Dynamic Origin-Destination Demand Flow Estimation under Congested Traffic Conditions

Chung-Cheng Lu
Department of Transportation and Logistics Management, National Chiao Tung University,
Hsinchu, 30050, Taiwan, jasoncclu@gmail.com

Xuesong Zhou (corresponding author)
Department of Civil and Environmental Engineering, University of Utah,
Salt Lake City, UT, 84112, USA. zhou@eng.utah.edu

Kuilin Zhang
Transportation Research and Analysis Computing Center, Energy Systems Division,
Argonne National Laboratory, Argonne, IL 60439, USA. kzhang@anl.gov

Abstract: This paper presents a single-level nonlinear optimization model to estimate dynamic origin-destination (OD) demand. The model is a path flow-based optimization model, which incorporates heterogeneous sources of traffic measurements and does not require explicit dynamic link-path incidences. The objective is to minimize (i) the deviation between observed and estimated traffic states and (ii) the deviation between aggregated path flows and target OD flows, subject to the dynamic user equilibrium (DUE) constraint represented by a gap-function-based reformulation. A Lagrangian relaxation-based algorithm which dualizes the difficult DUE constraint to the objective function is proposed to solve the model. This algorithm integrates a gradient-projection-based path flow adjustment method within a column generation-based framework. Additionally, a dynamic network loading (DNL) model, based on Newell's simplified kinematic wave theory, is employed in the DUE assignment process to realistically capture congestion phenomena and shock wave propagation. This research also derives analytical gradient formulas for the changes in link flow and density due to the unit change of time-dependent path inflow in a general network under congestion conditions. Numerical experiments conducted on three different networks illustrate the effectiveness and shed some light on the properties of the proposed OD demand estimation method.

Keywords: OD demand estimation; path flow estimator; Lagrangian relaxation; Newell's simplified kinematic wave theory.

1. Introduction

Time-dependent origin-destination (OD) demand matrices are fundamental inputs for dynamic traffic assignment (DTA) models to describe network flow evolution as a result of interactions of individual travelers. Moreover, many emerging intelligent traffic management applications call for reliable estimates of dynamic OD demand, in order to generate proactive, coordinated traffic information provision and flow control strategies based on reliable traffic state estimates. Transportation authorities and practitioners have long been concerned about the unavailability of high quality time-dependent OD demand estimates which limits the potential for DTA deployments to analyze and alleviate traffic congestion. In the past decades, a rich body of literature, to be presented as follows, has been devoted to the methods of estimating static or time-dependent OD demand tables. However, the development of theoretically sound and practically deployable approaches for time-dependent OD demand estimation, particularly under congested conditions, remains a critical and challenging problem that is attracting significant attention from transportation researchers.

1.1 Literature review

To capture congestion effects in traffic networks, many researchers have attempted to integrate equilibrium assignment into the static OD demand estimation process. Nguyen (1977) and LeBlanc and Farhangian (1982) incorporated link count observations into a variable demand user equilibrium (UE) assignment program as equality side-constraints so that the estimated link flows can reproduce observed link counts. Fisk (1989) combined the maximum entropy model with an UE assignment program to construct a bi-level mathematical programming problem. Yang et al. (1992) and Florian and Chen (1995) further presented a more flexible bi-level framework to estimate consistent OD demand, where the upper level is a generalized least squares (GLS)-based OD estimation model and the lower level is an UE assignment program.

Extending the concepts and solution methodologies of the static OD estimation problem, Cascetta et al. (1993) proposed a GLS estimator for dynamic OD demand in a general network. A simplified assignment model was used in their study; that is, path choice fractions are first calculated using a route choice model and then the resulting path flows are propagated to link flows based on link travel times. Tavana (2001) proposed a bi-level GLS optimization model and an iterative solution framework to estimate dynamic OD demand, while seeking to maintain internal consistency between the upper-level demand estimation problem and the lower-level DTA problem. Along this line, Zhou et al. (2003) and Zhou and Mahmassani (2006) extended this bi-level dynamic OD estimation approach to utilize multi-day traffic counts and automated vehicle identification (AVI) data, respectively. Van der Zijpp (1997) also addressed the dynamic OD demand estimation problem using data from induction loops and automated vehicle identification (AVI) equipment in the network. Based on a least-square modeling approach, Bierlaire and Crittin (2004) proposed an algorithm for sparse least squares that is computationally efficient for real-time estimation and prediction of dynamic OD demand tables. Zhou and Mahmassani (2007) developed a structural state space model for real-time traffic origin-destination demand estimation and prediction in a day-to-day learning framework.

In light of the above review, a typical bi-level dynamic OD demand estimation model needs to iteratively solve two optimization subproblems, namely upper-level and lower-level problems. The upper-level problem is the constrained ordinary/generalized least squares (OLS/GLS) problem, with time-dependent OD flows as decision variables. This approach aims to minimize the following two deviation functions: (i) the deviation between observed and estimated link flows over all time intervals, and (ii) the deviation between the target or historical demand and estimated demand matrices. The lower-level problem is the UE DTA problem, which determines a time-dependent network flow pattern that satisfies dynamic user equilibrium (DUE) conditions. However, it has been widely recognized that, under congested conditions, the mapping between demand inflow from the origin and link measurements is not a linear relationship as observed in the static case.

Yang (1995) provided two heuristic solution approaches for solving the general bi-level OD estimation problem, the iterative estimation-assignment (IEA) algorithms and sensitivity-analysis based algorithms (SAB). Tavana (2001) suggested that the IEA algorithm still provides solutions to the Cournot-Nash game, rather than the Stackelberg game in a bi-level program, because the upper-level optimization model in IEA does not consider the dependence of link-flow proportions on the OD flows. Based on the SAB approach, an alternative nonlinear least squares formulation was proposed by Tavana (2001) to explicitly consider the changes of link flow proportions due to the adjustments in dynamic OD flows. Additionally, numerical derivatives of link flow proportions with respect to OD flows are obtained from a mesoscopic DTA simulation program (Jayakrishnan et al., 1994; Mahmassani, 2001). On the other hand, a standard SAB algorithm needs to approximate the derivatives through simulation for each OD pair and each time interval at every iteration, which is computationally intensive, especially for large-scale networks. Recently, Balakrishna et al. (2008) and Cipriani et al. (2011) introduced gradient approximation methods within a simultaneous perturbation stochastic approximation (SPSA) framework in order to reduce the number of simulation runs when calculating numerical derivatives or gradients. Artificial intelligence techniques have recently

been adopted to the context of dynamic OD demand estimation to replace iterative procedures. For instance, Kattanand Abdulhai (2006) proposed a non-iterative approach to dynamic OD demand estimation based on a machine-learning technique using advanced parallel evolutionary algorithms. More recently, Huang et al. (2012) developed an approach to estimate travel demand in large-scale microscopic traffic simulation models based on a Guided Genetic Algorithm with a distributed implementation to improve computational efficiency and reduce memory requirements.

Intending to develop an internally consistent approach for the dynamic OD demand estimation problem, single-level path flow estimators (PFE) have been proposed for the static OD estimation problem; e.g., the linear programming PFE by Sherali and Park (2001) on estimating UE path flows, and the nonlinear programming PFE by Bell et al. (1997) on estimating stochastic UE path flows. Recently, Nie and Zhang (2008) formulated a novel single-level formulation based on variational inequalities (VI), which utilizes the dynamic link-path incidence relationships in a generic projection-based VI solution framework. Based on a single level reformulation, Lundgren and Peterson (2008) proposed a descent heuristic, which is an adaptation of the projected gradient method, to the dynamic OD demand estimation problem. The search direction was determined by approximating the Jacobian matrix representing the derivatives of the link flows with respect to a change in the OD-flows. By adapting the analytical approach of Ghali and Smith (1995) for evaluating the local link marginal travel times, Qian and Zhang (2011) further incorporated the travel time gradients into the single-level OD estimation framework proposed by Nie and Zhang (2008), in order to utilize travel time measurements, while linear mappings between OD flow and link flows are still assumed in their upper-level OD estimation problem. In addition, Nie et al. (2005), Nie and Zhang (2010), and Shen and Wynter (2011) integrated the integral term of UE formulation (Beckmann et al., 1956) with the measurement deviations (in terms of the GLS objective function) to develop alternative single-level path flow formulations for static OD estimation, and their models can be viewed as a special case of UE assignment with elastic demand.

To fully capture the nonlinear dependency between dynamic OD demand and heterogeneous traffic measurements, such as link flow, density and travel time, in a general network, it is necessary to incorporate a reliable dynamic network loading (DNL) component into the OD demand flow estimation process. The existing DNL methodologies are classified into two major groups: analytic approach and simulation-based approach. The former includes three types of formulations that have the potential for deriving theoretical insights: mathematical programming, optimal control and VI. Earlier studies (e.g., Merchant and Nemhauser, 1978; Carey, 1987; Friesz et al., 1989) used link/node exit flow constraints to propagate traffic flows and link performance functions to determine path travel costs. While using well-defined exit constraints and cost functions makes it theoretically possible to establish the mathematical properties of solutions (e.g., existence and uniqueness), these analytical models suffer from several limitations in terms of representing the dynamics and complexity of real-world traffic flow systems. For example, the models are still subject to difficulty in ensuring the first-in-first-out (FIFO) queueing discipline and capturing spillback queues.

In the pioneering works of capturing shock waves and congestion (i.e., queue formation, spillback and dissipation), Lighthill and Whitham (1955) and Richards (1956) proposed the kinematic wave (KW) theory, which rigorously describes traffic flow dynamics by integrating flow conservation constraints and traffic flow models to a set of partial differential equations (PDE). Because it is difficult to obtain analytical solutions for these PDEs, many researchers have presented various finite-difference approximations to solve the equations numerically. Based on a triangular flow-density relation, Newell proposed a simplified KW model (Newell, 1993) to better keep track of shock waves and queue propagation using cumulative flow counts on links. By discretizing a link into a set of homogeneous unit cells, Daganzo (1994, 1995) presented a cell transmission model that adopts a "supply- demand" or "sending- receiving" framework to model flow dynamics between two successive discretized cells in the link.

1.2 Overview and contribution of proposed method

This paper contributes to the growing body of literature on dynamic OD demand estimation as follows:

- Instead of working on the commonly-used OD flow variables, this study presents a new path flow-based optimization model for jointly solving the complex OD demand estimation and UE DTA problems. This model simultaneously minimizes (i) the deviation between measured and estimated traffic states, and (ii) the deviation between aggregated path flows and target OD flows, subject to a dynamic user equilibrium (DUE) constraint, which is reformulated using an equivalent gap function. Working in this path-flow dimension, our formulation can directly aggregate estimated path flows to obtain final OD flow patterns, and obviate explicit dynamic link-path incidences, as opposed to the majority of previous studies.
- By dualizing the difficult DUE constraint into the objective function, this research proposes an effective Lagrangian relaxation-based solution framework. The relaxed problem can be viewed as a simultaneous route and departure time user equilibrium (SRDUE) problem with elastic demand The final solution is a set of path flows

satisfying "tolled user equilibrium" (Lawphongpanich and Hearn, 2004), where the deviation with respect to traffic measurements can be viewed as an additional penalty for over-estimated or under-estimated path flows. By incorporating heterogeneous real-world measurements in the objective function, such as link densities from video surveillance and road side detectors, the proposed estimation model fully utilizes available information to reflect route choices in a congestion network.

- A DNL model that encapsulates Newell's simplified KW model in a mesoscopic traffic flow simulation framework is proposed to describe congestion phenomena, such as queues formation, spillback, and dissipation. Explicitly using the cumulative arrival and departure curves, Newell's traffic flow model provides a rigorous mathematical formulation to realistically represent traffic dynamics and capture the impact of shock waves on various macroscopic traffic measures.
- Based on the proposed DNL model, this research derives analytical, local gradients of different measurement types, such as link flow and density, with respect to path inflows. This valuable gradient information not only considers the dependences of link flow/density changes on the OD/path flow, but also allows for computing feasible descent directions in an efficient gradient-projection-based method embedded in the Lagrangian relaxation-based solution framework.

While the general relaxation scheme and some common features (e.g., path flow variables) are shared by our work and some previous research (e.g., Nie and Zhang, 2008), our approach differs significantly from the existing methods in the following aspects.

- The proposed dynamic OD demand estimation model is a GLS-based bi-level formulation that adopts a gap function to measure the deviation from the DUE conditions, while previous research used complementary equations or VIs to describe optimality or equilibrium conditions of the problem.
- The proposed Lagrangian relaxation-based algorithm dualizes the gap function-based DUE constraint into the objective function, and solves the single-level relaxation problem by reducing the difference between the upper and the lower bounds. On the other hand, previous research developed different (e.g., column generation) algorithms to solve the VI-based single-level model.
- Based on the proposed DNL model, this research derives analytical, local gradients of different measurement types with respect to path inflows. These gradients are different from the link and path marginal travel time or cost typically used in system optimal DTA problems. The proposed algorithm explicitly utilizes the gradient information to compute feasible descent directions in an efficient gradient-projection-based method embedded in the Lagrangian relaxation-based solution framework. Although previous research had applied basic projection algorithm (BPA) and extra-gradient method (EGM) to the dynamic OD demand estimation problem, the issue of gradient derivations has not been explicitly addressed.

This paper is organized as follows. The single-level time-dependent path flow estimation models and the mesoscopic DNL model based on Newell's simplified KW model are delineated in Section 2. Section 3 presents the Lagrangian relaxation-based solution algorithm. Section 4 describes the derivation of the gradients of typical traffic measurements. Numerical studies on a simple corridor and a real-world network with time-dependent sensor data are presented in Section 5. Section 6 concludes this study.

2. Single-level time-dependent path flow estimation models

2.1 Notation and problem statement

Notation and variables for the proposed single-level dynamic OD demand estimation framework are listed as follows.

Set:

A: set of links
W: set of OD pairs
P: set of paths
S: set of links with sensors, $S \subseteq A$
H_d: set of discretized departure time intervals
H_o: set of discretized observation time intervals

Index:

t: index of simulation time intervals, $t = 0,\ldots,T$. This paper refers to any particular time interval t as the time t.
τ: index of departure time intervals, $\tau \in H_d$
w: index of OD pairs, $w \in W$
p: index of paths for each OD pair, $p \in P$
l: index of links, $l \in A$

Input parameters of link attributes

length(l): length of link l

nlanes(l): number of lanes of link l

$k_{jam}(l)$: jam density on link l

$n^{max}(l)$: number of vehicles that can be stored on link l, $n^{max}(l) = k_{jam}(l) \times nlanes(l) \times length(l)$

$v_f(l)$: free-flow speed on link l

$w_b(l)$: backward wave speed on link l

FFTT(l): free-flow travel time on link l, $FFTT(l) = length(l) / v_f(l)$

BWTT(l): backward wave travel time on link l, $BWTT(l) = length(l) / w_b(l)$

$cap^{out}(l, t)$: maximal outflow capacity of link l at time t

$cap^{in}(l, t)$: maximal inflow capacity of link l at time t

Traffic measurements inputs

$\bar{q}(l,t)$: observed number of vehicles passing through an upstream detector on link l during observation interval t

$\bar{k}(l,t)$: observed density on link l during observation interval t

$\bar{d}(w)$: target demand, which is the total traffic demand for OD pair w over a planning horizon

Estimation variables

$r(w, \tau, p)$: estimated path flow on path p of OD pair w and departure time interval τ

$c(w, \tau, p)$: estimated path travel time on path p of OD pair w and departure time interval τ

$\pi(w, \tau)$: estimated least path travel time of OD pair w and departure time interval τ

$q(l, t)$: estimated number of vehicles passing through an upstream detector on link l during observation interval t

$k(l, t)$: estimated density on link l during observation interval t

$d(w, \tau)$: estimated demand of OD pair w and departure time interval τ

Given a road network G with a set of nodes and a set of links, the input data to the dynamic OD demand estimation problem include a set of time-dependent observation data (e.g. flows and densities) on a subset of links with sensors for a set of discretized observation time intervals, H_o, and target OD demand matrices for a set of departure time intervals, H_d. The problem aims to find a vector of time-dependent path flows $\boldsymbol{r} = \{r(w, \tau, p), \forall w, \tau, p\}$ such that the estimated flow pattern can both match the time-dependent observation data and satisfy the DUE conditions (i.e., for each OD pair w and each departure time interval τ, the travel time of a path with a positive path flow is the same and equal to $\pi(w, \tau)$, while the travel time of a path with zero flows is larger than $\pi(w, \tau)$). Then, the time-dependent OD demand estimates can be obtained by aggregating path flows for each OD pair and each departure time interval.

2.2 A path flow-based nonlinear program

Given sensor data (i.e. observed link flows and densities) and target (aggregated historical) OD demands, the proposed single-level time-dependent path flow estimation model is a nonlinear program with the path flows $r(w, \tau, p), \forall w, \tau, p$ and least path travel times $\boldsymbol{\pi} = \{\pi(w,\tau), \forall w, \tau\}$ as the decision variables. Denote $\boldsymbol{c} = \{c(w, \tau, p), \forall w, \tau, p\}$, $\boldsymbol{q} = \{q(l, t), \forall l, t\}$ and $\boldsymbol{k} = \{k(l, t), \forall l, t\}$. The nonlinear program is presented as follows.

P1: Nonlinear program

$$\text{Min } Z = \beta_d \sum_w \Big[\sum_{\tau \in H_d} \sum_p r(w,\tau,p) - \bar{d}(w)\Big]^2 + \sum_{l \in S} \sum_{t \in H_o} \Big\{\beta_q [q(l,t) - \bar{q}(l,t)]^2 + \beta_k \big[k(l,t) - \bar{k}(l,t)\big]^2\Big\} \quad (1)$$

Subject to

$$(\boldsymbol{c}, \boldsymbol{q}, \boldsymbol{k}) = DNLF(\boldsymbol{r}), \quad (2)$$

$$g(\boldsymbol{r}, \boldsymbol{\pi}) = \Sigma_w \Sigma_\tau \Sigma_p \{r(w,\tau,p)[c(w,\tau,p) - \pi(w,\tau)]\} = 0, \quad (3)$$

$$c(w,\tau,p) - \pi(w,\tau) \geq 0, \forall w, \tau, p, \quad (4)$$

$$\pi(w,\tau) \geq 0, \forall p \in P(w, \tau)\}, \forall w, \tau \quad (5)$$

$$r(w,\tau,p) \geq 0, \forall w, \tau, p. \quad (6)$$

The objective funciton, Eq.(1), minimizes the weighted sum of the deviation between estimated time-dependent OD demands (or aggregated path flows) and target demands and the deviation between estimated and observed link flows and densities, where β_d, β_q and β_k are the weights reflecting different degrees of confidence on target OD demands and observed link flows and densities, respectively. In a GLS framework, those weights can be viewed as the inverses of the variances of the distinct sources of measurements. In the objective function, OD demand $d(w,\tau)$ is substituted by its corresponding aggregated time-dependent path flows, so the OD demand and path flow balance constraints are not included in the model, P1.

Eq.(2) states that estimated path travel times (***c***), and link flows (***q***) and densities (***k***) are obtained by a given DNL function of path flows, *DNLF*(***r***). This research proposes the DNL model based on Newell's simplified KW model (Newell, 1993). The details of the DNL model are presented in Sections 2.3 and 2.4 of the current paper.

The gap function-based constraint in Eq.(3) is adopted to satisfy the DUE conditions in the proposed model, whereas other existing single-level OD demand estimation models incorporated VI (e.g., Nie and Zhang, 2008), or Beckmann's integral term (e.g., Shen and Wynter, 2011) to represent the DUE conditions. The non-negative gap function $g(\boldsymbol{r}, \boldsymbol{\pi})$ is used to measure the deviation from the DUE conditions, and it vanishes when the DUE conditions are satisfied (e.g., Lu et al., 2009). Eqs.(4) and (5) are definitional constraints of the least travel time $\pi(w,\tau)$ for each (w,τ). Eq.(6) are non-negative constraints for path flow variables.

2.3 Modeling queue spillback through Newell's simplified kinematic wave model

The proposed time-dependent path flow estimation models, P1, encapsulats the DNL model based on Newell's simplified KW model (Newell, 1993) to describe traffic congestion propagation (e.g., queue build-up, spillback, and dissipation) in a road traffic network. As shown in Fig. 1, Newell's model is concerned about three state variables on each link *l*: (i) cumulative flow count *A*(*l*, *t*) for vehicles moving into link *l* through the upstream node, (ii) cumulative flow count *V*(*l*, *t*) for vehicles waiting at the vertical queue of the downstream node of link *l* at time *t*, and (iii) cumulative flow count *D*(*l*, *t*) for vehicles moving out of link *l* through the downstream node.

Let *x* be the location along the corridor, and *N*(*x*, *t*) the cumulative flow count at location *x* and time *t* of a link. The change of *N*(*x*, *t*) along a characteristic line (wave) is represented as follows.

$$dN(x,t) = \frac{\partial N}{\partial x}dx + \frac{\partial N}{\partial t}dt = qdt - kdx. \quad (7)$$

A wave represents the propagation of a change in flow and density along the roadway, and the wave speed is the slope of the characteristics line $w = \frac{\partial q}{\partial k} = \frac{dx}{dt}$. Along the movement of a wave, we substitute $dt = \frac{dx}{w}$ into Eq.(7), so that we can link the difference of cumulative flow counts together through

$$dN(x,t) = qdt - kdx = \left(-k + \frac{q}{w}\right)dx. \quad (8)$$

For the triangular shaped flow-density relation with constant forward and backward wave speeds, it is easy to verify that, when the speed of the forward wave is v_f, the general cumulative flow count updating formula, Eq.(7), reduces to $-k + \frac{q}{v_f} = -k + k = 0$. Under congested traffic conditions with a constant backward wave speed w_b, we have $-k + \frac{q}{w_b} = -k_{jam}$, and this equation can be rewritten as

$$dN = \left(-k + \frac{q}{w_b}\right)dx = -k_{jam}(a) \times length(a) \times nlanes(a). \quad (9)$$

Eq.(9) is used to describe how a backward wave travels through the link. As shown in Fig. 1, when a queue spills back from the downstream to the upstream, the arrival and departure cumulative flow counts at two ends of a link (at timestamps *t* and time *t*–*BWTT*(*a*)) need to ensure a constant difference of d*N* = k_{jam}(*l*) ×*length*(*l*) ×*nlanes*(*l*), and the capacity restriction is propagated throughout the link using a time duration of *BWTT*(*l*) = *length*(*l*) / w_b(*l*).

[place Fig. 1 about here]
[place Fig. 2 about here]

The space-time plot in Fig. 2 further demonstrates how Newell's model uses cumulative counts to model the forward and backward waves in describing queue phenomenon on two links *a* and *b*. Exactly at time *t*–1, the tail of the queue at the downstream link *b* propagates to its upstream node. That is, if the cumulative outflow count at a lagged time stamp *t* − *BWTT*(*b*) −1 on link *b* equals to the cumulative inflow count at time *t*–1, then a queue spillback occurs as a backward wave is able to propagate through the congested time-space "mass" of the link. Compared to the conventional vertical queue model, where the inflow rate of a link is not constrained by the available physical length, the inflow rate in Newell's model is governed by the discharge flow of a link through the backward wave propagation process. Furthermore the number of vehicles on a lane should not be greater than k_{jam}×*length*(*b*). As a result, the proposed model is able to detect and capture queue spillbacks to upstream link(s).

2.4 Traffic flow dynamics constraints and states updating

The main building block of the DNL model is a set of demand, path and link flow balance constraints. Firstly, the demand flow balance constraint for each (w,τ) pair over its path set is as follows.

$$\sum_{p \in P(w,\tau) \text{and } \Phi(t,\tau)=1} r(w,t,p) = q(w,\tau), \forall w,\tau, \quad (10)$$

where $\Phi(t,\tau) = 1$ indicates that time interval *t* is inside departure time interval τ. Then, the DNL needs to load path

flow $r(w, t, p)$ to the first link of path $p \in P(w, \tau)$.

$$A(p, l(p,1), t) = r(w,t,p), \forall p \in P(w,\tau), w, t. \quad (11)$$

where $l(p, 1)$ represents the first link of path p, and $A(p, l, t)$ is the cumulative number of vehicles that are assigned to path p and have arrived at the upstream end of link l in time t. Cumulative arrival counts are moved from the beginning of link l to its vertical queue for each p, l, and t. That is,

$$V(p,l,t) = A\big(p, l, t - FFTT(l)\big), \forall p, l, t, \quad (12)$$

where $V(p, l, t)$ is the cumulative number of vehicles that are assigned to path p and have waited at the vertical queue of link l at time t.

The cumuative depature flow is less then the the cumulagive number of vehicles in the vertical queue on a link.

$$D(p,l,t) \le V(p,l,t), \forall p, l, t, \quad (13)$$

where $D(p, l, t)$ is cumulative number of vehicles that are assigned to path p and have departed from link l at time t. The cumulative departure flow count of link l is transfered to the cumulative arrival count on the next link l+1 along the path p,

$$A(p, l+1, t) = D(p,l,t), \forall p, l, t. \quad (14)$$

The balance constraint at the super sink for each OD pair is

$$D(p, L, T) = r(w,t,p), \forall w, p, t. \quad (15)$$

where L is the last link of path p and T is the end of the planning horizon. By summing up all cumulative arrival, vertical queue and departure counts across diffent paths, we can obtain cumulative link arrival, vertical queue and departure counts as follows.

$$A(l,t) = \sum_p A(p,l,t), \forall l, t. \quad (16)$$

$$V(l,t) = \sum_p V(p,l,t), \forall l, t. \quad (17)$$

$$D(l,t) = \sum_p D(p,l,t), \forall l, t. \quad (18)$$

In addition to the above set of flow balance constraints, when vehicles are discharged from the queue, Eq.(19) is used to satisfy the outflow capacity constraints and capture the effect of queue spillback.

$$D(l,t) - D(l,t-1) = \min\{V(l,t-1) - D(l,t-1), q^{max}(l,t), cap^{out}(l,t)\}, \forall l, t. \quad (19)$$

where $V(l, t-1) - D(l, t-1)$ is the number of vehicles waiting at the vertical queue of link l at time t. To determine the maximum outflow capacity $cap^{out}(l, t)$ in Eq.(19), we consider a corridor with two consecutive links l and l+1, where a lane drop bottleneck at the downstream end of link l+1 causes congestion and the queue spills back to link l (i.e., link l+1 is fully congested while link l is partially congested). Under congested conditions, $cap^{out}(l, t)$ is determined by the bottleneck inflow rate of link l+1.

$$cap^{out}(l,t) = cap^{in}(l+1, t). \quad (20)$$

As shown in Fig. 3, the inflow capacity $cap^{in}(l+1, t)$ is defined in terms of the difference of the cumulative arrival flow counts between two consecutive time stamps $t-1$ and t as follows,

$$cap^{in}(l+1,t) = A^{max}(l+1,t) - A(l+1, t-1), \quad (21)$$

where (according to the backward wave propagation constraint in Eq.(7))

$$A^{max}(l+1,t) = D(l+1, t - BWTT(l+1)) + k_{jam}(l+1) \times length(l+1) \times nlanes(l+1). \quad (22)$$

Finally, the queue spillback constraint based on Newell's simplified kinematic wave mode is

$$A(l,t) \le D\big(l, t - BWTT(l)\big) + k_{jam}(l) \times length(l) \times nlanes(l), \forall l, t. \quad (23)$$

Fig. 3 shows that if the cumulative outflow count at a lagged time stamp $t - BWTT(l+1) - 1$ on link l+1 equals to the cumulative inflow count at time $t-1$, then a queue spillback occurs as a backward wave propagates through the congested link.

[place fig. 3 about here]

The numerical computation scheme of the above DNL model can be implemented as an event-based simulation process which does not keep track of vehicles' positions on links at each time stamp. Specifically, if a vehicle j is moved into link l at its arrival time $Arr(j, l)$, the time entering the vertical queue at the stop bar is $t_{vq} = Arr(j, l) + FFTT(l)$. When the simulation time clock advances to t_{vq}, if the out-flow link capacity at time t_{vq} is still available for vehicle j to exit from the vertical queue, then this vehicle moves to the next link of its path and its departure time stamp $Dep(j, l) = t_{vq}$; otherwise vehicle j has to wait in the vertical queue for the available outflow capacity. As this proposed model strictly satisfies the first-in-first-out (FIFO) constraint, if a vehicle j in the beginning of the vertical queue is blocked due to unavailable outflow capacity, then vehicles behind j in the queue will be also blocked as well. To model complex geometric features, such as short left-turn bays on a multi-lane facility, one needs to decompose a link into multiple connected cells with each cell satisfying the FIFO constraint.

To describe traffic flow dynamics with multiple OD pairs and paths at merge and diverge junctions, many

existing DNL models (e.g. Jayakrishnan et al., 1994; Daganzo, 1995; Tampere et al, 2011) proposed various modeling approaches to determine outflow capacities for merging vehicular flows moving out from incoming links to a link, or outflow proportions for diverging vehicular flow moving out from a link to its outgoing links. In this paper, the available outflow capacity of an incoming link at a merge is assumed to be proportional to the number of lanes on that link. Similar to a mesoscopic DNL model (Jayakrishnan et al., 1994), each vehicle carriers its own OD and path information in our proposed model, so the outflow proportions from a link are indirectly determined by the path attributes associated with vehicles waiting at that link's vertical queue, which is implemented as a FIFO queue that explicitly enables the FIFO constraint to be satisfied.

3. Solution algorithm

This section describes the Lagrangian relaxation-based heuristic for solving the single-level time-dependent path flow estimation model, presented in Section 2. We propose the following heuristic solution method to efficiently obtain good solutions for problem instances on road networks of practical sizes. The heuristic integrates Lagrangian relaxation and column generation methods to solve the time-dependent path flow estimation model, **P1**. The gap function constraint Eq.(3) is relaxed to the objective function Eq.(1) with a non-negative Lagrange multiplier λ. The resulting Lagrangian subproblem is given as follows.

P2: $\text{Min}_{\boldsymbol{r},\boldsymbol{\pi}}\, L(\boldsymbol{r},\boldsymbol{\pi},\lambda) = Z + \lambda\{g(\boldsymbol{r},\boldsymbol{\pi})\}$ (24)

Subject to constraints (2), (4), (5) and (6),

where $\boldsymbol{r}$ and $\boldsymbol{\pi}$ are the vectors of path flows and least path times respectively. For a given λ, the solution to **P2** provides a lower bound to **P1**. The Lagrangian dual problem is given as follows.

P3: $\text{Max}_{\lambda}\text{Min}_{\boldsymbol{r},\boldsymbol{\pi}}\, L(\boldsymbol{r},\boldsymbol{\pi},\lambda)$ (25)

Subject to $\lambda \geq 0$.

The heuristic consists of two major algorithmic steps: at each iteration *n*, (i) given a Lagrange multiplier $\lambda^{(n)}$, find an optimal path assignment $\boldsymbol{r}^{(n)}$ and least path travel times $\boldsymbol{\pi}^{(n)}$ by solving the Lagrangian subproblem, **P2**, and (ii) given a vehicle path assignment $\boldsymbol{r}^{(n)}$ and least path travel times $\boldsymbol{\pi}^{(n)}$, update the Lagrange multiplier $\lambda^{(n+1)}$ by using the following rule.

$$\lambda^{(n+1)} = \max\{0,\ \lambda^{(n)} + \alpha^{(n)}\{\Sigma_w\Sigma_\tau\Sigma_p\, r(w,\tau,p)[c(w,\tau,p) - \pi(w,\tau)]\}\}, \quad (26)$$

where $\alpha^{(n)}$ is the step size for updating the Lagrange multiplier.

Accordingly, this heuristic has two loops (see Fig. 4). The outer loop is for updating the Lagrange multiplier using the rule described in Eq.(26). For each outer loop iteration *n* (i.e., corresponding to a given Lagrange multiplier $\lambda^{(n)}$), a column generation-based approach is used to solve the Lagrangian subproblem **P2**. This approach forms an inner loop for solving a DUE assignment problem under a restricted feasible solution space. In each inner loop iteration *m*, a time-dependent shortest path algorithm (Ziliaskopoulos and Mahmassani, 1993) is adopted to generate time-dependent least time paths and to augment the restricted path set. In light of the time-dependent shortest path algorithm, the least path travel times $\boldsymbol{\pi}^{(m)}$ are obtained to satisfy the constraints Eq. (4) and (5), thus, these definitional constraints of the least travel times can be dropped in solving the restricted Lagrangian subproblem. To solve the restricted subproblem, a gradient-projection-based descent direction method (Lu et al., 2009) is used to update path flows $\boldsymbol{r}^{(m+1)}$, while maintaining the feasibility of non-negativity constraints Eq.(6). Specifically,

$$r(w,\tau,p)^{m+1} = \text{Max}\left\{0, r(w,\tau,p)^m - \gamma^{(m)}\left[\beta_d\nabla h^d(r)\big|_{r=r^{(m)}} + \beta_q\nabla h^q(r)\big|_{r=r^{(m)}} + \beta_k\nabla h^k(r)\big|_{r=r^{(m)}} + \lambda^{(n)}\nabla g(r,\pi)\big|_{r=r^{(m)}}\right]\right\} \quad (27)$$

where $\gamma^{(m)}$ is the step size, and the gradients, which consist of the first-order partial derivatives with respect to a path flow variable $r(w,\tau,p)$, can be derived as follows.

$$\nabla h^d(\boldsymbol{r}) = \frac{\partial\left\{\sum_{\tau\in H_d, p\in P}\left(r_{(w,\tau,p)}\right) - \bar{d}_{(w)}\right\}}{\partial r_{(w,\tau,p)}} = 2\left(\sum_{\tau\in H_d}\sum_{p\in P} r(w,\tau,p) - \bar{d}(w)\right) \quad (28)$$

$$\nabla h^q(\boldsymbol{r}) = \frac{\partial\sum_{l\in S, t\in H_o t}[q_{(l,t)}(r) - \bar{q}_{(l,t)}]}{\partial r_{(w,\tau,p)}} = 2\sum_{t\in H_o}\sum_{l\in S}\left\{[q(l,t) - \bar{q}(l,t)] \times \frac{\partial q(l,t)(r)}{\partial r_{(w,\tau,p)}}\right\} \quad (29)$$

$$\nabla h^k(\boldsymbol{r}) = \frac{\partial\sum_{l\in S, t\in H_o t}[k_{(l,t)}(r) - \bar{k}_{(l,t)}]}{\partial r_{(w,\tau,p)}} = 2\sum_{t\in H_o}\sum_{l\in S}\left\{[k(l,t) - \bar{k}(l,t)] \times \frac{\partial k(l,t)(r)}{\partial r_{(w,\tau,p)}}\right\} \quad (30)$$

$$\nabla g(\boldsymbol{r},\boldsymbol{\pi}) = \frac{\partial g(r,\pi)}{\partial r_{(w,\tau,p)}} = c(w,\tau,p) - \pi_{(w,\tau)} + r(w,\tau,p)\frac{\partial c_{(w,\tau,p)}}{\partial r_{(w,\tau,p)}} \quad (31)$$

Estimated link flows, densities, and link/path travel times and the corresponding partial derivatives, namely $\nabla h^q(r)$, $\nabla h^k(r)$ and $\partial c(w,\tau,p)/\partial r(w,\tau,p)$ are obtained from the DNL model presented in Section 2.

The steps of this algorithm are presented as follows.

Algorithm 1: Lagrangian relaxation-based heuristic

Step 1: Initialization

Step 1.1: Set outer loop iteration counter $n = 0$. Input a historical demand table, $\bar{d}(w)$, $\forall w$, its corresponding temporal distribution profile, and time-dependent link observations (densities, and flows). Initialize $\lambda^{(n)}$ to a positive value, such as 1.0.

Step 1.2: Perform DUE traffic assignment with time-dependent OD demand matrix, $d^{(n)}$, based on $\bar{d}(w)$, $\forall w$ and a temporal profile, and build a feasible path set $\hat{P}$.

Step 1.3: According to the assignment results (i.e., estimated link densities, flows, and travel times), compute the upper bound of the optimal solution to the primal problem **P1** as follows:

$$Z^{UB} = \sum_{l\in S}\sum_{t\in H_o}\{\beta_q[q(l,t) - \bar{q}(l,t)]^2 + \beta_k\left[k(l,t) - \bar{k}(l,t)\right]^2\} + \lambda^n\{g(\boldsymbol{r},\boldsymbol{\pi})\}. \quad (32)$$

Note that $\sum_w\left[\sum_{\tau\in H_d}\sum_p r(w,\tau,p) - \bar{d}(w)\right]^2 = 0$, as the DUE assignment loads the target demands.

Step 2: [Inner loop] Solve the Lagrangian subproblem, **P2**, to find the optimal path flows corresponding to $\lambda^{(n)}$.

Step 2.1: Set inner loop iteration counter $m = 0$. Read estimated link flows, densities, path flows, path travel times and corresponding gradients $\nabla h^d(r)$, $\nabla h^q(r)$ and $\nabla h^k(r)$ from the last outer iteration n.

Step 2.2: Identify the least time paths and determine $\boldsymbol{\pi}^{(m)}$.

Step 2.3: Calculate the gradients according to Eqs.(28)-(31).

Step 2.4: Determine the step size $\gamma^{(m)}$ according to the method of successive averages (MSA), $\gamma^{(m)} = 1/(m+1)$.

Step 2.5: Update path flows $\boldsymbol{r}^{(m+1)}$, according to Eq.(27).

Step2.6: Load path flow assignment $\boldsymbol{r}^{(m+1)}$ to the DNL model based on Newell's simplified KW model, to obtain estimated link densities and flows, and path travel times.

Step 2.7: Convergence checking for solving the Lagrangian subproblem, **P2** (i.e., the inner loop). Update the objective function, $L(\boldsymbol{r}^{(m+1)},\boldsymbol{\pi}^{(m+1)},\lambda^{(n)})$. If $m < M_{max}$ (maximum number of inner loop iterations) or $L(\boldsymbol{r}^{(m+1)},\boldsymbol{\pi}^{(m+1)},\lambda^{(n)})$ is improved, then $m = m+1$, and go to Step 2.2; otherwise, go to Step 2.8.

Step 2.8: Update the tightest lower bound, Z^{LB}, as the following:

$$Z^{LB} = \text{Max}\{Z^{LB}, \beta_d h^d(\boldsymbol{r}^{(m)}) + \beta_q h^q(\boldsymbol{r}^{(m)}) + \beta_k h^k(\boldsymbol{r}^{(m)}) + \lambda^{(n)}[g(\boldsymbol{r}^{(m)},\boldsymbol{\pi}^{(m)})]\}. \quad (33)$$

Step 3: Update the upper bound

Step 3.1: Construct a time-dependent OD demand matrix by aggregating the time-dependent path flows obtained from the solution to the Lagrangian subproblem **P2**, for each OD pair and each departure time interval.

$$d(w,\tau) = \Sigma_{p\in P}\, r(w,\tau,p),\ \forall w,\ \tau. \quad (34)$$

Step 3.2: Perform DUE traffic assignment for the constructed OD demand matrix $\boldsymbol{d} = \{d(w,\tau),\ \forall w,\ \tau\}$. According to the assignment results (i.e., estimated link densities, flows, and travel times), obtain the equilibrium path flows and update the tightest upper bound as:

$$Z^{UB} = \text{Min}\{Z^{UB}, \beta_d h^d(\boldsymbol{r}^{(m)}) + \beta_q h^q(\boldsymbol{r}^{(n)}) + \beta_k h^k(\boldsymbol{r}^{(n)}) + \lambda^{(n)}[g(\boldsymbol{r}^{(n)},\boldsymbol{\pi}^{(n)})]\}. \quad (35)$$

Step 4: Convergence checking for solving the Lagrangian dual problem (i.e., the outer loop)

If $n > N_{max}$ (maximum number of outer loop iterations) or $Z^{UB} - Z^{LB} < \phi$ (a preset threshold), then stop; otherwise, set $n = n + 1$.

Step 5: Path generation. Compute least time path for each departure time and OD pair using the time-dependent shortest path algorithm, developed by Ziliaskopoulos and Mahmassani (1993), and add newly generated paths to the feasible path set $\hat{P}$.

Step 6: Update Lagrange multiplier λ

Step 6.1: Obtain estimated link flows, path flows, and path travel times from the final iteration of the last inner loop (i.e., solving **P2** in Step 2). Update the gap value, $g(\boldsymbol{r}^{(n)}, \boldsymbol{\pi}^{(n)})$ defined in Eq.(3), using these estimates.

Step 6.2: Determine the step size, $\alpha^{(n)}$, by MSA, i.e. $\alpha^{(n)} = 1/(n+1)$.

Step 6.3: Update Lagrange multiplier, $\lambda^{(n+1)}$, according to Eq.(26). Return to Step 2 (the inner loop).

Typically, a Lagrangian solution framework requires obtaining exact solutions to relaxed subproblems. It should be remarked that, analyzing existence and uniqueness of solutions to the DUE problem for multiple OD pairs are very challenging, and the gradient-based algorithm through Eqs. (27)-(31) cannot guarantee that the relaxed (nonlinear) problem P2 is solved to its optimality. Thus, when no global optimum solution is available for P2, the proposed overall Lagrangian solution algorithm is still a heuristic method in nature.

[place fig. 4 about here]

4. Evaluation of partial derivatives with respect to path flow perturbation

Solving the proposed single-level dynamic OD estimation model requires the evaluation of the partial derivatives with respect to time-varying path flows, i.e., $\frac{\partial q_{(l,t)}(r)}{\partial r_{(w,\tau,p)}}$, $\frac{\partial k_{(l,t)}(r)}{\partial r_{(w,\tau,p)}}$,, and $\frac{\partial c_{(w,\tau,p)}(r)}{\partial r_{(w,\tau,p)}}$ in Eqs.(29)-(31). These partial derivatives represent the marginal effects of an additional unit of path inflow on link flow and density and path travel time. This section delineates the evaluation of these partial derivatives due to path flow perturbation in a congested network, based on cumulative link inflow and outflow curves. The following notation is used throughout this section.

L : the number of links on the path
l: link index l=1, 2, …L
t_l': the time when an additional unit of perturbation flow arrives at link l
t_l'': the time when an additional unit of perturbation flow departs from link l
t_l^{qs}: the time when the queue starts to form on link l
t_l^B: the time when the queue vanishes on link l
$t_l^A := t_l^B - FFTT(l)$
t_l^{q*}: the time when the queue on link l starts to spillback to its upstream link l–1
n^A : cumulative arrivals at time t_l^A
$n^{'}$: cumulative arrivals at time $t_l^{'}$

4.1 Evaluation of link partial derivatives on a congested link

In this study, the link partial derivatives are referred to as the changes in link flow and density due to an additional unit of link/path inflow. In their pioneering work, Ghali and Smith (1995) presented an analytical approach to evaluate (local) link marginal travel time (or delay) on a congested link, based on link cumulative flow curves. An illustration of the approach is depicted in Fig. 5. In the figure, the two solid lines represent the cumulative arrival and departure curves, while the dashed line represents the cumulative vertical queue. The (outflow) capacity of the link is c. The key result of their approach is that the link marginal delay equals the grey area. While link and path marginal delays were investigated in the context of system optimal DTA (e.g., Ghali and Smith, 1995; Peeta and Mahmassani, 1995; Shen et al., 2007; Qian and Zhang, 2011), to the authors' current knowledge, the study on the partial derivatives of link flow and density and path travel time with respect to path flow was nonexistent.

[place fig. 5 about there]

(For example, the queue starts at t_l^{qs}= 7:00 AM, an additional unit of vehicle, n'= 1000, enters link l at time $t_l^{'}$= 7:10AM, leaves the link at time $t_l^{''}$= 7:50AM, the queue dissipates on link l at t_l^B = 8:30AM, and t_l^A = 8:15AM, assuming FFTT(l) = 15 min.)

The following propositions can be directly induced from Fig. 5 for deriving the marginal effects on link flow (inflow and outflow) and density.

Proposition 1: Under *free-flow* conditions, an extra unit of flow arriving at the upstream end of link l at time t_l' results in the following: (i) the link inflow and outflow increase by 1 at times t_l' and t_l'', respectively, and the flow rates at other time intervals do not change; (ii) the link density increases by 1 from t_l' to t_l''; (iii) the individual travel times are not changed, and $t_l'' = t_l' + FFTT(l)$.

Proposition 2: Under *partially* congested conditions and constant link (outflow) capacity c, an extra unit of flow arriving at the upstream end of link l at time t_l' results in the following: (i) the link inflow and outflow increase by 1 at times t_l' and t_l^B, respectively, and the flow rates at other time intervals do not change; (ii) the link density increases by 1 from t_l' to t_l^B; (iii) the flows arriving between t_l' and t_l^A experience the additional delay 1/c, because it takes 1/c to discharge this perturbation flow.

Proposition 3: Under *fully* congested condition, an extra unit of flow arriving at the upstream end of link l at time t_l' would not change the link inflow, outflow, and capacity. Both link inflow and outflow remain the same at the maximum link flow rate, while the link density is equal to the jam density. The flows arriving between t_l' and t_l^A experience the additional delay 1/c, because it takes 1/c to discharge this perturbation flow.

With the proposed DNL model based on Newell's simplified KW model, we can adapt Eq.(9) to detect if the queue spills back to the upstream end of link l using the following equation.

$$A(l,t) < D(l, t - BWTT(l)) + k_{jam}(l) \times length(l) \times nlanes(l). \quad (36)$$

Specifically, if the above strict inequality holds, then the queue has not propagated back to the upstream end of the link (or the link is partially congested). Otherwise, the link is fully congested and the queue propagates throughout the link along the backward wave line, as illustrated in Fig. 5. Note that, under fully congested conditions, the prevailing link density $k(l)$ at time t might be smaller than $k_{jam}(l)$.

A common pitfall for deriving the partial derivative of link density under congested conditions, is to record the increase in density by 1 from t_l' to t_l'' (e.g., from 7:10AM to 7:50AM with the duration of 40-min in Fig. 5). Proposition 2, induced from Fig. 5, clarifies that the actual change in link density would last until the queue vanishes at time t_l^B, so the change in link density should cover a duration of 80-min from 7:10AM to 8:30AM in our example. Proposition 2 also indicates that the change in link outflow, due to an extra unit of flow arriving under congested conditions, occurs at the time t_l^B (e.g. 8:30AM), rather than t_l'' (e.g., 7:50AM). Note that, by multiplying the (individual) extra delay, $1/c$, by the number of vehicles arriving between t_l' and t_l^A (i.e., $n^A - n'$), we can obtain that, if t_l' is between t_l^{qs} and t_l^{A}, the local link marginal delay is equal to $t_l^B - t_l'$, which is the sum of the traversal time of the perturbation flow, $t_l'' - t_l'$, and the additional delay imposed by the perturbation on others is equal to $t_l^B - t_l''$. This evaluation approach for link marginal travel times was also adopted by, Shen et al. (2007), and Qian and Zhang (2011) in the context of system optimal dynamic traffic assignment.

4.2 Evaluation of the impact of path flow perturbation on two sequential links

In static transportation networks, the path travel time marginal, which is the partial derivative of the total system-wide travel time, can be obtained as the sum of the link marginals of a path's constituent links. However, in dynamic and congested transportation networks, this additivity assumption may lead to a severe deficiency in the evaluation of path marginals, as indicated by Shen et al. (2007). They showed that it is necessary to explicitly trace the propagation of path flow perturbation in evaluating path marginal travel times and proposed an evaluation method of path marginals. Thus, this study evaluates the impact of path flow perturbation (i.e., the partial derivatives) in the individual link-time level by tracing the changes in link flow and density on a sequence of links (i.e., a path) and over different time intervals, due to the addition of unit flow to a path. Qian and Zhang (2011) conducted a similar analysis for individual path marginal travel times.

Let's first consider a freeway or an arterial segment with two sequential links, without merges and diverges, say link l–1 and link l. Under congested conditions, there are three basic cases of interest, when the additional unit of flow arrives at this segment.

(i) There is a bottleneck on the downstream link l and the queue on link l does not spill back to link l–1; that is, link l–1 is in free-flow condition while link l is partially congested.

(ii) There is a bottleneck on the downstream link l and the queue on link l spills back to link l–1; that is, link l–1 is partially congested while link l is fully congested.

(iii) There is a bottleneck on each of the two links, and the two bottlenecks are independent, assuming that both links are sufficiently long so that the queue in the downstream does not spill back to the upstream. This is in fact the case in which both links l–1 and l are partially congested.

For cases (i) and (iii), Proposition 2 can be applied to determine the marginal effects of the additional unit of flow on link flow and density. For case (ii), there are two possible scenarios. As depicted in Fig. 6 (a), one scenario is that the additional unit of flow does not encounter the queue on link l–1, so it can enter link l at time $t_l^{'} = t_{l-1}^{''} = t_{l-1}^{'} + FFTT(l-1)$, when link l is partially congested, i.e.,$t_l^{'} < t_l^{q*}$ or $t_l^{'} > t_{l-1}^{B}$. The other scenario, as depicted in Fig. 6(b), is that the perturbation flow encounters the queue on link l–1, i.e. $t_{l-1}^{qs} < (t_{l-1}^{'} + FFTT(l-1)) < t_{l-1}^{B}$, so it cannot enter link l until time $t_l^{'} = t_{l-1}^{''} = t_{l-1}^{B}$. Note that, $t_{l-1}^{qs} = t_l^{q*}$, the time at which the queue on link l start to spill back to link l–1. While it is possible to detect that the queue spills back based on Eq.(36), tracing the propagation of the perturbation flow in the case of queue spillback is still very difficult. In order to strike a balance between computational efficiency and numerical accuracy, this study applies Proposition 3 to the analysis of link marginal effects on the current link l under fully congested states, and does not trace back the flow perturbation from link l to link l–1.

Next, we discuss the evaluation of partial derivatives of two sequential links in merge or diverge junctions using the illustrative network depicted in Fig. 7. Assume that link b is sufficiently long or consists of several links, so the two junctions are distantly separated. The additional perturbation flow goes through links $a1$, b, and $c1$ in order. In the merge junction, the inflow capacity of link b is assumed to be distributed according to the number of lanes of the incoming links (i.e., links $a1$ and $a2$), so the perturbation flow from link $a1$ to link b does not affect the density and flow on link $a2$. Therefore, the analysis results presented above for two sequential links (i.e., Cases (i), (ii), and (iii)) can be directly employed to determine the path marginal effects of link flow and density.

[place fig. 6 here]

When the perturbation flow arrives at the diverge junction, if there is a bottleneck active on link $c1$ or a bottleneck active on each of the two links, b and $c1$, the results discussed above for two sequential links can be applied to evaluate the path marginal effects of interest. Besides, a bottleneck active on link $c2$ will not affect the evaluation of the path partial derivatives on links b and $c1$, unless the queue on link $c2$ spills back to link b. In this case (i.e., there is no inflow capacity on link $c2$ to accommodate the incoming flow), due to the FIFO constraint, the perturbation flow cannot enter link $c1$ until the vehicles in front of it in the vertical queue of link b get discharged. We may consider this case as there is a bottleneck at the downstream end of link b, and adopt the results in Proposition 2 to determine the link/path partial derivatives.

[place fig. 7 about here]

Lastly, we illustrate the evaluation of the partial derivative of path travel time with respect to an unit of additional flow in the same time interval, i.e., $\frac{\partial c(w,\tau,p)}{\partial r(w,\tau,p)}$ in Eq.(31), based on the case of two partially congested links (i.e., case (iii)), depicted in Fig. 8. Note that the analysis result can be applied to the other two cases (i and ii). Let link l–1 be the first (congested) link of a path under evaluation. Consider a vehicle *veh* that enters link l–1 at time $t_{l-1}^{'}$ and leaves at time $t_{l-1}^{''}$. According to Proposition 2, the extra delay experienced by this vehicle on link l–1 is 1/c, due to the perturbation flow entering the link in the same time interval, where c is the capacity of link l–1. Then, *veh* enters link l at time t_l^{in} and leaves at time t_l^{out}, but the impact of path flow perturbation on link l occurs at a later time stamp $t_l^{'} = t_{l-1}^{B} > t_l^{in} = t_{l-1}^{''}$. Because $\frac{\partial c(w,\tau,p)}{\partial r(w,\tau,p)}$ refers to the marginal effect of the perturbation flow on the path travel time, $c_{(w,\tau,p)}$, of vehicle *veh*, $c_{(w,\tau,p)}$ is affected by the perturbation flow only on link l–1. The perturbation flow, which departs in the same time interval τ as vehicle *veh***,** will not affect the travel time of *veh* after link l–1. Therefore, the partial derivative can be approximated as $\frac{\partial c(w,\tau,p)}{\partial r(w,\tau,p)} = 1/c$.

[place fig. 8 here]

4.3 Computational method for evaluating partial derivatives due to path flow propagation

Based on the above analyses, we now present Algorithm 2, which traces the propagation of an additional unit of flow and calculates the gradients in Eqs.(29)-(30) of a given path at departure time τ. Let t_l^s and t_l^e be the starting and end times of (event) impact period on link l, respectively. Under the event-based mesoscopic traffic simulation framework, t_l^e can take on one of the three values: $t_l^s + FFTT(l)$, t_l^B, or t_l^{q*}, corresponding to the free-flow condition, congested without queue spillback (or partially congested), and congested with queue spillback (or fully congested), respectively.

Algorithm 2: Computational method for path partial derivatives

Initialize link entry time t_l^s = departure time τ, and link index l = 1.

Do while link index $l \le L$

Given t_l^s, determine the end time of impact period t_l^e as follows.

Step 1: Set the tentative entrance time to the vertical queue as $t_l^{temp} = t_l^s + FFTT(l)$.

Step 2: (Determining the current traffic condition)

Step 2.1: If there is no vertical queue at time t_l^{temp} and t_l^{temp} is not equal to the beginning of the queue (t_l^{qs}), then the perturbation flow enters link l under uncongested conditions: perform Step 3; otherwise, time t_l^{temp} is under congested conditions.

Step 2.2: Next, if there is no queue spillback between time $t_l^s + FFTT(l)$ and t_l^B, then perform Step 4; otherwise, perform Step 5.

Step 3: (Free-flow condition)

Step 3.1: $t_l^e = t_l^s + FFTT(l)$.

Step 3.2: Update link inflow, outflow and travel time partial derivatives according to Proposition 1.

Step 3.3: If link l has the measurements (e.g., $\overline{q}$ and $\overline{k}$,), then cumulate $\nabla h^q(r)$ and $\nabla h^k(r)$, according to Eqs.(29) and (30).

Step 3.4: Set $l = l$+1 and $t_l^s = t_{l-1}^e$.

Step 4: (Congested condition without queue spillback)

Step 4.1: Set $t_l^e = t_l^B$.

Step 4.2: Update link partial derivatives according to Proposition 2 for impact period $[t_l^s, t_l^e]$.

Step 4.3: If link l has the measurements (e.g., $\bar{q}$ and $\bar{k}$,), then cumulate $\nabla h^q(r)$ and $\nabla h^k(r)$, according to Eqs.(29) and (30).

Step 4.4: Set $l = l+1$ and $t_l^s = t_{l-1}^e$.

Step 5 (Congested condition with queue spillback):

Step 5.1: Find the queue spillback time (in terms of link entrance time) t_l^{q*}, and set $t_l^e = t_l^{q*}$.

Step 5.2: Update link partial derivatives according to Proposition 3 for impact period $[t_l^s, t_l^e]$.

Step 5.3: If link l has the measurements (e.g., $\bar{q}$ and $\bar{k}$,), then cumulate $\nabla h^q(r)$ and $\nabla h^k(r)$, according to Eqs.(29) and (30).

Step 5.4: Keep l unchanged, set $t_l^s = t_l^e$, go back to Step 4.

End Do

Note that, because of the limitation of using local (instead of global) link partial derivatives and the lack of an exact approach to keep track of the propagation of perturbation flows, the gradients obtained using Algorithm 2 are still numerical approximates. Thus, the proposed solution algorithm, based on the approximate gradients, is a heuristic that cannot be guaranteed to find (globally) optimal solutions.

5. Numerical experiments

Three sets of numerical experiments were conducted on three different networks to systematically evaluate the performance of the proposed dynamic OD demand estimation algorithm under different scenarios. The first network is a simple two-link corridor with steady state travel time functions. The second network is a simple freeway corridor with time-dependent sensor data. The last network is a real-world transportation network with time-dependent sensor data.

5.1 Experiments on a simple two-link corridor with steady state travel time function

In the first set of experiments, we aim to examine the convergence pattern of the proposed algorithm on a simple corridor with a single OD pair connected by two parallel links (or paths). A simple linear travel time function, Eq.(37), is used in performing the traffic assignment which loads a total peak-hour demand, 8000 vehicles/hour (or veh/hr) to those two paths. Then, the resulting UE assignment results, shown in Table 1, are used as the ground-truth condition to evaluate the path flow estimation performance under various testing conditions.

$$T(l) = FFTT(l) + r_l / cap(l), \tag{37}$$

where $T(l)$ and $FFTT(l)$ are the travel time and free-flow travel time on link/path l, respectively. r_l and $cap(l)$ are the flow volume and capacity of link/path l, respectively.

[place Table 1 here]

First, we start with an initial path flow distribution that loads 3000 veh/hr to each link. The ground-truth demand of 8000 veh/hr is set as the target demand, and the error-free flow counts (r_1 = 5400, r_2 = 2600) are used as the observations. In this case, the objective function of problem P1 can be simplified as follows.

$$\text{Min } L(r) = \beta_d f(r) + \beta_q h^q(r) + \lambda[g(\boldsymbol{r}, \boldsymbol{\pi})] \tag{38}$$

For simplicity, the weight parameters β_d = 1, and β_q = 1, which means that the decision maker has the same level of confidence on target demand and flow observations. By further considering the Lagrange multiplier, λ = 1, the first order partial derivative of the objective function with respect to path flow on path 1 is

$$\partial L(r) / \partial r_1 = 2(r_1 + r_2 - 8000) + 2(r_1 - 5400) + T(1) - \pi + r_1/cap(1), \tag{39}$$

where the minimum path travel time π is considered as an exogenous variable for this restricted optimization problem, and it can be obtained as $\pi = \text{Min}\{T(1), T(2)\}$ in each iteration. The second order partial derivative for path flows r_1 and r_2 are 4 + 2/cap(1) and 4 + 2/cap(2), respectively. Similar to the Newton-type algorithm, if the inverse of the second order gradient is used as the step size in Eq.(26), then the step size γ^m in inner iteration m for updating the flow on path l can be approximated as 1/4, as 2/cap(l) reduces to a very small value. This leads to the following approximate gradient-projection-based flow updating formula,

$$r_l^{m+1} = max\left\{0, r_l^m - 1/4 \times \frac{\partial L(r)}{\partial r_l}\right\} \tag{40}$$

Fig. 9 and Fig. 10 demonstrate the convergence patterns of the proposed path flow estimation algorithm in the first 20 inner iterations. We can observe that, after 3 or 4 inner iterations, the total estimated demand is quickly adjusted to a level very close to the ground-truth demand, while the equilibrium processes of path flow distribution and path travel times are relatively slow.

[place fig. 9 here]

[place fig. 10 here]

The following experiment was conducted to examine the impact of different values of the Lagrange multiplier λ on the solution quality. We consider two different criteria for evaluating the solutions, (i) the total gap $g(\boldsymbol{r}, \boldsymbol{\pi})$ that measures the distance from the UE conditions and (ii) the value of $\beta_d f(r) + \beta_q h^q(r)$ that measures the total deviation from the target demand and observed path flows. In this experiment, a slightly biased target demand with 7000 veh/hr and flow observations (r_1 = 5500, r_2 = 2500) are adopted. The weight parameters remain as β_d = 1, and β_q = 1, but the Lagrange multiplier λ was varied from 0.1 to 10 to obtain different solutions. By solving the optimization problem in Eq.(38) using Microsoft's Excel Solver, we can construct the resulting trade-off plot between user equilibrium gap and total deviation, as shown in Fig. 11. Essentially, a larger Lagrange multiplier/penalty associated with the gap function leads to a solution closer to the UE conditions at the expense of an increased total deviation from the target demand and the observations.

It should be noted that, in order to iteratively search for the maximum value of the Lagrangian dual problem, one can use the rule, described in Eq.(26), to determine the step size $\alpha^{(n)}$ in an outer iteration n for updating the Lagrange multiplier $\lambda^{(n+1)}$. However, because the gap function value could vary significantly, e.g., from 0 to 14000 in the above example, we can use a simple line search method to update the step $\lambda^{(n+1)}$ in this experiment to obtain stable improvement in the search process.

[place fig. 11 here]

As described in Step 2 of Algorithm 1, in each outer iteration n, the estimated OD demand and path and link flows (corresponding to a particular value of the Lagrange multiplier) are used to compute the lower bound (LB) of the primal problem, P1, based on Eq.(33). On the other hand, the upper bound (UB) is obtained by solving the corresponding UE problem. Then, in Step 3, the solution with the smallest gap between the UB and LB values is selected as the final solution. As shown in Fig. 12, in this experiment, λ = 7 gives the solution with the smallest gap, which corresponds to a total demand of 7353 veh/hr, a path flow distribution of (r_1 = 5012, r_2 = 2341), and equilibrium travel time of 53.4 min. In this final solution, the resulting relative Lagrangian gap value = (UB−LB)/LB= 0.17%, indicating that a close-to-optimal solution is found to the original primal problem, although the total demand, 7353 veh/hr, is still slightly different from the ground-truth OD demand, 8000 veh/hr, due to the inherent error in the target demand, 7000 veh/hr.

Note that, when varying the Lagrange multiplier from 0.1 to 7, we obtain several similar solutions, with the total demand ranging from 7334 to 7353 veh/hr. This demonstrates that the proposed solution framework is able to provide a theoretically rigorous method to quantify the solution quality, generate alternative solutions, and finally identify the optimal solution that can minimize the total measurement deviation and satisfy the UE conditions.

[place fig. 12 here]

Table 2 shows the estimation results under different degrees of information availability, for example, partial vs. complete sensor coverage, slightly biased vs. error target demand, as well with or without travel time observations. In general, more information from either target demand or measurements leads to a demand/path flow estimate closer to the ground-truth value.

[place table 2 here]

5.2 Experiments on a freeway corridor with time-dependent sensor data

In the second set of experiments, we test the performance of the proposed algorithm on a freeway corridor with time-dependent real-world sensor data. As shown in Fig. 13, the freeway corridor of interest is a 2-mile section of I-210 westbound, located in Los Angeles, CA. This corridor includes three on-ramps and one off-ramp. In the network representation constructed for the proposed DNL model, we ignore the HOV lane and only consider 4 general purpose lanes on the freeway. Traffic speed, flow count and occupancy are measured at 5-mins intervals on freeway and ramp links.

[place fig. 13 here]

[place fig. 14 here]

In this simple corridor, each OD pair only has a single path, so it does not involve a complex flow equilibration process required for multiple alternative paths. As a result, our focus is on demonstrating how the proposed gradient-based adjustment algorithm adjusts the incoming demand pattern to capture the observed queue formation,

propagation and dissipation. We first describe the details for preparing the input data for the dynamic OD demand estimation problem under consideration.

(Traffic Measurements) The traffic measurements were obtained from the PErformance Measurement System (PeMS) database (Varaiya, 2002). This study considers a planning horizon between 6AM and 10AM, and uses the flow count and density data on April 28, 2008 as the measurements, while the density data are constructed based on the observed occupancy data on the same day and estimated average vehicle length. Due to possible sensor malfunction, clean ramp sensor data are not always available for the entire planning horizon, so we use the historical average time series to fill out the data gaps on-ramps.

(Traffic flow model) Using the historical data from April 22 to 25, 2008 (i.e., 4 weekdays) during the early peak hours between 6 AM to 10AM, we calibrate the triangular traffic flow-density model for each link. For instance, Fig. 14 shows the flow-density relationship for sensor *b* with jam density K_{jam}=120 veh/ml/lane, free-flow speed V_f = 76 mph, backward wave speed w = 20 mph, and a maximum capacity of 1900 veh/hour/lane.

(Boundary outflow capacity) One of the critical boundary conditions in this OD estimation problem is the time-dependent maximum outflow discharge rate (i.e. capacity) on the downstream sensor *d*, which is a result of the queue spillback from links further downstream and cannot be captured internally in the study network. Thus, we use the historical average flow data to construct the link outflow capacity on station *d* and set it as the fixed input for the DNL model.

(Historical demand) Based on the historical flow counts on the boundary sensors, namely, the stations on freeway location *a* and all on-ramps, we estimate the total origin volume and then apply an estimated destination split to setup a historical OD demand table, where the destination split assigns the majority of flow toward the freeway destination *d*. This historical demand table serves as the target demand table in the experiment.

(Initial demand table) To evaluate if and how the gradient-based algorithm can adjust a biased, initial OD demand table toward the target demand table, we start with an OD demand table with only 70% of historical demand volume.

Fig. 15 shows the observed time series of the traffic speeds at stations *b*, *c* and *d*. After 6:30 AM, the traffic congestion from downstream location *d* is propagated to upstream sensors *b* and *c*, leading to a slow-moving queue on freeway mainline segments between 7:00AM and 9:00AM. The traffic congestion starts to dissipate afterwards, but the overall speed is still significantly below the free-flow level. The observed flow pattern at entrance station *a* can be viewed as the total origin demand flow from station *a* to different destinations. The root mean squared absolute errors (RMSEs) of the estimation results for speed measurements and for link count measurements are 13.2 miles per hour and 159.1 vehicles per hour per lane, respectively. Fig. 16 shows the estimated and observed flow patterns on entry link *a*, and the corresponding average relative estimation errors are less than 10%. This indicates that the proposed flow estimation algorithm can adjust a biased, initial demand pattern to match the target demand volume at the entry point. The estimated space speeds and the observed point speeds at station *c* are plotted in Fig. 17, which demonstrates that the DNL model is able to accurately reproduce the queue spillback phenomenon along the corridor.

[place fig. 15 here]

[place fig. 16 here]

[place fig. 17 here]

5.3 Experiments on large real-world traffic networks

The third set of numerical experiments was performed on two real-world networks. The first is a subarea network within the Portland, Oregon metropolitan area, which includes 858 nodes, 2000 links, and 208 OD zones shown in Fig. 18. We first loaded a subarea OD demand matrix generated from a recent study (Kittelson et al., 2011) as the "historical" OD demand, and obtain 5-min flow counts from 14 loop detectors and peak-hour flow counts converted from 392 annual average daily traffic (AADT) counts. To assess the estimation performance of the proposed approach, we used the root mean squared absolute error (RMSE) of peak-hour link volume as the estimation criterion. Fig. 19 shows a decreasing pattern of the estimation error as the iterative algorithm proceeds. The final RMSE value of peak-hour link volume is about 55.6 veh/link, corresponding to 15% error compared to the total link volume and R^2 = 0.91. The average travel time gap reduces to 0.33 min per vehicle on a network with an average travel time of 7 min, which demonstrates that the proposed algorithm is able to produce accurate estimation results for this medium-scale network.

[place fig. 18 here]
[place fig. 19 here]

The second real-world network is the Triangle Regional Model network, North Carolina, USA. This large-scale regional network has 2,389 zones, 20,259 links and about 2,000 signalized intersections. Provided by the local metropolitan planning agency, the morning peak-hour demand matrix has about 1.06 million vehicles, covering a time period from 6AM to 10AM. There are about 16 and 14 sensors, respectively, on freeway and arterial links, producing a total of 120 hourly link count observations. The experiments were performed on a PC with 16 GB memory and 8-core processors running at 2.70 GHz. We implemented the proposed path flow adjustment algorithm in the open-source dynamic traffic assignment package DTALite (Zhou et al., 2012), which has a mesoscopic traffic network loading model based on Newell's simplified kinematic wave theory. The starting and ending timestamps of vehicles on each link are used to construct link-based cumulative arrival and departure count curves, which are further used to track the queue spillback phenomena and shock wave propagation. The running time of this light-weight simulator is about 2 min and 45 sec per iteration for the Triangle Regional Model network. When incorporating the additional path flow adjustment process, the average running time increases to 5 min and 3 sec per iteration. The iterative sequential adjustment converged after 140 iterations, which take a total of 12 hours of CPU time. The scatter plots in Figs. 20-21 and additional three MOEs, namely R-squared, total estimated vs. observed flow ratio. The average absolute link flow deviations are 435.15 and 212.21 vehicles per hour per link, respectively, on freeway and arterial links, which further demonstrate the effectiveness of the proposed method on both freeway and arterial links of the large-scale network.

[place fig. 20 here]
[place fig. 21 here]

6. Concluding Remarks

With the particular focus on providing consistent and efficient time-dependent path flow estimators, this study proposes a single-level dynamic OD demand estimation model, without using link proportions, based on the mathematical program with complementary (or equilibrium) constraints approach. The Lagrangian relaxation-based heuristic which dualizes the gap function-based DUE constraint is developed to solve the proposed models. In each iteration, the tightest upper bound is updated according to the DUE assignment results of the OD demand table constructed from the estimated path flows, while the tightest lower bound is determined by the (gradient projection-based) path flow adjustment process. The algorithm proceeds until the gap between the upper bound and the lower bound is minimized.

The mesoscopic DNL model, based on Newell's simplified KW theory, is proposed to realistically estimate network performances for the DUE assignment result of an estimated OD demand table. Furthermore, based on the DNL model, we derive partial derivatives of link flow, density and path travel time with respect to path flow perturbations, which are essential to determine the feasible descent direction in the gradient-projection-based algorithm. The proposed OD demand estimation model utilizes a wide variety of traffic measurements available from traffic sensor networks, and circumvents the difficulty of providing complex mapping matrices between OD demand flows and those measurements in most of the existing dynamic OD demand estimation methods. The seamless integration between the path flow estimation and DTA models provides an effective and efficient way of utilizing heterogeneous data sources. The application to the three test networks demonstrates the effectiveness and performance of the proposed models under different network and data availability conditions.

In addition to examining its performance on large-scale networks in our future studies, the presented joint assignment and flow estimation model has considerable potential for generalizing the modeling framework into the field of real-time traffic state estimation and prediction. This would require further investigation into numerous issues, such as calibrating the maximum queue discharge rates which critically affect flows on downstream links, utilizing emerging end-to-end AVI travel times collected from Bluetooth readers, and accommodating possible modeling errors and behavioral heterogeneity in the DUE assignment. Moreover, it will be interesting to conduct computational experiments for comparing the performance of the proposed approach with that of other traditional approaches from the literature. Our subsequent research will also conduct a wide range of sensitivity tests on different parameters in our OD estimation method, and evaluate the impact of inaccurate input data on the estimation results.

Acknowledgement

The second author of this paper is partially supported through a FHWA project titled "An Open-Source Dynamic Traffic Assignment Tool for Assessing the Effects of Roadway Pricing and Crash Reduction Strategies on Recurring and Non-Recurring Congestion". Special thanks to anonymous reviewers and our colleagues Anxi Jia and Nagui Rouphail at the North Carolina State University for their constructive comments. The work presented in this paper remains the sole responsibility of the authors.

References

Balakrishna, R., Ben-Akiva, M., Koutsopoulos, H. N., 2008. Time-dependent origin-destination estimation without assignment matrices, in Transport Simulation: Beyond Traditional Approaches, Chung, E. and Dumont, A.-G. (eds), EPFL Press, 201-213.

Beckmann, M., McGuire, C. B., Winsten, C. B., 1956. Studies in the economics of transportation. Yale University Press, New Haven, CT, USA.

Bell, M., Shield, C., Busch, F. and Kruse, G., 1997. A stochastic user equilibrium path flow estimator, *Transportation Research Part C* 5(3/4), 197–210.

Bierlaire, M. and Crittin, F., 2004. An efficient algorithm for real-time estimation and prediction of dynamic OD tables. *Operations Research* 52(1), 116–127.

Carey, M., 1987. Optimal time-varying flows on congested networks. *Operations Research* 35(1), 56-69.

Cascetta, E., Inaudi, D. and Marquis, G., 1993. Dynamic estimators of origin-destination matrices using traffic counts. *Transportation Science* 27(4), 363-373.

Cipriani, E., Florian, M., Mahut, M., Nigro, M., 2011. A gradient approximation approach for adjusting temporal origin–destination matrices. *Transportation Research Part C* 19(2), 270-282.

Daganzo, C.F., 1994. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory, *Transportation Research Part B* 28(4), 269-287.

Daganzo, C.F., 1995. The cell transmission model Part II: Network traffic. *Transportation Research Part B* 29(2), 139-154

Fisk, C. S., 1989. Trip matrix estimation from link traffic counts: the congested network case. *Transportation Research Part B* 23(5), 331-336.

Florian, M., & Chen, Y., 1995. A coordinate descent method for the bi-level D-D matrix adjustment problem. *International Transactions in Operational Research* 2(2), 165–179.

Friesz, T.L., Luque, J., Tobin, R.L., Wie, B.-W., 1989. Dynamic network traffic assignment considered as a continuous time optimal control problem. Operations Research 37(6), 893-901.

Ghali, M.O. and Smith, M.J., 1995. A model for the dynamic system optimum traffic assignment problem. *Transportation Research Part B* 29(3), 155-170.

Huang, S., Sadek, A.W. and Guo, L., 2012. A computational-based approach to estimating travel demand in large-scale microscopic traffic simulation models. Accepted for publication in *ASCE Journal of Computing in Civil Engineering*. (DOI: http://dx.doi.org/10.1061/(ASCE)CP.1943-5487.0000202).

Jayakrishnan, R., Mahmassani, H. S., Hu, T.-Y., 1994. An Evaluation Tool for Advanced Traffic Information and Management Systems in Urban Network. *Transportation Research Part C*, 2(3), 129-147.

Kattan, L. and Abdulhai, B., 2006. Non iterative approach to dynamic traffic origin/ destination estimation using parallel evolutionary algorithms. *Transportation Research Record* 1964, 201–210.

Kittelson, W., Rouphail, N., Williams, B. Zhou, X. 2011. Analyzing operational improvements as an alternative to traditional highway construction. *Journal of Transportation Research Board*. 2223, 18-25.

Jayakrishnan, R., Mahmassani, H. S., Hu, T.-Y., 1994. An Evaluation Tool for Advanced Traffic Information and Management Systems in Urban Network. *Transportation Research Part C*, 2(3), 129-147.

Lawphongpanich, S., Hearn, D., 2004. An MPEC Approach to Second Best Toll Pricing. *Mathematical Programming* 101(1), 33-55.

Leblanc, L.J. and Farhangian, K., 1982. Selection of a Trip Table which Reproduces Observed Link Flows. *Transportation Research Part B* 16(2), 83-88.

Lighthill, M.J., Whitham, G.B., 1955. On kinetic wave II: a theory of traffic flow on crowded roads. *Proceedings of the Royal Society of London*, Series A 229(1178), 317- 345.

Lu, C.-C., Mahmassani, H. S., Zhou, X., 2009. Equivalent gap function-based reformulation and solution algorithm for the dynamic user equilibrium problem. *Transportation Research Part B* 43(3), 345-364

Lundgren, J. T. and Peterson, A. 2008. A heuristic for the bilevel origin–destination-matrix estimation problem. *Transportation Research Part B* 42(4), 339–354.

Mahmassani, H. S., 2001. Dynamic network traffic assignment and simulation methodology for advanced system management applications. *Networks and Spatial Economics*, 1(3/4), 267-292.

Merchant, D. and Nemhauser, G., 1978. A model and an algorithm for the dynamic traffic assignment problem. *Transportation Science* 12(3), 183-199.

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

Nguyen, S., 1977. Estimating an OD matrix from network data: A network equilibrium approach, Publication No. 87, Centre de Recherche sur les Tranports, Universite de Montreal, Montreal, Quebec.

Nie, Y, Zhang, H. M., Recker, W. W., 2005. Inferring origin-destination trip matrices with a decoupled GLS path flow estimator. *Transportation Research Part B* 39(6), 497–518

Nie, Y. and Zhang, H.M., 2008. A variational inequality approach for inferring dynamic origin-destination travel demands. *Transportation Research Part B* 42(7), 635-662.

Nie, Y. and Zhang, H.M., 2010. A relaxation approach for estimating origin-destination trip tables. *Networks and Spatial Economics* 10(1), 147-17.

Peeta, S., Mahmassani, H. S., 1995. System optimal and user equilibrium time-dependent traffic assignment in congested networks. *Annals of Operation Research* 60, 81-113.

Qian, Z. and Zhang, H.M., 2011. Computing individual path marginal cost in networks with queue spillbacks. *Transportation Research Record* 2263, 9-18.

Richards, P.I., 1956. Shock waves on the highway. *Operations Research* 4(1), 42-51.

Shen, W., Nie, Y., Zhang, H. M., 2007. On path marginal cost analysis and its relation to dynamic system-optimal traffic assignment. In: Allsop, R.E., Bell, M.G.H., and Heydecker, B.G. (Eds), *Transportation and Traffic Theory*, Elsevier, 319-352.

Shen, W., and Wynter, L., 2011. A new one-level convex optimization approach for estimating origin-destination demand. IBM Research Report, RC25147.

Sherali, H.D., and Park, T., 2001. Estimation of dynamic origin-destination trip tables for a general network. *Transportation Research Part B* 35(3), 217-235.

Tampère, C.M.J., Corthout, R, Cattrysse, D., Immers, L.H. 2011. A generic class of first order node models for dynamic macroscopic simulation of traffic flows. *Transportation Research B* 45(1), 289-309.

Tavana, H., 2001. Internally-consistent estimation of dynamic network origin-destination flows from intelligent transportation systems data using bilevel optimization. Ph.D. Dissertation, The University of Texas at Austin.

Van der Zijpp, N. J., 1997. Dynamic OD-matrix estimation from traffic counts and automated vehicle identification data. *Transportation Research Record* 1607, 87–94.

Varaiya, P., 2002. California's performance measurement system: improving freeway efficiency through transportation intelligence. TR News 218, 18-24.

Yang, H., Sasaki, T., Iida, Y. and Asakura, Y., 1992. Estimation of origin-destination matrices from link traffic counts on congested networks. *Transportation Research Part B* 26(6), 417-434.

Yang, H., 1995. Heuristic algorithms for the bilevel origin-destination matrix estimation problem. *Transportation Research Part B* 29(4), 231-242.

Zhou, X., Qin, X. and Mahmassani, H. S., 2003. Dynamic origin-destination demand estimation using multi-day link traffic counts for planning applications. *Transportation Research Record* 1831, 30-38.

Zhou, X. and Mahmassani. H. S., 2006. Dynamic OD demand estimation using automatic vehicle identification data. *IEEE Transactions on Intelligent Transportation Systems* 7(1), 105- 114.

Zhou, X. and Mahmassani. H. S., 2007. A structural state space model for real-time traffic origin-destination demand estimation and prediction in a day-to-day learning framework. *Transportation Research Part B* 41(8), 823–840.

Zhou, X., Lu, C.-C., Zhang, K., 2012. Dynamic origin-destination demand flow estimation under congested traffic conditions: a general framework. (#12-4339) The 91st Annual Meeting of Transportation Research Board, Washington, D.C., USA.

Ziliaskopoulos, A. K., Mahmassani, H. S., 1993. Time dependent shortest-path algorithm for real-time Intelligent Vehicle Highway System applications. *Transportation Research Record* 1408, 94-100.

List of Tables

Table 1 User equilibrium traffic assignment results on the two-link corridor

Path	FFTT (min)	Capacity (veh/hr)	Assigned Flow (veh/hr)	Travel Time (min)
Path 1	20	3000	5400	56
Path 2	30	3000	2600	56

Table 2 Estimation results under different degrees of information availability

Information Availability				Estimation Result			
Volume observations on path 1 only	Volume observations on path 2 only	Error-free target demand,8000 veh/hr	Error-free travel time on path 1	Flow on path 1	Flow on path 2	Total estimated demand	Equilibrium travel time (min)
X				5051.7	2367.8	7419.5	53.7
	X			4967.7	2311.8	7279.4	53.1
X	X			5011.8	2341.2	7353.0	53.4
X	X	X		5387.9	2592.0	7979.9	55.9
X	X	X	X	5401.1	2600.7	8001.8	56.0

List of Figures

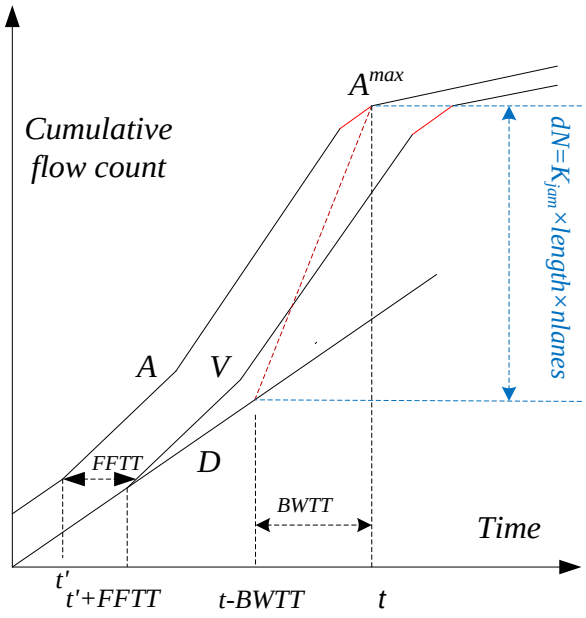

Fig. 1 Illustration of cumulative arrival and departure curves $A(t)$ and $D(t)$, and the shifted arrival curve $V(t)$ (link index l is removed for simplicity)

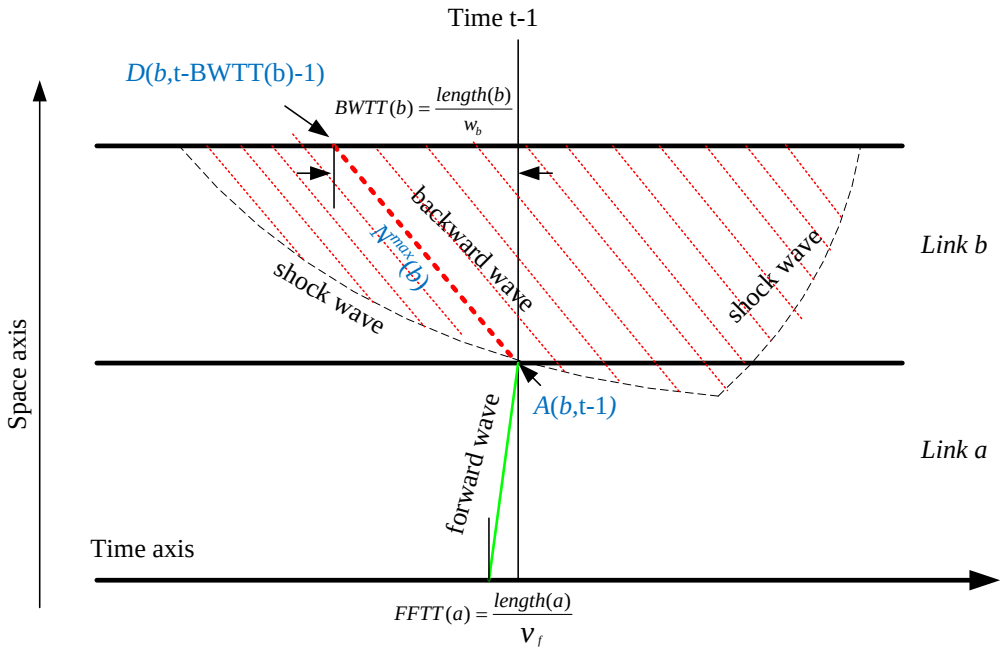

Fig. 2 Illustration of forward and backward wave representation in Newell's simplified KW model

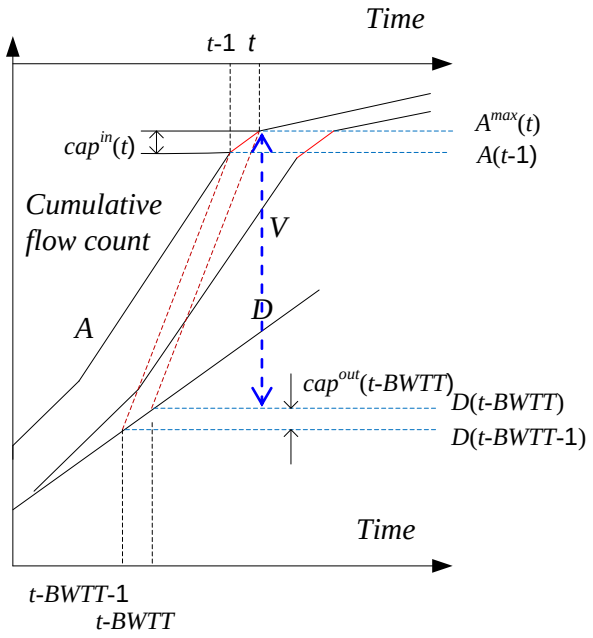

Fig. 3 Illustration of queue spillback and propagation of outgoing flow constraint from the downstream end at time *t–BWTT* to the upstream end at time *t*

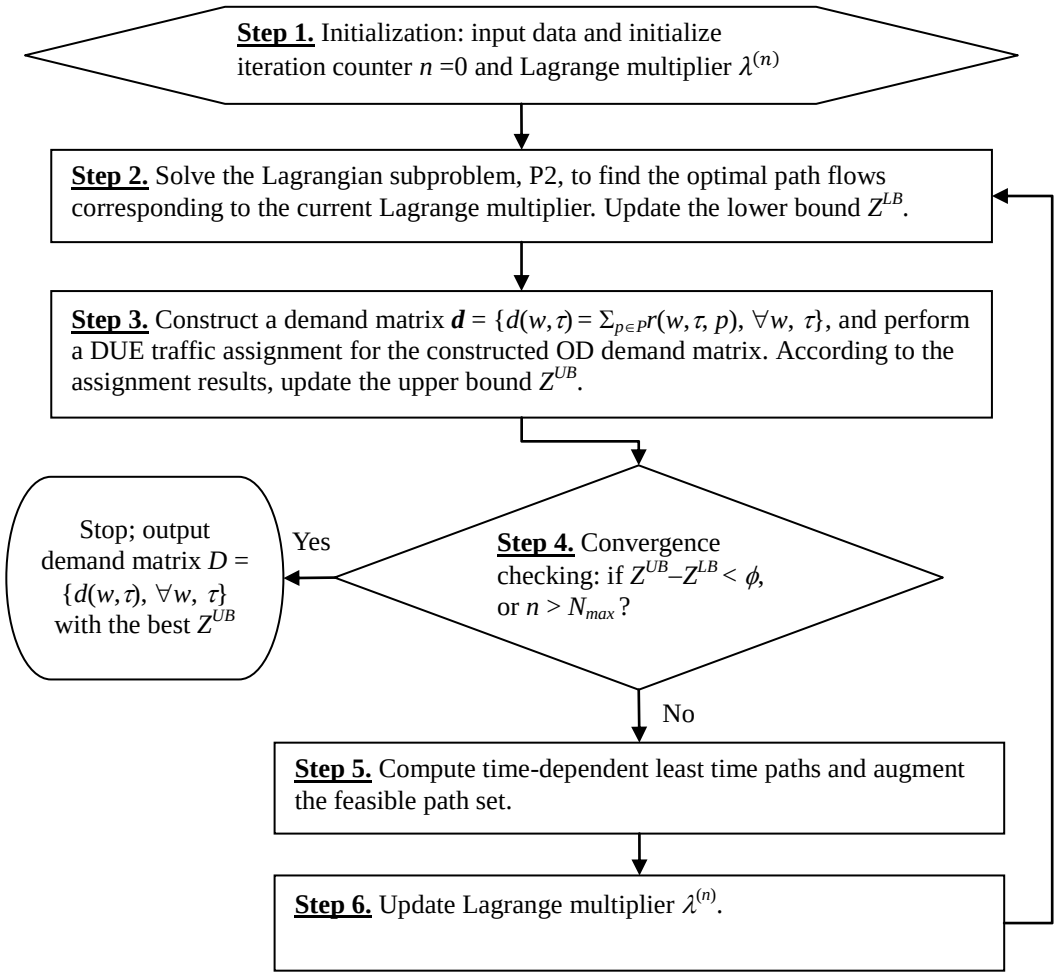

Fig. 4 Flowchart of the Lagrangian Relaxation-based heuristic

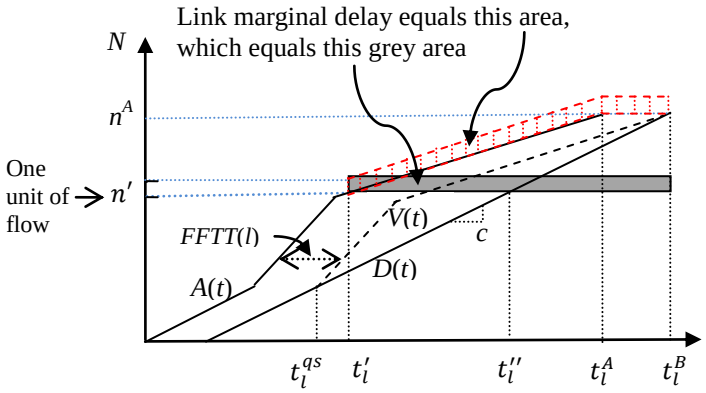

Fig. 5 Illustration of link marginal delay on a congested link

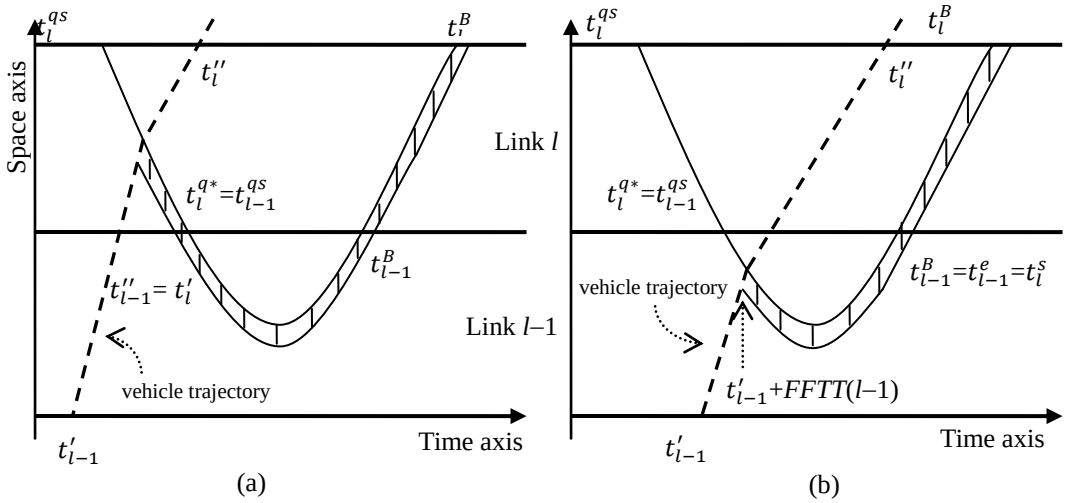

Fig. 6 Link marginal analysis for the case of queue spillback

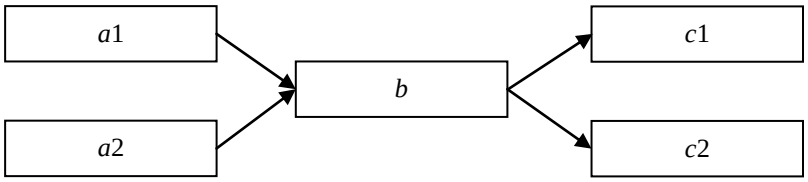

Fig. 7 An illustrative network with merge and diverge junctions

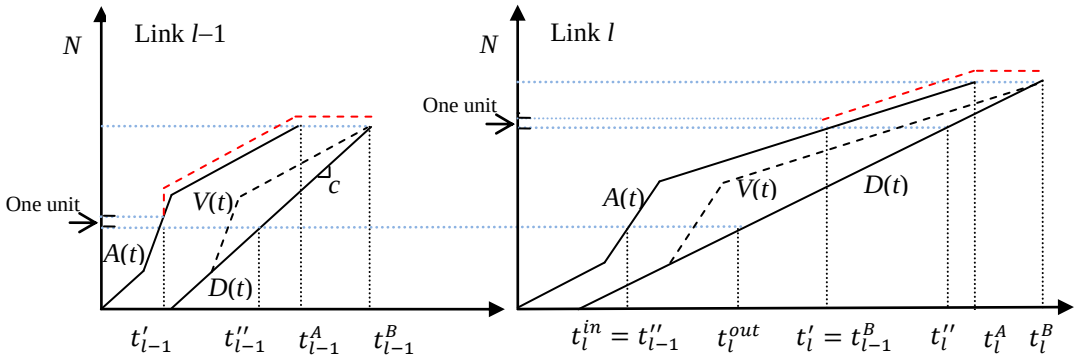

Fig. 8 Illustration of the scenario with two partially congested links

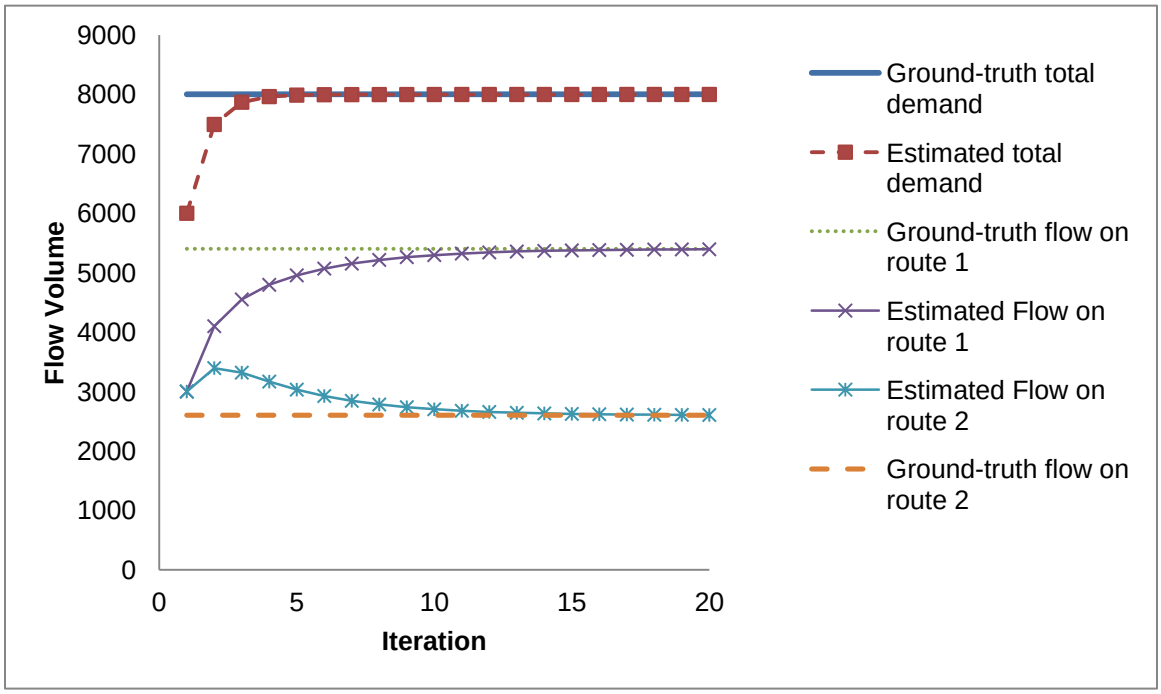

Fig. 9 Path flow volume convergence pattern as a function of inner iteration number

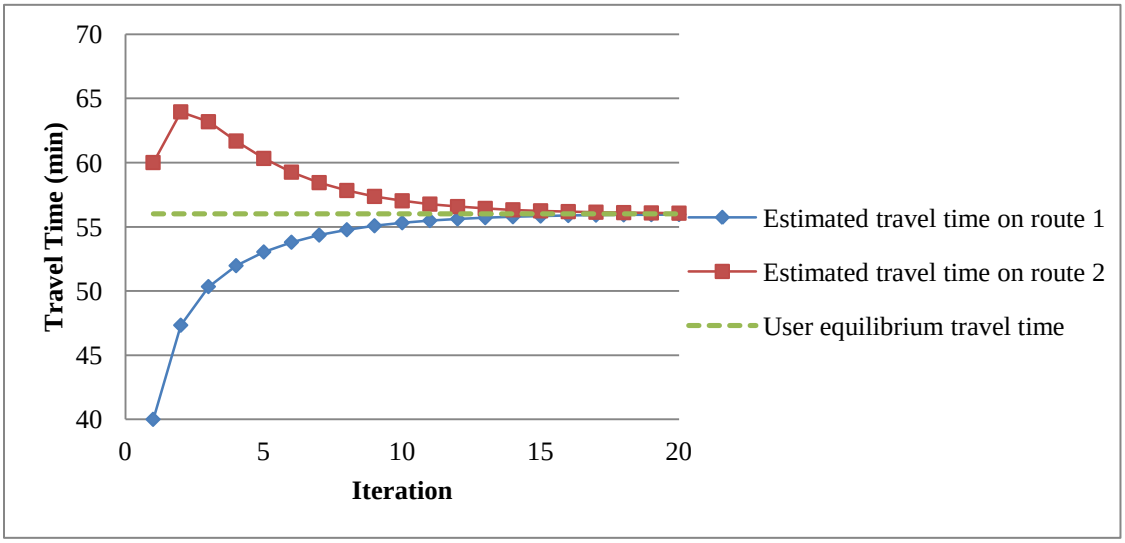

Fig.10 Path travel time convergence pattern as a function of inner iteration number

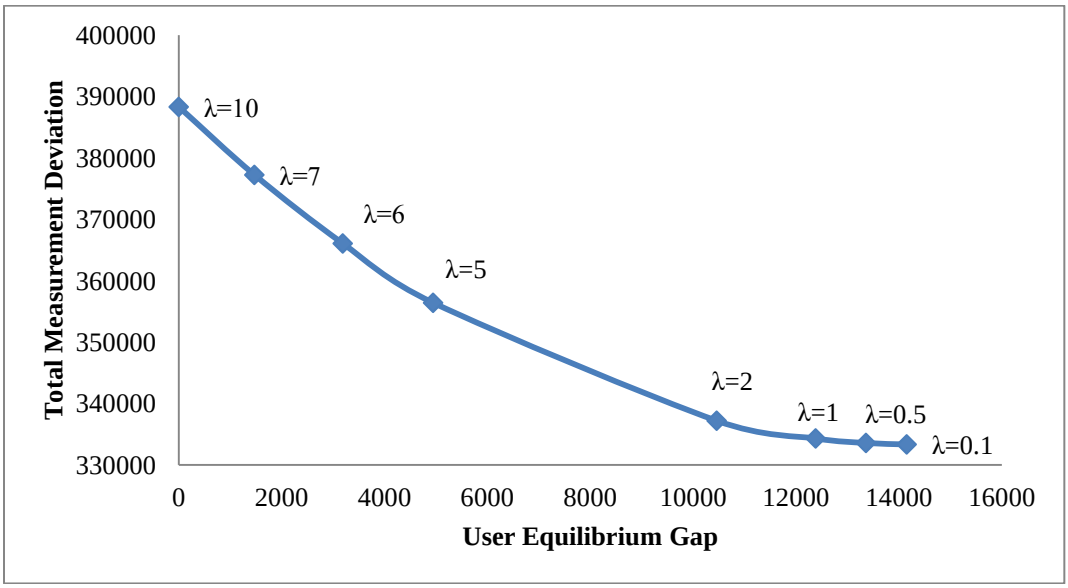

Fig. 11 User equilibrium gap vs. total deviation under different weights on the gap function

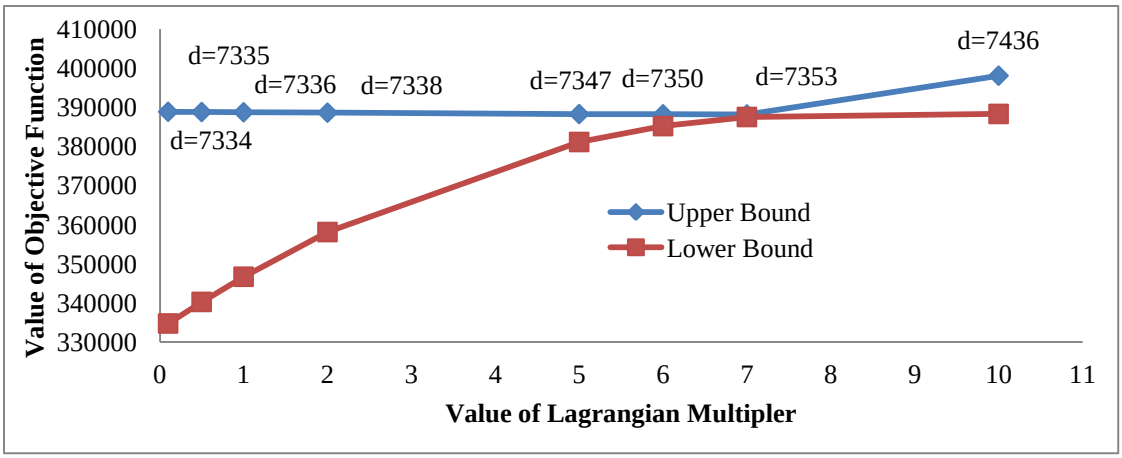

Fig. 12 Upper bound and lower bound of objective function as a function of Lagrangian multiplier

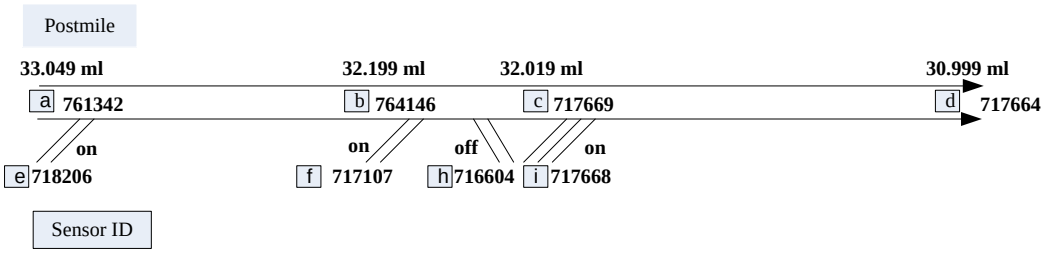

Fig. 13 Network representation of a section of I-210 West bound corridor

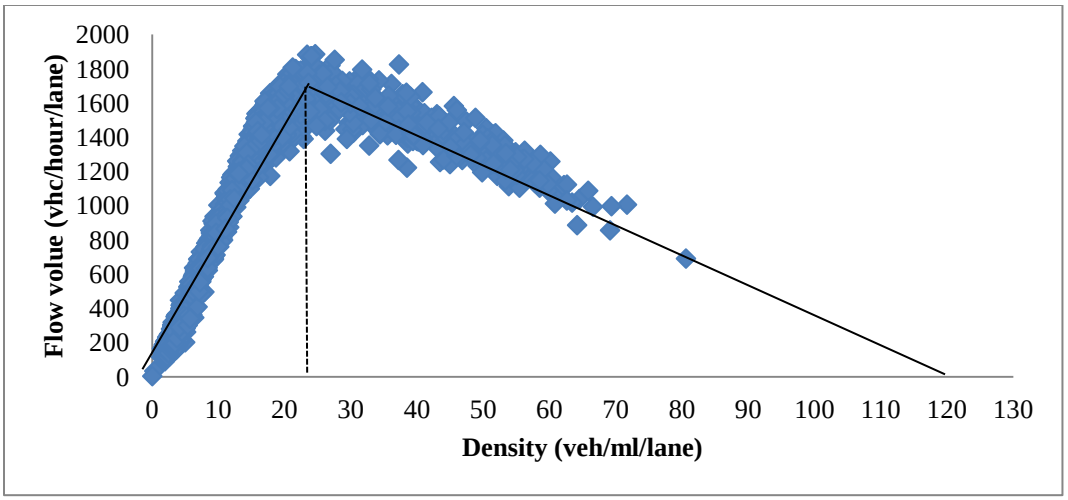

Fig. 14 Triangular relationship between flow and density at sensor *b*

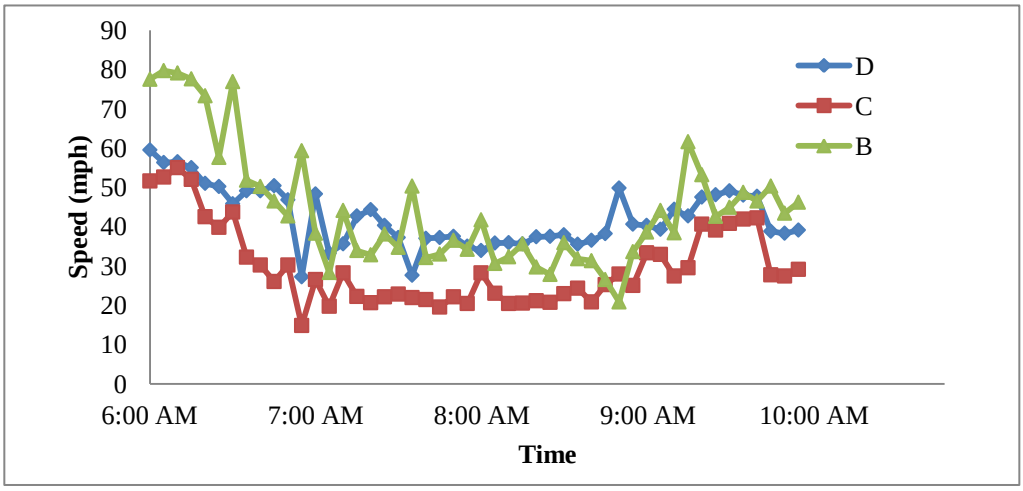

Fig. 15 Observed speed time series on freeway stations (demonstrating a congested period from 7:00AM to 9:00AM)

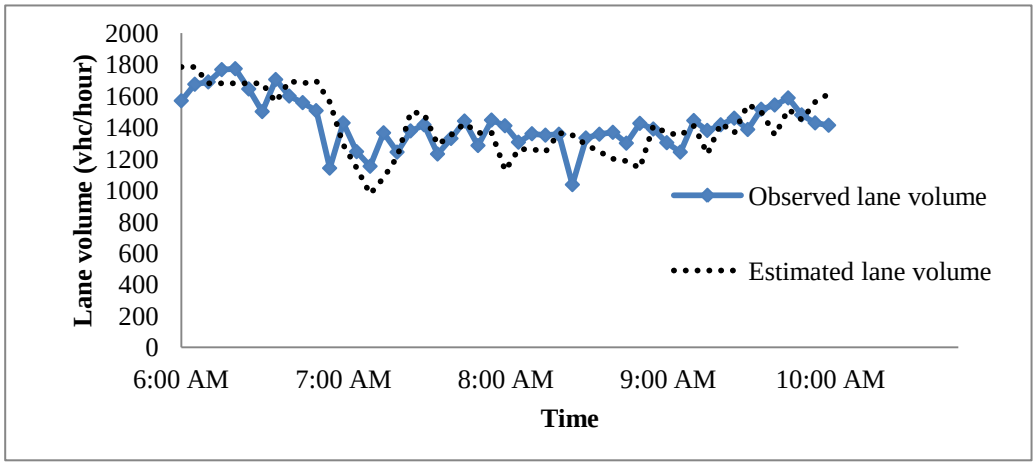

Fig. 16 Observed lane volume on station *a* vs. estimated lane volume on entrance link

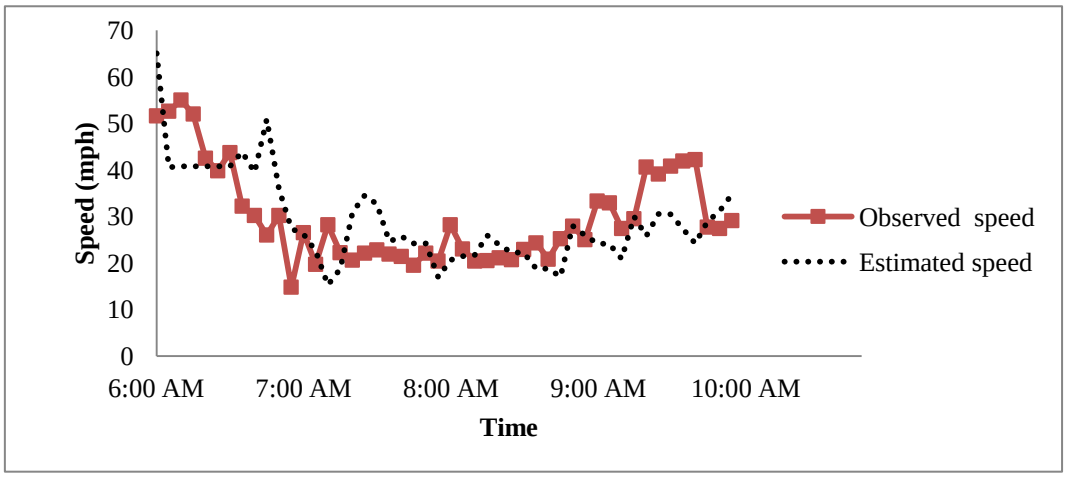

Fig. 17 Observed point mean speed at station *c* vs. estimated space mean speed on the link from off-ramp *h* to station *c*

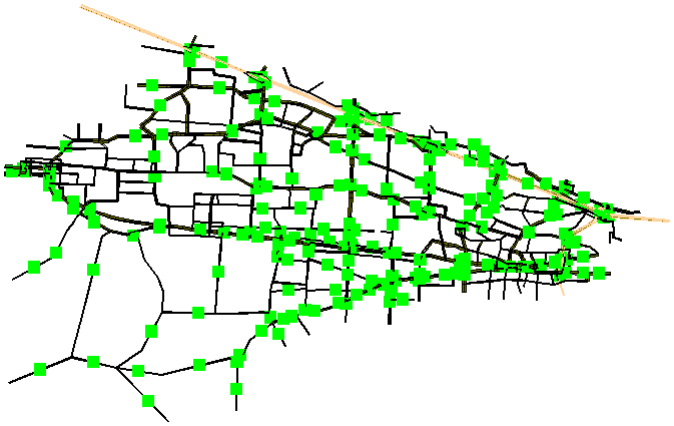

Fig. 18 Portland Beaverton subarea network with 392 AADT counting stations

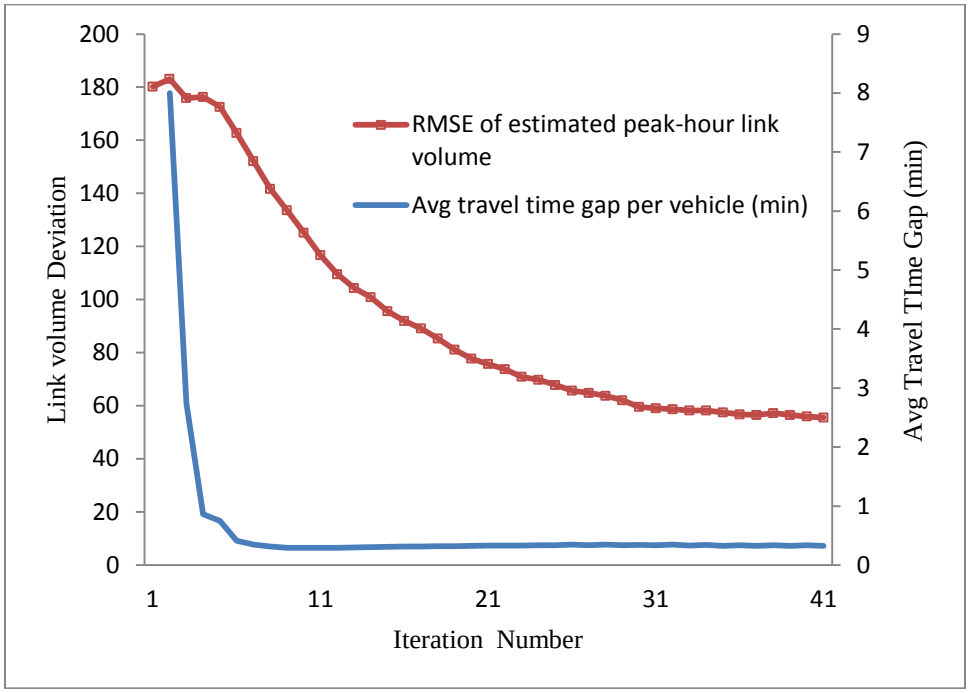

Fig. 19 MAE of the estimated link density as a function of iteration

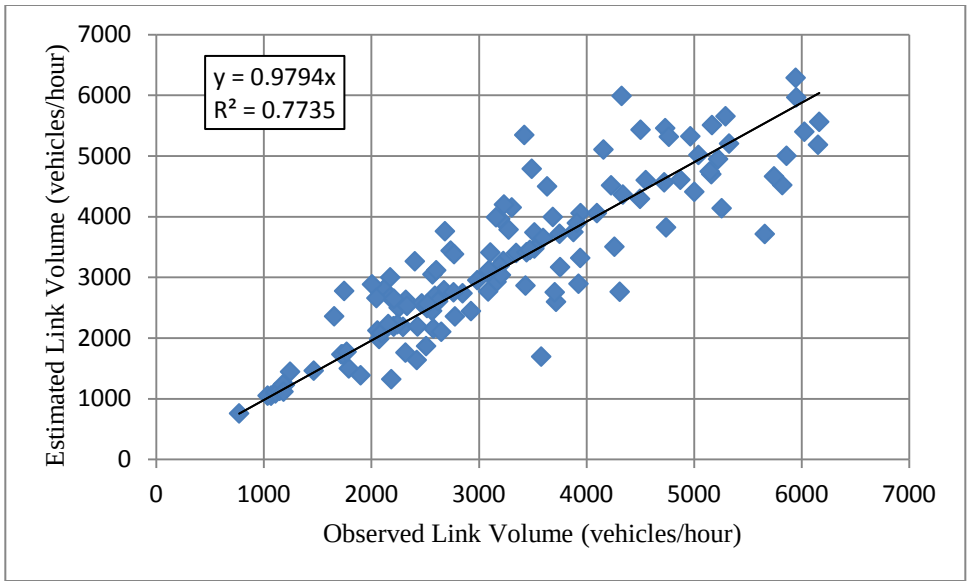

Fig. 20 Observed vs. Estimated Link Volume on Freeway Links on the Triangle Regional Model, NC network

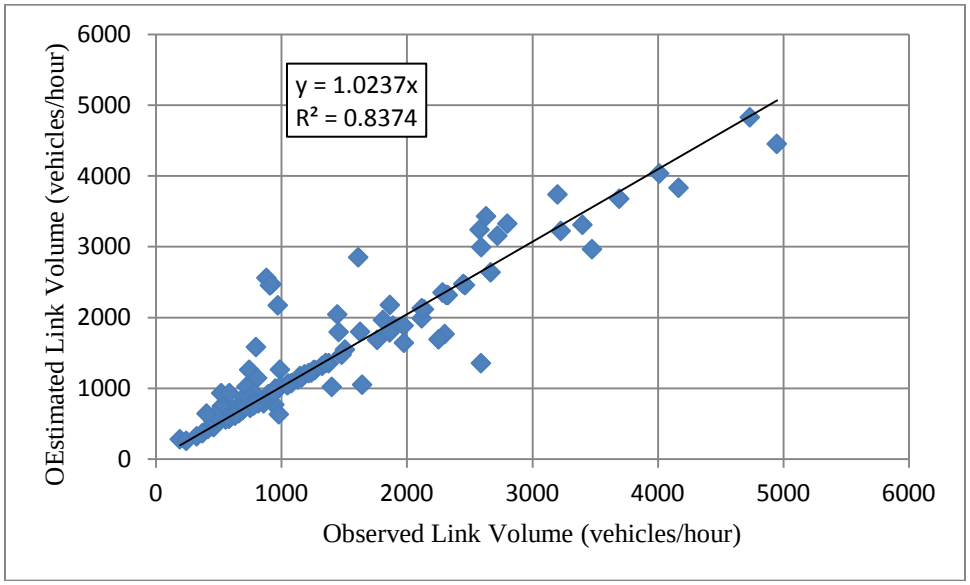

Fig. 21 Observed vs. Estimated Link Volume on Highway and Arterial Links on the Triangle Regional Model, NC network