path 2, and we can see flows on path 2 change from different support points, while in a path model, path flows are fixed across support points. We omit the presentation of other path flow distributions for the sake of brievity. When an incident happens, affected links will have longer travel times. Furthermore, at different stages of an incident, the realized link travel times so far are different. For example, if we are at a point when an incident just begins, then link travel times along the time axis would be flat at normal values and then jump to higher values. If we are at a point when an incident just ends, then we would see a longer period during which link travel times are at high values. If we are at a point when an incident has ended for a while, then we would be able to see link travel times first increasing and then decreasing. To sum up, the message contained in the current realized link travel times makes us adaptive to incidents. A very distinctive feature of the flows in Figure 4 is the decrease around incident across all support points. As we arrange the graphs by the start time of incident, we can see a moving "pit" in path flow. This is more intelligent than deterministic path flows as in path model. Note that policy flows are deterministic. As a policy will manifest itself as different paths in different incident support points, path flows are random and we can talk about their distributions.

The OD travel time distribution and path 2 flow distribution of the policy model are very simliar to those of the online path model, respectively, and thus are not presented here. In fact, these two models are both based on routing policies, and it is just that the methods of generating optimal routing policies are different. We expect that results from the two models are not significantly different in our simple test network, due to the limited diversion nodes. Further computational tests on larger networks are desirable to study the differences between these two models.

Next we compare expected OD travel times from all the four models in Figure 5. Expected OD travel time is the major measure of effectiveness in our tests. We observe that the path model gives lower expected OD travel times than the base model, and the two adaptive models (online path model and policy model) provide further travel time savings. Figure 6 gives the time-dependent OD time standard deviations. Although travelers are minimizing expected travel time only, their travel time variances are also reduced by taking adaptive routing choices. This is due to the fact that their travel times are reduced in incident scenarios, and thus more smooth across support points.

We just discussed in detail the results for a specific test setting (incident probability $p = 0.9$). We are also interested in learning the behavior of the models when we vary the incident probability. On the other hand, in reality online information could be provided only to part of the travelers, thus it is desirable to study how the traffic conditions change as a function of market penetration of online information. We define a single measure of effectiveness (MOE) to be compared in the sensitivity analysis, which is the expected OD travel time averaged over the statistics collection period: 7:00am through 7:29am.

First we carry out the sensitivity analysis with respect to incident probability $p$. We vary $p$ from 0 to 1.0 by a step size of 0.1. The result is plotted in Figure 7. For each of the models, the average expected OD travel time increases as incident probability increases, but different model has different increasing rate. This increasing function seems intuitively correct, as a more likely incident increases the probability that a network is congested, and thus a higher expected travel time. We note that the policy model gives a higher value for $p = 0.9$ (216.06) than for $p = 1.0$ (216.00). We believe that this difference is too small to be significant, and are inclined to believe that they are the same.

The relationship for the base model is linear. The explanation is as follows. First, the path flows are the same for various incident probabilities, since the base model does not consider incidents at all. Then the OD travel time for each incident support point is calculated, and a weighted average is taken to obtain the expected OD travel time, where the weight is the probability of an incident support point. As we can see from the design of incident distribution, incident probabilities are linear functions of $p$. Therefore the expected OD travel time is also linear function of $p$. While in other three models, random incidents are considered in the equilibrium process and equilibrium path (policy) flows differ when $p$ differs. Therefore the relationship is in general nonlinear.

In general, the path model gives less expected travel time than base model, and the two adaptive models (online

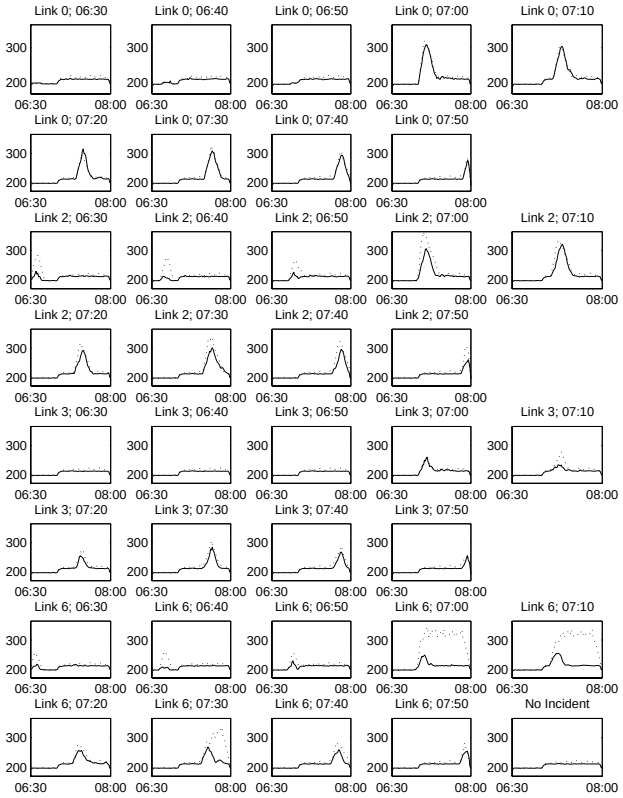



Figure 3: OD Travel Time Distribution of Online Path Model (X-Axis: Departure Time; Y-Axis: OD Travel Time (sec); $p = 0.9$)

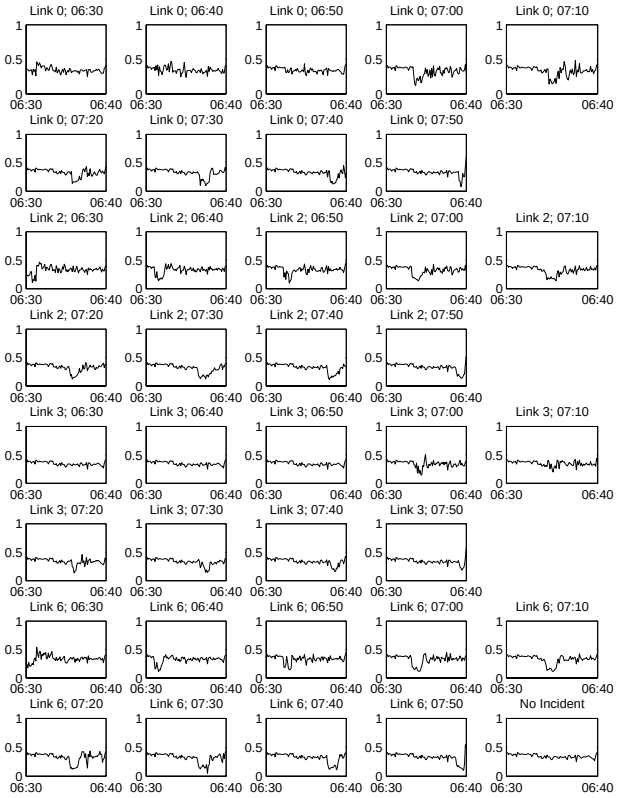



Figure 4: Path 2 Flow Distribution of Online Path Model(X-Axis: Departure Time; Y-Axis: Path Share; $p = 0.9$)

path model and policy model) give less expected travel time than the path model. The savings (path over base, and adaptive over path) increase as incident probability increases, both in absolute values and in relative percentage savings. The relative saving of the path model over the base model is in the range of $0 \sim 2.9\%$, and the relative saving of adaptive models over the path model is in the range of $0 \sim 4.4\%$. This increasing function suggests that values of both *a priori* and online information are more evident when traffic conditions are worse. This could be reasonable in reality when traffic conditions without incident are not too congested, as then there is enough room for diversion. This is actually the setting of our tests, as traffic is almost in free flow state with no incident. We expect that when a network is already quite congested without incident, this function might become flat after some point.

Next we carry out sensitivity analysis with respect to market penetration of online information. For a given penetration $k$ which is a value between 0 and 100%, we assign $k$ of the demand to take minimum expected travel time routing policies, while the remaining $1 - k$ of the demand to take minimum expected travel time paths. Equilibrium is sought by an MSA heuristic that updates the path splits and policy splits simultaneously. We have the result for $p = 0.1$ in Figure 8. The average expected OD travel time is at its largest value when market penetration of online information is zero. At that time, if one traveler is intelligent enough and take a routing policy rather than a path, he/she can save travel time. More and more of them find the benefits of online information, and they gain travel time savings and thus bring down the average expected travel time. However, in a congested traffic network, the changing of users' behavior changes the network-wide traffic conditions through interaction between supply and demand. As seen from the figure, the saving in travel time becomes less evident when penetration goes from 20% to 40% and from 40% to 60%. Later on, higher penetration actually does not bring any more savings. We see an increase in travel time from 60% to 100%. We then conclude that the savings gained from online information is larger when market penetration is lower. After some point, more online information could actually make things worse. Therefore the function of travel time saving against market penetration is not monotonic. Despite the varying effect of online information, travel time savings are always positive with online information, compared to no-online-information case. This analysis might only be valid for

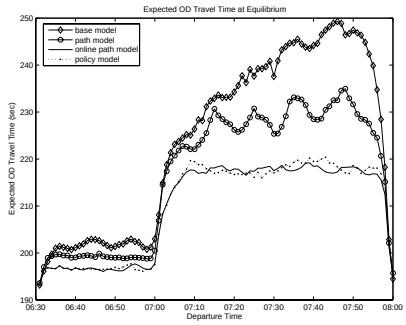



Figure 5: Expected OD Travel Time at Equilibrium of All Four Models ($p = 0.9$)

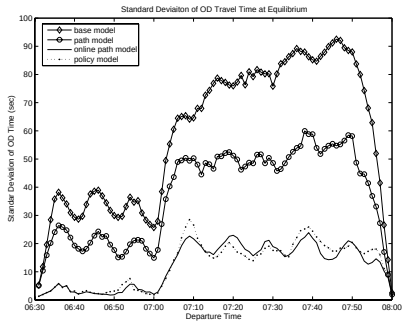



Figure 6: Standard Deviation of OD Travel Time at Equilibrium of All Four Models ($p = 0.9$)

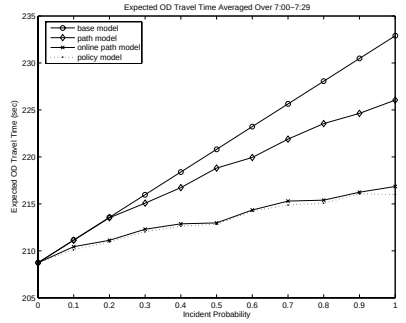



Figure 7: Average Expected OD Travel Times as Functions of Incident Probability

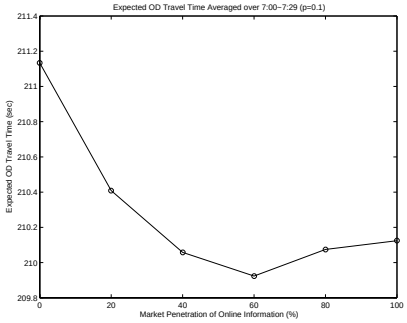



Figure 8: Average Expected OD Travel Times as Functions of Market Penetration ($p = 0.1$)

the test setting, and caution should be taken if one intends to generalize the result.

# 4 Conclusions

This paper establishes a policy-based dynamic traffic assignment model for the analysis of effects of online information in stochastic dynamic traffic networks. The distinctive feature of the proposed model is the ability to model travelers' adaptive routing choices based on online information. Computational tests are carried out in a hypothetical network, where random incidents are the source of stochasticity. System costs derived from four models with different information accessbility situations are compared. The adaptiveness to online information leads to less expected travel time and variance at equilibrium. The value of online information is an increasing function of the incident probability. Travel time savings are high when market penetrations are low. However, the function of travel time saving against market penetration is not monotonic.

# References

[1] S. Gao and I. Chabini. Optimal routing policy problems in stochastic time-dependent networks. *Transportation Research Part B*, 40(2):93–122, 2006.

[2] Y. Hamdouch, P. Marcotte, and S. Nguyen. A strategic model for dynamic traffic assignment. *Networks and Spatial Economics*, 4:291–315, 2004.

[3] P. Marcotte, S. Nguyen, and A. Schoeb. A strategic flow model of traffic assignment in static capacitated networks. *Operations Research*, 52(2):191–212, 2004.

# STABILITY OF NETWORK FLOWS WITH BOUNDED RATIONAL ROUTE CHOICE

*Shoichiro Nakayama*
*Department of Civil Engineering, Kanazawa University, Japan, snakayama@t.kanazawa-u.ac.jp*

## Abstract
In this study, it is assumed that a traveler is bounded rational, and chooses a route based on the satisficing principal, and also that the traveler is willing to travel the faster route and that there is a tendency for part of a flow to switch to a route with shorter travel time. On that basis, a model of the day-to-day route adjustment process (or mode choice) is formulated as a difference equation system. The model is applied to a simple network, in which on origin-destination pair is connected by a road and by mass transit. The difference equation system in this study is described as a discretized model of the projected dynamical system presented by Zhang & Nagurney (Trans. Res., 30B, 1996) for a simple network. Analysis is carried out to investigate the stability/instability of network flow and to determine the conditions for stability/instability. Further, the question of whether or not chaos can exist in network flow is considered.

## 1. Introduction
Network equilibrium models have become the main thrust of advances in the field of traffic network analysis. Following Wardrop's (1952) proposal of the network equilibrium model, Beckman et al. (1956) formulated the optimization problem for network equilibrium. Daganzo & Sheffi (1977) modeled a user stochastic equilibrium model. These days, dynamic network equilibrium has become the focus of much research. Thus, network equilibrium is being further and further elaborated. However, a couple of fundamental issues of network equilibrium remain in dispute. Is equilibrium really reached? What are the required conditions for equilibrium? Aren't the assumptions for equilibrium — that travelers are rational and have perfect knowledge — too idealistic? In this study, light is shed on the twin problems of whether or not network equilibrium is reached and what conditions are required for stability of network flow. Possible methods of examining these problems are to analyze the stability of a determined equilibrium or to investigate the day-to-day dynamics of actual network flow. From the viewpoint of traffic management, it is particularly important to appreciate the conditions under which network flow is stable or unstable.

Quite a number of researchers have studied the stability or day-to-day dynamics of network flow. A review of their findings is given in the next section, but roughly speaking, the majority of such studies imply that network flow may fail to converge to equilibrium under certain conditions. These findings indicate that it is essential to examine network flow instability and that it is of great importance to do so. Studies of the mechanism by which networks become unstable and the properties of network flow fluctuations have not uncovered very much in this regard.

In general, traffic flow in a real network fluctuates. What causes these fluctuations? Are they caused only by exogenous variations, such as noisy origin-destination demand? As shown in studies of non-linear dynamics, fluctuations that appear to be probabilistic oscillations actually occur endogenously like chaos. Examining whether or not network flow can behave chaotically is one important point.

So far, there has been little theoretical research on chaos in network flows. Cantellare & Cascetta (1995) touched upon the chaos of network flow. Nevertheless, there is plenty of room for the study of chaotic network flow, including the problem of whether chaos exists. From the standpoint of forecasting travel times or traffic volumes, it is very important to know whether or not network flow is chaotic; if network flow is found to be chaotic, it is possible to forecast flows or travel times over the short term, but it is theoretically impossible over the long term. This is a direct result of sensitivity to initial conditions in chaotic systems. It is impossible to observe the infinitely small values of network flows without error. And even if the initial error is extremely small, it increases exponentially over time.

In this study, it is assumed that a traveler is bounded rational. The traveler is also assumed to be willing to travel the faster route, that is, there is a tendency for part of a flow to switch to a route with shorter travel time. A model for the process by which a route choice is adjusted day-to-day is formulated as a difference equation system. Then this model is applied to a simple network for an analysis of stability/instability of network flow and an investigation of stability conditions. Further, it is examined whether or not chaotic flow can exist and the conditions for chaotic oscillations in the network if it can exist.

## 2. Review

Models of the day-to-day dynamics or stability of network flow can be classified into two groups according to whether they model the system in continuous time or discrete time. The former are mainly differential equation systems, and most are formulated as difference equations or simulation models. The difference between use of continuous time and discrete time is important, as described below.

Zhang & Nagurney (1996) and Nagurney & Zhang (1997) formulated day-to-day dynamics (the route choice adjustment process) as a projected dynamical system, as first introduced by Dupuis & Nagurney (1993). The model developed by Zhang & Nagurney (1996) is the starting point for this discussion. The basic concept is very simple: if the route cost exceeds the equilibrium cost, the flow increases, whereas if the equilibrium cost exceeds the route cost, the flow decreases. The traffic flow on route $r$ changes at a rate proportional to the difference between the minimum travel cost $\lambda_\omega\{d(\boldsymbol{f})\}$ for the OD pair $\omega$ at equilibrium and the route cost $c_r(\boldsymbol{f})$, where $d(\boldsymbol{f})$ is the travel demand, $\boldsymbol{f}$ is the vector of route flows, $c_r(\boldsymbol{f})$ is the cost of route $r$, and route $r$ connects OD pair $\omega$. They formulated the adjustment process as:

$$\dot{f}_r = \begin{cases} \lambda_\omega(\boldsymbol{f}) - c_r(\boldsymbol{f}) & \text{if } f_r > 0 \\ \max\{0, \lambda_\omega(\boldsymbol{f}) - c_r(\boldsymbol{f})\} & \text{if } f_r = 0 \end{cases} \tag{1}$$

where $f_r$ is the flow on route $r$. They established global asymptotic stability according to Wardrop's equilibrium under the assumptions that the cost performance functions (or travel time functions) and inverse travel demand functions, $\lambda\{d(\boldsymbol{f})\}$, are continuous and strictly monotonic.

Smith (1984) studied the stability of Wardrop's equilibrium with fixed travel demand. He assumed that traffic flows switch from route $r$ to route $s$ at rate $f_r \max\{0, c_r(\boldsymbol{f}) - c_s(\boldsymbol{f})\}$. By compiling output flows to other routes and input flows from other routes, the dynamics of the flow on route $r$ are represented as:

$$\dot{f}_r = -\sum_s f_r \max\{0, c_r(\boldsymbol{f}) - c_s(\boldsymbol{f})\} + \sum_s f_s \max\{0, c_s(\boldsymbol{f}) - c_r(\boldsymbol{f})\} \tag{2}$$

He showed that his dynamical system converges to Wardrop's equilibrium using the Lyapunov theorem (e.g. Hahn, 1963) if the travel cost function is continuously differentiable and monotonic and there are no explicit capacity restrictions.

Friesz et al. (1994) described the route choice adjustment process model in continuous time as follows:

$$\begin{aligned} \dot{u}_\omega &= \rho_\omega\left[\max\left\{0, u_\omega + \zeta\left(d_\omega(\mathbf{u}) - \sum_r f_r\right)\right\} - u_\omega\right] \\ \dot{f}_r &= \sigma_r\left[\max\left\{0, f_r - \xi\left(c_r(\mathbf{f}) - u_\omega\right)\right\} - f_r\right] \end{aligned} \tag{3}$$

where $d_\omega(\mathbf{u})$ is travel demand for the OD pair $\omega$ and $\zeta$, $\xi$, $\rho_\omega$, and $\sigma_r$ are positive (constant) parameters.

In these three studies, the authors presupposed that the traveler is rational and willing to take a route with lower cost (or shorter travel time). They formulated the adjustment process in continuous time. The models seem to be general and reasonable, so some believe that network flow will converge to the Wardrop equilibrium if sufficient time elapses under the same traffic conditions. However, the day-to-day dynamics of a traffic network in fact constitute a discrete time system due to repeated daily trips. Certain problems might arise if a discrete time system is formulated as a continuous time system. An illustration of this is the famous paradox called "Achilles and the tortoise": imagine a race between the swift Achilles and a slow tortoise where Achilles is pursuing the tortoise.

Achilles gives the tortoise a head start. The Greek philosopher Zeno argued that Achilles can never overtake the tortoise, because he must reach a point that the tortoise has already passed, whereas the distance between Achilles and the tortoise will always be divisible. Logically the tortoise will always be ahead, because no point can be reached before the previous point has been reached. The effect is that there can be no motion at all, and Achilles, on these assumptions, can never overtake the tortoise. Thus, in some cases, a discrete time model and a continuous time model for the same system give a completely different picture.

Most studies in discrete time show that the system may fail to converge to a (fixed-point) equilibrium, while the continuous time studies described above guarantee stability and equilibrium. In many models in discrete time, a traveler assumed to take the route that has the minimum perceived travel time, and this is generally calculated as weighted average of experienced travel times (Horowitz, 1984; Canterella & Cascetta, 1995; Emmerink et al., 1995; Nakayama et al., 1999; Watling, 1999).

There is one discrete time model that guarantees convergence to equilibrium. Kobayashi (1994) formulated travelers' learning based on Bayes' theorem, and showed that the system will converge to an equilibrium state that Kobayashi (1994) and Kobayashi & Tatano (1996) call the rational expectation equilibrium. The term "rational expectation" comes from economics (Muth, 1965; Sheffrin, 1996), and at rational expectation equilibrium the distribution of actual travel times in the network is identical to the distribution of travelers' expected (or recognized) travel times.

Models by simulation approach provides enriched findings relating to day-to-day dynamics, including an elaborate learning process resulting from the flexibility of the modeling. Nakayama & Kitamura (2000) assumed that travelers learn inductively, according to the theory of cognitive psychology (e.g. Holland et al., 1986); that is, they inductively learn how to choose routes based on experience. Their model system is basically a production system (Newell & Simon, 1972), which is a compilation of if-then rules that are revised by applying a genetic algorithm (Goldberg, 1989; Holland, 1975). It is found that the average network flow converges very close to the Wardrop equilibrium (the variance is not zero, but is small enough) in a case where the traveler has a high capacity for information processing. On the contrary, in a case where the ability of a traveler to process information is low, the variance of network flow is large and it cannot be said to converge to equilibrium. If the traveler rarely switches routes capriciously, network flow may converge to a deluded equilibrium (Nakayama et al., 1999) as described below, which is far from the Wardrop equilibrium or stochastic user equilibrium. Deluded equilibrium may develop into frozen equilibrium, where most or all travelers habitually choose the same route. Note that a habitual traveler (automatically) continues to take the same route without taking other possible routes into account.

Nakayama et al. (1999) indicated that travelers sometimes have excessively false perceptions (with systematic bias). Nakayama et al. call this phenomenon "delusion", and deluded travelers form a stable condition known as "deluded equilibrium". The point is that travelers with false understanding can form a stable system, and there is a possibility that a condition such as deluded equilibrium exists.

## 3. Formulation

In this study, the very simple network shown in Fig. 1 is investigated. The network has one pair Origin-Destination (OD) pair, connected by a road link and a public transit route. The service level of the public transit system is assumed to be independent of the number of users, and travel time is assumed constant. An implicit presupposition is that travel demand is elastic, and the study concentrates wholly on cases that are departure-time independent.

Let $x$ denote the traffic volume on the road link ($x \geq 0$), $t_c$ or $t_c(x)$ the travel time over the road link, and $t_t$ the travel time by mass transit. Let us assume that $t_c(x)$ is a strictly monotonically increasing function of $x$. As described above, $t_t$ is constant. At equilibrium, the following equalities and inequalities hold:

$$t_c \begin{cases} = t_t & if\ x > 0 \\ \geq t_t & if\ x = 0 \end{cases} \tag{4}$$

where $x \geq 0$. Clearly, the equilibrium is unique. When $t_c(0) > t_t$, there is no flow on the road at equilibrium. This

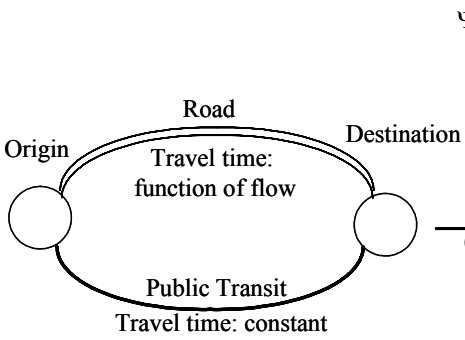



**Fig. 1 The network in this study**

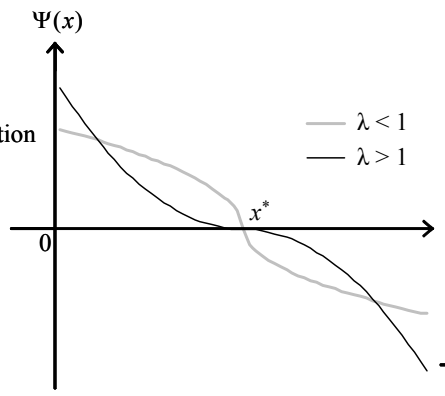



**Fig. 2 The function $\Psi(x)$**

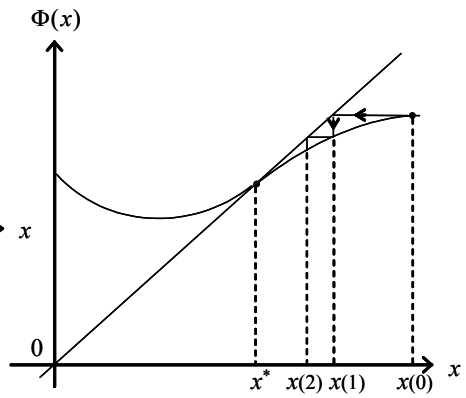



**Fig. 3 The function $\Phi(x)$**

state with no road traffic is of no interest. Hereafter, we investigate the system only for $t_c(0) < t_t$.

The concept behind the flow dynamics (or the adjustment process) is that flow along the road link increases if road travel time is less than that by public transit, while it decreases if the travel time exceeds that by public transit. This may be seen as natural, and is similar to the projected dynamical system. These dynamics can be expressed by the following difference equation:

$$x(i+1) - x(i) = \Psi[x(i)] \tag{5}$$

$$\Psi(x) = \begin{cases} \max\left\{-\eta\left[t_c(x) - t_t\right]^{\lambda}, -x\right\} & \text{if } t_c(x) - t_t > 0 \\ \eta\left[-t_c(x) + t_t\right]^{\lambda} & \text{if } t_c(x) - t_t \le 0 \end{cases} \tag{6}$$

where $x(i)$ is the flow on day $i$, and $\eta$ and $\lambda$ are positive parameters. The former parameter, $\eta$, represents the sensitivity (or speed) of the adjustment. If the latter parameter, $\lambda$, is less than 1.0, the adjustment function, $\Psi$, is degressive in respect of the difference of the travel times. Namely, travelers react more sensitively when the travel time difference is small than when it is large in the case of the same travel time difference. If $\lambda$ is more than 1.0, travelers react more sensitively when the difference is large. Fig. 2 illustrates the function, $\Psi(x)$.

For illustration in the next section, let us define the function $\Phi(x)$ as follows:

$$\Phi(x) = x + \Psi(x) = \begin{cases} x + \max\left\{-\eta\left[t_c(x) - t_t\right]^{\lambda}, -x\right\} & \text{if } t_c(x) - t_t > 0 \\ x + \eta\left[-t_c(x) + t_t\right]^{\lambda} & \text{if } t_c(x) - t_t \le 0 \end{cases} \tag{7}$$

The dynamics can then be rewritten as the following equation using $\Phi(x(i))$:

$$x(i+1) = \Phi[x(i)] \tag{8}$$

Fig. 3 is an example of equation (7). Here, the diagonal line is $x = \Phi(x)$ and $x^*$ is the equilibrium point. To reveal the orbit of $x(0)$ graphically, first locate $x(0)$ on the horizontal axis. The horizontal line through $(x(0), \Phi[x(0)])$ crosses diagonal line $x = \Phi(x)$ at point $(\Phi[x(0)], \Phi[x(0)]) = (x(1), x(1))$. By applying the same process with $x(1)$ in place of $x(0)$, we obtain point $(x(2), x(2))$. Repeating the same process, the orbit of $x(0)$ is produced.

The above system of difference equations appears to be a discretized version of the projected dynamical system by Zhang & Nagurney (1996) for the network shown as Fig. 1. Note that this difference equation system is elaborated further than Zhang & Nagurney (1996) in some respects. The parameters $\eta$ and $\lambda$ are newly introduced. Parameter $\eta$ was left as a future task by Zhang & Nagurney (1996).

Here let's consider behavioral principal of Eq. (5) and (6). Assume that the traveler has a threshold for mode choice (or route choice). Let $\theta_j^i$ denote the threshold of the $j$th traveler. If the road travel time is greater than $\theta_j^i$, that is, $t_c > \theta_j^i$, the $j$th traveler switch from the road to the public transit. In the case that the number of travelers is continuous and that the threshold is uniformly distributed, the dynamics is expressed as Eq. (5) and (6) with $\lambda = 1$. Fig. 4

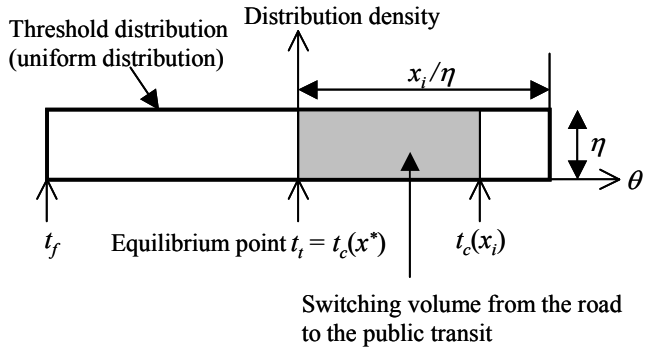



**Fig. 4. The threshold and switching volume**

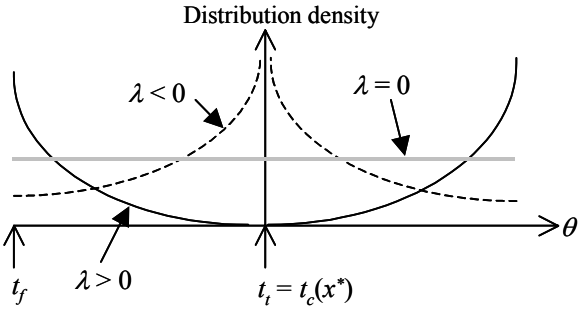



**Fig. 5. The threshold distribution**

illustrates this in the case of $t_c > t_t$. The gray part in the figure means $t_c > \theta_f^i$, and the area of the gray part switch from the road to the public transit. The maximum of switching volume is $x(i)$, and in this case, the volume of the road is 0. The distribution of the threshold depends on the parameter $\lambda$ as Fig. 5 shows.

The threshold can be interpreted as satisfaction principal proposed by Simon (1957). Simon (1947) proposed the concept of bounded rationality. He also suggested that people with bounded rationality adopt satisficing principal rather than optimizing one. Even if they do not choose the optimum choice, they do not switch their choices in the case that they are satisfied with their choices. They are probably satisfied with the choices whose difference from the optimum one is not so large. Simon called this satisficing principal. In this study the threshold can be interpreted as the satisficing domain. Therefore, the macroscopic dynamics of Eq. (5) and (6) is a model of the mode choice adjustment with bounded rationality.

## 4. Stability/Instability

Let $\Phi'(x)$ denote the differential of $\Phi(x)$, namely $d\Phi/dx$, and $\Phi^n(x)$ the $n$th iterate of $x$ for $\Phi$. The derivative of $\Phi(x)$ at equilibrium, $\Phi'(x^*)$, can be classified into three cases for the analysis of stability/instability:

$$\Phi'(x^*) = \begin{cases} +\infty & if\ 0 < \lambda < 1 \\ 1 - \eta \cdot \lambda \cdot t_c'(x^*) & if\ \lambda = 1 \\ 1 & if\ \lambda > 1 \end{cases} \tag{9}$$

When $0 < \lambda < 1$, $|\Phi'(x^*)|$ is greater than 1 and the (unique) equilibrium is not locally stable. Naturally, then, the equilibrium is not globally stable. When $\lambda > 1$, the equilibrium is locally stable, but is not asymptotically stable. When $\lambda = 1$, in the case that $\eta$ is sufficiently large, it is clear that the equilibrium is not (locally) stable because $|\Phi'(x^*)| > 1$. At that time, the step size, $x(i+1) - x(i)$, is very large. In the case that $\eta$ is sufficiently small but not 0, the equilibrium is locally stable because $\eta$, $\lambda$, and $t_c'(x^*) > 0$ ($\eta$ is sufficiently small). In this case, the equilibrium might also be globally stable because the step size, $x(i+1) - x(i)$, is so small that $x$ approaches the equilibrium gradually. On the other hand, when $0 < \lambda < 1$, the equilibrium is not stable and the flow does not converge even if $\eta$ is small. This could imply that network flow does not converge (to the equilibrium) and is not stable when travelers react more sensitively to small cost differences than to large.

In this study, the BPR-type function given below is generally adopted as the travel time function of the road, $t_c(x)$:

$$t_c(x) = t_f \left\{ 1 + \alpha \left( \frac{x}{C} \right)^{\beta} \right\} \tag{10}$$

where $t_f$ is the free-flow travel time of the road, $C$ is the road's capacity, and $\alpha$ and $\beta$ are positive parameters. The standard values of parameters $\alpha$ and $\beta$ are 0.15 and 4.0, respectively. If the travel time function follows this BPR function, the traffic volume on the road at equilibrium, $x^*$, is:

$$x^* = C\left[\frac{1}{\alpha}\left(\frac{t_t}{t_f} - 1\right)\right]^{\frac{1}{\beta}} \tag{11}$$

because $t_t > t_f$ as described above.

Let us examine stability/instability for the case where the BPR travel time function with $\lambda = 1$. $\Phi'(x^*)\big|_{\lambda=1}$ is expressed as:

$$\Phi'(x^*)\Big|_{\lambda=1} = 1 - \frac{\alpha^{1/\beta} \cdot \beta \cdot \eta \cdot t_f}{C}\left(\frac{t_t}{t_f} - 1\right)^{\frac{\beta-1}{\beta}} \tag{12}$$

In the case of the standard values of $\alpha$ and $\beta$, $\Phi'(x^*)\big|_{\lambda=1}$ is approximately $1 - 2.489\ \eta\ t_f(t_t/t_f - 1)^{3/4}/C$. If $-1 \le 1 - 2.489\ \eta\ t_f(t_t/t_f - 1)^{3/4}/C \le 1$, $\Phi(x^*)\big|_{\lambda=1}$ is stable. That is, the condition for stability of $\Phi$ with $\lambda = 1$ in the standard BRP function can be expressed as:

$$\eta \le \frac{0.803C}{\left[t_f\left(t_t - t_f\right)^3\right]^{\frac{1}{4}}} \tag{13}$$

## 5. Day-to-Day Dynamics and Chaos

In this section, the dynamics and the presence of chaos in the dynamical system is examined. To illustrate the definition of chaos, let us introduce the Li-Yorke theorem (1975).

**Theorem (Li-Yorke).** *Let $J$ be an interval and let $F$: $J \to J$ be continuous. Assume there is a point $a \in J$ for which the points $b = F(a)$, $c = F^2(a)$, $d = F^3(a)$, satisfy*

$$d \le a < b < c \quad (\text{or } d \ge a > b > c)$$

*Then*

T1: *for every $k = 1, 2, \ldots$ there is a periodic point in $J$ having period $k$.*

T2: *there is an uncountable set $S \subset J$ (containing no period points), which satisfies the following conditions*:

(A) *For every $p, q \in S$ with $p \ne q$,*

$$\limsup_{n\to\infty}\left|F^n(p) - F^n(q)\right| > 0$$

*and*

$$\liminf_{n\to\infty}\left|F^n(p) - F^n(q)\right| = 0$$

(b) *For every $p \in S$ and periodic point $q \in J$,*

$$\limsup_{n\to\infty}\left|F^n(p) - F^n(q)\right| > 0\,.$$

If conditions T1 and T2 are satisfied, $F$ is said to be chaotic in the sense of Li-Yorke. Condition T1 means that $F$ can have infinitely many periodic points and condition T2 represents sensitivity to initial conditions.

Let us use an example to illustrate chaotic behavior in network flow. The settings for this example are as follows. The travel time by mass transit, $t_t$, is 30 (minutes). The travel time by road follows the BPR function. Regarding the performance of the road, its capacity, $C$, is 1,000 (pcu/hour), the free-flow travel time, $t_f$, is 20 (min.), and parameters $\alpha$ and $\beta$ are standard; that is, $\alpha = 0.15$ and $\beta = 4$. The flow along the road at equilibrium is approximately 1,351.2. Hereafter, we examine the flow with $\lambda = 1$ as an example because it is the most fundamental.

As described in equation (9), the flow is locally unstable if approximately $\left|1 - \eta \cdot t_c'(x^*)\right| = \left|1 - 0.0296\eta\right| > 1$, namely $\eta > 67.56$. This implies that the flow oscillates when $\eta > 67.56$ under the above settings. What kind of oscillation is it? Let us investigate the condition for chaos in the flow dynamics. Set $a$ at 0. $b = \Phi_\eta(a) = 10\,\eta$. $c = \Phi_\eta(b) = \max\{0,\ 20\,\eta - 3\times10^{-8}\,\eta^5\}$. $d = \Phi_\eta(c) = \max\{0,\ 30\,\eta - 3\times10^{-8}\,\eta^5 - 3\times10^{-12}\,\eta\,(20\,\eta - 3\times10^{-8}\,\eta^5)^4\}$. Fig. 4 shows $b$, $c$, and $d$, which are the functions of $\eta$. As stated above, a system is chaotic in the sense of Li-Yorke when $d \le a <$

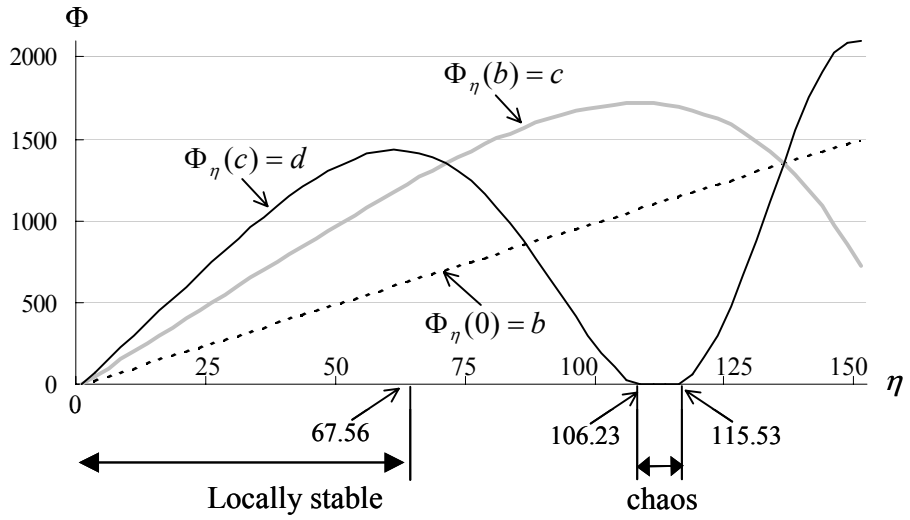



**Fig. 6 The chaotic domain in the sense of Li-Yorke**

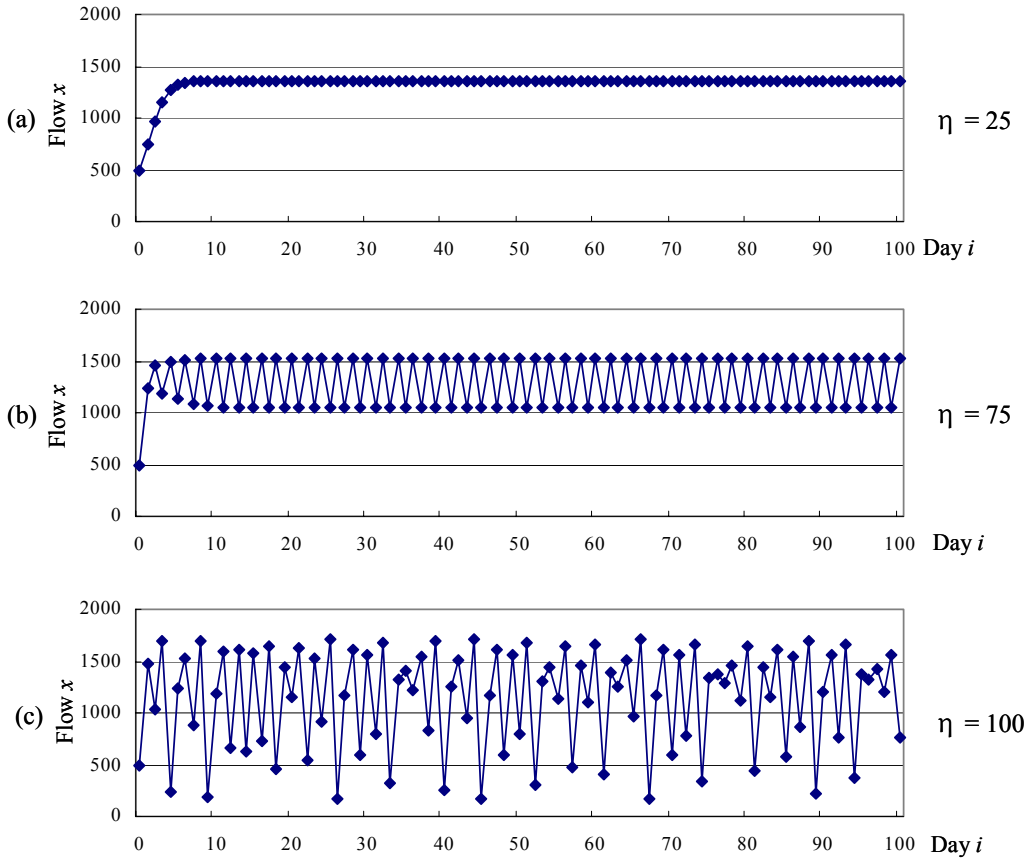



**Fig. 7 (a) (b) (c) The flow dynamics**

$b < c$. From Fig. 6, when $\eta$ is in the range from 106.23 to 115.53, $d \leq a < b < c$ is satisfied. Therefore, the flow is chaotic in the sense of Li-Yorke when $\eta$ is at least in the range from 106.23 to 115.53. Note that this condition is not necessary and sufficient, but it is sufficient.

Thus, Φ is chaotic in the sense of Li-Yorke under the above conditions. Since parameter $\eta$ represents the reactions of travelers, especially, the sensitivity (or speed) of the reaction, the implication is that network flow may be chaotic under certain conditions of traveler reaction and behavior. In other words, we cannot completely exclude the possibility of chaotic network flow. A future task is to examine whether or not there exists real-world network flow that is chaotic.

Figures 7(a) through 7(c) illustrate the dynamics of road flow, $x(i)$, until day 100, when $\eta$ is 25, 75, and 100, respectively. The horizontal axis and the vertical axis represent day, $i$, and flow, $x$, respectively, and the initial value of $x$, $x(0)$, is 500 in each figure. In Fig. 7(a), the flow converges to equilibrium. In Fig. 7(b), the flow oscillates with a 2-period cycle. The flow behavior is very complex in Fig. 7(c).

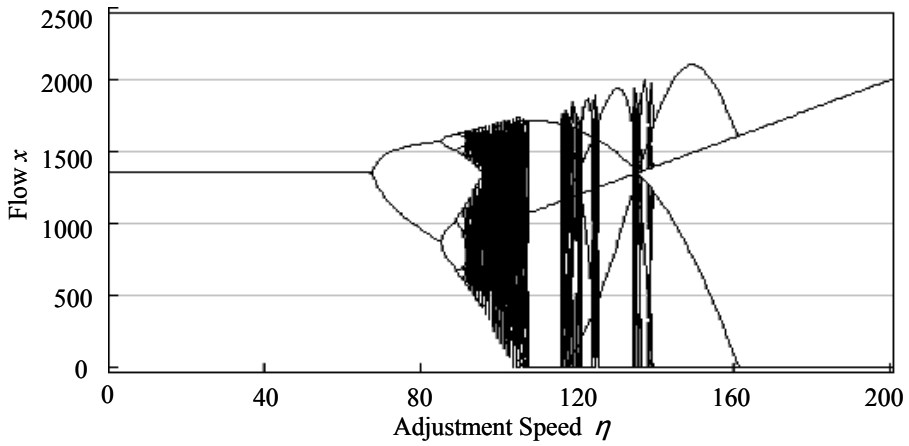



**Fig. 8 The bifurcation diagram on $\eta$**

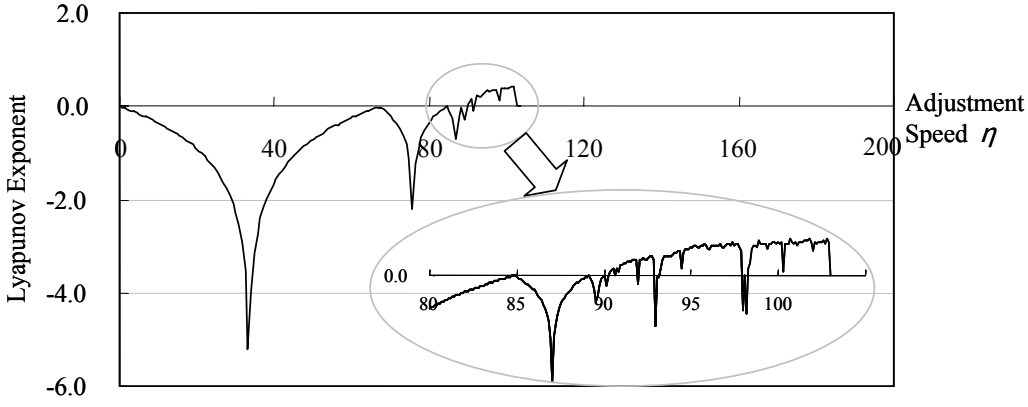



**Fig. 9 The Lyapunov exponent of Φ**

Fig. 8 presents a bifurcation diagram for $\eta$. The diagram is a schematic representation of behavior of the system as it varies with the value of $\eta$. Fig. 8 is obtained by computer, with each $x(i)$ ($i$ = 5001, 5002,…, 5250) plotted in the range $0 < \eta < 200$. For $0 \le \eta \le 67.5$, the line represents the point $(\eta, x^*)$ since $x^*$ attracts the iterates of all $x$. When $67.6 \le \eta \le 85.1$, there are two curves, and the iterates of all $x$ that are not eventually fixed are attracted to the 2-period cycle. We find that the bifurcation point is located near 67.5. This coincides with the domain of local stability as mentioned in the previous section. As described in equation (9), the road flow is locally unstable if approximately $\left|1-\eta\cdot t_c'(x^*)\right| = \left|1-0.0296\eta\right| > 1$, namely $\eta > 67.56$. The two curves bifurcate, and the iterates of $x$ are attracted to the 4-period cycle. It seems that the various attraction cycles with $2^k$-periods appear in sequence ($k$ = 0, 1, 2,…).

In the introduction, it was pointed out that sensitivity to initial conditions is important in forecasting travel time or flow. A quantitative measure of sensitivity to initial conditions is the Lyapunov exponent. This is the averaged rate of divergence (or convergence) of two neighboring trajectories. According to one interpretation, behavior is defined as chaotic when the (maximum) Lyapunov exponent is more than 0. The Lyapunov exponent, $L$, of function $F$ is defined as:

$$L = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \ln\left|F'(x(i))\right| \tag{14}$$

If the Lyapunov exponent is greater than 0, the orbit is unstable and chaotic. The orbit is attracted to a stable fixed point or a stable periodic orbit if the exponent is less than 0. The orbit is a neutral fixed point (or an eventually fixed point) if the exponent is exactly 0. Fig. 9 presents the Lyapunov exponent of the above system Φ. The settings for calculation of the Lyapunov exponent are that $N$ (in equation (14)) is 500 and $x(0)$ is 500. In Fig. 7, the Lyapunov exponent for $\eta$ between approximately 103 and 200 is not depicted because the differential of function Φ at 0 cannot be defined. When $\eta$ is in the range from 91 to 103, the behavior is generally chaotic in view of sensitivity to initial conditions with some exceptions. Therefore, the behavior in Fig. 5(c) is chaotic.

## 6. Conclusions

In this study, it is assumed that a traveler is willing to switch to a faster route, and a model of the day-to-day route adjustment process (or mode choice) is formulated as a difference equation system. The model is applied to a simple network, in which on origin-destination pair is connected by a road and by mass transit. The difference equation system in this study is described as a discretized model of the projected dynamical system presented by Zhang & Nagurney (1996) for a simple network. Analysis is carried out to investigate the stability/instability of network flow and to determine the conditions for stability/instability. Further, the question of whether or not chaos can exist in network flow is considered.

The results of the study can be summarized as follows: 1) the condition for stability of network flow is that the traveler is not very sensitive to differences in travel time (or cost) between road and mass transit when the flow is close to equilibrium. In other words, the flow does not converge to equilibrium and is unstable when travelers react more sensitively when the cost difference is small than when it is large; 2) network flow becomes chaotic under certain conditions of traveler reaction, especially sensitivity (or speed) of reaction, in the example. These results are obtained for a simple network example, but they are still very interesting and highly suggestive. As Intelligent Transportation Systems (ITS) move forward and become more widespread, information on travel times and route guidance systems will become very well organized. Sensitivity to travel times may become extremely high. In this case, paradoxically, there is possibility that network flows will become unstable under the influence of ITS. Furthermore, the fact that chaotic flow can exist implies that forecasting travel times and flows in a network may be impossible no matter how much progress is made with ITS. Network flows cannot be observed with infinite accuracy, so if this is the case long-term forecasts of flows cannot be made due to sensitivity to initial conditions that is the nature of chaos. In this situation, a very small error increases exponentially as time goes by.

As a future work, this model and its analysis should be expanded to a more general network. The question of whether or not there exists network flow in the real world that is chaotic should be examined.

## References


1. Akcelik, R. (1978) A New Look at Davidson's Travel Time Function, *Traffic Engineering and Control*, Vol. 19, pp. 459-463.
2. Bureau of Public Roads (1964) *Traffic Assignment Manual*, Urban Planning Division, U.S. Department of Commerce, Washington D.C.
3. Beckmann, M., C. B. McGuire, and C. B. Winsten (1959) *Studies in the Economics of Transportation*, Yale University Press, New Haven.
4. Canterella, G. E. and E. Cascetta. (1995) Dynamic Process and Equilibrium in Transportation Networks: Towards a Unifying Theory, *Transportation Science*, Vol. 29, No. 4, pp. 305-329.
5. Cascetta, E. A (1989) Stochastic Process Approach to the Analysis of Temporal Dynamics in Transportation Networks, *Transportation Research*, Vol. 23B, 1989, pp. 1-17.
6. Davidson, K. B. (1966) A Flow Travel Time Relationship for Use in Transportation Planning, *Proceedings of Australian Road Research Board*, Vol. 3(1), pp. 183-194.
7. Daganzo, C. F. and Y. Sheffi (1977) On Stochastic Model of Traffic Assignment, *Transportation Science*, Vol. 11, pp. 253-274.
8. Dupuis and Nagurney (1993) Dynamical Systems and Variational Inequalities, *Annals of Operations Research,* Vol. 44, pp. 9-42.
9. Emmerink, R. H. M., K. W. Axhausen, P. Nijkamp, and P. Rietveld (1995) Effects of Information in Road Transport Networks with Recurrent Congestion, *Transportation*, Vol. 22, pp. 21-53.
10. Friesz, T. L., D. Bernstein, N. J. Mehta, R. L. Tobin, and S. Ganjalizadeh. (1994) Day-to-day Dynamic Network Disequilibria and Idealized Traveler Information Systems, *Operations Research*, Vol. 42, pp. 1120-1136.
11. Goldberg, D. G.: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Pub. Co., Reading, Massachusetts, 1989.

12. Hahn, W. (1963) *Theory and Application of Lyapunov's Direct Method*, Prentice-Hall, Englewood Cliffs, N. J.
13. Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, University of Michigan Press, Ann Arbor.
14. Holland, J. H, Holyoak, K. J., Nisbett, R. E., and Thagard, P. R. (1986) *Induction—Processes of Inference, Learning, and Discovery*, MIT Press, Cambridge.
15. Horowitz, J.L. (1984) The Stability of Stochastic Equilibrium in a Two-Link Transportation Network, *Transportation Research*, vol.18B, pp.13-28.
16. Kobayashi, K. (1994) Information, Rational Expectations, and Network Equilibria—An Analytical Perspective for Route Guidance Systems, *The Annals of Regional Science*, Vol. 28, pp. 369-393.
17. Kobayashi, K. and H. Tatano (1996) Traffic Network Equilibria with Rational Expectations, *Interdisciplinary Information Sciences*, Vol. 2, pp. 189-198.
18. Li, T. Y. and J. A. Yorke (1975) Period Three Implies Chaos, *American Mathematical Monthly*, Vol. 8, pp. 985-992.
19. Muth, J. F. (1961) Rational Expectations and the Theory of Price Movements, *Econometrica*, Vol.29, No.3, pp.315-335.
20. Nakayama, S. and R. Kitamura A Route Choice Model with Inductive Learning, *Transportation Research Record*, 2000 (forthcoming).
21. Nakayama, S., R. Kitamura, and S. Fujii (1999) Drivers' Learning and Network Behavior: A Dynamic Analysis of the Driver-Network System as a Complex System, *Transportation Research Record*, No. 1676, pp. 30-36.
22. Nagurney, A. and D. Zhang (1997) Projected Dynamical Systems in the Formulation, Stability Analysis, and Computation of Fixed-Demand Traffic Network Equilibrium, *Transportation Science*, Vol. 31(2), pp. 147-158.
23. Newell, A. and Simon, H. A. (1972) *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, New Jersey, 1972.
24. Sheffrin, S. M. (1996) *Rational Expectations*, 2nd ed., Cambridge University Press, Cambridge.
25. Simon, H.A. (1947) *Administrative Behavior: A Study of Decision-Making Process in Administrative Organization*, Macmillan, New York.
26. Simon, H.A. (1957) *Models of Man: Social and Rational*, John Wiley & Sons, New York.
27. Smith, M. J. The Stability of Dynamic Model of Traffic Assignment: An Application of Method of Lyapunov, *Transportation Science*, Vol. 18, No. 3, 1984, pp. 245-252.
28. Wardrop J. G. (1952) Some Theoretical Aspects of Road Traffic Research, *Proceedings of the Institution of Civil Engineers*, Part II, Vol. 1, pp.325-378.
29. Watling, D. (1999) Stability of the Stochastic Equilibrium Assignment Problem: A Dynamical Systems Approach, *Transportation Research*, Vol. 33B, pp. 281-312.
30. Zhang, D. and A. Nagurney (1996) On the Local and Global Stability of a Travel Route Choice Adjustment Process, *Transportation Research*, Vol. 30B, pp. 245-262.

# DYNAMIC SIMULATION-BSAED MODEL OF URBAN TAXI SERVICES

*Ziqi SONG: The University of Hong Kong, China ziqi@hkusua.hku.hk*
*C.O.TONG: The University of Hong Kong,China cotong@hku.hk*

## Abstract

In urban areas taxi service is an important para-transit mode that serves as a complement to public transportation. Early aggregate taxi models do not take into account network topology. Our simulation based network model simulates taxi drivers' individual learning process and use dynamic traffic assignment (DTA) as a basis to represent real-time network conditions. It is an attempt to model urban taxi service more realistically and to investigate the impacts of taxi drivers' learning behaviours on the whole taxi service system. Two indicators are used to measure system performance, total taxi vacant time and total customer waiting time. A numerical example is provided at the end to demonstrate this model.

## 1 Introduction

In Asian countries where private cars are not widely owned, taxi provides a convenient and speedy option other than public transport. Due to the complex interactions between demand and supply as well as lack of knowledge of taxi drivers' behaviours, taxi research has been overlooked comparing with other transport modes. Many economists have studied the regulation and fare structure of urban taxi services; however, generally they treated the taxi system as an aggregate, abstract model without considering the network topology of the road network. Detailed reviews on aggregate model can be found in Yang and Wong (1998). The first network model for taxi operations was developed by Yang and Wong (1998). Their model deals with the impact of fleet size on the level of service and taxi utilization for a given and fixed customer demand OD pattern. Wong et al. (2001) extended the network model into a bi-level optimization of two equilibrium problems in congested road network. However, these models are based on static equilibrium and do not considering time-varying traffic patterns in the road network which could be quite significant in rush hours especially in congested road network. Most recently, Yang et al. (2005) and Kim et al. (2005) made a first attempt to model taxi services in a dynamic network. Our model follows a similar approach but a different taxi learning model is used. Furthermore, we intend to incorporate the taxi model within a simulation based dynamic traffic assignment (DTA) model.

## 2 Main assumptions of taxi model

With the rapid deployment of Intelligent Transport Systems (ITS) especially Advanced Traveller Information Systems (ATIS) providing real-time traffic information to drivers, taxi drivers will perceive the travel time quite close, if not identical, to the real travel time they may experience. However due to the randomness of costumers' time-varying demand (within day or day to day); it is still hard for taxi drivers to know where the passengers are. Most common situation is that the taxi drivers will update the information about passenger demand according to their own experience, which is a learning process. Based on the above logic, we use a predictive dynamic user equilibrium model to depict the travel time of taxi drivers and use an individual cognitive model to simulate the day to day learning process of taxi drivers.

Because of the diversity of taxi operation mode around the world, we also need to specify the taxi operation mode used in this study. There are mainly three methods of hiring a taxi: First, waiting taxis at taxi stands, which usually locate at places where large numbers of passengers are likely to be found. Second, taxis are often hailed on the street. This is the most convenient way of hiring a taxi, however, this mode causes taxis' stop-and-go phenomenon, which may cause severe traffic safety problems as well as reduce road capacity heavily. Therefore, this kind of taxi operation mode is being discouraged by regulations such as parking restriction and road segregation facilities in central part of urban area. Another commonly used system in U.S cities, Dial-and-ride system, is not so popular in Hong Kong because of the high supplements and

demands for taxis. In this study, we focus on only one taxi operation mode, waiting taxis at taxi stands, which is the most common situation in CBD of Hong Kong.

# 3 Network conditions based on DTA

Predictive dynamic user optimal conditions are based on actual travel costs experienced by travellers. One illustrative definition of PDUO can be found in Tong and Wong (2000). "If, for any travellers between any O-D pair leaving their origin at any instant ,the actual travel time that these travellers experienced on any used routes are equal and minimal; and the actual travel times that these travellers would experience on any unused routes are greater than or equal to the minimum actual travel time on used routes ". The DTA model used in this study has two major components: the pathfinders (PF) and the traffic simulator (TS).The PF finds the optimal path for each O-D pair, based on an assumed network loading pattern. The TS loads the O-D demand onto the network based on the results of the PF and thereby updates the network loading pattern. The PF and TS modules are iterated successively until a convergent solution is obtained. Tong and Wong (2000) described in detail the simulation-based model used here. After the completion of DTA part, we assume that the travel time is independent of taxi flows during the taxi drivers' day to day learning process. In future studies, we will also consider the interaction between taxis and normal vehicles. That means travel time of taxi will also be influenced by taxi flows. When taxi drivers' learning process finish and taxi performance tends to be stable, we will use the new combined flow (taxis plus normal cars) to calculate new travel time again by looping back to the DTA part. After several iterations, the system will reach a consistency.

# 4 Behaviours of taxi drivers

Taxi drivers are seeking for passengers in the road network whenever it becomes empty and try to minimize their idling time, the summation of travel time to new taxi stands and expected waiting time at the taxi stands. For a particular taxi driver j, this idling time of choosing a certain stand i can be expressed as follow:

$$U_j(i) = Travel\_time_j(i) + E\_Waiting_j(i) + \varepsilon_{ji} \text{ , for any } j \in J, i \in I,$$

This utility of certain stand i is a random variable due to random arrival of passengers and variations in perceptions of taxi drivers. This random variable is assumed to follow Gumbel distribution. Then the possibility of taxi driver choosing a certain taxi stand i can be expressed as a multinomial Logit model:

$$P_j(i) = \frac{\exp\left[-\beta U_j(i)\right]}{\sum_{m \in I} \exp\left[-\beta U_j(m)\right]} \text{ , for any } j \in J, i \in I,$$

where I is the total set of taxi stands that can be chosen from and J is set of all taxi drivers. $\beta$ is a nonnegative parameter reflecting the information that an individual taxi driver perceive about the passenger demand and taxi service in the whole network. Small $\beta$ stands for high stochastic phenomenon. While large $\beta$ means more perfect information that an individual taxi driver obtains. When $\beta$ goes to infinity, taxi drivers tend to do "all or nothing" choice between alternative taxi stands, which is a deterministic case.

We have mentioned that taxi drivers update the passenger information according to their own experience. An individual cognitive model is adopted here to represent the day to day learning process. By repeatedly making decision, individuals are forming expected values of the attributes through acquiring knowledge from their memories. In our study, taxi drivers accumulate their knowledge about the waiting time at each stand through the waiting time they experienced previously. Two issues should be noted here: Firstly, how do taxi drivers store their knowledge about a certain stand? An individual taxi driver can either update the expected waiting time whenever the driver passes the stand once or each day the driver will average the waiting time experienced at the particular stand and get one value to stand for that day's experience. We think the second way is more realistic and adopt this in our study. Because a particular taxi driver may get passengers from a certain stand several times during a day or even during the rush hours. Taxi driver will not be so sensitive to the difference of waiting time in the taxi stand within the day. In reality, it is also impossible for a taxi driver to remember all the waiting time whenever they get a passenger from a certain stand after several days. Therefore, we average the waiting time of a certain stand i for taxi driver j within a certain day d. Mathematically it can be expressed as follows:

$$Waiting\_average_{j,i,d} = \frac{\sum_{t \in T} Waiting\_time_{j,i,d}}{T}, \text{for any } j \in J, i \in I, d \in D,$$

where T is the set of total number of times for a particular taxi j gets passengers from stand i within day d and D is set of days. Secondly, how do taxi drivers build their expected waiting time based on previous experiences? Based on the literature review of learning model (Arentze and Timmermans, 2005), a commonly used learning model in transportation literature is:

$$Q_n(a) = \lambda r_n + (1-\lambda) Q_{n-1}(a), \quad \text{for any } n \in N,$$

where $Q_n(a)$ is the expected value of action a at time n. Parameter $\lambda$ is typically interpreted as a measure of habit strength. If it is assumed that $\lambda$ is a constant, it follows that:

$$Q_n(a) = (1-\lambda)^n Q_0(a) + \sum_{p=1}^{n} \lambda (1-\lambda)^{n-p} r_p, \text{for any } n \in N,$$

We can observe that expected value of the last day is a weighted average of the expected value of the first day and the past values. The weight given to past values depends on how long ago it was experienced. More specifically, more recent experience takes more weight than the older ones Sometimes it is called an exponential recency-weighted average. If we interpret the above rules in our taxi model content, the expected waiting time for taxi j at taxi stand i of day D can be expressed as:

$$E\_Waiting_{j,i,d} = \lambda Waiting\_average_{j,i,d} + (1-\lambda) E\_Waiting_{j,i,d-1}, \text{for any } j \in J, i \in I, d \in D,$$

Another natural way of formulating taxi drivers' expected waiting time is giving each day's waiting time average the same weight. Thus,

$$E\_Waiting_{j,i,d} = \frac{\sum_{p=1}^{d} Waiting\_average_{j,i,p}}{d}, \text{ for any } j \in J, i \in I, d \in D,$$

Both ways of estimating expected waiting time are applicable and will be used in the numerical tests, though only the latter rule has been implemented due to its simplicity.

In order to deal with the time-varying passenger demand, taxi driver may also separate the whole day or rush hours into several time slots. Ideally, when the number of these discrete time slots goes larger, taxi drivers can perceive time-varying demand more precisely. However, it is not realistic for taxi drivers to memorize and differentiate too many time slots. On the other hand, more time slots take more computational efforts in computer simulation .Therefore, limited time slots are used in this study. It should also be noted that dividing a day into many time slots means that all the variables mentioned above need to add one more dimension time slot S, because time variables perceived by taxi drivers in different time slots do not intervene with each other.

## 5 Taxi service performance index

As we have mentioned, we are interested in how taxi service performance evolves through taxi drivers' day to day learning process. Before we looking into this problem, sound taxi performance indexes should be defined. In Yang et al. (2002), they used taxi availability (as measured by expected customer waiting time) to describe the demand side of the taxi service. Passengers will consider taxi availability as well as fares to determine their mode choice. On the supply side, they used another indictor, taxi utilization (as measured by expected fraction of time a taxi is occupied). Taxi companies will look at the taxi utilization to determine whether they are profitable. In a similar manner, we also use two indictors to reflect taxi service performance. From the passengers' point of view, we introduce total waiting time of all passengers during the simulation period to reflect their satisfactory index. The smaller the total waiting time is the higher taxi service they perceive. Total waiting time consists of two parts: the waiting time of people who have gotten a taxi and leave the taxi stand and the waiting time of people who are still in queue at the stand by the end of the simulation period. From the taxi company's point of view, the total taxi time of being occupied or total taxi vacant time are what they concern. Actually the summation of these two is just the total time of taxi operation. In our study, we use total taxi vacant time as the performance index of taxi companies. For those taxi companies, they definitely wish to decrease total taxi vacant time because that means occupied time will increase and their revenues will goes up too, vice versa. Total taxi vacant time also consists of two

components: Taxis' waiting time in queues at the stands and the travel time in the network without a passenger. It is noted that similar taxi performance indictors can also be found in Kim et al. (2005).

Another remark about Yang et al. (2002) is that they treated the taxi market as a demand-supply equilibrium, which means for a certain fleet size of taxi and fare structure there will be a corresponding taxi demand. However, in our study, at least at this stage, we do not consider that elastic demand and supply situations. Taxi demand and fleet size are treated as extraneous variables.

# 6 A case study

We use a case study to demonstrate the evolution of taxi service performance under learning behaviours of taxi drivers. Using two scenarios, we identify the relationship between taxi drivers' learning behaviours and taxi service performance and give some discussions on under what circumstances learning process helps taxi drivers not only to improve their operation efficiency but also provide better services to passengers. The case study uses a four nodes' network and two of them are taxi stand node (T1 and T2) and the other two are destination nodes (D3 and D4). All links are two directional links. The network structure is showed in Figure1 below.

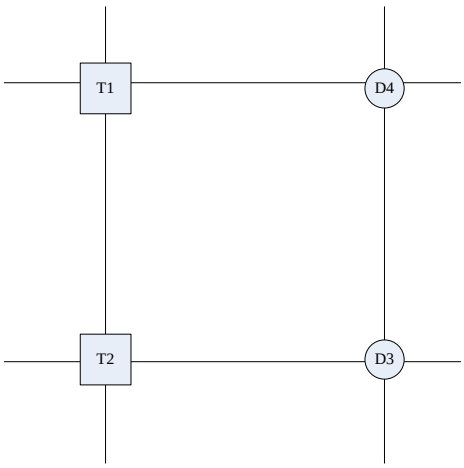



Figure 1

The parameter $\beta$ used in the Logit model can be calibrated with real world data (Wong et.al 1999). In this case study, $\beta$ is assumed as 1 for all scenarios. The length of time slots which is the duration of each discrete section of a day or rush hours is set to be 15 minutes for all taxi drives. From realistic point of view, duration of time slot should be large enough for a taxi driver to differentiate. We think 15 minutes is a proper one. In further studies, we can use SP-survey to calibrate this parameter.

Scenario one:

We assume two taxi stands get different taxi demand. Constant demand is used in this scenario. The taxi demand of taxi stand 1 is 1 person per minute (60 persons /h), while taxi stand 2 gets 3 persons per minute (180 persons/h). Taxi fleet size is assumed to be 10 vehicles which are evenly distributed at the beginning of each day. The total simulation period is one hour each day for 100 days. We plotted the total taxi vacant time and total customer waiting time against the days in Figure2. In order to see the trends of these two parameters more clearly, they are plotted in the same graph and total taxi vacant time is scaled up 100 times to get a better illustration.

From the simulation experiment, it is clear that in the first few days total taxi vacant time drops substantially and then tends to be stabilized with some fluctuations. So does the total customer waiting time. Several conclusions can be made from the observation. First, taxi vacant time and people waiting time are changing in the same fashion, which indicates that taxi service performance are improving from operators' as well as users' point of view. Second, taxi drivers' learning process does help them improve taxi service performance. Third, the fluctuations observed after being stabilized are caused by stochastic decision elements.

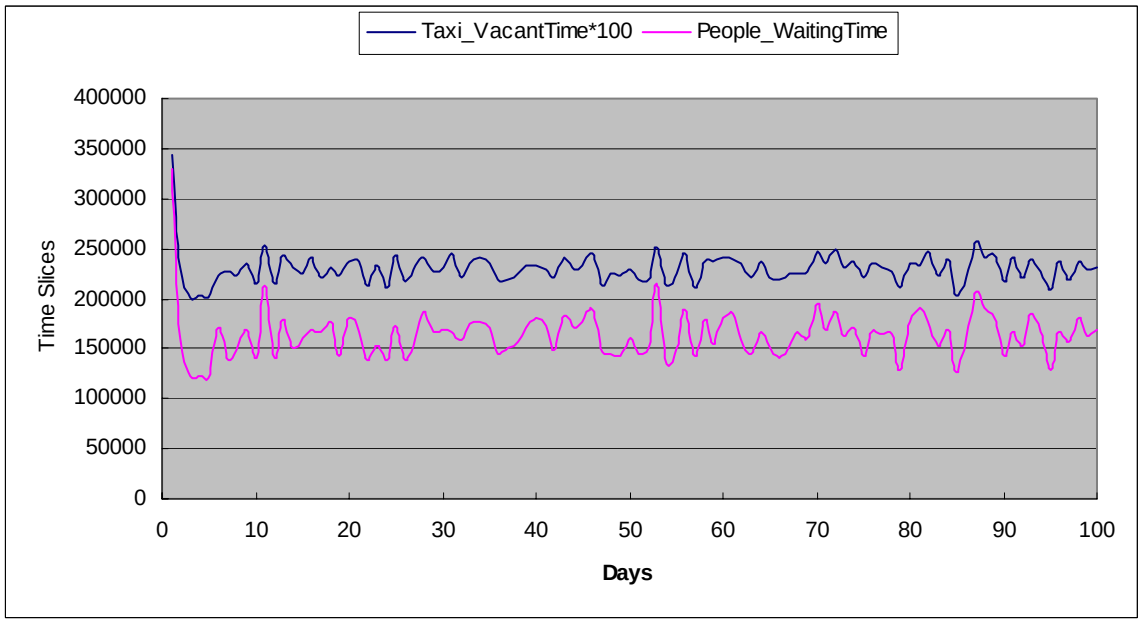



Figure 2

Scenario two:

In order to illustrate the influence of taxi drivers' learning process, we tested another scenario. All the settings of the network and taxis are the same but the taxi passenger demand at the stands. We assume two taxi stands get even demand, both of stands 1 and 2 get 2 persons per minute (120 persons/ h).Intuitively taxi drivers can not differentiate these two stands through learning process because for all taxi drivers perceive the two stands are just the same. Therefore, in this scenario day to day learning process should not help. The simulation experiment just verified the idea. We plotted the total taxi vacant time and total customer waiting time against the days in Figure3 for scenario 2.

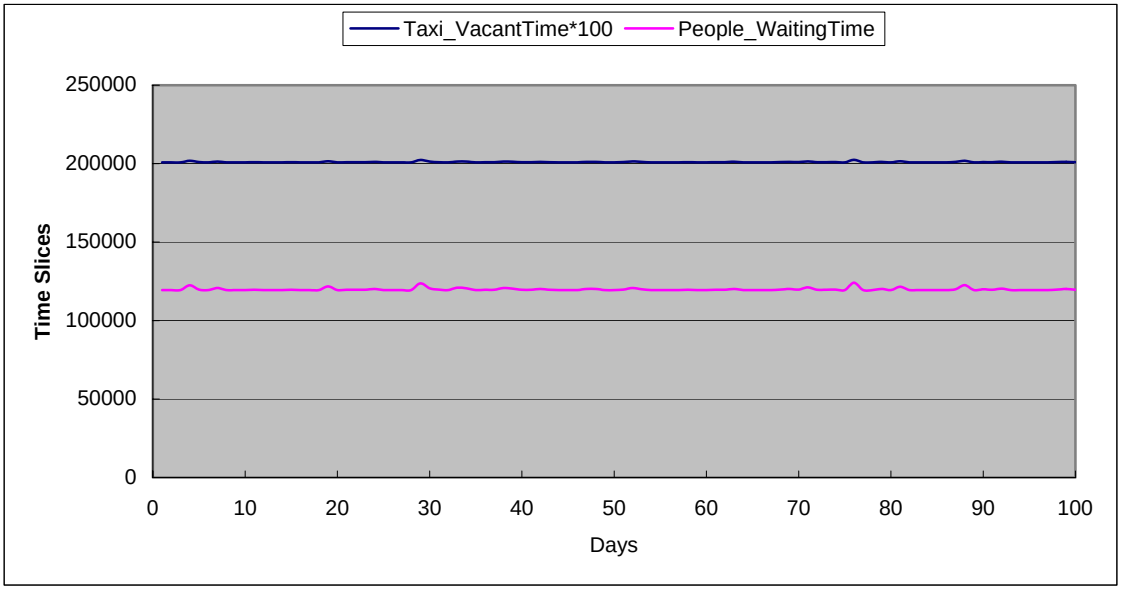



Figure 3

We can observe that neither the total vacant time nor the total customer waiting time is decreasing from day to day, which means taxi drivers can not be better off from the day to day learning process. However, we should note that this scenario will not happen in the real world. Every taxi stand has its own passenger arrival pattern and even for the same stand passenger arrival rate is also time-varying. In rush hours, this kind of variation could be quite significant, that is why we introduce discrete time slots in the taxi drivers' behaviour part to cope with the time-varying people arrival patterns.

# 7 Findings and future research

In this study, we introduced a framework of dynamic taxi service model and simulated taxi drivers day-to-day learning process. In embracing the dynamic elements in taxi services and taxi drivers' day-to-day learning behaviour, our taxi model made it possible to research taxi service in more realistic situations as well as in different spatial road networks. From the simulation experiments, we got the conclusion that taxi drivers can improve their services and operation efficiency through their day to day learning process in real world situations.

Our research on the dynamic taxi model is still continuing. Firstly, we will test more complicated scenarios. On the taxi demand side, time-varying demand needs to be introduced. On the taxi supplement side, different fleet size needs to be tested to see its influence on taxi performance. We will also use a larger size road network to test the stability of the model. Secondly, we will taxi and private car's interactions. We understand that in some urban areas, taxis can take a great portion of total traffic volume. How to simulate the taxi-car interaction in the DTA model needs to be further investigated. Thirdly, we will develop different taxi drivers' behaviours models. For examples, different cognitive rules or multi-class taxi driver model can be formulated so that urban taxi services can be modelled more accurately.

# References

Arentze ,T .and Timmermans, H. 2005, 'Modelling learning and adaptation in transportation contexts', *Transportmetrica,* vol.1,No.1, pp.13-32.

Kim, H., Oh, JS., Jayakrishnan, R. 2005, 'Effect of taxi information system on efficiency and quality of taxi services', *Transportation research record,* No.1903, pp.96-104.

Wong, K. I., Wong, S. C., Yang, H. 2001, 'Modeling urban taxi services in congested road networks with elastic demand', *Transportation Research*, vol. 35B, pp.819 – 842.

Wong, K.I., Wong, S.C., Yang, H., 1999, 'Calibration and validation of a network equilibrium taxi model for Hong Kong', In: *Proceedings of the Fourth Conference of Hong Kong Society for Transportation Studies*, Hong Kong, December 4, 1999, pp. 249 – 258.

Yang, H. and Wong, S.C. 1998, 'A network model of urban taxi services', *Transportation Research*, vol.32B, pp.235-246.

Yang, H., Wong, S. C. and Wong, K. I. 2002, 'Demand-supply equilibrium of taxi services in a network under competition and regulation', *Transportation Research,* vol.36B, pp.799–819.

Yang, H., Ye, M., Tang, W.H.C., and Wong, S.C. 2005, 'A Multiperiod Dynamic Model of Taxi Services with Endogenous Service Intensity', *Operations Research*, vol. 53, No. 3, pp. 501 – 515.

Tong C.O. and Wong S.C. 2000, 'A predictive dynamic traffic assignment model in congested capacity-constrained road networks' *Transportation Research,* vol.34B, pp. 625-644.

# COMBINING DTA APPROACHES FOR STUDYING ROAD NETWORK ROBUSTNESS

*Minwei Li: Delft University of Technology, The Netherlands m.li@tudelft.nl*
*Henk Taale: Delft University of Technology, The Netherlands h.taale@tudelft.nl*
*Henk J. van Zuylen: Delft University of Technology, The Netherlands h.j.vanzuylen@tudelft.nl*

## Abstract
In this paper a DTA model with two components is described: a user equilibrium (UE) model and an en-route model. The UE model is called MARPLE (Model for Assignment and Regional Policy Evaluation) that uses an iterative process to achieve equilibrium (deterministic or stochastic) (Taale *et al.*, 2004). In each iteration a network loading model is used to determine travel times. MARPLE en-route is developed based on the MARPLE model, which uses one simulation of the network loading model, starting with the equilibrium assignment results. It updates the path set and path costs after each evaluation interval during the simulation. Travellers will update their path choice according to the instantaneous path costs at the end of each interval using some heuristic rules.
A systematic framework for the robustness study of road networks is built up by combining both DTA approaches, in which the results of UE approach are used as references and en-route approach is used to simulate the network response for non-recurrent and short-term disturbances. The results for a hypothetical network show that for evaluating the network performance after such disturbances, the en-route assignment approach based on UE assignment results shows its capability and advantages in appropriately representing dynamic drivers' route choice behaviour when facing unfamiliar or unexpected situations on the route.

## 1 Introduction
Network robustness, defined as the ability of a system to continue to operate correctly under a wide range of operational conditions, and to fail gracefully outside of that range (Gribble, 2001), has been widely developed in large-scale networks such as electronics and internet. It also became an important topic for transport networks. In that context robustness can be considered as the ability of the system to keep a certain capacity level to handle traffic demand under abnormal situations. Dynamic Traffic Assignment (DTA) models play an important role in almost all the network robustness studies, because they take into account the reaction of drivers concerning route choice. Two approaches of DTA, user equilibrium (UE) assignment and en-route assignment, are separately implemented for different categories of network robustness and/or reliability studies. Basically, UE assignment models are used by many researchers when considering random changes in supply or/and demand of a transportation network. En-route assignment models are normally used to evaluate the effectiveness of certain traffic management schemes or measures for emergency situations or a short-term disturbance in the network. But, so far, little work has been done to develop a combined DTA model, to realize both UE and en-route assignment approaches, with the aim to be able to do a complete network robustness study.

A main task of network robustness studies is to assess whether an existing transport network system is susceptible to random failures (i.e. severe accidents) and destructive events (i.e. earth quake or terrorist attack). More important, we would like to know which parts (so-called hot or weak spots) of the network are most fragile, or vulnerable to the external disturbances, so that both infrastructure and control schemes could be improved in such a way that the deterioration of the network caused by those disturbances is mitigated.

Network robustness is rather new in the transportation domain and a limited amount of literature references could be found, such as Chiu and Mahmassani (2002) and Kaysi *et al.* (2003). Most of the methods implemented in these studies are borrowed from network reliability studies, which is in fact a quite different concept from robustness. Network reliability is defined as the probability of a device or a system performing according its purpose adequately for the period of time intended under the operating conditions encountered (Henley and Kumamoto, 1981; Wakabayashi and Iida, 1992), which means that reliability studies are generally concerned with probabilities only. And reliability problems are rooted in the uncertainty of traffic conditions. Existing reliability studies of road networks are mainly categorised according to three aspects: connectivity reliability, travel time reliability, and capacity reliability. In most of these studies, stochastic

user equilibrium (SUE) assignment models are implemented representing choice behaviour, especially route choice behaviour of travellers, to get the values of some chosen performance measures. This is done in the work of Bell and Iida (1997), Chen *et al.* (1999), Chen *et al.* (2002) and Du and Nicholson (1997).

However, user equilibrium is an ideal situation for any road network, which never appears in reality due to many uncertainties in both demand and supply. So, it is mainly meaningful for network planning purposes. But for the network robustness problem that focuses more on the evaluation of network performance and assessment of its ability to handle unpredictable incidents, this equilibrium assumption is no longer suitable, especially when congestion, as the result of those unpredictable incidents, is non-recurrent and short-term. In order to achieve more accurate and realistic values of network performance measures after the occurrence of such disturbances, appropriate dynamic traffic assignment models must be developed to realise more accurate description or simulation of the choice behaviour of travellers.

The objective of this paper is to develop methods for the evaluation of robustness of a road network, thereby forming a systematic framework for comprehensive network reliability and robustness studies. The paper is organised as follows. Section 2 describes the features and differences of user equilibrium assignment approach and en-route assignment approach, highlights the importance of en-route assignment approach in network robustness studies. Section 3 introduces the proposed en-route assignment approach based on an existing macroscopic traffic assignment model MARPLE. Section 4 provides a simulation-based systematic framework for network reliability and robustness studies, founded on both the UE assignment and en-route assignment approaches. In Section 5, the framework proposed in this paper is illustrated with a simple network. Section 6 summarises and analyses the results.

# 2 Equilibrium assignment and en-route assignment

A traffic assignment model, especially a dynamic traffic assignment (DTA) model, is the core of any model based reliability and robustness study of transportation networks. A DTA model typically describes route choice by an assignment sub-model, and the way in which traffic propagates through a network by a network loading sub-model. ‘Dynamic’ does not only mean that the demand between origins and destinations (OD) varies in different discrete time slices, but also covers the dynamics of the network situation (e.g. capacity) and the queuing behaviour, due to all kinds of variability, and the changes in behaviour of travellers. A realistic DTA model should be able to capture "over-capacity" queuing, because it follows the trajectories in time and space of the vehicles. Basically, two distinct approaches exist to model route choice and network loading in DTA: equilibrium assignment and en-route assignment.

## 2.1 Equilibrium assignment

Wardrop (1952) was the first to propose the following condition for a deterministic user equilibrium (DUE): for each OD pair, the costs of the paths actually used are equal, and they are less than or equal to the costs of each unused path (Wardrop's first principle). It assumes that each traveller has perfect information and chooses a route that minimises his/her travel time or travel costs, such that all travellers between the same OD have the same travel time or cost. A consequence of the DUE principle is that all used paths for each OD pair have the same minimum costs. Unfortunately, this is not a realistic description of loaded and congested traffic networks (Slavin, 1996).

The stochastic user equilibrium (SUE) was (amongst others) detailedly illustrated by Daganzo and Sheffi (1997). They defined the equilibrium state of traffic flow on a network as a stochastic user equilibrium when every user chooses his/her path such that his/her perceived travel time or cost between origin and destination is minimal. But perceived travel time or cost on a link varies randomly across users. The SUE problem is of the probit type if the perceived travel time or cost follows a normal distribution and of the logit type if it follows a Gumbel distribution.

In the equilibrium assignment problem, only pre-trip path choice and iterative process are considered. It consists of two main components: a method to determine a new set of time-dependent path flows given the experienced path travel times in the previous iteration, and a method to determine the actual travel times that result from a given set of path flow rates. The algorithm furthermore requires a set of initial path flows, which are normally determined by assigning all vehicles to the shortest free-flow paths for each OD.

### 2.2 En-route assignment

In the en-route assignment problem, the routing mechanism consists of successive executions of a set of behavioural rules, which determine how drivers iteratively react to information received en-route. Information may be available at discrete points in time, discrete points in space, or continuously in both space and time. Some information may only be available to a certain class of vehicles. Typically, the information strategy is an exogenous input. Drivers' responses to information can be modelled by some heuristic rules that may involve one or more parameters, such as the 'penetration rate' or the 'compliance rate' that is the fraction of drivers that react on the information. Another input to this problem is a suitable pre-trip assignment. In many cases, a static all-or-nothing (AON) assignment is used for this purpose. An en-route assignment thus only requires running a single dynamic loading of the demand onto the network over the time period of interest – apart from the assignments need to determine the initial route choice.

### 2.3 Roles of both assignment approaches in the robustness study of road networks

If an equilibrium assignment model is available, it is possible to find the equilibrium traffic pattern in a transportation network, taking into account all kinds of uncertainties. Network robustness studies use these patterns, as well as certain aggregated network performance measures, to perform comparisons and analyses. But with the equilibrium assignment approach it is not possible to represent the network situation under irregular and non-recurrent incidents, such as accidents. This means that DTA models with equilibrium assignment function only are not suited for the most common reliability and robustness studies: the analysis of the impact of disturbances on the traffic network. But it can be used in the network planning domain to analyse the impact of repeatable and long-term network changes, like introducing new measures of intelligent transportation system (ITS) or adding a new link to the network.

On the other hand, according to the features of the en-route assignment approach, it can be used for the analysis of unrepeatable and short-term incidents, such as accidents or a natural disaster, because it can update the perception of travellers to the network status and make new path choices for them. If a DTA model is only capable of en-route assignment, it is necessary to find the exogenous input, especially the pre-trip assignment, from other simulation tools or by other means.

At the same time, it is logical that for the robustness study of road networks, a reference status, which could be treated as the situation when the network serves under normal traffic demand, is needed. Since the aim of (dynamic) equilibrium assignment is to achieve an 'ideal' long-term equilibrium status for a chosen transportation network, its result could be used as the basic scenario, i.e. reference, for en-route assignment.

## 3 Framework for robustness studies

Based on the features of both assignment approaches and requirements of robustness studies of road networks, a simulation-based two-stage systematic framework is designed by integrating both an equilibrium assignment model and en-route assignment model in each stage (Figure 1). In this framework, the equilibrium assignment model is a macroscopic model named MARPLE (Taale *et al.*, 2004). The en-route assignment model is developed based on MARPLE, by using successively the network loading model and the route choice model for every pre-defined discrete interval. In Figure 2, differences between equilibrium approach and en-route approach are illustrated. In Stage One, as shown in the left part of the framework, only the equilibrium assignment for the basic situation is carried out. The following output for different functions is obtained:

1. Being the reference state of the network, including the following equilibrium (*) network performance indicators:
   a. TTT* total travel time during whole simulation period [veh•h];
   b. TTD* total travel distance during whole simulation period [veh•km];
   c. TD* total delay during whole simulation period [veh•h];
   d. NAS*(t) equilibrium dynamic network average speed within period t [km/h], defined as

$$NAS^*(t) = \frac{\sum_a \upsilon_a^*(t) f_a^*(t)}{\sum_a f_a^*(t)}, \tag{1}$$

where $v_a$*(t) is the average link speed and $f_a$*(t) is the link flow of link $a$ during period $t$ in the equilibrium situation;

e. $NL$*(t) equilibrium network load within period t [veh•h], defined as

$$NL^*(t) = \sum_a f_a^*, \tag{2}$$

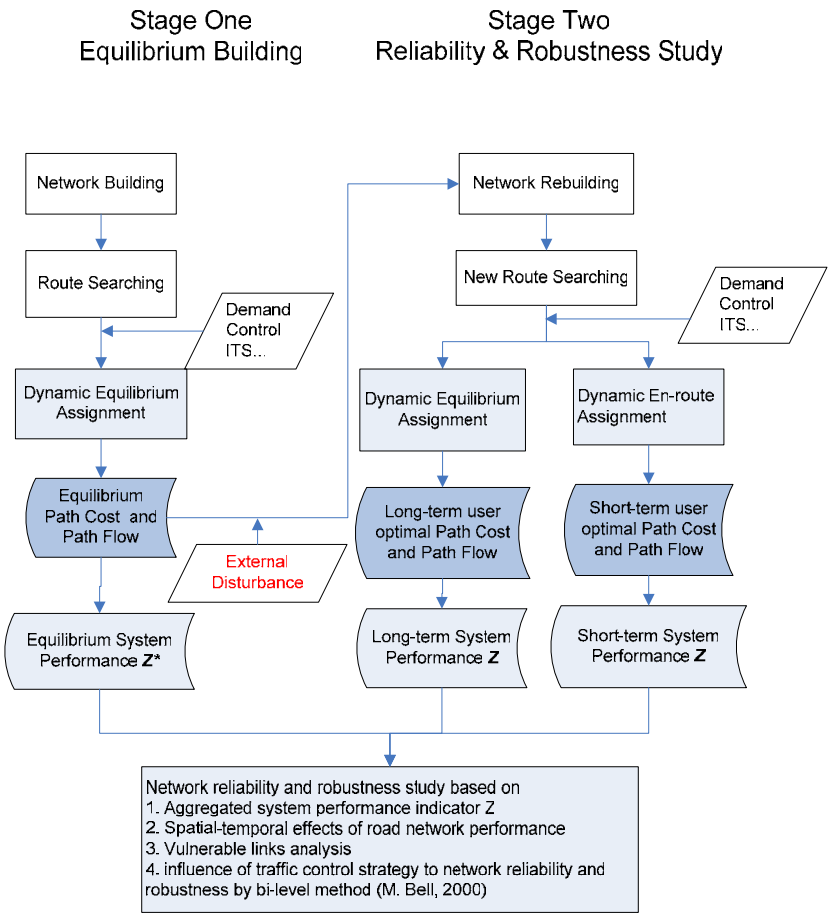



*Figure 1: Systematic framework for robustness studies of road network*

2. Being the initial assignment input for the en-route assignment in Stage Two, including the following assignment variables:
   1. $R^{od}$ route set for OD pair (o,d);
   2. $q_k^{od}$ average departure rate during time interval $k$ for OD pair (o,d) [veh/h];
   3. $u_k^r$ flows assigned to route $r$ during time interval $k$ [veh/h];
   4. $\tau_k^r$ average travel time of route $r$ during interval $k$ [min].

After obtaining the equilibrium assignment results for the network, random disturbances on individual links will be introduced in Stage Two. Depending on the duration (long-term or short-term) of the disturbance, either the equilibrium assignment approach or the en-route assignment approach could be used for the simulation respectively. The five indicators (a-e) that are mentioned above will be calculated and compared to the reference values to evaluate the network performance. Specially, we use the loading multiplier $\mu$ in equation (3) as the main robustness indicator, which can reflect the aggregate network load for the whole simulation period. The so-called hot spots in the network are those arcs with the smallest loading multiplier.

$$\mu = \sum_t NL(t) / \sum_t NL^*(t). \tag{3}$$

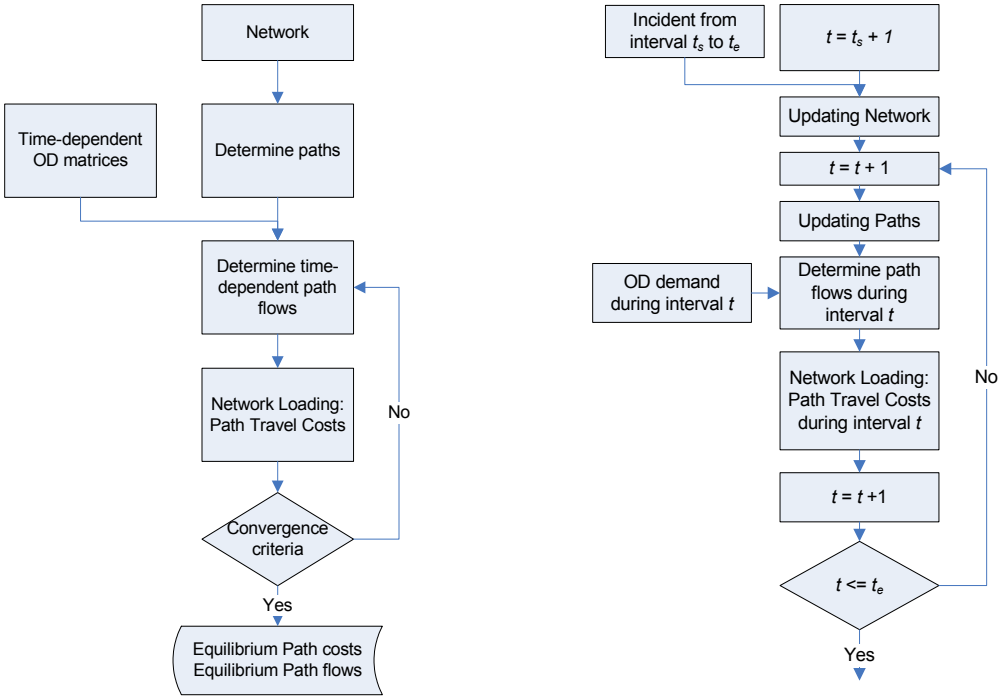



*Figure 2 General differences between equilibrium approach (left) and en-route approach (right)*

# 4 Case study

To demonstrate the framework described in the previous chapter and related assignment models, we applied it to a simple, hypothetical network as shown in Figure 3. The network consists of 10 nodes, 11 one-directional links, and three OD pairs, in which origin 2 (O2) and destination 2 (D2) represent a town centre.

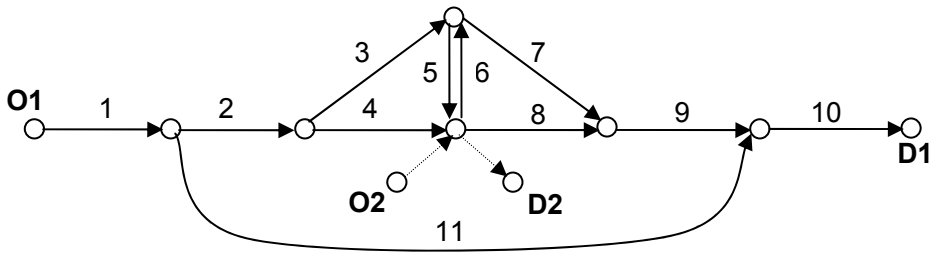



*Figure 3: Test network*

*Table1: Link characteristics*

| Link No. | No. of lanes | Length (km) | Capacity (veh/h) | Desired speed (km/h) |
|---|---|---|---|---|
| 1 | 4 | 4 | 8600 | 120 |
| 2 | 4 | 4 | 8600 | 120 |
| 3 | 3 | 4 | 6450 | 120 |
| 4 | 1 | 2 | 1800 | 50 |
| 5 | 1 | 2 | 1800 | 50 |
| 6 | 1 | 2 | 1800 | 50 |
| 7 | 3 | 4 | 6450 | 120 |
| 8 | 1 | 2 | 1800 | 50 |
| 9 | 4 | 4 | 8600 | 120 |
| 10 | 4 | 4 | 8600 | 120 |
| 11 | 1 | 16 | 1800 | 50 |

In Table 1, link characteristics, including number of lanes, length, desired capacity, and desired speed are listed. Link 3 and 7 form a faster, but longer motorway, and link 11 is a parallel, slower arterial. Links 4, 5,

6, and 8 are urban links with lower speed and capacity, which are the connectors between the motorway and the town centre. There are in total 7 routes available for all the OD pairs, whose detailed information and peak-hour demand of each OD pair are listed in Table 2. Route 3 is not used under normal conditions.

*Table 2: Route information and peak-hour demand for the OD pairs*

| (O,D) | Route No. | Link Sequence | Length (*km*) | Free flow travel time (*sec*) | Peak-hour demand (*veh/h*) |
|---|---|---|---|---|---|
| (O1,D1) | 1 | 1-2-3-7-9-10 | 24 | 720 | 6000 |
| | 2 | 1-2-4-8-9-10 | 20 | 768 | |
| | 3 | 1-11-10 | 24 | 1392 | |
| (O1,D2) | 4 | 1-2-4 | 10 | 384 | 1500 |
| | 5 | 1-2-3-5 | 14 | 504 | |
| (O2,D1) | 7 | 8-9-10 | 10 | 384 | 1500 |

In our preliminary case study, simple scenarios are designed. In each scenario, one and only one link, except entrance link 1 and exit link 10, is blocked during the peak hour. In this hour, the capacity of the chosen link is set as the certain ratio (from 0.0 to 1.0) of the designed capacity. Before and after the peak hour, the network is working with the designed capacity and speed. In all the scenarios, we assume that 70% of the travellers will get instant and perfect knowledge about the network conditions and will update their paths, while other 30% will stay on their pre-defined paths.

For each scenario, a 3.5-hour demand profile as shown in Figure 4 is used, representing a 'warming-peak-cooling' loading procedure. The last half an hour is designed with zero demand in order to clear up the network and take into account the influence of queuing delays completely.

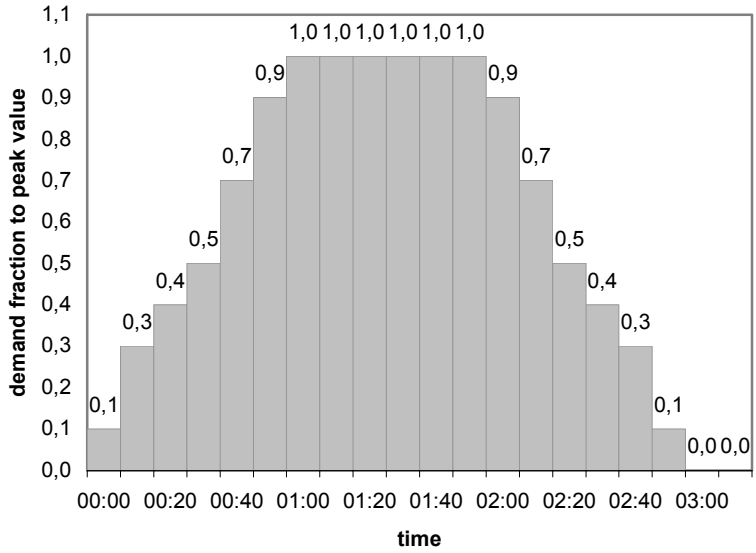



*Figure 4: Profile of the demand ratio (related to peak-hour value)*

### 4.1 Aggregated network performance measures

As noted earlier, three aggregated network performance indicators are used for the network robustness study in table 3. Compared with the equilibrium situation (when link 11 is blocked), the network performance deteriorated at different levels. The values in the parenthesis are the ratios to the equilibrium values. The highest delays appear when motorway links are blocked, such as link 2, 3, 7 and 9. For the arterial links, the influence of off-ramps (link 4 and 5) is much higher than on-ramps (link 6 and 8). Some of the total travel distance results give out relative values less than 1, which are the scenarios when the total delay is enormously high. This is caused by some remaining traffic, still present in the network at the end of the simulation. So for these cases, the actual values of the indicators must be higher than those in the table.

*Table 3: Aggregated network performance indicators when links are blocked*

| Blocked link | Total travel time (*veh•h)* | Total travel distance (*veh•km*) | Total delay (*veh•h*) |
|---|---|---|---|
| 2* | 10256.83 (1.59) | 349208.02 (1.01) | 4119.47 (13.23) |
| 3* | 11771.86 (1.82) | 330207.10 (0.96) | 5634.50 (18.09) |
| 4 | 8785.83 (1.36) | 349700.57 (1.01) | 2648.47 (08.50) |
| 5 | 6712.87 (1.04) | 346888.87 (1.01) | 575.51 (01.85) |
| 6 | 6474.63 (1.00) | 343449.95 (1.00) | 337.27 (01.08) |
| 7* | 12091.59 (1.88) | 320616.49 (0.93) | 5954.23 (19.12) |
| 8 | 7253.05 (1.12) | 349336.07 (1.01) | 1115.69 (03.59) |
| 9* | 11666.83 (1.81) | 334251.18 (0.97) | 5529.47 (17.76) |
| 11 (equilibrium) | 6448.78 (1.00) | 344757.57 (1.00) | 311.42 (01.00) |

* means route 3 is used in that case

### 4.2 Dynamic network performance analysis

Since en-route assignment is just a one-shot procedure, using the time-dependent indicators calculated in equations (1) and (2), the dynamics of the network performances are more clearly described in Figure 5. In both figures of figure 4, the equilibrium scenario and the scenarios with link 3 blocked, link 4 blocked and link 9 blocked are presented. When a link is blocked, its capacity drops to zero. The left sub-figure in figure 4 demonstrates how the average network speed changes. It is obvious that after blocking link 3 and link 9 network speed deteriorates much more and the negative impact lasts longer than other scenarios. In the right sub-figure of figure 4, curves of instantaneous network capacity also illustrate the different influences of blocking different link. For the scenario that link 3 is blocked, although the blockage is removed after the peak hour (interval 72) and the link capacity returns to its desired value, it takes about one hour (till interval 100) for the total network capacity to recover to the normal value, which indicates the remarkable after effect level of the incident.

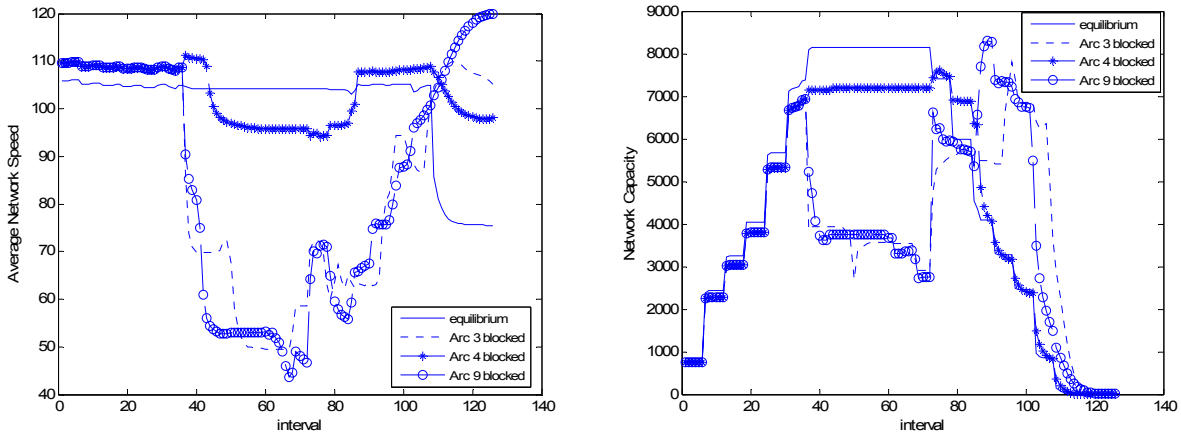



*Figure 5: Dynamic network performance: network average speed (left) and network capacity (right)*

### 4.3 Sensitivity analysis

Sensitivity information provides the change of the network loading multiplier calculated by equation (3) with respect to the changes of service level of each link capacity. Level of service is defined as the remaining capacity of the link after the incident on the link. Figure 6 shows the results of total network loading multiplier results the whole simulation period with different level of services in the scenarios when the capacity on links 3, 7, 8 or 9 drops. In general, when the level of service increases, the network capacity also increases. The curve of link 9 drops first and also the fastest if the service level decreases in those curves, because link 9 is a common link for both OD pairs (O1, D1) and (O2, D1). So to a certain extent, link 9 is the hot spot in this network.

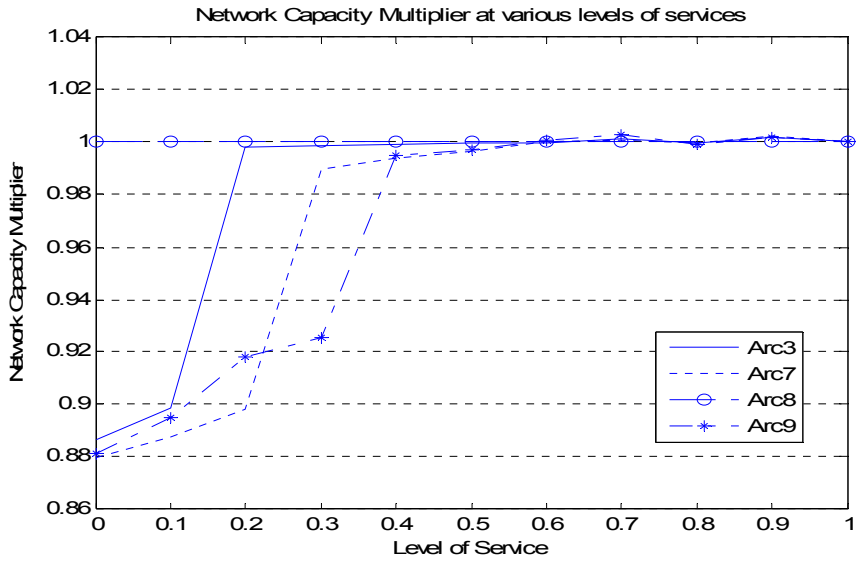



*Figure 6: Changes of Network Loading Multiplier (NLM) in relation with different level of service*

# 5 Conclusions and discussion

In this paper, we presented a new simulation-based systematic framework for the study of robustness in road networks and tested its feasibility by evaluating several network performance indicators for the network robustness. It is the first time that both equilibrium assignment and en-route assignment approaches are integrated in the study of robustness of road networks. This systematic framework can be considered as a complete structure. The reason is simple: the equilibrium assignment model can represent normal daily situations as the reference situation and represent the abnormal situation with long-term disturbances, while the en-route assignment model can represent network performance for abnormal situations after non-recurrent and short-term disturbances. Neither model can be neglected due to the different travel behaviour under different conditions. In addition, several time-dependent network performance indicators, next to some common aggregated indicators, have been developed as the results of both DTA approaches. A simple hypothetical network, which represents a typical city network with a motorway and parallel low-level path bypass, has been used for testing. Numerical results for this network demonstrated the feasibility of the robustness evaluation procedure. Some general remarks on the robustness of such kind of network can be made:

1. Common links that are used by multiple OD pairs have more influence on the network performance;
2. Disturbances on off-ramps have more deterioration effects to the network, because they will immediately influence the motorway traffic and cause high delay;
3. Time-dependent indicators, i.e. average network speed and network capacity, can clearly describe the change of the network performance, as well as the robustness of the network.

Thought the framework for the robustness evaluation of a degradable road network is comprehensive, the current study is still in a preliminary stage. We outline a few potential research topics that need to be explored further.

- Examine the sensitivity of network robustness of demands. In the study, we only tested the change of service level. The study of network robustness with different demand levels is of great practical importance for the situation that big events are held in the city, which attracts more traffic than normal.
- Calibrate the route choice models in the framework, especially the choice models in the en-route assignment. It must be done together with other calibration processes, such as OD estimation. The calibration procedure proposed by Chu et al. (2004) might be included in the future study.
- Incorporate robustness constraints to the network design problem. Some researchers, such as Yin et al. (2004) and Zhang and Levinson (2004), have introduced the concept of robustness to network design and upgrade. But due to the simplicity of their static assignment models, the understanding of the disturbances involved and their impact on the robustness performance of a road network is not suitable for describing the dynamics. Thus, one possible formulation to integrate robustness in the network

design problem is to maximise the network reserve capacity subject to meeting a pre-specified service standard (e.g., capacity multiplier).

## References

Bell, M.G.H., (2000) A Game Theory Approach to Measuring the Performance Reliability of Transport Networks, *Transportation Research Part B*, Vol. 34, pp. 533-545

Bell, M.G.H., Iida, Y., (1997) Network Reliability, *Transportation Network Analysis*, England, John Wiley & Sons, West Sussex, pp. 179-192

Chen, A., Yang, H., Lo, H.K., Tang, W., (1999) A Capacity Related Reliability for Transportation Networks, *Journal of Advanced Transport*, Vol. 33(2), pp. 183-200

Chen, A., Yang, H., Lo, H., Tang, W., (2002) Capacity Reliability of a Road Network: An Assessment Methodology and Numerical Results, *Transportation Research Part B*, Vol. 36, pp. 225-252

Chiu Y. C., and Mahmassani H.S., (2002) Hybrid Real-time Dynamic Traffic Assignment Approach for Robust Network Performance, *Transportation Research Record*, No. 1783: Transportation Network Modelling, pp. 89-97

Chu L.Y., Liu H.X., Oh J.S., Recker, W., (2004) A Calibration Procedure for Microscopic Traffic Simulation, In: *CD-ROM of 83rd Annual Meeting of Transportation Research Board*, January 11-15, Washington, D.C., No. 4165

Daganzo, C.F., Sheffi, Y., (1997) On Stochastic Models of Traffic Assignment, *Transportation Science*, Vol. 11, No. 3, pp. 253-274

Du, Z.P., Nicholson, A., (1997) Degradable Transportation Systems: Sensitivity and Reliability Analysis, *Transportation Research Part B*, Vol. 31, pp. 225-237

Gribble, S.D., (2001) Robustness in Complex Systems, In: *Proceedings of the 8th Workshop on Hot Topics in Operation Systems (HotOS-VIII)*, May, Elmau/Oberbayern, Germany

Henley, E.J., Kumamoto H., (1981) *Reliability Engineering and Risk Assessment*, NJ, Prentice-Hall, Englewood Cliffs

Kaysi I.A., Moghrabi M.S., and Mahmassani H.S., (2003) Hot Spot Management Benefits: Robustness Analysis for a Congested Developing City, *Journal of Transportation Engineering*, pp 203-211

Slavin, H., (1997) An Integrated, Dynamic Approach to Travel Demand Forecasting, *Transportation*, Vol. 23, pp. 313-350

Taale, H., Westerman, M., Stoelhorst, H. and van Amelsfort, D., (2004) Regional and Sustain-able Traffic Management in The Netherlands: Methodology and Applications, In: *Proceedings of the European Transport Conference 2004*, October 4-6, Strasbourg, France, Association for European Transport

Wakabayashi, H., Iida, Y., (1992) Upper and Lower bounds of Terminal Reliability of Road Networks: An Efficient Method with Boolean Algebra, *Journal of Natural Diaster Science*, Vol. 14, pp 29-44

Wardrop, J.G., (1952) Some Theoretical Aspects of Road Traffic Research, In: *Proceedings of the Institute of Civil Engineers*, Part II, pp. 325-378

Yin Y.F., Madanat S., Lu X.Y., (2005), Robust Improvement Schemes for Road Networks Under Demand Uncertainty, In: *CD-ROM of 84rd Annual Meeting of Transportation Research Board*, January 9-13, Washington, D.C.

Zhang L., Levinson, D., (2004) Investing for Robustness and Reliability in Transportation Networks, In: *Proceedings of 2nd International Symposium on Transport Network Reliability*, August 20-24, Christchurch & Queenstown, New Zealand, pp. 160-166

# Dynamic Activity-Travel Networks: A Unified Framework to Model Transportation Demand-Supply Interactions

Gitakrishnan Ramadurai: Rensselaer Polytechnic Institute, USA ramadg@rpi.edu

Satish Ukkusuri: Rensselaer Polytechnic Institute, USA ukkuss@rpi.edu

Abstract: In this paper, activity location, time of participation, duration, and route choice decisions are jointly modeled in a single unified framework referred to as Activity-Travel Networks (ATNs). The proposed framework is motivated by the following objectives: (a) to capture transportation demand-supply dynamics, (b) to capture activity demand-supply dynamics, and (c) to develop a framework for testing alternative behavioral mechanisms in urban transport models. ATNs use a network representation where nodes are activity centers that are joined by travel links. Each route in the network represents a set of travel and activity arcs. Therefore, choosing a route results in the choice of activity location, duration, time of participation and travel route. Equilibrium behavior requires that each individual choose the activity-travel sequence that provides the maximum utility and that all individuals from an origin participating in the same set of activities have the same utility irrespective of the route chosen. The dynamic user equilibrium (DUE) conditions based on the above behavior are then presented. An equivalent variational inequality problem is obtained. A solution method based on route-swapping algorithm is proposed. ATNs offer an unified framework to meet the objectives listed above; however, proof of existence of DUE solutions and efficiency of algorithms to solve large-sized networks are unresolved.

## 1. Introduction.

Urban transport modeling process involves several dimensions of individual choice including activity participation, location, time of participation, duration, choice of mode, route, etc. Two critical characteristics in modeling urban transport are: i) each individual makes choices so as to maximize his/her benefit, however, ii) the choice environment is dynamic and interactive. Dynamic because the supply side, represented by opportunities to participate in activities and transportation system accessibility, varies over time (both within-day as well as day-to-day). And interactive since decisions require use of shared resources: every individual's decisions affects and are affected by every other individual's decisions (for example travel time experienced when only one individual uses a road is different from that experienced when several individuals travel on the same road at the same time.)

Broadly, urban transport models can be divided into two dichotomous categories: one, descriptive statistical and econometric models of travel choice and, the other, network equilibrium models based on mathematical programming formulations and prescriptive behavior. The former category of models estimate demand for travel (hereon referred to as demand models). The latter category of models equilibrates demand and supply over a transportation network (network assignment models).

Descriptive demand models honor the first characteristic - utility maximization, but ignore the dynamic and interactive nature of the choice environment (at best they incorporate static supply characteristics). Network assignment models, on the other hand, directly incorporate the interactions and dynamics of the choice environment however at the expense of a sufficiently rich behavioral specification on individual choice behavior. As a result, demand modelers focus on the dimensions of activity participation, location, duration, time of participation, and choice of mode while network modelers focus on determining route choices and, in some cases, departure time choices. In practice, demand and network models are employed in conjunction and in some cases include a feedback mechanism between the two.

However, integrated models of both demand and supply dimensions have been developed recently (Lam and Huang (2003); Zhang et al. (2005);Kim et al. (2006)). Lam and Huang (2003) develop a

dynamic equilibrium model considering activity location, route, and departure time dimensions. But, their framework ignores the duration of activity participation. Capturing activity duration is essential to to understand the effect of activity scheduling on traffic congestion. An integrated work activity scheduling and departure time choice model in a network with bottleneck congestion is developed by Zhang et. al. (2005). However, their paper is concerned a single activity only. A logical extension is to consider multiple activities and activity chaining decisions. This is the focus of the paper by Kim et. al. (2006). They present an activity chaining model formulated from the perspective of a time use problem with budget constraints. Their model includes a dynamic traffic assignment simulation model to obtain network travel times and an iterative day-to-day dynamic process where activity chains are updated based on the network travel times computed in previous iteration. Whether such an iterative procedure results in consistent solutions and the performance of the solutions compared to more holistic frameworks are interesting research questions that merit attention.

The current study, extends above studies by integrating activity location, time of participation, duration, and route choice decisions in a single unified framework referred to as Activity-Travel Networks (ATNs). The basic description of ATNs and the motivation for developing a unified framework is presented in the next section.

## 2. ATN Representation and Motivation

ATNs use a network representation where nodes are activity centers that are joined by travel links. Activities are represented by arcs that both originate and terminate in the same node (activity centers). Each activity arc is characterized by a unique activity type and duration. An activity-travel sequence for an individual can be represented as a route that includes both travel and activity arcs. The model time frame may be set arbitrarily. However, all individuals at the beginning of the model start from 'home' and must participate in a predefined set of activities. All activity-travel sequences that traverses the set of activity arcs in which an individual participates in are considered feasible sequences. Consistent with rational behavior, each individual is assumed to choose the activity-travel sequence that provides the maximum generalized utility. However, modeling the network dynamics at an individual level would complexify the problem. Therefore, we treat all individuals who participate in a given set of activities as similar. We accordingly modify the behavioral framework to be consistent with Wardrop's (Wardrop (1952)) equilibrium framework. The behavior rule adopted is 'each individual chooses the activity-travel sequence that provides the maximum utility and all individuals from an origin participating in the same set of activities have the same utility irrespective of the route chosen.

The framework is presented in a discrete-time setting. Durations of arc-traversal for travel arcs is always assumed to be a function of flow, while for activity arcs it is independent of flow in some cases. For example, the duration of work activity could be assumed to be fixed irrespective of number of employees working at the location, but duration of a shopping activity could depend on the number of shoppers. Further, restrictions on minimum and maximum activity participation durations can be easily incorporated.

The proposed framework is motivated by the following considerations: (a) to capture transportation demand-supply dynamics by jointly modeling activity location, time of participation, duration, and route choice, (b) to capture activity demand-supply dynamics in addition to transportation demand-supply dynamics, and (c) to develop a framework for testing alternative behavioral mechanisms for urban transport models. We discuss each of these with examples below.

### 2.1. Transportation demand-supply dynamics

The importance of transportation demand-supply dynamics can best be illustrated through the following example. Consider a hypothetical scenario in the double-diamond network shown in figure 1. The network consists of 8 nodes: Home node (H), Work node (W), four Non-work activity centers (N1-N4), and two intermediate nodes (I1 and I2). The nodes are connected by 12 arcs: 3, 4, 10, and

11 are the activity arcs and the rest are travel arcs. Let us call the diamond with the home node as the residential neighborhood diamond (R-diamond) and the other as business neighborhood diamond (B-diamond). The total demand for travel from home to work is 100 individuals; all individuals drive alone to work. Further, 50 individuals drive directly from home to work while 50 individuals make a stop to participate in a non-work activity en route to work. All individuals have to arrive at work at the same time, (say) $T$. It is assumed that $T$ is sufficiently large to allow all individuals to reach work. All travel arcs have a capacity of 50 vehicles per time unit and free-flow traversal time of one time unit, while the duration of non-work activity participation (which is also the time for traversal of activity arc) is two time units. The utility of participating in the non-work activity is 100 utils (let us call the unit of measuring utility as utils) while the utility of travel on an arc is -5*(travel time) utils. As mentioned earlier, the travel arcs have fixed capacities: at free-flow a travel arc traversal would fetch -5 utils, while a queuing delay by one time unit would result in a payoff of -10 utils.

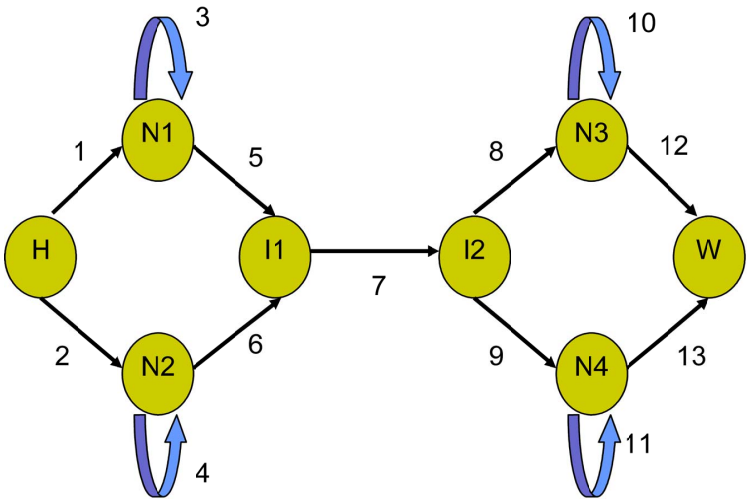



Figure 1

There exist two possible activity-chain sequence here: i) Home to Work, and ii) Home to Non-work activity to Work. The former can be accessed via four different paths while the latter has eight paths - four paths each that visit a non-work activity center in R-diamond and B-diamond. Since the utility of participating in the non-work activity in all four nodes is the same, based on traditional models of utility maximization, they attract equal amount of traffic. Therefore such an assignment model would result in each of the eight paths that pass through the non-work activity having a flow of 50/8. The corresponding total free-flow traversal time is 7 time units (therefore start time is $(T-7)^{th}$ time unit). For the individuals who drive straight to work, the traversal time is 5 time units; the corresponding flow is divided among the four paths (50/4). However, link 7, with a capacity of 50 vehicles per time unit, has an upstream demand of 75 vehicles at the start of $(T-3)^{th}$ time unit. This leads to delay by one time unit for 25 individuals and a loss in overall utility of 125 utils (assuming there is no late arrival penalty).

On the other hand, if the traffic dynamics is also incorporated in the assignment model, we would obtain a solution where none of the individuals visit the non-work activity center in the R-diamond. In this case, there is no delay for any of the individuals and the total overall utility is 125 utils more than the previous case. The reason for the difference in utilities is the limited capacity of link 7. Ignoring the traffic flow dynamics, could lead to sub-optimal assignment patterns. Therefore, it is important to consider transportation demand-supply dynamics.

### 2.2. Activity demand-supply dynamics

A second motivation for adopting ATN framework is the ability to incorporate activity demand-supply dynamics such as capacity restrictions in shopping mall check-out counters. Current models that ignore activity demand-supply dynamics could over-estimate trip-chaining of shopping activity by commuters or under-estimate non-peak hour shopping trips. If in the example above, the non-work activity centers located in the B-diamond had the following modified utility specification: 100 utils if flow on arc is less than or equal to 15 individuals, 75 otherwise; then, the corresponding destination choice and traffic assignment model would result in 15 individuals choosing to participate in the non-work activity in B-diamond while 10-individuals choose the R-diamond. However, all individuals with similar activity-travel sequence do not have same generalized utilities. This is a case where an equilibrium solution does not exist. We discuss this problem in detail in section 5.1.

### 2.3. Testing alternative behavioral mechanisms

The behavioral framework adopted in traditional frameworks such as four-step models has been a combination of random utility maximization (RUM) and Nash equilibrium. The individual level disaggregate models are usually based on the RUM framework. Here, every individual is presented with a set of alternatives and s/he chooses the alternative with the best utility/payoff. By construction, RUM framework does not consider the interaction effects: that is the effect of every individual's choice on payoff of other individuals. On the other hand, the network route assignment models equilibrate choices till Nash equilibrium is reached. This equilibrium is defined such that no individual can unilaterally change his/her choice and improve payoff (or generalized cost/utility). Several interesting questions arise.

- Why should the equilibrium modeling be restricted to route choice alone treating location, time, duration, and mode choice decisions as exogenous?
- Are the solutions obtained at equilibrium considering the above multiple dimensions different from that obtained in prior framework? Which captures behavior better?
- Are the demand model parameter values obtained by considering supply characteristics as exogenous different from those obtained when jointly estimating the parameters and supply characteristics (this is similar to simultaneous equations model but estimates of supply characteristics are obtained from solving mathematical programs instead of regression equations).

To answer the questions raised above, results from a unified model must be compared with traditional models. The Activity-Travel network proposed here provides one such model framework.

## 3. Conceptual Framework

### 3.1. Definitions and Notation

h: index for household.

$i_h$: index for individuals in household h.
$i_h \in 1, 2, \ldots, I_h$.

$G = \{\nu, \alpha\}$ is the activity-travel network, where $\nu$ is the set of nodes and $\alpha$ is the set of arcs.
$\alpha \ni \{\alpha^T, \alpha^A\}$ correspond to the set of travel and activity arcs.

$A_{i_h}$: Set of activities individual $i_h$ participates in.

$A_{i_h}^{trav}$: Set of travel activities for individual $i_h$.

$A_\bullet$: Set of activities for all individuals residing in zone represented by node n.

The elements of the above sets are characterized by attributes that denote their 'state'. We represent the set of characteristics as $\Omega[.]$. The characters shown in **bold** are assumed to be 'known' while the *italics* indicate characteristics that need to be determined in our present framework.

$\Omega_\nu[\mathbf{i}_\nu, \mathbf{e}_\nu, X_\nu]$: corresponding to incoming arcs, egress arcs, and accessibility measures.

$\Omega_{\alpha^T}[(\mathbf{m}, \mathbf{n}), (f, TT(f))]$: source node, sink node, flow, and travel-time.

$\Omega_{\alpha^A}[(\mathbf{n}, \delta), (U, f)]$: activity-center node, duration, and utility of traversing the arc, flow.

$\Omega_{A_{i_h}}[t_s, \delta, n]$: activity start time, duration, and location node.

$\Omega_{A^{trav}_{i_h}}[t_s, \delta, o, d, \mu, \rho]$: travel start time, duration, origin, destination, mode, and route.

$\Omega_h[.]$: Characteristics of the household h such as type of household, number of vehicles.

$\Omega_{i_h}[.]$: Characteristics of individual such as age, gender, employment status.

### 3.2. Relationships: ATN Framework

We use two types of functional relationships, $\Phi$ and $\Psi$, to capture the various complex relationships between the above variables. $\Phi$ functions are direct functional maps from $\Re^{mn}$ to $\Re^m$ (for example, regression equations), while $\Psi$ functions represent more complex relationships such as mathematical programs with equilibrium constraints, fixed-point problems, and variational inequalities.

Several different frameworks arise based on the relationships assumptions among the above sets and their characteristics. The set of relationships below represent the framework adopted in this paper. In particular, we assume $\Phi_1$, the set of activities that an individual participates in, as known; the focus of this paper is on the two $\Psi$-functions only.

$$A_{i_h} = \Phi_1(\Omega_h, \Omega_{i_h}, \Omega_\nu, \hat{\Omega}_\alpha) \tag{1}$$

$$\{\Omega_{A_\bullet}, \Omega_{A^{trav}_\bullet}\} = \Psi_1(\Omega_{\alpha^T}, \Omega_{\alpha^A}, \Omega_\nu) \tag{2}$$

$$\{\Omega_{\alpha^T}[f, TT(f)], \Omega_{\alpha^A}[U, f]\} = \Psi_2(\Omega_{A_\bullet}, \Omega_{A^{trav}_\bullet}) \tag{3}$$

The reader would note that the $\Phi$ function is at an individual level while the $\Psi$ functions are at a network or zonal level. Also, the $\Phi$ functions are similar to disaggregate demand models while $\Psi_2$ is similar to an aggregate network assignment model. Given the complexity of the $\Psi$ functions they are not modeled at an individual or disaggregate level (however, a quasi-disaggregate approach can be adopted by grouping the individuals into user classes). The reader would also note that, the $\Phi$ function includes estimates of $\Omega_\alpha$ denoted as $\hat{\Omega}_\alpha$.

$\Phi_1$ determines the set of activities that an individual participates in. This could depend on household and individual characteristics, activity center location and accessibility characteristics, (estimate of) transportation and activity supply characteristics, and also the set of fixed activities the individual participates in.

The $\Psi$ functions represent complex relationships between arc (both travel and activity arc) characteristics and characteristics of activities (both fixed and flexible). The two relationships $\Psi_1$ and $\Psi_2$ together represent a fixed-point problem given by (4) below.

$$\{\Omega_{\alpha^T}[f, TT(f)], \Omega_{\alpha^A}[U, f]\} = \Psi_2(\Psi_1(\Omega_{\alpha^T}, \Omega_{\alpha^A}, \Omega_\nu)) \tag{4}$$

## 4. Operational Framework

Two critical issues to operationalize the ATN framework are flow propagation dynamics and utility function specification.

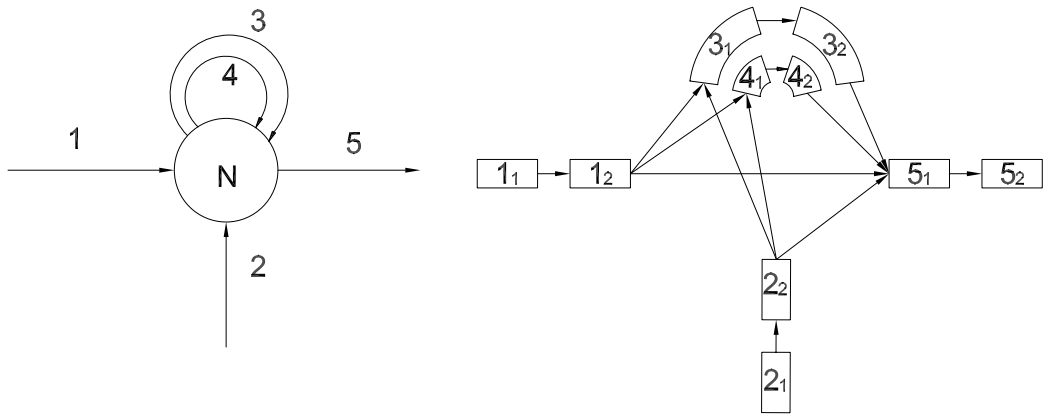
### 4.1. Dynamics of Flow Propagation

Traffic flow has been modeled at different levels in the past. The most realistic models are disaggregate microsimulation models where behavior of each vehicle on the network is modeled explicitly. On the other hand, macroscopic models describe traffic flow based on relationships between speed, flow, and density. Though microscopic models are more accurate they require greater computation time and lack analytical solutions.

Macroscopic models, on the other hand, can be modeled as side constraints or provide approximate, quick solutions and are more suited for analytical DTA models. Macroscopic models can be further divided into exit flow models, point queue models, and physical queue models. In this study, we adopt the network level simulation adaptation of the cell transmission model (CTM) (Daganzo (1994), Daganzo (1995)) to capture the dynamics of traffic flow propagation. The CTM is capable of capturing the effect of physical queues (for a discussion on differences between point queues and physical queues the reader may refer to Szeto and Lo (2006)). Also, we formulate the ATN problem as a variation inequality (VI) problem. Existing VI solution techniques are based on heuristic searches and require several iterations of network loading step. Therefore, microsimulation models may not be a computationally feasible option. We omit the details of the traffic flow model here. The reader is referred to Lo (1999) and Lo and Szeto (2002) for a detailed discussion.

The flow propagation on activity arcs are based on a simple point-queue model. Every activity has a fixed duration and a flow capacity but no density constraints. The activities can also be represented in the cell transmission framework with a simple modification to the cell occupancy update step: flow entering an activity arc must remain in the arc for a period that is at least equal to the duration of the activity represented by the arc.

An example of the cell transmission model representation of activities at a node is shown in Figure 2.




Arcs 1 and 2 are incoming (into Node N) travel arcs.
Arcs 3 and 4 are activity arcs. Arc 5 is an outgoing travel arc.

The equivalent cell-based representation of arcs is shown on the right.
$1_1$ and $1_2$ are two cells representing arc 1.

Figure 2

### 4.2. Utility Function Specification

Another critical step in the ATN framework is the utility function specification.
Let,

$A^c$ be the set of all possible activity combinations.

$R^a_{od}$ : Set of routes from origin o to destination d containing activity arcs $\alpha^A$ such that they traverse all activities in activity combination $a \in A^c$. r is a route that belongs to the set $R^a_{od}$. Each route r represents a set of travel and activity arcs. Therefore choosing a route r, results in the choice of activity location, duration, time of participation and travel route.

$U^{a,r}_{od}$ denote utility derived by individuals departing from o and reaching d, participating in activity chain $a \in A^c$ using route r.

$h^{a,r}_{od}$: Path flow from o to d, participating in activity chain combination $a \in A^c$ using route r.
The temporal dimension in dynamic traffic assignment models (such as departure or arrival time index) is not associated with the above definitions since all individuals are always traveling on the network or participating in an activity.

Similar to Lam and Huang (2003), we assume an additive specification for the above utility expression.

$$U^{a,r}_{od} = U^a(r) - U^{trav}(r) \tag{5}$$

$U^a(r)$ is the utility derived from participating in activity combination $a \in A^c$ and is a function of route r. $U^a(r)$ can be represented as the sum of utilities derived from traversing each activity arc $\alpha_a$ in route r.

$U^a(r) = \sum_{\forall \alpha^A \in r} U_\alpha(r, f)$
where, f is the flow in activity link. In general, utility derived from activity participation may be assumed to be a function of type and duration of activity, time of participation, location of activity with respect to the origin/destination of flow on route r, and the total flow on activity link $\alpha^A$.

$U^{trav}(r) = \beta * TT(r)$ is the disutility from travel on route r.
where, $\beta$ is a parameter to convert travel-time into utility units and TT(r) is the total travel time on route r.

The focus of the present paper is not on estimating utility function form or parameters. We may assume reasonable functional forms and parameter values to illustrate the ATN framework. However, accurate estimation of utility function form and parameters is an important issue that needs further investigation in the future.

## 5. Mathematical Formulation of ATNs

### 5.1. Dynamic User Equilibrium Conditions

We can now express the DUE conditions as follows:

$$U^{a,r}_{od} = \begin{cases} = \overline{U}^a_{od} \text{ if } h^{a,r}_{od} > 0 \\ \\ \leq \overline{U}^a_{od} \text{ if } h^{a,r}_{od} = 0 \end{cases} \quad \forall \text{o,d,} a \in A^c, \text{and r}(: r \in R^a_{od}) \tag{6}$$

Subject to the condition that flow on network should satisfy demand. This is expressed as:

$$\sum_{\forall r \in R^a_{od}} h^{a,r}_{od} = \sum_{\forall i_h \in (o,d)} \zeta^a_{i_h} \quad \forall a \in A^c, o, d \tag{7}$$

where,

$$\zeta^a_{i_h} = \begin{cases} 1 \ldots \text{ if activity combination } a \in A_{i_h} \\ 0 \ldots \text{ otherwise} \end{cases}$$

$\overline{U}^a_{od}$ is the maximum utility derived by individuals departing from o and reaching d, participating in activity combination $a \in A^c$ using route r.

DUE conditions, however, are not always satisfied in capacitated networks (Szeto and Lo, 2006). Because for DUE to hold a packet of flow that departs at the same time from the same origin should reach the destination intact. However in capacitated networks it is possible that packets of flow are broken because of the lack of available capacity downstream. Also, discontinuities in travel time or utility functions could also result in non-existence of solutions (Szeto and Lo (2006)). Any discrete-time model exposes itself to the above drawback. Further study is required to understand the properties of DUE in discrete-time capacitated network models.

### 5.2. Equivalent variational inequality formulation

The above DUE conditions can now be formulated as an equivalent VI problem.

$$\sum_{\forall a \in A^c} (\mathbf{h}^a - \hat{\mathbf{h}}^a)^T U^a(\mathbf{h}) \geq 0 \quad \forall \mathbf{h}^a \in H^a \text{ and } \forall a \in A^c \tag{8}$$

where,
$H^a$ is the set of feasible route flows traversing all activities in activity combination $a$, given by (7),
$\mathbf{h}^a$ is the vector of route flows $\in H^a$,
$\hat{\mathbf{h}}^a$ is the vector of route flows that satisfy the DUE condition in equation 6, and
$U^a$ is a vector whose each element is given by $U_{od}^{a,r} - \overline{U}_{od}^{a}$.

### 5.3. Solution Approach

The utility derived from traversing the activity-travel sequence represented by route r, expressed as the sum of utility derived from participating in activities and the disutility from travel, is assumed to be a monotone decreasing function of flow on route r. Therefore, a route-swapping algorithm (Szeto and Lo, 2006; Lam and Huang, 2003; Nagurney and Zhang (1997)) may be adopted to obtain solutions to the VI problem shown in (8). The detailed algorithm is presented below:

Step 0: Initialize.
Set iteration counter i = 0.
Choose an initial feasible vector of flows $\mathbf{h}(i)$.

Step 1: Computation.
Load flow $\mathbf{h}(i)$ and compute travel times TT(r) using the Cell-based transmission model.
Compute utilities $U_{od}^{a,r}$ using (5) $\forall r,\ o,\ d$ .
Set $U_{od}^{a} = \max_{\forall r \in R_{od}^a} U_{od}^{a,r} \quad \forall o,\ d.$

Step 2: Update flows.
Set $\hat{R}_{od}^a = r \in R_{od}^a : U_{od}^{a,r} = U_{od}^a$.
For ever activity combination $a \in A^c$,
$h_{od}^{a,r}(i+1) = \max(0, h_{od}^{a,r}(i) + \rho h_{od}^{a,r}(i)(U_{od}^{a,r} - U_{od}^{a})) \ \forall r \in R_{od}^a \backslash \hat{R}_{od}^a$
$\Sigma_h = \sum_{\forall r \in R_{od}^a \backslash \hat{R}_{od}^a} (h_{od}^{a,r}(i) - h_{od}^{a,r}(i+1))$
$h_{od}^{a,r}(i+1) = h_{od}^{a,r}(i) + \frac{\Sigma_h}{|\hat{R}_{od}^a|} \quad \forall r \in \hat{R}_{od}^a$
$\rho$ is a scale parameter.

Step 3: Check for convergence.
if $\sum_{\forall r \in R_{od}^a} (|h_{od}^{a,r}(i+1) - h_{od}^{a,r}(i)|) < 2\epsilon$, then terminate. $\epsilon$ is a convergence tolerance value.
else, i = i + 1; Go to Step 1.

Demonstration of the above algorithm on test networks is a direction of on-going research. Efficiency of the above algorithm to solve large-sized networks has to be established. It is expected that the efficiency would depend on utility function specification and model of flow dynamics. The effect of these factors on convergence and on existence of DUE solutions are interesting topics for further research.

## 6. Summary and Further Work

In this paper, a formulation based on Activity-Travel network representation, capable of accommodating both traffic and activity demand-supply dynamics is presented. The motivation for adopting ATNs are (a) to capture transportation demand-supply dynamics by jointly modeling activity location, time of participation, duration, and route choice, (b) to capture activity demand-supply dynamics in addition to transportation demand-supply dynamics, and (c) to develop a framework for testing alternative behavioral mechanisms. The importance of transportation and activity demand-supply interactions where demonstrated using examples.

The dynamic user equilibrium conditions were then presented and the equivalent variational inequality problem obtained. A solution method based on route-swapping algorithm is proposed. Several open issues merit further investigation: first, we need to derive the properties such as solution existence and uniqueness of the variational inequality problem. Second, numerical or analytical results on convergence properties of solution algorithms need to be developed. Also, other specialized algorithms that leverage the particular structure of the problem needs to be explored. This would depend, among other factors, on the utility function specification and the traffic flow dynamic model. Finally, a holistic framework that includes data collection and simultaneous estimation of activity and travel utility function parameters needs to be explored to completely realize the potential of the proposed framework.

## References

Daganzo, C. F.: 1994, The cell transmission model: A simple dynamic representation of highway traffic, *Transportation Research Part B* **28**, 269–287.

Daganzo, C. F.: 1995, The cell transmission model, part ii: Network traffic, *Transportation Research Part B* **29**, 79–93.

Kim, H., Oh, J.-S. and Jayakrishnan, R.: 2006, Activity chaining model incorporating time use problem and its application to network demand analysis, *Proceedings of the 85th Transportation Research Board Meeting, Washington D.C.* .

Lam, W. H. and Huang, H.-J.: 2003, Combined activity/travel choice models: Time-dependent and dynamic versions, *Network and Spatial Economics* **3**, 323–347.

Lo, H. K.: 1999, A dynamic traffic assignment formulation that encapsulates the cell transmission model, *In A.Ceder (ed.) Transportation and Traffic Theory, Pergamon, Oxford* pp. 327–350.

Lo, H. K. and Szeto, W.: 2002, A cell-based variational inequality formulation of the dynamic user optimal assignment problem, *Transportation Research Part B* **36**, 421–443.

Nagurney, A. and Zhang, D.: 1997, Projected dynamical systems in the formulation, stability analysis, and computation of fixed-demand traffic network equilibria, *Transportation Science* **31**, 147–158.

Szeto, W. and Lo, H. K.: 2006, Dynamic traffic assignment: properties and extensions, *Transportmetrica* **2**, 31–52.

Wardrop, J.: 1952, Some theoretical aspects of road traffic research, *Proceedings of the Institute of Civil Engineers, Part II* pp. 325–378.

Zhang, X., Yang, H., Huang, H.-J. and Zhang, H. M.: 2005, Integrated scheduling of daily work activities and morningevening commutes with bottleneck congestion, *Network and Spatial Economics* **39**, 41–60.

# A FIRST APPROACH TO DYNAMIC FREQUENCY-BASED TRANSIT ASSIGNMENT

*Jan-Dirk Schmöcker and Michael G H Bell, Centre for Transport Studies, Department of Civil and Environmental Engineering, Imperial College London, U.K.*
*Fumitaka Kurauchi, Department of Urban Management, Graduate School of Engineering, Kyoto University, Japan*

**Abstract**
This paper discusses an approach to transit assignment considering especially the following public transport properties: a) Transit vehicles have a finite capacity and there might be times of the day when the demand exceeds this capacity; b) demand is changing even within peak-hours; and c) in networks without published timetables passengers consider multiple routes and their actual route choice depends only on which vehicle arrives first (the so-called 'common line' effect). Central to the approach is the introduction of a probability that passengers are not able to board the transit line they wish to ride if this service does not have sufficient available space for all demand from this station during some time of day. This "fail-to-board probability" is set in such a way that all the available space is used but all demand exceeding the available capacity remains on the platform. To reflect changes in the fail-to-board-probability over time, time intervals are considered. Further, trips that can not be completed in one time interval are carried over to the next period. It is assumed that those passengers who failed to board attempt to continue their journey from the same platform in the next time interval.

## 1. Introduction

In road traffic the consideration of dynamic assignment is of importance because of changing supply and demand characteristics over the modelling period of interest. There is often a complex interaction between travel time of links and user route and departure time choice if the network becomes congested.

The consideration of dynamic aspects is also an issue for transit assignment. The travel time of buses which share the road space with private cars is also depending on the congestion. However, also for rail (and buses operating on separate bus lanes) dynamic aspects need to be considered leading to the same assignment problems as for road traffic. The changing perception of travel cost over time for these modes is also due to capacity shortage; in this case however this does not primarily lead to changes in link travel times but to on-board crowding causing discomfort to passengers. In particular some passengers might not get a seat. Further to this, in some situations the overcrowding might get so severe that some passengers will not get on-board leading to an increase in travel time equal to the service headway (if passengers do not consider boarding other lines).

In recent years schedule-based approaches to transit assignment have been further developed to capture dynamic aspects of transit assignment (see Wilson and Nuzzolo, 2004, for a summary). This work does however not overcome the need to also consider dynamic assignment for networks where the services operate with a high frequency (possibly without a timetabled schedule) and where passengers make their route choices by deciding for one line (or a group of lines) without opting for a particular service of a line.

In the following an approach to frequency-based transit assignment is discussed considering capacity constraints explicitly as well as time intervals in order to model the build-up of congestion over time. A "fail-to-board probability" **q** is introduced which is set in such a way that all the available space is used but all demand exceeding the available capacity remains on the platform. It is assumed that those passengers who failed to board attempt to continue their journey from the same platform in the next time interval (including a chance that they might fail to board again at this platform). Further, "long trips" that can not be completed in one time interval are carried over to the next period. These passengers are also assumed to continue their journey from the last node they reached in the previous time interval.

## 2. Other approaches to dynamic capacity constrained transit assignment

In general there two main approaches to transit assignment need to be distinguished. The classic, rather macroscopic frequency-based approach assigns passengers to lines. Compared to this, the major advantage of schedule-based approaches is that vehicle loadings can be predicted for specific runs. 'This approach allows us to take into account the evolution in time of both supply and demand, as well as run loads and level of service attributes' (Nuzzolo, 2003). Since Tong and Richardson (1984), scheduled-based assignment has been gaining in popularity. Theoretical advances as well as case studies have been published, for example Wilson and Nuzzolo (2004). However, schedule-based models are not advantageous in every situation as discussed in Schmöcker and Bell (2005). One might prefer to use a frequency-based approach for a number of reasons. Firstly, the models require less detailed input data and a less detailed network representations. This also often leads to advantages in run time. Frequency-based models might therefore be preferred for the strategic modelling with large scale networks. Secondly, if passenger arrivals and/or vehicle departures include some random element, the common line problem is easier to handle with frequency-based models. Thirdly, schedule-based models assume First-In-First-Out behaviour. Mingling among passengers who are already waiting for a long time and those who have just arrived, which happens at least to some degree on long platforms, can however be more easily modelled by frequency-based models.

Frequency-based modelling constitutes the classical approach as it is simpler, requiring less input data and less computational power. Advanced frequency-based route choice models consider *strategies* as introduced by Spiess and Florian (1989). These modelling techniques have been developed to reflect the choice passengers face in a public transport network where a number of lines would bring a passenger to his destination. Often a passenger at a stop has a choice of lines, referred to as *common lines*, which will take him directly or indirectly to his destination. The lines may differ in their attractiveness, perhaps due to the travel time to the destination, the number of changes, the probability of seat availability, etc.

In their seminal paper Spiess and Florian further suggest the "effective frequency" approach as a way of reflecting the priority of those on board over boarders in the competition for space. This approach is followed up by De Cea and Fernández (1993). The travel costs of transit arcs are assumed to be constant but the travel costs of waiting links are dependent on the link flows. With increasing flow the waiting time is strictly monotone increasing. The effective frequency is then defined as the inverse of the waiting time. The idea behind this is that with more buses arriving full, the waiting time will increase, because it is harder to get onto the next vehicle. If the vehicle arrives empty the effective frequency at this station is equal to the nominal frequency. But a service might still carry more passengers than it has capacity because only for line flows reaching

infinity the waiting cost becomes infinity and hence the effective frequency zero. The vehicle capacity acts more like a "practical capacity" which is part of the waiting cost function. De Cea and Fernández (1993) acknowledge that "practical capacities" do not solve the problem of overcrowding. However, they argue that this approach might be sufficient for strategic planning where one is not concerned about passenger loads for specific runs but is looking at the demand for the service over a longer time period.

Cominetti and Correa (2001) therefore instead use a formulation of the effective frequency $f_i'$ based on queuing theory: If $v_i \rightarrow \bar{v}_i$ then $f_i' \rightarrow 0$ and $w_i \rightarrow \infty$ where $\bar{v}_i$ is the saturation flow and $w_i$ is the waiting time on the link. With this formulation of "strict capacities", line loads will not exceed capacity. The assignment algorithm assumes that passengers travel on shortest hyperpaths and Dijkstra's algorithm is used for finding shortest hyperpaths. As in De Cea and Fernández, the distribution of flows across links of a hyperpath is proportional to the effective frequency of this link. Cepeda et al (2005) continue the work of Cominetti and Correa (2001) show that a local equilibrium of the assignment is also the global equilibrium. Therefore it is possible to use a gap function that becomes zero if the solution is equal to the optimal one. The gap function is based on Wardrop's first principle and shows the difference between the cost of all passengers using the cheapest path and the cost for all passengers with the current assignment. Based on these findings they propose a heuristic solution algorithm using the *Method of Successive Averages* that solves congested transit assignment for large-scale networks. Several case-studies in major cities around the world are briefly described demonstrating the validity of this approach.

In summary, the effective frequency approach has been shown to successfully model large scale overcrowded networks. It might however be criticised for the assumption of a continuously increasing cost function, which means that passengers are deterred from boarding a line even if the flow is below capacity. It also does not allow an explicit estimation of the number of passengers that are not able to board the vehicle. In the following an approach is presented which seeks to overcome these shortcomings. Further, none of the above approaches consider dynamic effects. As the case study will show, providing sufficient capacity for the demand of the extended morning peak might still lead to severe congestion during the peak of the peak. Overcrowding within short time intervals present safety risks for the whole service and lead to delays and hence reduced overall capacity during long periods of time.

## 3. Network description and notation

Figure 1 illustrates the nodes and arcs used for the network description by illustrating one station. Besides the origin and the destination there are four node types. The *stop node* represents the platform at which passengers wait for the service to arrive. At the *alighting node* passengers who stay on-board and those who alight at this station are separated whereas at the *boarding node* those passengers staying on-board from previous nodes and those passengers boarding anew are mingled. To consider the capacity constraints, *failure nodes* are introduced. In this approach passengers are penalised by not being transferred further downstream in this time interval but have to attempt to board again in the subsequent time interval.

There are nine arc types which connect the above described nodes. *Line arcs* (LA) correspond to transit lines and connect a boarding node with the alighting node of the next station downstream. *On-board wait arcs* (OBWA) are used by passengers not alighting at this station. *Access walking*

*arcs* (AWA) and *egress walking arcs* (EWA) connect the origin with the stop node and the stop node with the destination respectively. There is exactly one access arc and one egress arc in every hyperpath. *Alighting arcs* (AA), connect alighting and stop nodes, and *boarding demand arcs* (BDA), connect a stop node with a fail node. In order to reflect the reduced waiting time if passengers include common lines the waiting time is not associated with the boarding demand arc or the *boarding arc* (BA) but instead with the stop node. For those passengers who get onto the service, i.e. do not fail at the failure node, there is no additional cost, meaning that there is no cost associated with *boarding arcs*. The amount of passengers exceeding the available capacity is transferred back to the stop node via the *failure arcs* (FA). Finally, *transfer arcs* (TA) represent walking between platforms.

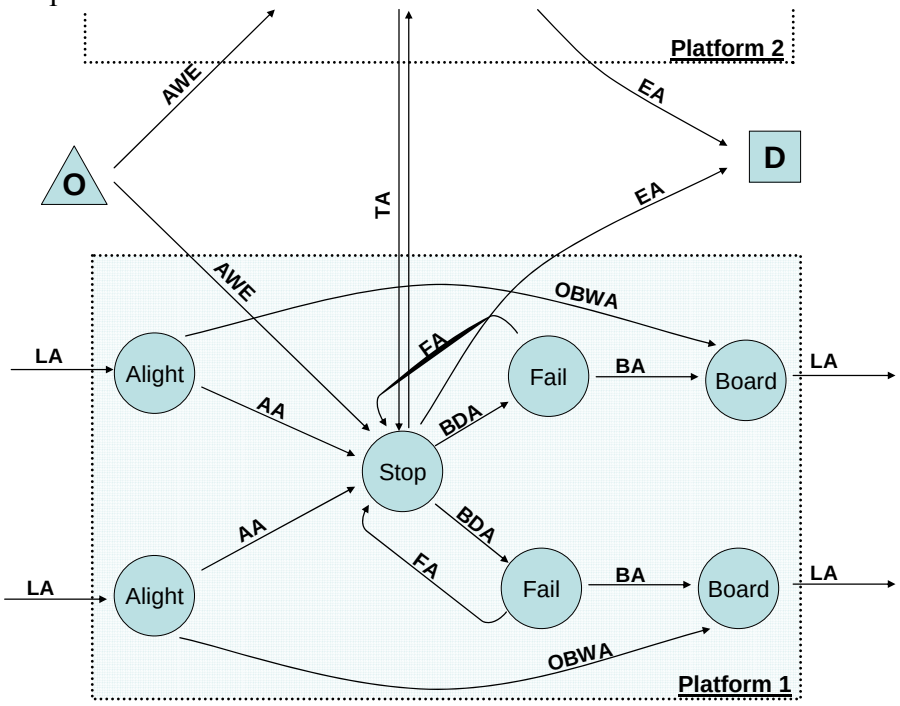



**Figure 1** Nodes and Arcs at a station with multiple platforms and lines

Further, following notation will be used for the network description:

| | |
|---|---|
| $L$: | Set of transit lines（with $l \in$ L) |
| $cap_l$: | Capacity of a service on line $l$ |
| $f_l$: | Frequency of transit line $l$ |
| $U_l$: | Set of platforms served by line $l$ (with $u \in U_l$) |
| $I$: | Set of nodes (with $i \in I$) and |
| $S$: | Set of stop nodes (with $s \in$ S and $s_u$ denoting a stop node at platform $u$) |
| $B$: | Set of boarding nodes |
| $E$: | Set of failure nodes (with $e \in E$) |
| $A$: | Set of arcs (with $a \in$ A) |
| $c_a$: | Cost of travelling on arc $a$ connecting nodes $i$ and $j$ |

In opposite to above variables flow and path choice dependent variables are changing in the time intervals. However, because in the "quasi-dynamic" model formulation these variables always refer to the current time interval $t$ (unless indicated), a superscript $t$ is for simplicity not added.

## 4. Cost function and hyperpath search

In the search for the shortest route to each destination, the common line problem is considered. Therefore, not choosing shortest path $p$, but choosing a set of a paths called hyperpath $h$ minimises the travel time or generalised costs. In order to search for $h$ the arc transition probabilities $\pi_{ah}$ need to be defined. For every node $i$ the arc transition probabilities are non-negative and satisfy:

$$\sum_{a \in A_{hi}} \pi_{ah} = 1, \ \forall i \in I_h \tag{1}$$

In (1) and in all following equations the subscript $h$ indicates that the set of arcs or nodes concerns only those included in the hyperpath. At failure nodes following relationship holds:

$$\pi_{ah} = \begin{cases} 1 - q_i & if\ In(a) \in B \\ q_i & otherwise \end{cases}, \ \forall i \in E_h \tag{2}$$

where $q_i$ is the fail-to-board probability at node $i$.Since there are always exactly two arcs leading out of a failure node (one boarding arc and one failure arc (2) also fulfils (1). The fail-to-board probability $q_i$ causes an expected delay of $TID*q_i$ for passengers traversing node $i$ in this time interval where $TID$ is the duration of the time interval. The passenger must further consider that there is a probability that he also fails to board in the subsequent time interval. Therefore the probability of failing to board in the current time interval but being able to board in time interval $t$+1, $t$+2, $t$+3,… is $q(1-q), q^2(1-q), q^3(1-q), ...$ respectively which causes an expected delay of

$$TID * (q(1-q) + 2q^2(1-q) + 3q^3(1-q) + ...) = TID \frac{q}{1-q} \tag{3}$$

Therefore the generalised cost of travelling on hyperpath $h$, $g_h$, becomes

$$g_h = \sum_{a \in A_h} \alpha_{ah} c_a + \sum_{i \in S_h} \beta_{ih} w_{ih} - \theta \sum_{i \in E_h} \beta_{ih} \left(\frac{TID \cdot q_i}{1-q_i}\right) \tag{4}$$

with $\alpha_{ah}$ as the probability of using arc $a$ and $\beta_{ih}$ as the probability of traversing node $i$ when travelling on hyperpath $h$. $w_{ih}$ is the expected waiting time at node $i$ under consideration of all lines that are included in $h$. Further, $\theta$ is the person's value of likely delays through overcrowding. If one would weight the first two terms of the generalised cost (on-board travel time and waiting time) with a parameter $\gamma$ for "normal travel time" it is probably realistic to assume $\gamma < \theta$ because of "delay frustration". Therefore in this case where $\gamma$ is omitted a value $\theta > 1$ is reasonable. The hyperpath search algorithm used in this approach is the same as the one described in Kurauchi et al (2003) which follows the one suggested by Nguyen and Pallotino (1988). In the same way as in Kurauchi et al (2003) it is easy to show that with cost function (4) the hyperpath costs are node separable which means Bellman's principle applies. Therefore the hyperpath search and the resulting transition probabilities can be formulated destination-specific, i.e. it is not necessary to formulate the optimal transition probabilities OD or hyperpath specific.

## 5. Network loading

Let $\Pi_d = [\pi_{ij}]_d$ denote the transition probability matrix for trips destined to $d$ on a previously calculated set of hyperpaths (because Bellman's principle applies we can omit subscript $h$ in this

section). Further $y_{id}$ is defined as the demand from node $i$ to destination $d$ (in time intervals $t > 1$ a starting point with $i \notin O$ is also possible). Then for a static formulation of the network loading the vector of passengers traversing intermediate nodes $i$, $\mathbf{v}_d$, can be obtained by (5) following Bell (1995). In (5) and following equations the dash indicates matrix transposition.

$$\mathbf{v}_d^{static} = (\mathbf{I} + \Pi_d + \Pi_d^2 + \Pi_d^3 + \Pi_d^4 + ...)' \mathbf{y}_d^{static} = \left( \left[ \mathbf{I} - \Pi_d \right]^{-1} \right)' \mathbf{y}_d^{static} \quad (5)$$

For a dynamic formulation one needs, however, to consider that passengers might not traverse all nodes in the same time interval. For this let us define the matrix $\Delta_d$ with elements $\delta_{ij}$, which takes the value of 1 if node $j$ is reachable from node $i$ in one time interval and otherwise zero. $\Delta_d$ is clearly dependent on *TID*, further the subscript $d$ is required because the travel time between nodes $i$ and $j$ is dependent on the set of hyperpaths used to $d$. Then the passenger volume that could reach node $j$ can be expressed as $\Delta' \mathbf{y}_d$ and the node volumes of passengers destined to $d$ can be written as:

$$\mathbf{v}_d = \left( \left[ \mathbf{I} - \mathbf{\Pi}_d \right]^{-1} \right)' \Delta_d' \mathbf{y}_d \quad (6)$$

Similar to the transition probabilities let us define the arcs also in terms of head and tail nodes, i.e. $a = (i,j)$ with $i := Out(a)$ and $j := In(a)$. Then it follows that the arc volumes $x_{ij}$ can be obtained from the node volumes as in (7):

$$\mathbf{x} = \sum_d \mathbf{x}_d = \sum_d \left( diag(\mathbf{v}_d) \cdot \mathbf{\Pi}_d \right) \quad (7)$$

The passengers who cannot reach their destination within one time interval are re-assigned in the following time intervals. In this study, all passengers are assigned at the beginning of the time interval and will be re-assigned from the furthest tail node they can reach within the time interval. To formulate this, let us define $\mathbf{K}^r_d$ with elements $[\kappa^r_{ij}]$ as an arc connecting $(i,j)$ satisfying following criteria:

$$\kappa_{ij}^r = \begin{cases} 1 & \text{if } tt_{ri} \le TID \text{ and } tt_{rj} > TID \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where $r$ is the start node of trips in this time interval, which can be an origin but also another node if the trip is continued over several time intervals. Note that because of common lines there might be several final nodes from each start node $r$ that can be reached within one time interval. An algorithm to find the travel time matrix $\mathbf{TT}_d := [tt_{ij}]_d$ showing the travel times between nodes $i$ and $j$ is described in Schmöcker (2006). With this definition of $\mathbf{K}^r_d$ indicating the "final nodes reachable in one time interval" the number of passengers who could not finish their trip and who have to be reassigned in the following time interval are calculated as in (9):

$$\mathbf{mu}_d = \sum_r \left( \left[ \mathbf{I} - \mathbf{\Pi}_d^* \right]^{-1} \right)' \mathbf{K}_d^{r\,\prime} \mathbf{y}_d \quad (9)$$

The star in (9) indicates that $\mathbf{mu}_d$ is calculated only after the equilibrium solution has been found. Figure 2 shows the flow chart of the dynamic assignment. It illustrates that the equilibrium for one time interval is calculated before moving on to the next time interval. Therefore the unfinished trips are added to the OD demand of the next time interval as formulated in (10):

$$\mathbf{y}_d^{t+1} \leftarrow \mathbf{y}_d^{t+1} + \mathbf{mu}_d \quad (10)$$

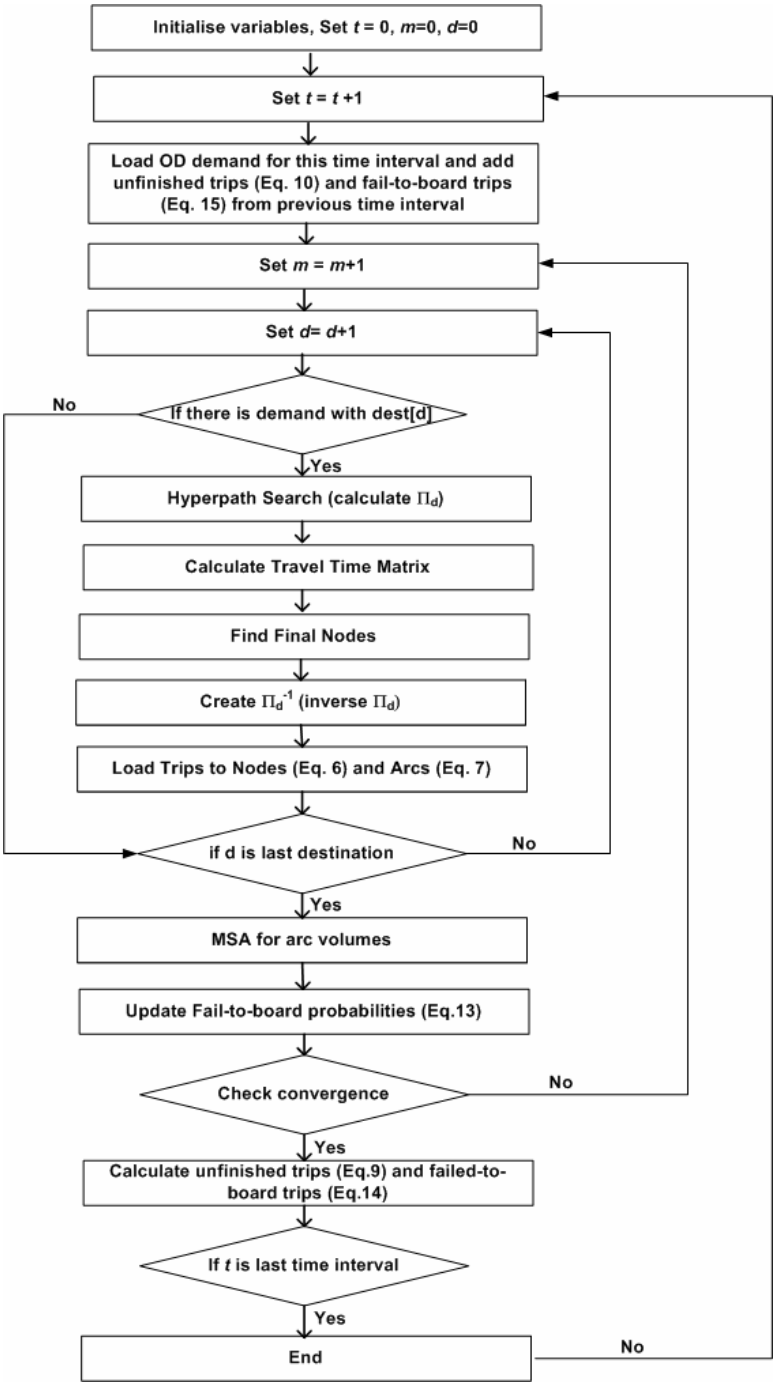



**Figure 2** Flow Chart of the quasi-dynamic assignment procedure

## 6. Ensuring capacity constraints and reassignment of passengers who failed-to-board

In (2) the transition probabilities are dependent on the failure probabilities **q** which have to be estimated. For each line arc (11) must be satisfied, where $x_{LA_{ul}}$ is the flow on an arc of line $l$ leaving platform $u$ and $cap_l$ is the The flow on a line arc consists of those staying on board $x_{OB}$ plus those wishing to board $x_{BDA}$. Those staying on board have priority over those wishing to board as they have boarded the service earlier. This means that the fail-to-board probability $q_{ul}$ needs to be adjusted in such a way that (12) is satisfied.

The adjustment is done by (13), which implies that at the equilibrium either $q_{ul}$ is non-zero or constraint (11) does not need to be enforced (there are still spaces available on the service).

$$cap_l \geq x_{LA_{ul}} \quad \forall u \in U_l, l \in L \tag{11}$$

$$x_{LA_{ul}} = x_{OB_{ul}} + x_{BA_{ul}} = x_{OB_{ul}} + (1 - q_{ul}) x_{BDA_{ul}}, \quad \forall u \in U_l, l \in L \tag{12}$$

$$q_{ul} := 1 - \max\left(0, \min\left(\frac{\left(cap_l - x_{OB_{ul}}\right)}{x_{BDA_{ul}}}, 1\right)\right), \quad \forall u \in U_l, l \in L. \tag{13}$$

Similar to unfinished trips, passengers who fail to board are assumed to continue their journey in the subsequent time interval. $mq_{d\,s_u}$ is the sum of those who failed to board at the failure nodes of lines $l$ that are served at the same platform $u$. As illustrated in Figure 1, each platform has one stop node and passengers are re-assigned from this stop node so that the number of passengers starting their journey from $s_u$ in the subsequent time interval can be calculated with (14):

$$mq_{d\,s_u} = \sum_{l \in L_u} q^*_{ul} x^*_{d,BDA_{ul}} \quad \forall l \in L, \forall u \in U \tag{14}$$

where $d$ is the destination of the travellers and as in (9) the star indicates an equilibrium value, i.e. $\mathbf{mq}_d$ is calculated after the equilibrium is found. Also in the same way as for unfinished trips $\mathbf{mu}_d$, the passengers who failed to board are added to the origin demand before the assignment in the next time period.

$$\mathbf{y}_d^{t+1} \leftarrow \mathbf{y}_d^{t+1} + \mathbf{mq}_d \tag{15}$$

## 7. Numerical Example

In order to illustrate the above approach, it is demonstrated with the small network shown in Figure 4. The demand is assumed to be 100 passengers travel from each station to each station downstream (600 passengers in total, split into 100 passengers for OD pairs: A→B, A→C, B→C, B→D, C→D). In order to illustrate the treatment of the excess demand it is assumed that there is no new demand after one hour. However, the simulation period is extended to 3 hours to let all passengers arrive at their destination.

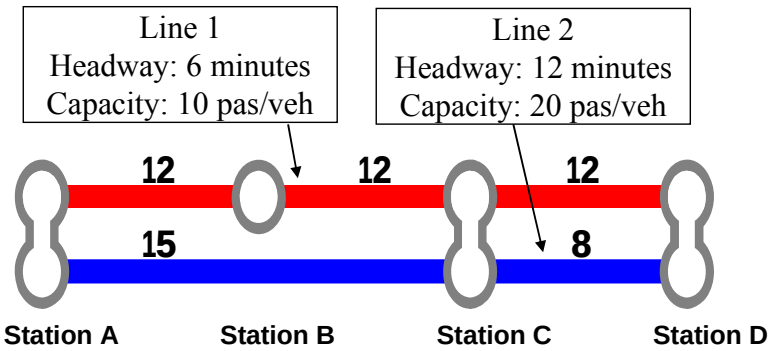



**Figure 4** Small example network

The three hour simulation period are modelled with *TID* equal to 60min, 30min and 15min. In case of *TID* = 30min (15min) it is assumed that the demand is equally spread between the two

(four) intervals of the first hour. In order to allow a better comparing of these results with the static results $\theta$ is set to 0.

In Table 1 the fail-to board probability and the number of passengers failing to board during the first hour are identical to the results gained with the static model. Those passengers that failed to board are then reassigned in the second and third hour. In the second time interval the demand at Station B is still higher than the available capacity ($q_B$=0.45) so that 63.9 passengers only arrive during the third time interval at their destination even though they started their journey during the first time interval. The line loads in Table 1 confirm that Line 1 leaving Station B is still fully occupied and in the second time interval and is also still used in the third time interval.

Modelling shorter time intervals allows a more detailed analysis of the overcrowding in the network. Assignment with *TID* = 30 and especially with *TID*=15 reveal the gradual increase and later decrease of the congestion at Station B better. In case of *TID*=30 it can be seen that the overcrowding at B is worse after 30min ($q_{B,L1}^{t=2}$=0.82) compared to the first time interval ($q_{B,L1}^{t=1}$= 0.7). This effect is hidden if one models too long time intervals. Note further that in the model with *TID*=15 the fail-to-board probability $q_{C,L2}^{t=1}$= 0 in the first time interval in contrast to the 0.25 for the assignment with *TID* =30 and *TID* =60. This is because it takes more than 15min for passengers from A to arrive at C so that the model recognises that the first group of passengers from Station C do not compete for space with passengers from Station A.

**Table 1** *TID* = 60min a) Boarders and Fail-to-board probabilities, b) Line flows

| | | 1st hour | | 2nd hour | | 3rd hour | |
|---|---|---|---|---|---|---|---|
| Station | Line | q | Board | q | Board | q | Board |
| A | L1 | 0.4 | 100 | 0 | 63.3 | 0 | 0 |
| A | L2 | 0.25 | 100 | 0 | 36.7 | 0 | 0 |
| B | L1 | 0.7 | 60 | 0.45 | 76.7 | 0 | 63.9 |
| C | L1 | 0 | 66.7 | 0 | 5.6 | 0 | 0 |
| C | L2 | 0.25 | 25 | 0 | 2.8 | 0 | 0 |

| Line Arc | | 1st hour | 2nd hour | 3rd hour |
|---|---|---|---|---|
| L1 | A→B | 100.1 | 63.2 | 0.0 |
| L1 | B→C | 100.0 | 100.1 | 63.7 |
| L1 | C→D | 96.4 | 44.0 | 31.9 |
| L2 | A→C | 100.1 | 36.6 | 0.0 |
| L2 | C→D | 99.7 | 27.8 | 0.0 |

**Table 2** *TID* = 30min a) Boarders and Fail-to-board probabilities, b) Line flows

| | | 0-30min | | 30-60min | | 60-90min | | 90-120min | | 120-150min | | 150-180min | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Station | Line | q | Board | q | Board | q | Board | q | Board | q | Board | q | Board |
| A | L1 | 0.4 | 50 | 0.57 | 50 | 0.2 | 50 | 0 | 11.4 | 0 | 0 | 0 | 0 |
| A | L2 | 0.25 | 50 | 0.41 | 50 | 0 | 37.4 | 0 | 1.6 | 0 | 0 | 0 | 0 |
| B | L1 | 0.7 | 30 | 0.82 | 30.4 | 0.77 | 31.6 | 0.57 | 46.8 | 0.2 | 50 | 0 | 13 |
| C | L1 | 0 | 33.3 | 0.04 | 34.8 | 0 | 4.1 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | L2 | 0.25 | 12.5 | 0.27 | 13.2 | 0 | 2.1 | 0 | 0 | 0 | 0 | 0 | 0 |

| Line Arc | | 0-30min | 30-60min | 60-90min | 90-120min | 120-150min | 150-180min | 180-210min |
|---|---|---|---|---|---|---|---|---|
| L1 | A→B | 50.1 | 50.1 | 49.9 | 11.4 | 0.0 | 0.0 | 0.0 |
| L1 | B→C | 50.0 | 49.9 | 50.2 | 49.9 | 49.8 | 12.9 | 0.0 |
| L1 | C→D | 48.1 | 50.0 | 20.0 | 23.3 | 24.9 | 6.5 | 0.0 |
| L2 | A→C | 50.0 | 50.1 | 37.3 | 1.6 | 0.0 | 0.0 | 0.0 |
| L2 | C→D | 12.3 | 50.5 | 38.8 | 25.8 | 0.0 | 0.0 | 0.0 |

**Table 3** *TID* = 15min a) Fail-to-board probabilities, b) Line flows

| Station | Line | 0-15 | 15-30 | 30-45 | 45-60 | 60-75 | 75-90 | 90-105 | 105-120 | 120-135 | 135-150 | 150-165 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | L1 | 0.4 | 0.57 | 0.66 | 0.72 | 0.6 | 0.32 | 0 | 0 | 0 | 0 | 0 |
| A | L2 | 0.25 | 0.41 | 0.52 | 0.6 | 0.36 | 0 | 0 | 0 | 0 | 0 | 0 |
| B | L1 | 0.5 | 0.8 | 0.86 | 0.89 | 0.88 | 0.87 | 0.84 | 0.75 | 0.63 | 0.41 | 0 |
| C | L1 | 0 | 0.25 | 0.16 | 0.19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | L2 | 0 | 0.25 | 0.37 | 0.37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| Line | Arc | 0-15 | 15-30 | 30-45 | 45-60 | 60-75 | 75-90 | 90-105 | 105-120 | 120-135 | 135-150 | 150-165 | 165-180 | 180-195 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| L1 | A→B | 0.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 10.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| L1 | B→C | 0.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 17.7 | 0.0 |
| L1 | C→D | 0.0 | 16.7 | 25.0 | 25.0 | 25.0 | 13.1 | 7.7 | 7.9 | 8.3 | 11.1 | 12.5 | 12.5 | 8.8 |
| L2 | A→C | 0.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 16.0 | 1.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| L2 | C→D | 8.3 | 6.3 | 25.4 | 25.2 | 20.9 | 18.1 | 17.5 | 9.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

## 8. Concluding remarks

This paper presented a dynamic approach to frequency-based modelling in order to overcome some of the disadvantages so far experienced with frequency-based models if one wants to analyse capacity problems at a strategic level. The assignment procedure described is a dynamic one, the route choice procedure does however only partially take the dynamic effects into account: It is considered that the route choice might vary in different time intervals and that passengers might change their initial route choice if they encounter too much congestion. The limitation is that the initial route choice for trips spanning more than one interval cannot respond to conditions in future intervals. This is an area for further work, if one finds that passengers' risk averseness is significant. The advantage of the current simpler approach is however, that none of the variable values needs to be stored for more than one time-interval and a time-dependent and memory consuming notation of the variables can be avoided.

References


Bell, M G H (1995) Alternatives to Dial's logit assignment algorithm. *Transportation Research B*, **29B**, 287-295.

Cepeda, M.; Cominetti, R. and Florian, M. (2005) A transit network model with strict capacity constraints: Characterization and computation of equilibria. *Accepted for publication in Transportation B*

Cominetti, R and Correa, J (2001) *Common lines and passenger assignment in congested transit networks.* Transportation Science, **35**, 250-267.

De Cea, J. and E. Fernández (1993) *Transit assignment for congested public transport system: An equilibrium model.* Transportation Science, **27(2)**, 133-147.

Kurauchi, F., Bell, M.G.H, Schmöcker, J.-D (2003) Capacity Constraint Transit Assignment with Common Lines. *Journal of Mathematical Modeling and Algorithms,* **2(4)**, pp. 309-327.

Nguyen, S and Pallottino, S (1988) *Equilibrium traffic assignment for large scale transit networks.* European Journal of Operational Research, **37**, 176-186.

Nuzzolo, A. (2003). *Schedule-Based Transit Assignment Models.* In: Advanced Modeling for Transit Operations and Service; edited by William H.K. Lam and Michael G.H. Bell. Pergamon.

Schmöcker, J-D. and Bell, M.G.H. (2005) The build up of Capacity Problems during Peak Hour. A dynamic Frequency-based model. Presented at 2nd International Workshop on Schedule-Based Approach to Dynamic Transit Modelling (SBDTM), Ischia (Naples, Italy), May 29-30

Schmöcker, J.-D. (2006) *Dynamic Capacity Constrained Transit Assignment*. Ph.D. thesis, Imperial College London, U.K., April 2006.

Spiess, H and Florian, M (1989) Optimal Strategies: A new assignment model for transit networks. *Transportation Research*, **23B**, 83-102.

Tong, C.O. and Richardson, A. J. (1984). *Estimation of time-dependent origin-destination matrices for transit networks.* Journal of Advanced Transportation, **18**, 145-161.

Wilson, N H M and Nuzzolo, A (2004). *Schedule-Based Dynamic Transit Modeling. Theory and Applications*. Kluwer Academic Publishers.

# TOWARDS A HOLISTIC FREQUENCY-BASED TRANSIT ASSIGNMENT MODEL – THE STOCHASTIC PROCESS APPROACH

*Fitsum Teklu: Institute for Transport Studies, University of Leeds, UK, f.teklu02@leeds.ac.uk;*
*David Watling: Institute for Transport Studies, University of Leeds, UK, d.p.watling@its.leeds.ac.uk;*
*Richard Connors: Institute for Transport Studies, University of Leeds, UK, r.d.connors@its.leeds.ac.uk.*

## Abstract
When planning line frequencies and fares in multimodal transit networks, frequency-based transit assignment models that consider the differences in the modes, fares, and passengers' preferences are required. Stochastic Process (SP) models that seek an equilibrium probability distribution of route flows and costs considering daily variations in route costs have been developed for road traffic assignment, and were shown to give solutions that subsume UE and SUE solutions as particular cases. As a first step toward building an analytic model for multimodal, multi-user-class transit assignment that considers day-to-day dynamics, a SP model is presented in this paper. Monte Carlo simulation is used to model passengers' information acquisition and decision process, based on a micro-simulation model for passengers' and buses' movements. A Random Utility Model is adopted for passengers' route choice. Passengers' expectations of the route costs are updated each day based on the experiences of ghost objects made to travel the alternative routes without contributing to the costs. Applications of the model are presented.

## 1 INTRODUCTION
Public transport users in many cities have a choice of different transit modes with different level of service and cost attributes for the different social classes. For example, small minibuses (of 10-15 seats) operating together with bigger buses are common in some cities of developing countries. Forecasting patronage levels in such networks requires, as applicable, consideration of: parallel lines, possibly of different costs or modes; different and, sometimes, limited vehicle capacities; different fares and fare structures; passengers' preference towards different modes; and number and type of inter- and intra- modal transfers likely to be made.

Frequency-based transit assignment models forecast mean passenger route flows/proportions and costs for a transit system characterised by average line frequencies or headways (as opposed to time tables). Both UE and SUE models have been presented in the literature (see section 3) assuming passengers minimize their travel costs considering some of the issues listed in the preceding paragraph. These models are mostly formulated as fixed-point problems, that seek long-term steady state route costs and flows.

Stochastic process (SP) models are an area of growing research interest. They simulate the convergence of a system towards an equilibrium state taking into account passengers' information acquisition and decision-making processes with full consideration of temporal fluctuations and transient conditions, without imposing an equilibrium condition. In addition, they provide important variability statistics of their estimates which decision-makers could use to determine the risks associated with the different options they appraise.

With the intention of ultimately developing a model for day-to-day dynamics from which analytic approximations to passenger route costs and flows could be derived, this paper presents such an SP model. The following two sections present brief literature reviews of SP and frequency-based transit assignment models, respectively. Section 4 discusses the Monte Carlo simulation based framework for evaluating the SP model. Numerical applications are presented in the succeeding section after which conclusions are made in section 6.

## 2 STOCHASTIC PROCESS MODELS
SP models for route choice represent the day-to-day evolving interaction between transit system costs and passengers' information acquisition and choice processes, based on assumptions of passengers' choice updating behaviour and learning and forecasting mechanisms. In discrete-time process models for dynamic evolution of a system, passengers are assigned on the different routes based on the mean perceived costs and their distributions, giving a distribution of route flows. Passengers' experiences of different routes are used to

continuously update passengers' route perceptions. SP models that have the Markov property – which states the conditional probability distribution of future states of a SP, given the present state, depend only upon the current state – also have a unique stationary probability distribution of the equilibrium state, for a certain demand-supply system (e.g. Meyn and Tweedie, 1993). From the stationary distribution, comprehensive statistical descriptions of the equilibrium state of the system could be obtained.

SP models have been applied in road traffic assignment models. Cascetta (1989) introduced the approach and discussed sufficient conditions for the stationarity of the process. Under certain conditions, similar results are reported when comparing the mean flows obtained from the model with steady state assignment models based on the SUE paradigm; divergent results are observed when multiple equilibria exist. Cantarella & Cascetta (1995) propose conditions for the existence and uniqueness of stationary probability distributions. Taking advantage of the existence of such distributions, Watling (1996) investigated the use of SP models for some commonly encountered asymmetric problems for which solution uniqueness is not guaranteed. Using simple examples, these distributions are shown to have a single peak or multiple peaks based on the uniqueness of the solution and stability of different points of equilibrium. Nuzzolo et al. (2001) discuss the application a similar approach for schedule-based transit assignment models, where explicit time table information is used to describe network supply.

Monte Carlo simulation is the obvious approach to generate pseudorandom observations for such dynamic processes over some given time horizon, and to obtain estimates of the equilibrium mean and covariance matrix of (for example) route flows. Approximations to the mean and covariance matrix obtained from such simulation-based methods have been presented by Davis & Nihan (1993) and Hazelton & Watling (2004), respectively, for particular classes of such models.

Without imposing equilibrium conditions, applications of SP models have been shown to converge to unique stationary probability distributions for road traffic assignment problems under certain conditions. SP models are particularly attractive for modelling asymmetric problems with multiple equilibria. To the authors' knowledge, their use has not been investigated for frequency-based transit assignment models.

# 3 FREQUENCY-BASED TRANSIT ASSIGNMENT

This section gives a brief review of research in frequency-based transit assignment, concentrating on passenger route choice behaviours, issues of handling capacity constraints and multimodal networks, and highlighting some major developments and outstanding issues.

In recognition of parallel services that might be available to passengers, Spiess & Florian (1989) introduce the notion of optimal strategies that assumes passengers have a fixed subset of "attractive" lines chosen for every stop they might encounter on their trip, and (at each stop) board the first arriving bus from that set to arrive at their destination. The attractive lines are identified by assuming passengers only consider lines that minimize their expected travel time. Nguyen & Pallottino (1988) provided a graph theoretic framework for this work that represented a strategy as a hyperpath. De Cea & Fernandez (1989) define a route as a (fixed) sequence of transfer stops. The set of attractive lines is chosen between each pair of transfer stops to minimize expected travel time, to make up what are called "route-sections". This could be rather restrictive as it only considers transit lines that "visit" both termini of the route section in the attractive lines set. The two approaches are compared in De Cea et al. (1988).

To model the effect of vehicle capacities on passenger waiting times and route choice, De Cea & Fernandez (1993) propose BPR-type additional waiting cost term that depends on the passenger flow-to-capacity ratio. The so-called effective-frequency approach, presented in Cominetti & Correa (2001) and Cepeda et al. (2006), uses flow-dependent decreasing functions, instead of fixed nominal frequencies, to represent impacts on waiting costs. Both these approaches use continuously increasing waiting cost functions which might over-estimate user costs. Kurauchi et al. (2003) incorporate strict line capacity constraints through failure-to-board probabilities. A generalized cost specification that includes the risk of failing to board is used to consider passengers' risk-aversion to highly congested stops. Waiting time impacts are, however, not considered until capacity is exceeded. Hamdouch et al. (2004) assume passengers' route choice strategies specify ordered sets of attractive lines, where the probability of boarding a line is proportional to the residual capacity of the line and inversely proportional to the number of users that want to board the line. All the above methods aggregate all the runs of the line over the period they model and might not realistically

capture the variability of the waiting times, especially when the vehicle capacities and/or line frequencies are relatively small.

Research in multimodal transit assignment models that consider different modes in the same framework have been presented in the literature. Lozano & Storchi (2002) presented an expanded, multimodal version of the work by Nguyen & Pallottino (1988) which forecast shortest viable hyperpaths, eliminating paths that do not respect constraints on the sequence of used modes and exceed a pre-specified number of modal transfers. Lo et al. (2003) note the unrealistic type of intra-modal transfers and number of transfers that might arise from equilibrium formulations and the difficulty of modelling non-linear fare structures by equilibrium models. They suggest a transformation of the base network to a state-augmented network that avoids such problems. Nielsen (2000) presented a Probit-based SUE formulation for multimodal transit assignment where passengers' preferences for different modes and cost attributes are simulated along side the perceived costs. He highlights the need to define more robust rules for when and how different lines should be aggregated to represent correlations and covariance in overlapping routes' costs. Except for Nielsen (2000) which implemented BPR-type functions, the other works cited here have not considered capacity constraints.

Despite major developments in the issues listed in section 1, research into multimodal multi-user-class frequency-based transit assignment model that consider strict capacity constraints are scarce.

# 4 MODEL IMPLEMENTATION

In this section, the analytic model for the SP and SIMTRANSIT, the Monte Carlo simulation model built to implement the SP approach to multimodal frequency-based transit assignment, are presented. First, the micro-simulation model for the buses and passengers, which is the core for simTRANSIT based on which the SP model is evaluated, is presented.

## 4.1 Bus and Passenger Simulation Model

A micro-simulation type approach is used to model the buses' and the passengers' movements. This will allow separate simulation of the different lines' runs (i.e. without aggregating all the runs over the modelling period) to obtain better representation of the passengers' experiences. Using a simulation time step-length of $\omega$, the model processes bus arrivals, passenger arrivals, and bus departures sequentially; each initiating the different procedures briefly described as follows. Firstly, at each time step, due passengers and buses are generated based on assumed headway distributions. Then, alighting passengers are allowed to do so for every arriving bus, after which bus departure times would be assigned. Transferring passengers join the queue for their next service. Next, newly generated passengers are made to choose their route and queue accordingly, following the first-in-first-out rule. Finally, queuing passengers are made to board the departing bus if it is in their attractive lines set and has spare capacity. After allowing passengers to board it, the buses are assigned arrival times to the next stop in their itinerary. A constant stop-to-stop travel time is assumed in this study.

The micro-simulation approach adopted allows the direct observation of waiting times without having to specify difficult-to-calibrate BPR-type functions to represent effects of strict vehicle capacities and ensures buses are not loaded beyond their capacities. Different theoretical or empirical passenger and line headway distributions could also be modelled explicitly. In addition, bus dwell times at transit stops and different fare structures could also be easily implemented in such a framework.

## 4.2 Multimodal Route Choice Model

In this study, the network supply representation presented in De Cea & Fernandez (1993) is used. As noted in section 3, this approach assumes passengers' routes are defined by a sequence of transfer stops. Passengers are not allowed to change their choice of transfer stops once they set off on their journey. For the small network shown in Figure 2, all alternatives routes are enumerated allowing passengers to determine the active route sections, and hence the attractive lines in the network, as the SP evolves. Although they could easily be implemented in the approach presented in this paper, it should be noted that mode-specific costs are not considered.

A stochastic transit assignment model is proposed assuming passengers do not have an exact knowledge of the "true" travel times and perceive waiting differently. Let $\mathbf{U}$ = vector of user classes, $\mathbf{D}$ = vector of OD

pairs, and $\mathbf{R}_h$ = set of routes for OD pair $h \in \mathbf{D}$. For each route $r \in \mathbf{R}_h$ and $u \in \mathbf{U}$, passengers' perceived cost ($PC_{r,u}$) is the sum of the deterministic generalised cost ($GC_{r,u}$) and an error term ($\xi_{r,u}$) as:

$$PC_{r,u} = GC_{r,u} + \xi_{r,u}\ , \quad \forall r \in \mathbf{R}_h, \forall u \in \mathbf{U} \tag{1}$$

The deterministic user-class specific generalized cost is composed of in-vehicle travel cost ($T$), waiting cost ($W$), and fares ($F$), see equation (2).

$$GC_{r,u} = \gamma_{(t),u} \cdot T_r + \gamma_{(w),u} \cdot W_r + F_r \tag{2}$$

where, $\gamma_{(t),u}$ and $\gamma_{(w),u}$ are the values of travel time and waiting time, respectively, for user-class $u$.

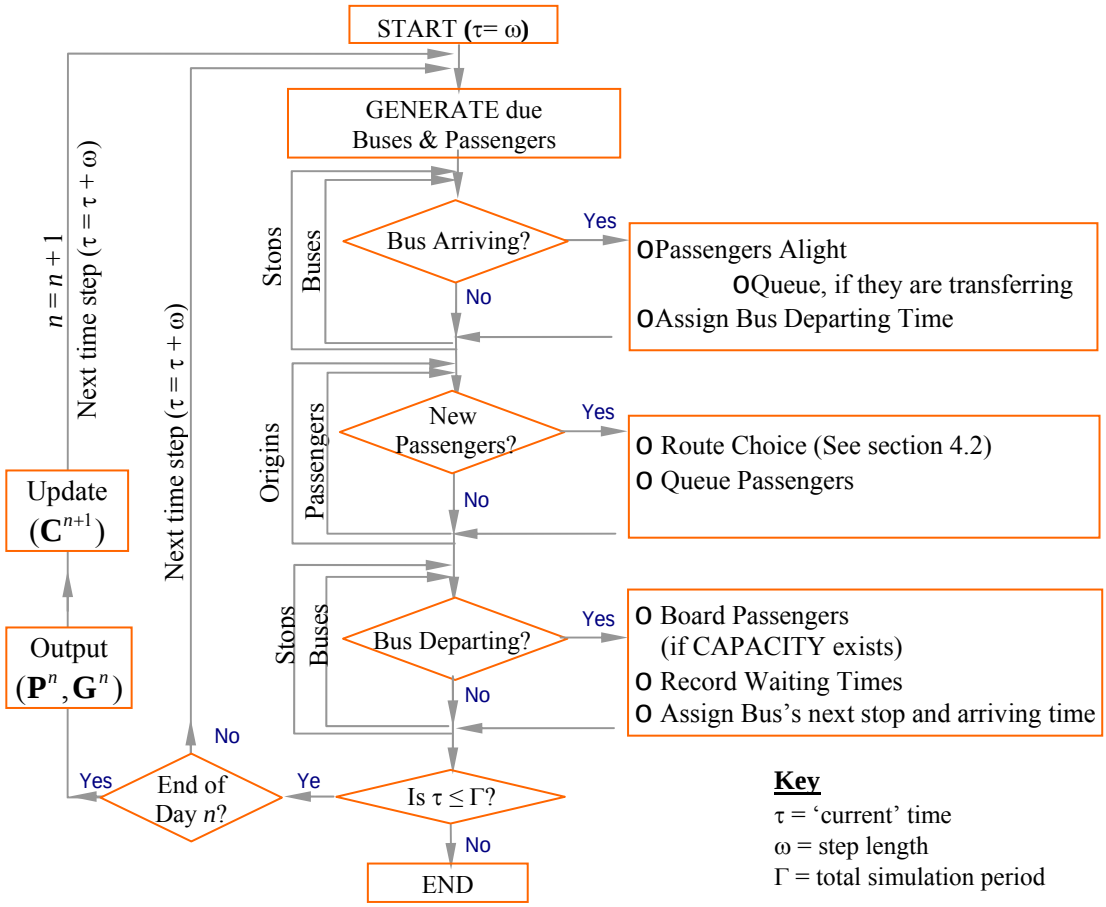



Figure 1 SIMTRANSIT Program Flow

To maintain realism, the model should consider the cost dependencies (in mean and variance) of the different route sections for each pair of transfer stops. For example, the waiting and travel costs of route sections that contain a particular line (the RED line in route sections B and D in Figure 2c) should be correlated. As a route section could be part of different routes, the costs of the routes that contain it should be correlated as well; e.g. routes 1 and 2 in Figure 2. These requirements negate the use of discrete choice models such as the Logit model which suffer from the IIA nuisance.

It is also necessary to obtain cost estimates for routes that would somehow not be sampled by the passengers. To obtain democratic cost sampling across the alternative routes, ghost objects are generated at a constant rate and made to travel the alternative routes. They experience the costs on all the routes, but do not contribute to them. These ghosts also extract the dependencies in the mean costs of route sections (due to shared lines) and routes (due to shared route sections). In simTRANSIT (Figure 1), the ghosts are treated in the same way as the passengers, except in the buses where they do not contribute to the buses filling up.

The variance in passengers' perceived costs is modelled using error terms starting from the lines, through the routes sections, up to the routes as follows. Let $\mathbf{K}_r$ = the vector of route sections for route $r \in \mathbf{R}_h$ and $\mathbf{A}_k$ =

the set of attractive lines that constitute the route section $k \in \mathbf{K}_r$. Then, the variance in the perceived cost $k \in \mathbf{K}_r$, ${\sigma_k}^2$, is given by equation (3). For uniformly and exponentially distributed passenger and line headways, respectively, the average uncongested waiting costs and frequency weighted average line travel cost is used as a basis for calculating ${\sigma_k}^2$. $\eta$ is a model parameter which can be interpreted as the variance of the perceived costs over a route section of unit cost. It is externally given and, in this paper, not assumed to be mode or user-class dependent. Admittedly, equation (3) assumes waiting times for lines that are included in multiple route sections are independent, which might not be realistic; this is an outstanding research issue which is not addressed in this study.

$$ {\sigma_{k,u}}^2 = \eta \cdot \left( \frac{\gamma_{(w),u} + \sum_{\forall l \in A_k} f_l \cdot \gamma_{(t),u} \cdot t_l}{\sum_{\forall l \in A_k} f_l} \right), \quad \forall k \in \mathbf{K}_r, \forall r \in \mathbf{R}_h, \forall h \in \mathbf{D}, \forall u \in \mathbf{U} \qquad (3) $$

where: $\eta > 0$, $f_l$ = frequency of line *l*, and $t_l$ = in-vehicle travel time on line *l*

Assuming the perceived costs of the non-overlapping route sections of the alternative routes are independent, the elements of the covariance matrix are given by:

$$ \mathrm{cov}(\xi_{r,u}, \xi_{s,u}) = \sum_{\forall k \in \mathbf{K}_r \cap \mathbf{K}_s} {\sigma_{k,u}}^2 \quad , \forall r, s \in \mathbf{R}_h \, , \forall h \in \mathbf{D} \qquad (4) $$

The variance in a routes' perceived costs are given by:

$$ {\rho_{r,u}}^2 = \mathrm{cov}(\xi_{r,u}, \xi_{r,u}) = \sum_{\forall k \in \mathbf{K}_r} {\sigma_{k,u}}^2 \quad , \forall r \in \mathbf{R}_h \qquad (5) $$

For each passenger, the error terms for the alternative routes are drawn from a normal distribution: $\xi_{r,u} \sim \mathrm{N}(0, {\rho_{r,u}}^2)$. The joint distribution of the perception errors across all routes is then multivariate normal, i.e. $MVN(0, \Omega)$ - where, $\Omega$ is the covariance matrix with elements $\mathrm{cov}(\xi_{r,u}, \xi_{s,u})$. In this study, it is assumed that the covariance matrix is constant through out the evolution of the SP.

### 4.3 Learning Process Model

The basic assumption behind SP models is that the state of the system on a particular day $n$ is random. The vector of the expected cost on the alternative routes, $\mathbf{C}^n$, which is based on the ghosts' experiences is chosen to describe the state of the system on day $n$. The route flow proportions vector $\mathbf{P}^n$ is based on $\mathbf{C}^{n-1}$. A Monte Carlo simulation model approach is followed to obtain estimates of $\mathbf{P}^n$'s and $\mathbf{C}^n$'s over a long period, from which the mean route flow proportions and costs are calculated.

The learning process model employs simple weighted averages to update $\mathbf{C}^{n+1}$ using the ghosts experiences from the previous day, $\mathbf{G}^n$. This captures the effect of passengers choices on day $n$ on the route costs based on which passengers' update their expected costs. For a particular route $r \in \mathbf{R}_h$, $G^n_{r,u}$ is calculated as a simple average of the experiences of all ghosts that used the route, as shown in equation (7). Using a "*learning parameter*" ($\phi$) as weight, $\mathbf{G}^n$ is used to update $\mathbf{C}^{n+1}$, for the next day, as shown in equation (6). Even though, past experiences will have increasingly lesser impacts as the process evolves, all previous experiences of will be influencing future decisions through a nested functional dependence.

$$ \mathbf{C}^{n+1} = \phi . \mathbf{G}^n + (1 - \phi) \cdot \mathbf{C}^n \qquad (6) $$

where, $0 < \phi < 1$

$$ G^n_{r,u} = \sum_{1}^{g^n_r} GC_{r,u} \Big/ g^n_r \qquad (7) $$

where, $g^n_r$ = number of ghosts finishing their journey on day $n$.

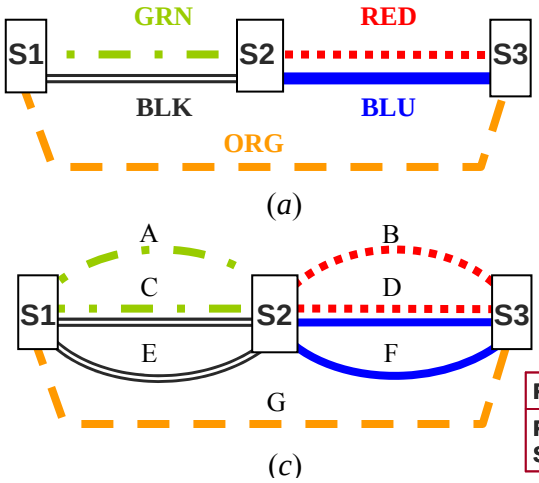
Figure 1 shows how the models described in this section are integrated to provide estimates of route flow proportions and costs considering the day-to-day variation in system costs. The learning process model

updates $\mathbf{C}^{n+1}$ and outputs $\mathbf{G}^n$ and $\mathbf{P}^n$ for the day that was just completed. This simulation is run continuously by increasing the simulation time by $\omega$ until a predetermined total period, $\Gamma$. In this model, a day is represented by one simulation hour to correspond with the demand and line frequency data.

Although expensive in terms of the time they require to give results, the Monte Carlo simulation provides a logical base upon which analytic models could be built. To that effect, the models and the micro-simulation type model presented in this section allow mathematical representations of the model to be written down.

# 5 NUMERICAL EXPERIMENTS

In this section some results of the proposed model are presented using a test network (Figure 2). The network has two modes and five lines connecting three stops. Lines GRN, RED, and ORG are of the "Expensive" (Exp) mode, which is relatively faster, more expensive and of a lesser capacity than the "Cheap" (Chp) mode, see Figure 2b for line specific average hourly frequencies (Freq), vehicle capacities (Cap), travel time, and fare details. The route sections that passengers consider to use between transfer stops are indicated in Figure 2c. For the single OD pair (S1 to S3) considered in this study, the alternative routes are enumerated in Figure 2d. So, a passenger choosing route "2" which is made up of route sections "A" and "D", waits for the GRN line at S1 and, when reaching S2, change to the first line of RED and BLU that arrives first.




(*a*)

| Line | Mode | Freq [/hr] | Cap [/veh] | Time [min] | Fare [Pence] |
|---|---|---|---|---|---|
| ORG | Exp | 20 | 15 | 20 | 200 |
| GRN | Exp | 15 | 15 | 10 | 100 |
| RED | Exp | 15 | 15 | 10 | 100 |
| BLK | Chp | 5 | 60 | 15 | 35 |
| BLU | Chp | 5 | 60 | 15 | 35 |

(*b*)




(*c*)

| Route | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Route Sections | A & B | A & D | A & F | C & B | C & D | C & F | E & B | E & D | E & F | G |

(*d*)

Figure 2 Test Network, Route Sections, and Alternative Routes

Two distinct user-classes are defined in this study, where user-class I represents passengers from a relatively lower social class from the more affluent user-class II. For user-class I the values of waiting and in-vehicle travel time are 3 and 2 pence/min, respectively; the corresponding values for user-class II passengers' are 10 and 5 pence/min, respectively. As mentioned earlier, passengers are assumed to be indifferent to lines of different modes, and have a multimodal attractive lines set from which the board the first arriving line. But passengers' mode related preferences, that are irrelevant to the route, could be readily incorporated in the generalized cost function (equation (2)) by defining mode specific constants for each user-class.

The line and passenger inter-arrival headways are assumed to follow exponential and uniform distributions, respectively. To maintain reality, the line headways are truncated to a maximum of 30 minutes (and the mean of the generating, untruncated distribution is adjusted accordingly to preserve the required mean of the truncated distribution). The learning parameter, $\phi = 0.1$ and the ghosts are generated every 5 minutes.

## 5.1 Demand Scenario 1

For this case, an average hourly S1 to S3 demand of 150 and 200 passengers is considered for user-classes I and II, respectively, giving a total demand-to-supply ratio of ~42%. The perceived costs variability parameter, $\eta = 0.2$.

To provide a basis for comparison the average costs of the 10 alternative routes are calculated using equation (2); see Table 1. The waiting time, in-vehicle travel times, and fare were calculated as:

$W_r = \sum_{\forall k \in r}\left(\sum_{\forall l \in A_k} f_l\right)^{-1}$, $T_r = \sum_{\forall k \in r}\left(\sum_{\forall l \in A_k} f_l \cdot t_l \Big/ \sum_{\forall l \in A_k} f_l\right)$, and $F_r = \sum_{\forall k \in r}\left(\sum_{\forall l \in A_k} f_l \cdot \tau_l \Big/ \sum_{\forall l \in A_k} f_l\right)$, respectively; $\tau_l$ is the fare paid on line *l*. As noted earlier $W_r$ does not consider the impact of passengers' failing to board full transit vehicles. From the table, it can be seen that, routes 9 and 10 are the least cost routes for user-classes I and II, respectively.

Table 1 Average generalized cost calculations without considering congestion impacts

| | Average Route Costs (pence) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Routes | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| User-class I | 264.0 | 247.3 | 233.0 | 247.3 | 230.5 | 216.3 | 233.0 | 216.3 | 202.0 | 249.0 |
| User-class II | 380.0 | 360.0 | 420.0 | 360.0 | 340.0 | 400.0 | 420.0 | 400.0 | 460.0 | 330.0 |

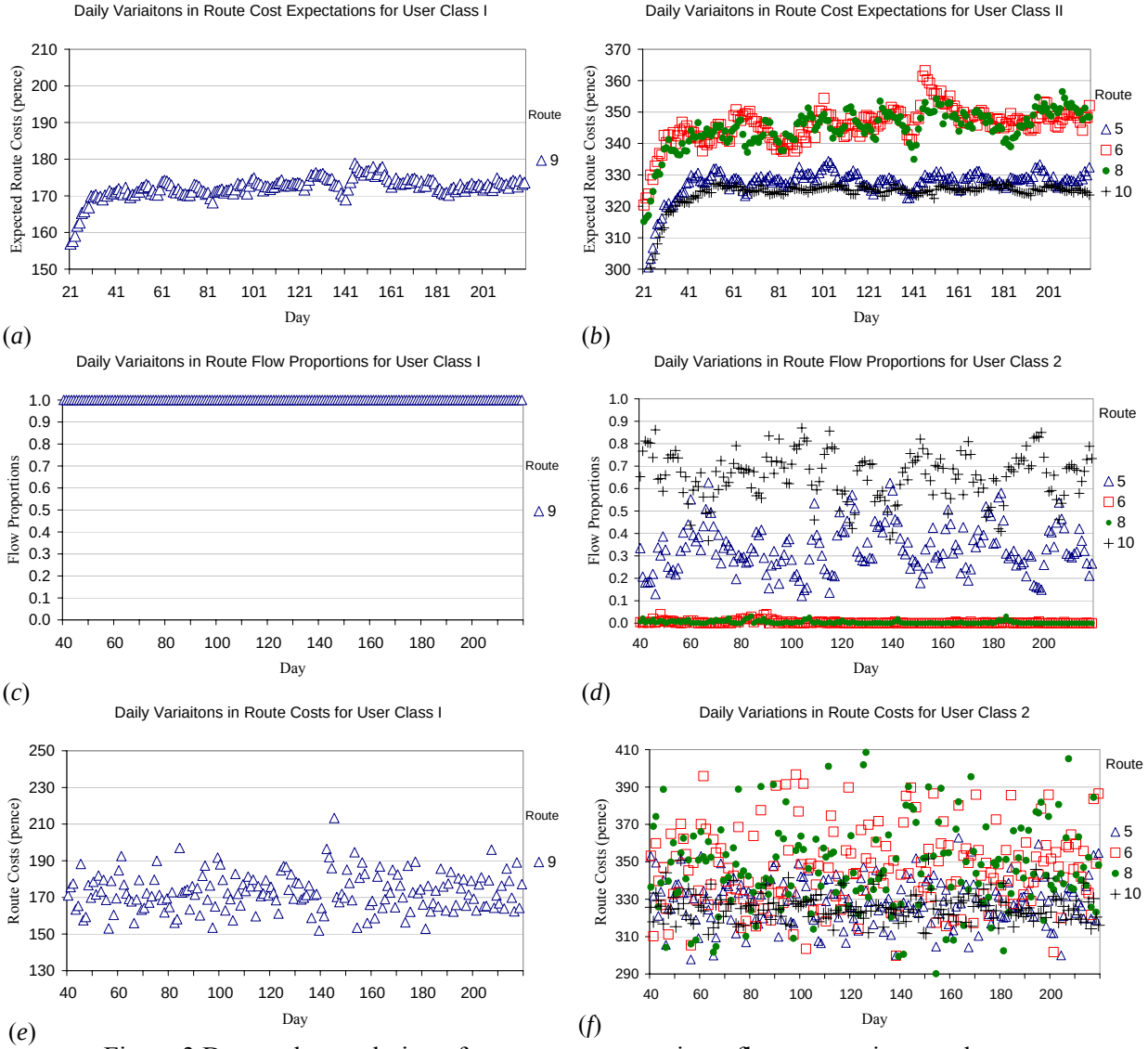



Figure 3 Day-to-day evolution of route cost expectations, flow proportions, and costs

After running simTRANSIT for a total of 220 days, the evolution of average daily route cost expectations ($GC^n_{r,u}$), route flow proportions ($P^n_{r,u}$), and route costs ($G^n_{r,u}$) for the different user-classes, are presented in Figure 3. Considering a burn-in period of 20 days, during which no outputs are given, only data from routes that are, on average, chosen by more than 0.1% of the passengers over the whole SP are plotted. Each day is represented by a point on the chart. It should be noted that with-in day dynamics is not considered in this study. As explained earlier, the average route cost expectations (Figure 3*a*-*b*), updated each day considering

average route cost experiences (Figure 3*e-f*), drive the passengers' route choice decisions (Figure 3*c-d*). Summary statistics of $P_{r,u}^{n}$ and $G_{r,u}^{n}$ are presented in Table 2, using two more seeds to check the impact of initial conditions.

Passengers' route cost expectations are observed going up until c. 40 days, the period it takes the SP to learn the "true" route costs. After that, they stabilize, varying over a smaller range. The summary statistics in Table **2** are thus calculated based on 160 observations starting from the 40th day. Passengers from user-class II are primarily split on routes 5 and 10, with the former being used by ~1/3 and the latter by ~2/3 of the passengers on average. Route 10 is direct, not requiring passengers to transfer. Route 5, which considers both modes in the attractive lines set, enables these passengers to compensate for the extra travel time due to mode Chp with the lower fare, and the lesser waiting time to offer competitive generalized cost. The generalized costs for these routes vary between 300 and 360 pence and seem to be in some sort of equilibrium; the average daily costs are 328.3 and 325.2 pence, respectively. Routes 6 and 8 are observed to be used by a very small proportion of the users (< 2 passenger/day) on average; this seems to be due to the value of $\eta$ used which makes passengers' perceive a more expensive route to be cheap (see Figure 3*b, d*). All passengers from user-class I use route 9 (the Chp mode) throughout, incurring an average daily generalized cost of 172.6 pence. Unlike passengers from user-class II, this group are not willing to pay higher fares for lesser journey times.

Using different random number generating seeds, Table 2 gives the daily route flow proportions and cost proportions and the corresponding standard errors considering day-to-day variations. The means from the three runs are given in the last row. Compared to the costs shown in Table 1, the costs obtained from simTRANSIT are on the lower side. A possible explanation for this is the truncation of the exponential line headway distribution to a maximum headway of 30 minutes. Especially for the Chp mode which has a low frequency, correcting for the truncation, to obtain the same mean headway, implies a slightly higher frequency. For composite route sections, this increases the probability of the Chp mode arriving first at the stop and have a relative use more than that implied by the $f_l \Big/ \sum_{\forall l \in A_k} f_l$ ratio, based on nominal frequencies. This leads to decreases in the fares and (relatively smaller) increases in the generalized in-vehicle travel costs for routes including the multimodal attractive lines set. As the ghosts are assumed to board full vehicles, so long as they are at the front of the queue, these might also decrease cost estimates for the different routes.

Table 2 Route costs and flow proportions summary using different random number generating seeds

| 1 | Route Flow Proportions (%) | | | | | Route Costs (pence) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| User-class | I | II | | | | I | II | | | |
| Route†† | 9 | 5 | 6 | 8 | 10 | 9 | 5 | 6 | 8 | 10 |
| Seed 1 | 100 (0) | 33.32 (10.70) | 0.39 (0.70) | 0.42 (0.65) | 65.80 (10.63) | 173.0 (10.1) | 328.9 (12.7) | 347.2 (21.9) | 346.8 (22.8) | 325.4 (7.3) |
| Seed 2 | 100 (0) | 37.33 (13.27) | 0.23 (0.49) | 0.86 (1.64) | 61.53 (13.60) | 172.9 (9.2) | 328.0 (13.2) | 348.4 (26.3) | 344.0 (20.7) | 325.4 (6.6) |
| Seed 3 | 100 (0) | 35.79 (11.16) | 0.17 (0.54) | 1.03 (1.38) | 62.99 (11.49) | 171.8 (8.9) | 327.9 (12.5) | 348.3 (23.0) | 342.3 (21.8) | 324.8 (6.4) |
| Average | 100 | 35.48 | 0.26 | 0.77 | 63.44 | 172.6 | 328.3 | 348.0 | 344.4 | 325.2 |

† The standard deviations are given in parenthesis, under the means.
†† For both user-classes, only routes with average flow proportions more than 0.1% are presented in this table.

## 5.2 Demand Scenario 2

Here, an average hourly demand of 250 & 350 is assumed for user classes I and II, respectively, giving a demand-to-supply ratio of ~73%. The model is run for 220 days using a higher $\eta$ of 0.3, assuming greater variability in passengers' perceptions in a more congested network.

After a burn-in period of 40 days, the day-to-day evolutions of the route costs expectations for the two user classes are given in Figure 4; summary statistics for route flow proportions and costs are given in Table 3, respectively. Compared to demand scenario 1, the passengers' cost expectations are less stable and vary of a

bigger range. The average route costs are relatively higher due to the waiting time impact of full vehicles (see Figure 5). Higher day-to-day variability in route costs and flow proportions is also observed as the lines are operating closer to capacity, being fully loaded on some days requiring passengers to revise their route cost expectations. As the first scenario, user-class I passengers predominantly use route 9, using the Chp mode. Routes 5 and 10 are used by ~97% of user-class II passengers, with routes 2 and 4 being used by ~3% that choose a sequence of multimodal and Exp mode route sections.

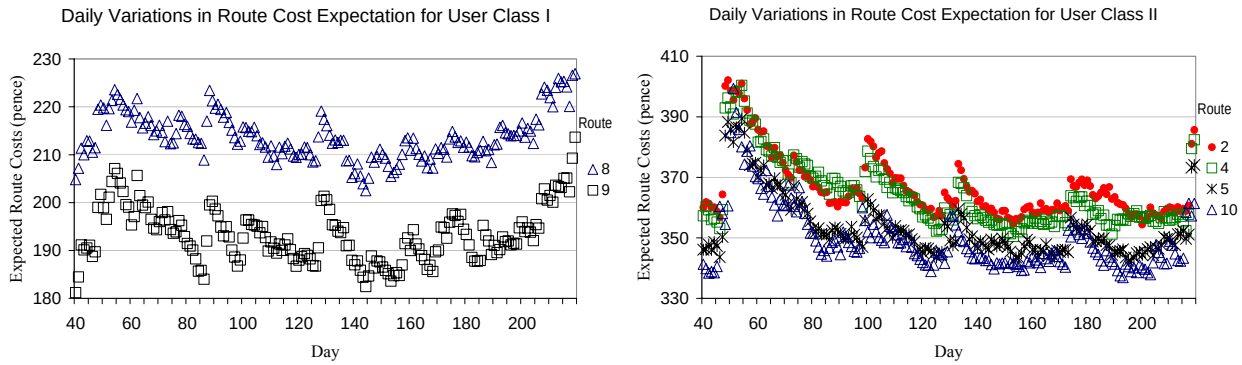



Figure 4 Day-to-day evolution of route cost expectations

Table 3 Route Costs and flow proportions summary† using different random number generating seeds

| | Route Flow Proportions (%) | | | | | | Route Costs (pence) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| User-class | I | | II | | | | I | | II | | | |
| Route†† | 8 | 9 | 2 | 4 | 5 | 10 | 8 | 9 | 2 | 4 | 5 | 10 |
| Seed 1 | 0.02 (0.07) | 99.98 (0.00) | 0.88 (1.20) | 2.39 (3.26) | 34.68 (16.21) | 61.95 (18.36) | 215.0 (22.4) | 195.0 (26.1) | 368.0 (40.0) | 365.3 (39.8) | 354.8 (40.0) | 350.7 (42.7) |
| Seed 2 | 0.05 (0.19) | 99.95 (0.00) | 0.85 (1.09) | 1.42 (1.44) | 34.23 (14.64) | 63.43 (15.92) | 211.4 (17.9) | 191.0 (20.3) | 365.8 (34.4) | 364.4 (33.0) | 353.0 (33.4) | 349.2 (42.5) |
| Seed 3 | 0.11 (0.62) | 99.89 (0.00) | 1.79 (2.16) | 1.43 (2.22) | 35.48 (18.14) | 61.36 (20.49) | 214.1 (22.0) | 193.3 (25.0) | 373.4 (53.4) | 374.8 (53.5) | 361.3 (52.1) | 357.9 (58.0) |
| Average | 0.06 | 99.94 | 1.17 | 1.75 | 34.80 | 62.25 | 213.4 | 193.1 | 369.1 | 368.2 | 356.4 | 352.6 |

† The mean and standard errors are given; the standard errors are given in parenthesis, under the means.
†† For both user-classes, only routes with average flow proportions more than 0.1% are presented in this table.

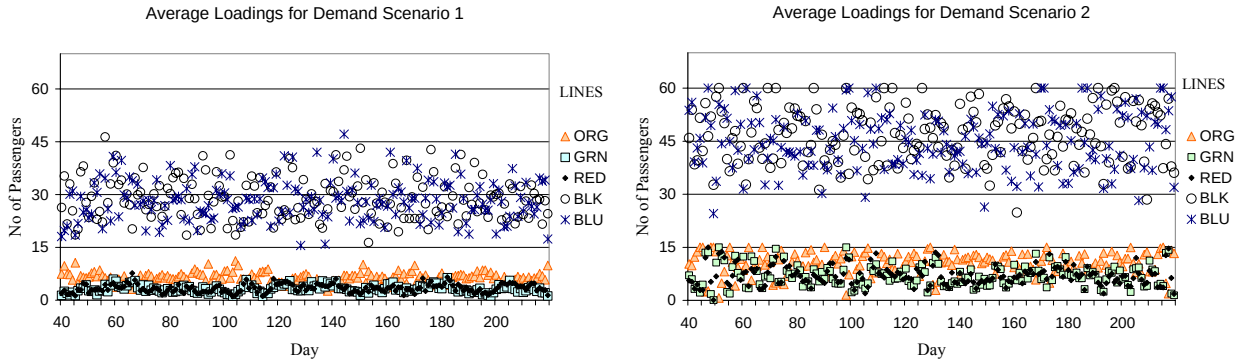



Figure 5 Day-to-day evolution of line loadings

# 6 CONCLUSIONS

In this paper a stochastic process approach for frequency-based transit assignment model considering day-to-day dynamics has been presented. The model is holistic in the sense that it is multimodal, considers the differences in behaviours of different social classes and is strictly capacity constrained. It was observed that the stochastic process is less stable with increased congestion. For the test network presented, higher variability in forecasted route flows and costs was observed with increased congestion. This raises an interesting research question of how big such variabilities are on bigger networks.

Further work on the route choice model needs to look at: implementing more general passenger route negotiating strategies, arising types and numbers of transfers, and consideration of mode specific constants. Applying realistic error components that consider the relationships in passengers' route cost perceptions starting from the lines up to the route level is an outstanding issue. As mentioned, the major purpose of this SP model is to provide a base for an analytic frequency-based transit assignment model for day-to-day dynamics. To that effect, studies into the stationarity of the SP and it having the Markov Property are important. In addition, sensitivity testing of the different model parameters is necessary.

# 7 ACKNOWLEDGEMENTS

This work is part of Fitsum Teklu's PhD research which is financially supported by Universities UK and Institute for Transport Studies, University of Leeds.

# 8 REFERENCES

CANTARELLA, G. E. & CASCETTA, E. (1995) Dynamic Processes and Equilibrium in Transportation Networks: Towards a Unifying Theory. *Transportation Science* 29**,** 305-329.

CASCETTA, E. (1989) A Stochastic Process Approach to the Analysis of Temporal Dynamics in Transportation Networks. *Transportation Research Part B,* 23**,** 1-17.

CEPEDA, M., COMINETTI, R. & FLORIAN, M. (2006) A Frequency Based Assignment Model for Congested Transit Networks with Strict Capacity Constraints: Characterization and Computation of Equilibria. *Transportation Research Part B,* 40**,** 437-459.

COMINETTI, R. & CORREA, J. (2001) Common Lines and Passenger Assignment in Congested Transit Networks. *Transportation Science,* 35**,** 250-267.

DAVIS, G. A. & NIHAN, N. L. (1993) Large Population Approximations of a General Stochastic Traffic Assignment Model. *Operations Research,* 41**,** 169-178.

DE CEA, J., BUNSTER, J. P., ZUBIETA, L. & FLORIAN, M. (1988) Optimal Strategies and Optimal Routes in Public Transit Assignment Models: an empirical comparison. *Traffic Engineering & Control***,** 520-526.

DE CEA, J. & FERNANDEZ, E. (1989) Transit Assignment to Minimal Routes: an Efficient New Algorithm. *Traffic Engineering & Control* 30**,** 491-494.

DE CEA, J. & FERNANDEZ, E. (1993) Transit Assignment for Congestion Public Transport Networks: an Equilibrium Model. *Transportation Science,* 27**,** 133-147.

HAMDOUCH, Y., MARCOTTE, P. & NGUYEN, S. (2004) Capacitated Transit Assignment with Loading Priorities. *Mathematical Programming B,* 101 205-230.

HAZELTON, M. & WATLING, D. (2004) Computation of Equilibrium Distributions of Markov Traffic Assignment Models. *Transportation Science,* 38**,** 331-342.

KURAUCHI, F., BELL, M. G. H. & SCHMÖCKER, J.-D. (2003) Capacity Constrained Transit Assignment with Common Lines. *Journal of Mathematical Modelling and Algorithms* 2**,** 309-327.

LO, H. K., YIP, C. W. & WAN, K. H. (2003) Modelling Transfer and Non-linear Fare Structure in Multi-Modal Network. *Transportation Research Part B,* 37**,** 149-170.

LOZANO, A. & STORCHI, G. (2002) Shortest viable hyperpath in multimodal networks. *Transportation Research Part B,* 36**,** 853-874.

MEYN, S. P. & TWEEDIE, R. L. (1993) *Markov Chains and Stochastic Stability,* London, Springer-Verlang.

NGUYEN, S. & PALLOTTINO, S. (1988) Equilibrium Traffic Assignment for Large Scale Transit Networks. *European Journal of Operational Research* 37**,** 176-186.

NIELSEN, O. A. (2000) A Stochastic Transit Assignment Model considering differences in passenger utility functions. *Transportation Research Part B,* 34**,** 377-402.

NUZZOLO, A., RUSSO, F. & CRISALLI, U. (2001) A Doubly Dynamic Schedule-based Assignment Model for Transit Networks. *Transportation Science,* 35**,** 268-285.

SPIESS, H. & FLORIAN, M. (1989) Optimal Strategies: A New Assignment Model for Transit Networks. *Transportation Research Part B,* 23**,** 83-102.

WATLING, D. (1996) Asymmetric Problems and Stochastic Process Models of Traffic Assignment. *Transportation Research Part B,* 30**,** 339-357.