Transportation Research Part B 94 (2016) 217–239

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Eco-system optimal time-dependent flow assignment in a congested network

Chung-Cheng Lu[a], Jiangtao Liu[b], Yunchao Qu[b,c], Srinivas Peeta[d], Nagui M. Rouphail[e], Xuesong Zhou[b,*]

[a] Department of Transportation and Logistics Management, National Chiao Tung University, Hsinchu, 300, Taiwan
[b] School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ, 85287, USA
[c] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, 100044, China
[d] School of Civil Engineering, Purdue University, West Lafayette, IN, 47907, USA
[e] Department of Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh, NC, 27695-7908, USA

ARTICLE INFO

Keywords:
Green transportation
Vehicular emission modeling
Eco-routing
Marginal emission
Multi-scale dynamic network loading

ABSTRACT

This research addresses the eco-system optimal dynamic traffic assignment (ESODTA) problem which aims to find system optimal eco-routing or green routing flows that minimize total vehicular emission in a congested network. We propose a generic agent-based ESODTA model and a simplified queueing model (SQM) that is able to clearly distinguish vehicles' speed in free-flow and congested conditions for multi-scale emission analysis, and facilitates analyzing the relationship between link emission and delay. Based on the SQM, an expanded space-time network is constructed to formulate the ESODTA with constant bottleneck discharge capacities. The resulting integer linear model of the ESODTA is solved by a Lagrangian relaxation-based algorithm. For the simulation-based ESODTA, we present the column-generation-based heuristic, which requires link and path marginal emissions in the embedded time-dependent least-cost path algorithm and the gradient-projection-based descent direction method. We derive a formula of marginal emission which encompasses the marginal travel time as a special case, and develop an algorithm for evaluating path marginal emissions in a congested network. Numerical experiments are conducted to demonstrate that the proposed algorithm is able to effectively obtain coordinated route flows that minimize the system-wide vehicular emission for large-scale networks.

1. Introduction

1.1. Motivation and objective

Highway vehicles have been the primary focus of environmental protection and transportation agencies to reduce greenhouse gas emissions, since they account for 72 percent of total transportation emissions (Greene and Schafer, 2003). Among various measures that have been considered for potential vehicular emission reduction (such as energy efficiency improvements, low-carbon alternative fuels, increasing the operating efficiency of the transportation system, and reducing travel),

* Corresponding author.
E-mail addresses: jasoncclu@gmail.com (C.-C. Lu), jliu215@asu.edu (J. Liu), quyunchao0613@gmail.com (Y. Qu), peeta@purdue.edu (S. Peeta), rouphail@ncsu.edu (N.M. Rouphail), xzhou74@asu.edu (X. Zhou).

http://dx.doi.org/10.1016/j.trb.2016.09.015
0191-2615/

eco-routing or green routing in route guidance provision is receiving increasing attention from the field of green transportation. The idea of green routing is to help drivers make greener choices about their routes by providing the most eco-friendly route in terms of minimum emissions. In a recent laboratory experiment conducted at the University of California at Berkeley, subjects were found to be willing to adjust their route choice behaviour to reduce emissions, exhibiting an average willingness to pay for emissions reduction, or value of green, of 15 cents per pound of CO_2 saved (Gaker et al., 2010, 2011).

Directing an individual vehicle to a green route can reduce its eco-cost or emission to the environment. However, without an effective system-wide coordination, independent drivers acting non-cooperatively would affect and even worsen traffic conditions and emissions. Instead, system optimal green routing or eco-routing policies that result in minimal total system emission may be of greater interest to public-sector environmental protection and traffic management agencies. The minimal total system emission serves as a benchmark to evaluate the benefits of practical traffic emission reduction measures. Moreover, the resulting green routing policies provide valuable insights for designing those measures. Therefore, this research intends to find time-dependent system optimal green routing policies (or path flow assignments) that minimize the network-wide vehicular emission, which is termed as the Eco-System Optimal Dynamic Traffic Assignment (ESODTA) problem.

Classical system optimal DTA (SODTA) models aim to direct all travelers to paths so as to minimize overall system travel time or cost (e.g., Ghali and Smith, 1995; Peeta and Mahmassani 1995a, 2001, 2006). A rich body of literature has been devoted to SODTA models and algorithms, and static traffic assignment (STA) models with environmental considerations have been proposed in a number of past studies (Szeto et al., 2012). Only recently have environment-related objectives and constraints been considered in the DTA context, as a few researchers started to recognize the need for incorporating the effect of traffic dynamics on estimating vehicular emissions so as to thoroughly consider the full set of interacting factors (e.g. Abdul Aziz and Ukkusuri, 2012; Zhou et al., 2015a; Ma et al., 2015).

As the development of ESODTA is still in its infancy, a number of critical but challenging issues need to be addressed to enable ESODTA in congested networks:

(i) The speed and time spent in the physical queue of vehicles are more appropriate traffic performance measure than the delay when evaluating vehicular emissions for vehicle-based DTA models in congested networks (Lawson et al., 1997). Thus, it is important for the underlying traffic flow model to clearly distinguish a vehicle's speed on a link in two regions: the free-flow speed v_F and the queueing speed v_Q (when moving in the queue at a lower speed). However, existing studies of DTA with emission considerations have not explicitly considered vehicles' trajectories in the congested region (physical queue). For instance, the pioneering CTM-based emission minimization SODTA models (e.g., Abdul Aziz and Ukkusuri, 2012) estimate emissions using cell-based average speed. Furthermore, in traditional point-queue models, vehicles move at free-flow speed to stop bar and wait there until discharging capacity is available.

(ii) The relationship between congestion/queueing measures (such as time spent in queue and delay) and emissions is fundamental to the development of ESODTA models in congested networks, but it was not rigorously analyzed in existing studies.

(iii) Marginal emission, the change in emission due to an additional unit of inflow, is critical to solution algorithms of simulation-based and analytical ESODTA models. Although marginal travel time (or delay) has been extensively studied in the literature, none of the existing studies has investigated emission-oriented marginal cost in congested networks. It is important to develop efficient approaches to evaluate link and path marginal emissions, as well as to study the relationship between marginal emission and marginal delay.

With the aim of expanding the boundary of SODTA from travel time-based models to emission-based models (i.e., ESODTA), this research addresses the above methodological challenges in enabling SODTA in region-wide emission applications.

1.2. Travel time-based and eco-cost-based traffic assignment models

Following the pioneering work of Merchant and Nemhauser (1978), various approaches have been proposed in the past decades to formulate and solve the travel time-based system optimal dynamic traffic assignment problem in ideal or general networks, such as mathematical programming (e.g., Merchant and Nemhauser, 1978; Carey, 1987, 1992; Ziliaskopoulos, 2000), optimal or convex control (e.g., Friesz et al., 1989; Lafortune et al. 1993; Wie et al., 1994), simulation-based optimization (Ghali and Smith, 1995; Peeta and Mahmassani, 1995a, b; Peeta and Zhou, 2006), game theory (Garcia et al., 2000), graphical method (Munoz and Laval, 2006; Shen and Zhang, 2009) and variational inequality (Shen et al., 2007a).

As extensions of the user equilibrium and system optimum principles, eco-cost-based, or emission-based, assignment principles have been adopted in a number of STA models. For instance, Benedek and Rilett (1998) presented the emission optimal principle which describes that travellers choose paths so as to minimize the total network emission, rather than total travel time. They also discussed an extension of the user equilibrium principle, the environmental equity principle, in which travellers are assigned in such a way that the amounts of emission on all selected routes are the same. Another line of research was to employ the multi-objective or multi-criterion approaches in traffic assignment models. For example, a multi-criterion system optimum model was proposed by Tzeng and Chen (1993), where the system optimum objective is the sum of total travel time for road users and air pollution for non-users. Nagurney et al. (1998, 2002) presented a

multi-class user equilibrium traffic assignment model in which each class of users was assumed to select a route with the least weighted sum of travel time, travel cost and emissions. Zhang et al. (2010a) developed a system optimal STA model with the objective being the weighted sum of travel time and emissions. They introduced a cell-based modelling approach for emission concentrations so that either the average or maximum emissions in a network can be considered in the optimization process. A comprehensive review of network equilibrium approaches addressing environmental concerns (e.g., emissions and noise) can be found in Szeto et al. (2012).

Despite the numerous aforementioned studies of STA with environmental considerations, very few DTA models have been developed for environment-related applications. Recently, Abdul Aziz and Ukkusuri (2012) integrated emission-based objective into the traditional travel time-based DTA framework, and developed a SODTA model with dual objectives. They formulated the problem as a nonlinear quadratic program which is readily solved by CPLEX. Using a light-weight emission estimator MOVE Lite developed by Frey and Liu (2013), Zhou et al. (2015a) presented a DTA model and its solution algorithm for a number of emerging emissions and fuel consumption related applications that require both effective microscopic and macroscopic traffic stream representations. Vallamsundar et al. (2016) integrated this DTA model into a comprehensive modeling framework for transportation-induced population exposure assessment.

1.3. Microscopic and macroscopic traffic flow models for emission estimation

In order to capture the impact of traffic congestion on the energy use and emissions output across different spatial scales (e.g., regions, corridors, segments, and intersections) and various temporal resolutions (e.g., second-by-second, peak hours, and entire day), it is essential for underlying traffic flow models to be able to capture traffic dynamics and describe congestion phenomena (e.g., queue formation, spillback, and dissipation). Microscopic traffic simulation models have been widely used to generate instantaneous speed and acceleration data required by emission models on a vehicle-by-vehicle and second-by-second basis (e.g., Bai et al., 2007; Boriboonsomsin and Barth, 2008; Mandavilli et al., 2008; Panis et al., 2006). However, microscopic traffic simulation is computationally expensive and typically requires a wide range of detailed geometric data and driving behavior parameters, which are difficult to calibrate. Mesoscopic traffic flow modeling may be a more viable approach to strike a balance between the model and computational complexities and the emission resolution.

In their pioneering works, Lighthill and Whitham (1955) and Richards (1956) (LWR) proposed the kinematic wave theory, which rigorously describes traffic flow dynamics by integrating flow conservation constraints, traffic flow models, and partial differential equations (PDEs). Based on a triangular flow density relation, two finite difference-based numerical schemes were proposed to solve the first order kinematic wave problem: (i) by extending deterministic queuing theory, Newell's simplified model (Newell, 1993a, b, c) keeps track of shock waves and queue propagation using cumulative flow counts on links; (ii) Daganzo's Cell Transmission Model (CTM; Daganzo, 1994, 1995) discretizes a link into many homogenous segments (i.e. cells), and adopts a "supply-demand" or "sending-receiving" framework to model flow dynamics between cells.

Abdul Aziz and Ukkusuri (2012) adopted the CTM as the traffic flow model underlying their SODTA model and derived link emissions based on the average speed inside a cell in a time interval (e.g. 60 seconds). Note that the average cell-speed approach may not effectively estimate time-dependent emissions, which are highly sensitive to second-by-second speed variations across different locations. In their work of modeling delay and emission for signalized intersections, Zhu et al. (2013) studied different dynamic traffic models for network loading that can produce the speed profile, including the car-following model, the point-queue model, the shockwave model, and the CTM, and applied them for both delay and emissions estimation.

Classical point-queue models assume that the link travel time consists of two parts: free-flow travel time and delay. Delay is a typical measure of the impact of congestion on travelers' time. However, for evaluating the congestion effect on vehicular emission and energy consumption, the more appropriate measure is the amount of time actually spent in queue (waiting time or time in queue), which is usually greater than the delay. To effectively measure the time and distance spent by vehicles in a queue, Lawson et al. (1997) proposed using the input-output (or queueing) diagram to determine the spatial and temporal extents of queue upstream of a bottleneck. They derived the relationship between delay and waiting time in queue and constructed the curve depicting the cumulative number of vehicles to have reached the back of the queue as a function of time.

Addressing the need to consistently incorporate different resolutions of traffic descriptions in a traffic flow model, Leclercq (2007) proposed a hybrid LWR model combining both macroscopic and microscopic traffic descriptions and defining simple interfaces to translate the boundary conditions when changing the traffic description. Recently, Zhou et al. (2015a) proposed a mesoscopic DNL model that seamlessly integrates Newell's simplified kinematic wave model (a macroscopic model) and simplified car-following model (a microscopic model) into a unified framework, to evaluate vehicle emission/fuel consumption impact of different traffic management strategies. The advantages of the mesoscopic approach in computational efficiency and in effectively describing free-flow and congested traffic states make it appealing for cross-resolution and multi-scale emission modelling in DTA applications. Zhang et al. (2013a), Ma et al. (2014), and Doan and Ukkusuri (2015) offered in-depth discussions on the use of point queue, spatial queue, cell transmission and continuous-time double queue models in SODTA. Recently, Ma et al. (2015) further developed an innovative traffic emission pricing model for dynamic traffic networks with single destinations, in which a first-best dynamic emission pricing scheme is examined to systematically consider both portions of free-flow travel time and shockwave travel times. As for the energy-efficient and emission-reduction operation model for the train flow in the railway/metro systems, many innovative models and algorithms

have been proposed by a variety of researchers in the literature (see Yang et al. 2015, 2016; Huang et al. 2016). Additionally, Yin et al. (2016) first proposed an approximate dynamic programming approach to solve the train-flow based rescheduling problem with the consideration of energy-efficient operation strategies.

1.4. Numerical solution methods for system optimal traffic assignment models

Both exact and heuristic methods have been developed in the literature to solve the SODTA problem on ideal networks or general networks with multiple origin-destination (O-D) pairs. Exact methods were mainly applied to solve the link-based SODTA problem formulated as mathematical programming or optimal control problems. For instance, Merchant and Nemhauser (1978) solved a piecewise linear version of their model by a one-pass simplex method. Wie et al. (1994) developed an augmented Lagrangian method in conjunction with the conjugate gradient method to solve the discrete time optimal control formulation of the problem.

On the other hand, heuristics based on some predefined averaging schemes (e.g., Magnanti and Perakis, 1997), such as the method of successive averages (MSA), have been used for solving path-based SODTA problems (Peeta and Mahmassani, 1995a, b; Peeta and Zhou, 2006; Shen et al., 2007a). The drawback of using MSA is that it uses an across-the-board step size for updating path assignments, so the degree to which the path flows deviate from optimality conditions is not taken into account for different O-D pairs and departure intervals. This may lead to a slow convergence or even failure to converge for some problem instances. (e.g., Mounce and Smith, 2007). To improve the convergence and the solution quality, Sbayti et al. (2007) proposed an efficient MSA-based implementation technique that uses a sorting technique in updating vehicle assignments based on a selected path travel attribute (e.g. travel time). Lu et al. (2009) developed a path-swapping method which was shown to outperform the MSA on several large network tests.

The link marginal delay (or travel time), which represents the change in delay due to an additional unit of link inflow, is critical to the solution algorithms of SODTA. Ghali and Smith (1995) presented an analytical approach to evaluate link marginal delays on a congested link, based on link cumulative flow curves. Peeta and Mahmassani, (1995a) developed a numerical method based on mesoscopic traffic simulation to evaluate link and path marginal travel times. Peeta and Zhou (2006) extended these concepts to the stochastic case under O-D randomness for multiple user classes. Shen et al. (2007a) 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 marginal delays (Qian and Zhang, 2011; Qian et al. 2012). Focusing on the connection from dynamic system optimum to dynamic user equilibrium conditions, Carey and Watling (2012) developed an analytical model for calculating the marginal costs and externalities based on simplified kinematic wave model. Lu et al. (2013) further examined the partial derivatives of link flow and density and path travel time with respect to an additional unit of perturbation flow. While link and path marginal delays were investigated in the context of SODTA, to the authors' current knowledge, there is no study that addresses link and path marginal emissions.

1.5. Overview of the paper

In Section 2, we present a generic agent-based model for the ESODTA problem, which aims to minimize total network-wide vehicular emission, and a simplified queueing model (SQM), which is able to clearly distinguish vehicles' speed in free-flow and congested conditions and effectively generate time-dependent speeds for multi-scale emission analysis. In Section 3, with the assumption of constant bottleneck discharge rates and without queue-spillback effect, we discuss an important property of link emission describing the relationship between emission and delay. Based on the SQM, an expanded space-time network is constructed to formulate the ESODTA with constant discharge rates as an integer linear model, which is solved by a Lagrangian relaxation-based algorithm.

In addition to the analytical ESODTA model presented in Section 3, Section 4 describes a simulation-based ESODTA model, where constant and time-dependent bottleneck discharge rates can be considered. A column generation-based algorithm, which consists of a mesoscopic dynamic network loading (DNL) model and a gradient projection-based descent direction method for updating time-dependent path assignments. We also derive a formula of link marginal emission which encompasses the link marginal travel time as a special case, and develop an algorithm for evaluating path marginal emissions in a congested network.

Section 5 presents two numerical examples for the ESODTA with constant bottleneck capacities: one based on the analytical formulation in Section 3 while the other based on the simulation-based model in Section 4. We compare the results of the ESODTA with those of the travel time-based SODTA and of the user equilibrium DTA (or UEDTA). Concluding remarks are given in Section 6.

2. Agent-based ESODTA model

The notations used to present the generic agent-based ESODTA model are defined as follows.

Indices

τ	Index of departure time interval
w	Index of OD pair
p	Index of path for time-dependent OD pair (w, τ)

i, j, k	Node index in physical network
$l, (i, j)$	Index of link $l=(i, j)$, $l = 1, 2, \ldots, L$
t	Index of simulation time interval
f	Index of agent (vehicle) with its departure time τ, OD pair w, and path p; $f=f(w, \tau, p)$

Sets

N	Set of nodes in the physical network
E	Set of road links in the physical network
F	Set of agents (vehicles)
$P(w, \tau)$	Path set of OD pair w and departure time interval τ

Parameters

$d(w, \tau)$	Number of vehicles for OD pair w and departure time interval τ
$\mathbf{d}$	Vector of time-dependent demands of all OD pairs
$v_F(l)$	Free-flow speed on link l, which is also the speed limit on link l
$v_Q(l, t)$	Queueing speed (or speed in queue) on link l at time interval t

Variables

$r(w, \tau, p)$	Number of vehicles on path p of OD pair w and departure time interval τ
$q(l, t)$	Time-dependent link flow of link l in time interval t
$\mathbf{q}$	Vector of link flows
$TT(f, l, t)$	Agent f's time-dependent travel time on link l in time interval t
$\mathbf{TT}$	Vector of agents' path travel times
$EC(f)$	Total path emission or eco-cost of agent f
$EC(f, l)$	Emission or eco-cost of agent f on link l
$EC(f, l, v(t))$	Instantaneous emission of agent f with (time-dependent) speed v on link l in time interval t
$EC(f, l, t', t'')$	Emission of agent f on link l with entering time t' and leaving time t''
$\mathbf{EC}$	Vector of path emission cost of all agents
TE	Total system emission cost

2.1. Model formulation

Consider a road network $G=(N, A)$ with a set of nodes N and a set of links E. Each link is denoted as a directed link $l=(i, j)$ from upstream node i to downstream node j. In the ESODTA, all (green) travellers are assumed to behave cooperatively in their route choices to minimize the total emission. For simplicity and with no loss of generality, three basic assumptions are made for this model: (i) departure time choices are not considered and time-dependent O-D travel desires are assumed to be given; (ii) link attributes, such as free-flow travel time, bottleneck discharge rate, density-flow relationship, are given to perform a dynamic network loading process and generate time-dependent vehicular trajectories; (iii) the emission cost function is given to estimate vehicular emissions. With these assumptions, the proposed ESODTA aims to determine the time-varying agent (path) flows which minimize the total network emission.

The agent-based ESODTA is presented as follows

Agent-based ESODTA

$$\text{Min } TE = \sum_f EC(f) \tag{1}$$

$$\text{Subject to } [\mathbf{q}, \mathbf{TT}, \mathbf{EC}] = \text{DNLE}(\mathbf{d}) \tag{2}$$

$$EC(f) = \sum_l EC(f, l) \tag{3}$$

$$\sum_{p \in P(w,\tau)} r(w, \tau, p) = d(w, \tau), \ \forall w, \tau \tag{4}$$

$$r(w, \tau, p) \geq 0, \ \forall w, \tau, \ p \in P(w, \tau) \tag{5}$$

The objective function, Eq. (1), minimizes the network-wide vehicular emission. Constraint (2) states that, given the time-dependent OD demand vector $\mathbf{d}$, the estimated path emissions of all agents, $\mathbf{EC}$, is obtained by a dynamic network loading (DNL) model for emission estimation, DNLE($\mathbf{d}$). Constraint (3) describes that the time-varying path emission $EC(f)$ of agent/vehicle f is assumed to be the sum of the emissions on its constituent links. Constraint (4) is the demand flow balance constraint for each OD pair w and each departure time interval τ. Constraint (5) requires non-negative path flows. Note that the above formulation is a generic ESODTA model in which different emission cost functions, $EC(f, l)$, and DNL models, DNLE($\mathbf{d}$), can be embedded for multi-scale emission optimization applications.

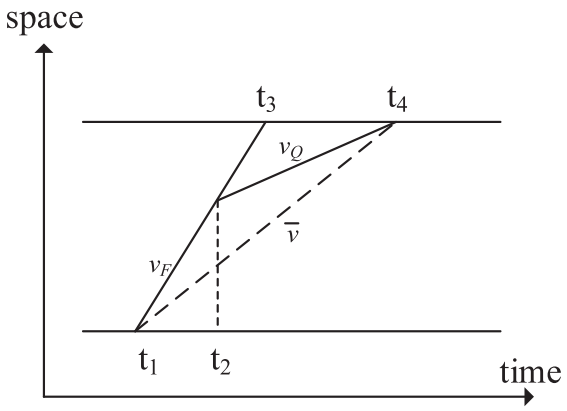


Fig. 1. Three mesoscopic methods for obtaining an agent's trajectory.

2.2. Dynamic network loading model for emission estimation

To enable multi-scale emission optimization applications, the ESODTA model relies on an effective DNL model (DNLE(**d**)) for obtaining high-resolution emission estimations. Essentially, given the network (link and node) data and time-dependent OD demands, the DNL model generates a set of vehicles along with their attributes (e.g., OD, departure time, type and age) that are loaded to the network, and performs traffic simulation to evaluate time-dependent link speed and acceleration/deceleration profiles, which are used in the emission estimation function for computing emission costs. The major difference between the DNLE(**d**) model for the ESODTA and classical DNL models for general DTA applications is that the former is used to generate detailed vehicle trajectories and time-dependent link speeds, whereas the latter's major outputs are link travel times and (aggregated) flows. As the context of DTA is extended from travel time-based models to emission or eco-cost based applications, it is necessary for the DNLE(**d**) model to output time-dependent speeds for estimating vehicular emissions.

Technically, both microscopic and mesoscopic traffic flow models can be embedded in DNLE(**d**) to construct detailed time-dependent speed profiles for emission estimation. Microscopic models have been widely used to generate vehicle emissions estimates by evaluating driving speed and acceleration characteristics/profiles on a vehicle-by-vehicle and second-by-second basis. Although microscopic traffic flow models, such as cellular automatic (CA) and car following models, are desirable for analyzing vehicular delays in congested conditions, microscopic simulation can be computationally intensive and typically requires a wide range of detailed geometric data and driving behavior parameters, which can be difficult to calibrate, especially for the purpose of producing high-fidelity emissions estimates.

To strike a balance between the model and computational complexities and the emission resolution, this research adopts the mesoscopic modeling approach, which adapts the simplified queueing model (SQM), proposed by Lawson, Lovell and Daganzo (1997), to generate detailed vehicle trajectories and time-dependent link speeds. The SQM, motivated by Newell's simplified kinematic wave model, is derived using the input-output diagram (or cumulative arrival and departure curves) and based on the assumption of constant bottleneck discharge rates. Note that the CTM dissects each link into multiple cells and assumes piecewise linear relationship between traffic flow and density for each cell. Actually, despite the distinct representations of queues in the two models, Daganzo (2006) proved that the vehicle trajectories predicted by cellular automata models match those predicted by Newell's simplified car-following model (Newell, 2002) and the simplified kinematic wave model with a triangular fundamental diagram (Newell, 1993a).

Recently, Abdul Aziz and Ukkusuri (2012) extended the CTM-based formulation of Ziliaskopoulos (2000) to develop (link flow-based) DTA models with emission considerations, in which the emission in each cell in a time interval (e.g., 60 seconds) was estimated using the cell-based average speed and the speed was determined from the speed-density relationship. In contrast, the ESODTA model proposed in our paper is a vehicle-based (or agent-based) model that aims to determine time-dependent system optimal green routing policies for the vehicles in a traffic network.

Because the speed and the time spent in queue of a vehicle are more appropriate traffic performance measure than the delay (or waiting time in a vertical queue) when evaluating vehicular emissions in congested networks (Lawson et al., 1997), we need to clearly distinguish a vehicle's speeds on a link in two regions: the free-flow speed v_F and the queueing speed v_Q (when moving in the queue at a lower speed) for estimating vehicular emissions, rather than relying on average speeds. The adopted SQM facilitates this modeling requirement and allows us to analytically derive the time spent in queue t_Q (including the moving and waiting times in the physical congested region) from the delay t_w. Fig. 1 depicts the trajectories of a vehicle on a link generated by three different methods based on mesoscopic traffic simulation that can be applied to obtain time-dependent travel times. The link index l is omitted for clarity. Assume that there is an agent entering the link at time t_1 and leaving the link at time t_4. The agent's trajectories described by the three methods are as follows. With Method 1 (point-queue model), the agent moves at free-flow speed v_F until it reaches the downstream node at time t_3, and then stops at

Table 1
Travel speeds and travel times obtained by the three methods under free-flow and congestion.

	Free-flow speed	Free-flow travel time	Speed in queue	Waiting time (time in queue)	Total travel time
Method 1	v_F	$t_3 - t_1$	0	$t_4 - t_3$	$t_4 - t_1$
Method 2	$\bar{v}$	$t_4 - t_1$	$\bar{v}$	0	$t_4 - t_1$
Method 3	v_F	$t_2 - t_1$	v_Q	$t_4 - t_2$	$t_4 - t_1$

this node until the capacity is available for discharging this agent at time t_4. With Method 2 (average link-speed model), the agent moves at an average speed $\bar{v}$ through the link. With Method 3 (SQM), the agent moves at a free-flow speed v_F until reaches the back of the queue at time t_2, and then at a slower speed v_Q before the bottleneck (e.g., Lawson et al., 1997).

Table 1 summaries the travel times and travel speeds obtained by the three methods under free-flow and congested conditions. Although the three methods output the same link travel time for the agent, the other four measures and the vehicle trajectory differ significantly in the three methods. Of particular concern are the waiting time (or time in queue) and speed in queue of the vehicle, which are critical for estimating emissions. The first method based on the point-queue model is able to approximate the delay of the vehicle at the bottleneck, but the time and speed in queue are not effectively represented by this method, which will underestimate the agent's emission on the link. The average speed obtained using the second method cannot truly reflect the speed changes of vehicles, leading to inaccurate estimates of emission in general. On the other hand, the SQM explicitly takes into account the physical queue in describing the vehicle trajectory under congested condition, resulting in a more reasonable estimate of emission for the vehicle.

Typically, the queueing speed v_Q is related to the bottleneck discharge capacity. Assume that the discharge capacity is constant and there is no queue-spillback effect, then the queuing speed is constant, which can be obtained by the flow-speed relationship curve. Consequently, the vehicular emission on this link is a linear function of the delay. This important property of link emission will be rigorously derived in Section 3. Moreover, based on Method 3, this study constructs the expanded space-time network to model the ESODTA.

The link emission cost represents the impacts of a vector of vehicular emissions (such as air pollutants: CO, NO and HC and greenhouse gases CO_2) on the environment and depends typically on the speed-based emission profile in a time-space cell. The total emission of agent f through link l, $EC(f, l, t', t'')$, is an integration of the instantaneous emission cost $EC(f, l, v(t))$ during the entrance time t' and the exit time t''.

$$EC\left(f, l, t', t''\right) = \int_{t'}^{t''} EC(f, l, v(t))dt. \tag{6}$$

Typically, vehicular emission is a function of vehicle speed and acceleration, while correction factors can also be applied to the function to take into account different vehicle types, roadway characteristics, driving patterns and weather conditions.

The emission costs per unit time of vehicles traveling in free-flow speed and congested speed can be determined using a speed-based emission model (e.g., Szeto et al., 2012). For example, Abdul Aziz and Ukkusuri (2012) followed a similar methodology to estimate regression models that correlate speed and CO emission from MOBILE 6.2. The CO emission rate (gm. of CO per vehicle per second) at the speed v (miles per hour) can be expressed as,

$$EC(v) = -0.064 + 0.0056v + 0.00026\ (v - 50)^2 \tag{7}$$

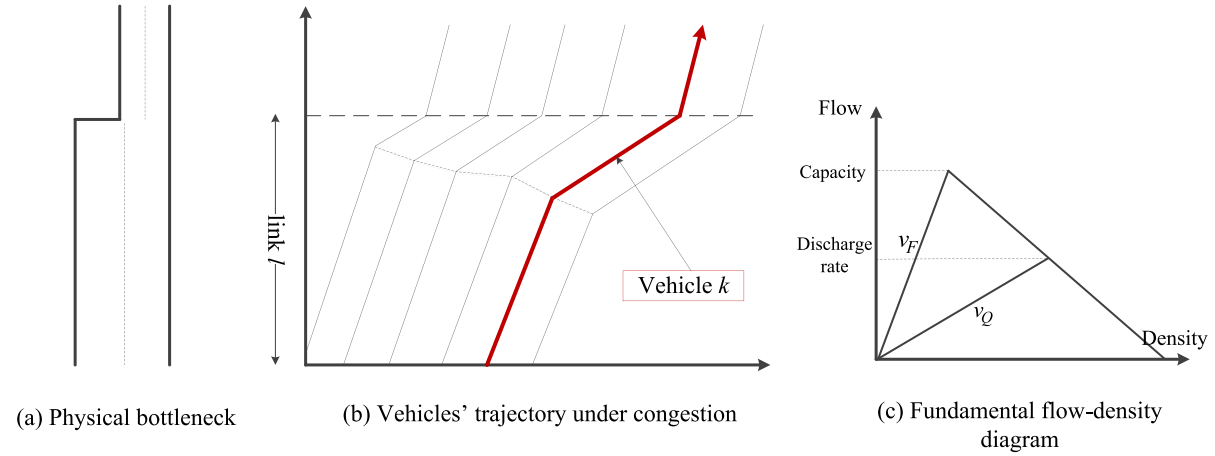
According to U.S. EPA, among the primary air pollutants, the top contributor that requires an air-quality standard is carbon monoxide (CO). Further, air quality data for 2008 reported by the EPA identifies metropolitan areas exceeding the CO emission thresholds set by National Ambient Air Quality Standards (NAAQS) (Zhang et al., 2010b). Thus, CO emission from on-road vehicles is a major issue that requires attention in transportation planning.

3. An integer linear programming model for the ESODTA

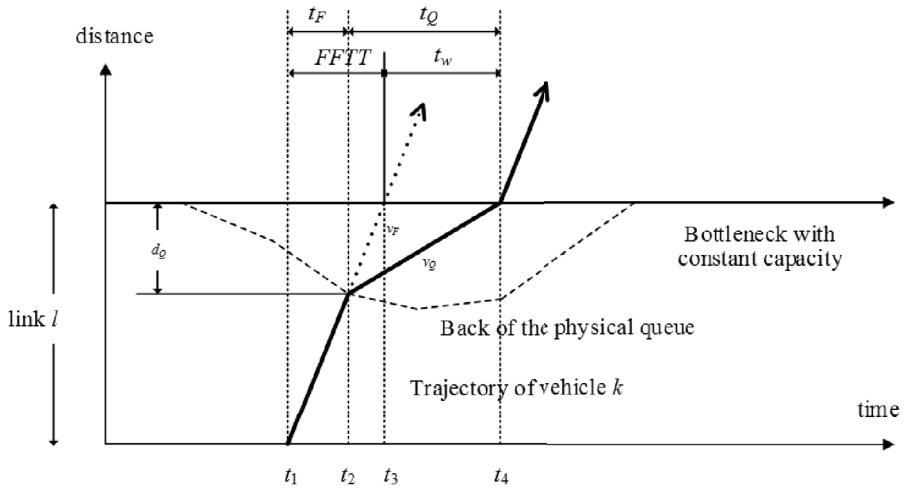
This section presents an integer linear programming model for the ESODTA with constant bottleneck discharge rates. We first discuss an important property of the link emission of an agent under congestion and describe the expanded space-time network, based on which the integer linear programming model for the ESODTA is developed. Then, we present the Lagrangian relaxation-based solution algorithm for solving the model.

3.1. Property of link emission under congestion

Consider the physical bottleneck due to lane reduction, as shown in Fig. 2(a). The bottleneck has a constant maximum discharge rate, c. The time-space diagram of Fig. 2(b) represents a set of vehicle trajectories approaching the bottleneck. We assume that a constant free-flow speed v_F holds for all uncongested traffic, and that whenever congestion occurs upstream of the bottleneck, vehicles traverse the queue at some (reduced) constant speed v_Q, where v_F and v_Q can be obtained from the fundamental flow-density diagram shown in Fig. 2(c). We also assume that for simplicity the speed changes from v_F to v_Q occur instantaneously. Because the acceleration and deceleration in state transitions are not considered, this part of



(a) Physical bottleneck (b) Vehicles' trajectory under congestion (c) Fundamental flow-density diagram



(d) Vehicle k's trajectory on link l with a bottleneck (adopted from Lawson et al., 1997)

Fig. 2. Vehicle trajectory on a link with a bottleneck.

emissions is not included in the proposed model, which is the limitation of the proposed approach. Fig. 2(d) is the time-space diagram which shows the trajectory of vehicle k moving through the bottleneck. The solid line depicts the actual trajectory of vehicle f on link l, while the dotted line represents the desired trajectory of the vehicle through the bottleneck.

The vehicle enters the link at time t_1, reaches the back of the queue (represented by the dashed curve) at time t_2, desires to leave the link at time t_3, and eventually leaves the link at time t_4 due to capacity restriction. Here, we denote the travel time under free-flow speed as $t_F = t_2 - t_1$, the travel time under queueing speed (i.e., time in queue) as $t_Q = t_4 - t_2$; the free-flow travel time $FFTT = t_3 - t_1$, and delay $t_w = t_4 - t_3$. The total travel time TT on link l can be described as follows:

$$TT = FFTT + t_w = t_F + t_Q. \tag{8}$$

It is obvious that the time in queue t_Q is greater than the delay t_w, because vehicles traveling at free-flow speed would naturally reach the back of the queue (which has physical length) before they would have reached the bottleneck with free-flow travel time $FFTT$ (without any obstruction). Fig. 2 shows that the delay varies with the distance traveled in the physical queue d_Q, as $d_Q = t_Q \cdot v_Q = v_F \cdot (t_Q - t_W)$, and the time in queue t_Q can be derived as follows:

$$t_Q = \frac{t_w \cdot v_F}{v_F - v_Q}. \tag{9}$$

Eq. (9) indicates that the time in the (physical) queue t_Q is a fixed multiple of the delay t_w (Lawson et al., 1997). According to Eq. (8) and Eq. (9), the free-flow travel time can also be derived as follows:

$$t_F = FFTT - \frac{t_w \cdot v_Q}{v_F - v_Q}. \tag{10}$$

With the assumption of constant discharge capacity, both v_F and v_Q are constant based on a specific triangular fundamental diagram. The link emission of agent f entering and exiting link l at times t_1 and t_4, respectively, consists of two parts: the emission when traveling at the free-flow speed v_F ($EC(f, l, v_F)$) and the emission when traveling at the queueing speed v_Q ($EC(f, l, v_Q)$), as follows.

$$EC(f, l, t_1, t_4) = EC(f, l, v_F) \cdot t_F + EC(f, l, v_Q) \cdot t_Q. \tag{11}$$

For clarity, the agent, link and time indices are omitted from the notations here. That is, $EC_F = EC(f, l, v_F)$, and $EC_Q = EC(f, l, v_Q)$, then Eq. (11) can be re-written as Eq. (12).

$$EC = EC_F \cdot t_F + EC_Q \cdot t_Q. \tag{12}$$

where EC_F and EC_Q can be determined using a function that describes the relationship between emission and speed, such as the emission function, Eq. (7).

Proposition 1. With a constant discharge rate, the link emission of a vehicle can be expressed as a linear function of its delay t_w.

Proof. The vehicular link emission $EC = EC_F \bullet t_F + EC_Q \bullet t_Q$, and according to Eqs. (9)–(12), we can obtain the emission of the vehicle as follows:

$$EC = \left(FFTT - \frac{t_w \cdot v_Q}{v_F - v_Q}\right) \cdot EC_F + \left(\frac{t_w \cdot v_F}{v_F - v_Q}\right) \cdot EC_Q. \tag{13}$$

Assume that $\lambda = \frac{EC_Q}{EC_F} > 1$, then vehicular link emission can be re-written as a linear function of the delay t_w.

$$EC = EC_F \cdot \left(FFTT + t_w \cdot \frac{\lambda \cdot v_F - v_Q}{v_F - v_Q}\right) \tag{14}$$

This completes the proof.

It is also important to note that if $EC_F = 1$ and $\lambda = 1$ (i.e., $EC_Q = EC_F$), then $\frac{\lambda \cdot v_F - v_Q}{v_F - v_Q} = 1$, and the link travel time is a special case of the link emission.

$$EC = EC_F \cdot \left(FFTT + t_w \cdot \frac{v_F - v_Q}{v_F - v_Q}\right) = FFTT + t_w = TT. \tag{15}$$

3.2. Expanded space-time network

Additional notations used in presenting the expanded space-time network and the ESODTA model with constant bottleneck discharge rates are defined as follows.

Sets

V	Set of vertices in the expanded space-time network
A	Set of arcs in expanded space-time network

Parameters

$O(f)$	Origin node of agent f
$D(f)$	Destination node of agent f
$DT(f)$	Departure time of agent f
$AT(f)$	Given assumed arrival time of agent f at the destination, and it is a given large value. The waiting cost on the destination node is 0.
$s_{i,j}$	Free-flow travel time of link (i, j), which is an integer multiple of one time interval
$t_{i,j}$	Travel time (free-flow or congested travel time) of arc (i, j)
$EC_F(l)$	Unit emission cost under constant free-flow speed $v_F(l)$ on link l
$EC_Q(l)$	Unit emission cost under constant queueing speed $v_Q(l)$ on link l
$EC_{i,j,t,t'}$	Emission cost of an agent traveling on arc (i, j) with entrance time t and exit time t'
$Cap_{i,j}$	Outflow capacity of arc (i, j)

Variables

$x^f_{i,j,t,t'}$	Binary decision variable indicating whether agent f travels on arc (i, j) with entrance time t and exit time t' ($x^f_{i,j,t,t'} = 1$), or not ($x^f_{i,j,t,t'} = 0$).

In order to obtain the v_Q on the congested links, the physical network needs to be modified to capture the queue based on the assigned link capacity. Three cases are considered in this study: (i) at the merge point, the assigned outflow capacities of upstream links are proportional to the downstream link's inflow capacity based on the number of lanes or the given

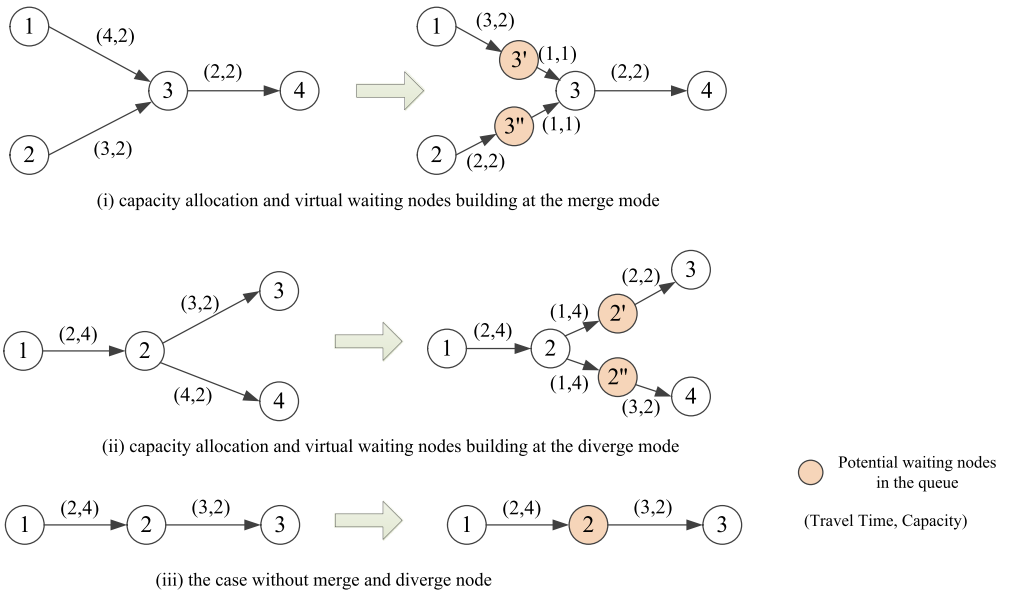


Fig. 3. Network modification for the capacity allocation and potential waiting nodes building.

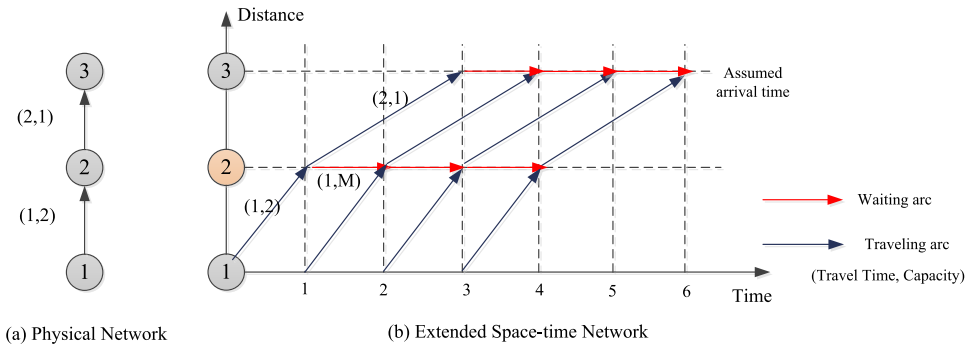


Fig. 4. A simple network and its corresponding space-time network.

upstream link's capacity values; (ii) at the diverge point, the assigned inflow capacities of downstream links is assumed to be the given outflow capacity of the upstream link; (iii) otherwise, the outflow capacity of the upstream link is equal to the inflow capacity of the downstream link. In case (i) and case (ii), virtual waiting nodes need to be created for the queue. One simple example is illustrated in Fig. 3 for the capacity allocation and virtual waiting nodes building.

Based on the modified physical traffic network, the expanded space-time network is constructed for the ESODTA with constant bottleneck discharge rates. Consider a space-time network, where V is the set of vertices and A is the set of arcs. Note that the physical network is represented by nodes and links, while vertices and arcs are defined to illustrate the expanded space-time network in this paper. A virtual waiting node i (including all destination nodes) is extended to a set of vertices (i, t) at each time interval t in the optimization horizon, $t = 1, 2, \ldots, T$. In addition, the origin nodes are not allowed for waiting. In the proposed space-time network representation, we consider two types of arcs as follows.

(1) Traveling arcs: A traveling arc represents the vehicle moves on a link (i, j) from time t to time t' in the physical network or from vertex (i, t) to vertex (j, t') in the space-time network, where $t' - t = s_{i,j}$. It is noted that the link free-flow travel time should be integer multipliers of one time interval.

(2) Waiting arcs: A waiting arc represents the vehicle waits at a virtual waiting node j (i.e., the vertical queue) for a time interval due to the capacity limitation of the corresponding travelling arc; that is, this arc is incident from vertex (j, t) and incident to vertex $(j, t+1)$. The arc travel time is 1 and arc capacity is set as a large constant, M. In addition, a large arrive time $AT(f)$ at the destination $D(f)$ for agent f will be assumed and the waiting cost is 0 on the destination nodes.

Fig. 4(a) depicts a simple physical network with three nodes, two links, one O-D pair (1, 3), and time-dependent travel demand input. By using dynamic network modeling processed described in Tong et al. (2015) and Li et al. (2015), the corresponding expanded space-time network is displayed in Fig. 4(b).

3.3. ESODTA mathematical formulation

Based on the expanded space-time network, the ESODTA problem with constant bottleneck discharge rates (ESODTA-constant discharge rates) can be formulated as follows.

$$\text{Minimize } Z = \sum_{f} \sum_{(i,j,t,t')} EC_{i,j,t,t'} \times x^{f}_{i,j,t,t'} \tag{16}$$

Subject to

$$\sum_{(i,j,t-t_{i,j},t)\in A} x^{f}_{i,j,t-t_{i,j},t} - \sum_{(j,k,t,t+t_{j,k})\in A} x^{f}_{j,k,t,t+s_{j,k}} = \begin{cases} -1 & \forall f \in F,\ j = O(f) \text{ and } t = DT(f) \\ 1 & \forall f \in F,\ j = D(f) \text{ and } t = T \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

$$\sum_{f\in F} x^{f}_{i,j,t,t+t_{i,j}} \le Cap_{i,j},\ \ \forall\left(i,j,t,t+t_{i,j}\right) \in A, t = 1,2,\ldots,T \tag{18}$$

$$x^{f}_{i,j,t,t'} = \{0,1\} \tag{19}$$

In the above formulation, the objective is to minimize the total vehicular emission in the network shown as Eq. (16). The flow balance constraints are presented as Eq. (17). Eq. (18) describes the outflow capacity constraints on traveling arcs. The constraints shown as Eq. (19) requires all decision variables to be binary.

The emission costs $EC_{i,j,t,t'}$ in the objective function are discussed as follows. According to Proposition 1, which is derived based on the SQM in Section 3.1, the link emission cost of one vehicle is a function of the delay $t_w(i,j)$ on link (i,j):

$$EC(i,j) = EC_F(i,j) \times \left(s_{i,j} + t_w(i,j) \times \frac{\lambda \times v_F - v_Q}{v_F - v_Q} \right) = \alpha(i,j) + \beta(i,j) \times t_w(i,j) \tag{20}$$

where $\alpha(i,j) = EC_F(i,j) \times s_{i,j}$ can be interpreted as the emission cost parameter on traveling arc (i,j,t,t') for each vehicle, and $\beta(i,j) = EC_F(i,j) \times \frac{\lambda \times v_{F_{i,j}} - v_{Q_{i,j}}}{v_{F_{i,j}} - v_{Q_{i,j}}}$ as the emission cost parameter on waiting arc for one time interval $(j,j,t,t+1)$ for each vehicle. This modeling technique and the property presented in Proposition 1 seamlessly integrate the SQM and the ESODTA formulation.

3.4. Lagrangian relaxation-based solution algorithm

This subsection presents the Lagrangian relaxation-based algorithm for solving the ESODTA model with constant bottleneck discharge rates. In this algorithm, the capacity constraints, shown in Eq. (18), are dualized to the objective function with the multipliers $\mu_{i,j,t,t'}$, $\forall(i,j,t,t')$, as follows.

$$\min Z^{LR} = \sum_{f} \sum_{(i,j,t,t')} (EC_{i,j,t,t'} + \mu_{i,j,t,t'}) \times x^{f}_{i,j,t,t'} - \sum_{(i,j,t,t')} \mu_{i,j,t,t'} \times Cap_{i,j} \tag{21}$$

The relaxation problem, which consists of the objective function Eq. (21) and constrains Eqs. (17) and (19), is a standard time-dependent shortest path problem for each agent, which can be easily solved by existing efficient algorithms (e.g., Ziliaskopoulos and Mahmassani, 1993). Thus, the objective function (21) of the relaxation problem is linear with variables $x^{f}_{i,j,t,t'}$. Specifically, according to Eq. (20), the link emission cost for an agent f includes two parts, and both $\alpha(i,j)$ and $\beta(i,j)$ are constant parameters. Moreover, the multipliers $\mu^{n}_{i,j,t,t'}$, $\forall(i,j,t,t')$ are constants which will be updated in each iteration of the algorithm.

The Lagrangian relaxation-based algorithm is described as follows:

Step 1: Initialization. Let iteration $n=1$, initialize the multipliers $\mu^{n}_{i,j,t,t'} = 0$.
Step 2: Solve the relaxation problem for each agent as the time-dependent least-cost path problem.
Step 3: Update the Lagrangian multipliers as follows:
The step-size updating is based on the method of successive averages (MSA): $\gamma^{n} = \frac{1}{n+1}$

$$\mu^{n+1}_{i,j}(t) = \max\left\{ 0, \mu^{n}_{i,j}(t) + \gamma^{n} \times \left(\sum_{f} x^{f}_{i,j,t,t+s_{i,j}} - Cap_{i,j} \right) \right\} \tag{22}$$

Step 4: Check termination condition: if $n < N_{max}$, then $n = n+1$ and return to step 2. Otherwise, stop the algorithm.

4. Solution algorithm for the simulation-based ESODTA

In Section 4.1, we present a solution algorithm for the simulation-based ESODTA. The proposed column generation-based ESODTA algorithm embeds: (i) the DNLE($\mathbf{d}$) to evaluate link and path emissions for the set of agents with assigned paths, and (ii) the gradient projection-based method to update the path assignment of the agents. Section 4.2 describes how to evaluate link marginal emissions based on Proposition 1 and simulated traffic conditions. The approach for evaluating path marginal emissions is presented in Section 4.3.

The DNLE($\mathbf{d}$) underlying the simulation-based ESODTA combines both macroscopic and microscopic traffic descriptions based on Newell's simplified kinematic wave model and simplified car-following model. Specifically, Newell's simplified kinematic wave model is employed in the dynamic mesoscopic traffic simulation package, DTALite, which outputs link arrival and departure times for each vehicle (Zhou and Taylor, 2014; Zhou et al., 2015a). Given the link arrival and departure times of individual vehicles, Newell's simplified linear car following model (Newell, 2002) is adopted to reconstruct the detailed vehicle trajectories which can be used to derive second-by-second vehicle speeds and accelerations. Then, the (time-dependent) speeds and accelerations are input to an emission model, such as the MOVES model (US EPA, 2009), to generate vehicular emissions.

Recall that the analytical ESODTA model, presented in Section 3, is formulated based on the SQM and Proposition 1. As mentioned in Section 2, the SQM is consistent with Newell's simplified kinematic wave model (Lawson et al., 1997; Daganzo, 2006), but focuses on the case of constant bottleneck discharge rates. To deal with the more general case of time-dependent bottleneck discharge rates, the proposed simulation-based ESODTA model adopt the aforementioned DNLE($\mathbf{d}$) to generate time-dependent speeds for emission estimation. Essentially, the simulation-based model provides a richer modeling capability for emission applications through the simulation-based DNLE($\mathbf{d}$), which generalizes the SQM.

Additional notations used in this section are defined as follows.

Sets or vectors

$P_m(w, \tau)$	Set of paths in iteration m for the agents of OD pair w and departure time τ
P_m	Set of paths in iteration m for all of the agents in the network; $P_m = \{P_m(w, \tau), \forall w, \tau\}$

Indices

n	index of inner loop iterations in the column generation-based algorithm
m	index of outer loop iterations in the column generation-based algorithm

Parameters

M_{max}	Maximum number of outer loop iterations
N_{max}	Maximum number of inner loop iterations

Variables

$r(w, \tau, p)^n$	Number of agents on path p of OD pair w and departure time τ in iteration n
$\mathbf{r}^n$	Path flow vectors (a feasible solution) in iteration n; $\mathbf{r}^n = \{r(w, \tau, p)^n, \forall w, \tau, p\}$
$EC(w, \tau, p)^n$	Total emission of the agents on path p of OD pair w and departure time τ in iteration n
$TE_w^{\tau}(\mathbf{r}^n)$	Total emission of the agents of OD pair w and departure time τ, evaluated at the feasible solution $\mathbf{r}^n$
$TE(\mathbf{r}^n)$	Total system emission, evaluated at the feasible solution $\mathbf{r}^n$

4.1. Solution algorithm

4.1.1. Column generation-based algorithmic framework

The column generation-based algorithm generates time-dependent least marginal emission paths as needed in the outer loop and solves a reduced (or restricted) ESODTA problem in the inner loop (e.g., Lu et al., 2009). The column generation-based approach operates as follows (see Fig. 5). In each outer loop iteration m, the time-dependent least-cost path algorithm, developed by Ziliaskopoulos and Mahmassani (1993), is applied to find the time-dependent least marginal emission path for each O-D pair and each departure time interval. New paths, if any, are added to augment the current subset of feasible paths, P_m, in iteration m. A gradient projection-based descent direction method, presented in Section 4.1.2, is then used to solve the reduced ESODTA problem defined on P_m. The algorithm terminates and outputs time-varying path flows obtained in the current iteration, if no new path is found or a preset convergence criterion is satisfied.

The gradient projection-based descent direction method proceeds iteratively and forms the inner loop in the column generation-based algorithmic framework. In each inner loop iteration n, the updated path flows $\mathbf{r}^n$ and the corresponding total system emission $TE(\mathbf{r}^n)$ and link marginal emissions are evaluated by the DNLE($\mathbf{d}$) model. If the difference between the objective values in two successive iterations (i.e., $TE(\mathbf{r}^n) - TE(\mathbf{r}^{n-1})$) is less than a preset threshold or a preset convergence criterion (e.g., $n = N_{max}$) is satisfied, the inner loop terminates and the algorithm returns to the outer loop.

4.1.2. Gradient projection-based descent direction method

The reduced ESODTA problem, defined by a subset of feasible paths P_m in outer loop iteration m, is solved by the gradient projection-based descent direction method in the inner loop to obtain least-emission path flows on the existing paths. With

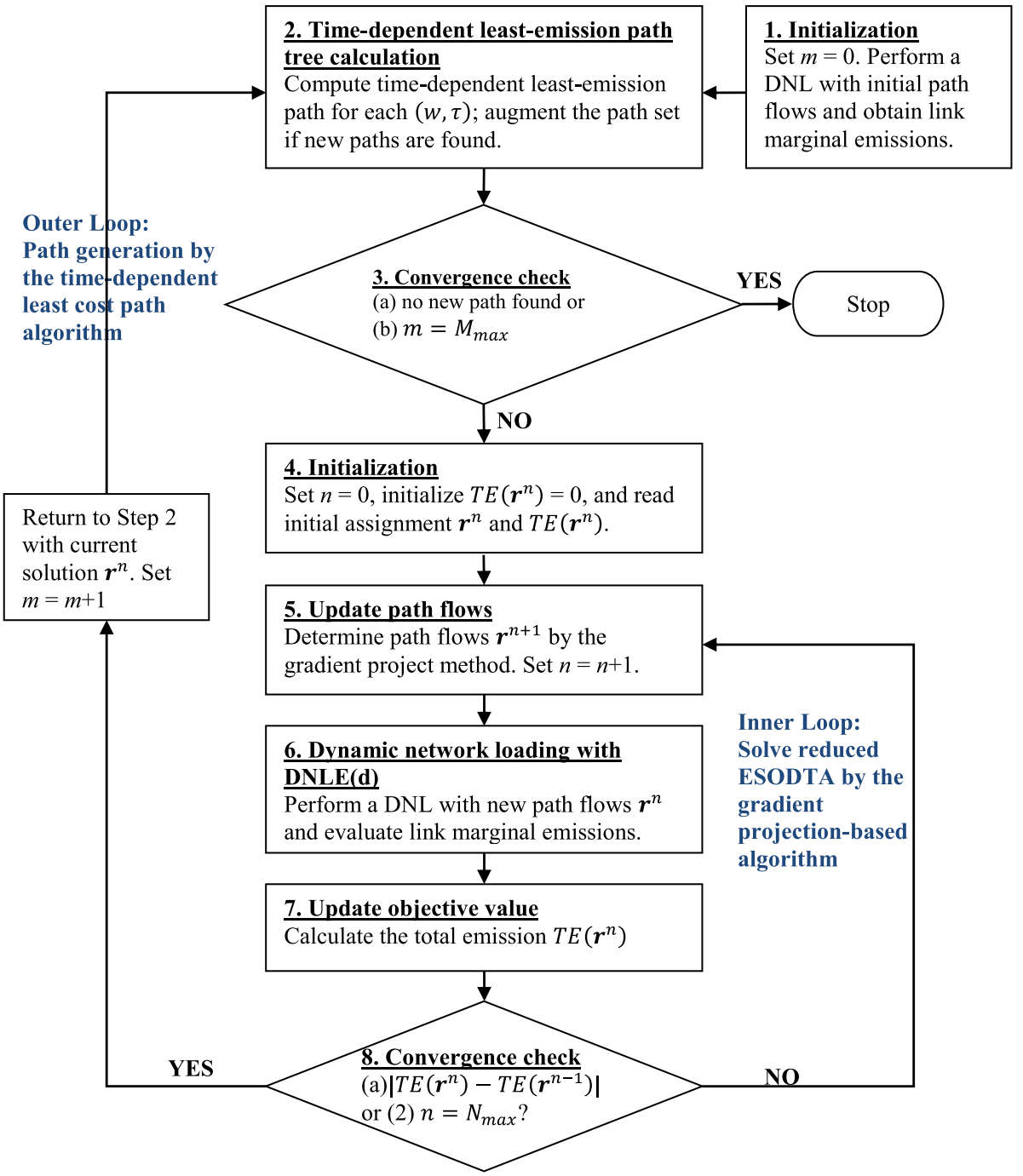


Fig. 5. Flow chart of the column generation-based ESODTA algorithm.

a feasible solution $\boldsymbol{r}^n$ in inner loop iteration n, the method adopts a search direction along the feasible descent direction based on the gradient, $\nabla TE(\boldsymbol{r}^n)$:

$$\boldsymbol{r}^{n+1} = Proj_{\Omega}[\boldsymbol{r}^n - s^n \times \nabla TE(\boldsymbol{r}^n)], \tag{23}$$

where $s^n \in (0, 1)$ is the step size in inner loop iteration n. $Proj_{\Omega}[u]$ denotes the unique projection of path flow vector u onto the feasible space Ω and is defined as the unique solution of the problem: $\text{Min}_{y \in \Omega}||u - v||$. Accordingly, the new iterate $\boldsymbol{r}^{n+1}$ is obtained by updating the current iterate $\boldsymbol{r}^n$ along the direction $-\nabla TE(\boldsymbol{r}^n)$ with a move size s^n.

To facilitate solving the ESODTA problem on large networks with multiple O-D pairs, the proposed method decomposes the original problem into many sub-problems, each of which corresponds to a (w, τ) pair, by assuming that cross-network marginal effects are negligible (e.g., Zhang et al., 2013b). Let $p*$ be the referenced least marginal emission path of a (w, τ) pair. Given a feasible solution $\boldsymbol{r}^n$, the flow balance conservation constraints Eq. (4) can be rearranged as follows:

$$r(w,\tau,p^*)^n = d(w,\tau) - \sum_{p\in P(w,\tau)\setminus p*} r(w,\tau,p)^n,\ \ \forall w,\tau. \tag{24}$$

Then, with this rearrangement, the objective function corresponding to a pair (w, τ) can be written as follows:

$$\begin{aligned} TE_w^\tau(\boldsymbol{r}^n) &= \sum_{p\in P(w,\tau)\setminus p*} r(w,\tau,p)^n \times EC(w,\tau,p)^n + r(w,\tau,p^*)^n \times EC(w,\tau,p^*)^n \\ &= \sum_{p\in P(w,\tau)\setminus p*} r(w,\tau,p)^n\left[EC(w,\tau,p)^n - EC(w,\tau,p^*)^n\right] + d(w,\tau)\times EC(w,\tau,p^*)^n. \end{aligned} \tag{25}$$

The first-order partial derivative of $TE_w^\tau(\boldsymbol{r}^n)$ with respect to a particular path flow $r(w, \tau, p)$ is

$$\begin{aligned} \nabla TE_{wp}^\tau(\boldsymbol{r}^n) &= \partial TE_w^\tau(\boldsymbol{r}^n)/\partial r(w,\tau,p) \\ &= EC(w,\tau,p)^n - EC(w,\tau,p^*)^n + \sum_{p'\in P(w,\tau)\setminus p*}\left[r\left(w,\tau,p'\right)^n\left(\eta_{wp'}^{\tau,n} - \eta_{wp*}^{\tau,n}\right)\right] + d(w,\tau)\times\eta_{wp*}^{\tau,n}. \end{aligned} \tag{26}$$

where $\eta_{wp}^{\tau,n} = \partial EC(w,\tau,p)^n/\partial r(w,\tau,p)$ denotes the path marginal emission which represents the change in path emission due to an additional unit of the path inflow, $r(w, \tau, p)$. Note that if within-path-set marginal effects are also ignored, then

$$\nabla TE_{wp}^\tau(\boldsymbol{r}^n) = EC(w,\tau,p)^n - EC(w,\tau,p^*)^n + r(w,\tau,p)^n\left[\eta_{wp}^{\tau,n} - \eta_{wp*}^{\tau,n}\right] + d(w,\tau)\times\eta_{wp*}^{\tau,n}. \tag{27}$$

Regarding the step size s^n, this study adopts a scheme of mixed step sizes as follows (Lu et al., 2009).

$$s^n = 1/m,\ \ \text{if}\, n = 0;\ \ s^n = 1,\ \ \text{otherwise}. \tag{28}$$

Recall that m is the (outer loop) iteration counter. The decreasing step size $s^n = 1/m$ follows the MSA. Based on Eqs. (23), (26), and (28), the gradient projection-based descent direction method derives the following path flow updating scheme in the inner loop of the algorithmic framework.

$$\begin{aligned} r(w,\tau,p)^{n+1} &= \text{Max}\left\{0,\ r(w,\tau,p)^n - s^n\times\nabla TE_{wp}^\tau(\boldsymbol{r}^n)\right\},\ \ \forall p\in P(w,\tau)\setminus p*, \\ r(w,\tau,p^*)^{n+1} &= d(w,\tau) - \sum_{p\in P(w,\tau)\setminus p*} r(w,\tau,p)^{n+1}. \end{aligned} \tag{29}$$

4.2. Link marginal emission evaluation

The gradient projection-based method requires evaluation of path marginal emissions, $\eta_{wp}^\tau,\ \forall w,\tau,p$, and their constituent link marginal emissions $\eta_l^t, \forall l, t$. This subsection first presents the evaluation of link marginal emissions under constant bottleneck discharge rates, and then we discuss how to approximate link marginal emissions under time-dependent bottleneck discharge rates using simulated link capacities. Additional notations used in this section are defined as follows.

w	backward wave speed
k_{jam}	jam density
$length(l)$	length of link l
$nlanes(l)$	number of lanes on link l
$A(l, t)$	cumulative number of vehicles that have arrived at link l at time t
$V(l, t)$	cumulative number of vehicles that have waited at the vertical queue of link l at time t
$D(l, t)$	cumulative number of vehicles that have departed from link l at time t
$q^{max}(l, t)$	maximum flow rate on link l at time t
$cap^{in}(l, t)$	inflow capacity of link l at time t
$cap^{out}(l, t)$	outflow capacity of link l at time t
$FFTT(l)$	free-flow travel time on link l; i.e., $length(l)/v_f$
$BWTT(l)$	backward wave travel time on link l; i.e., $length(l)/w$

Fig. 6 depicts the cumulative arrival $A(l, t)$, virtual arrival $V(l, t)$, and departure curves $D(l, t)$ for a congested link l, with a constant outflow capacity, $c(l)$. The queue starts at t_l^{qs} and dissipates at t_l^B on the link. Let t_l', t_l'', and t_l''' be the times when an additional vehicle (n_1) arrives at link l, joins the queue of the bottleneck, and leaves the link, respectively. According to Ghali and Smith (1995), the link marginal travel time due to the additional vehicle n_1 is equal to the gray area (i.e., $mTT(l, t_l') = t_l^B - t_l'$), which includes the travel time of the additional vehicle $(t_l''' - t_l')$ and its impact on the travel times of the vehicles behind the unit of flow $(t_l^B - t_l''')$. Note that if this additional vehicle does not encounter a queue, the link marginal travel time equals $FFTT(l) = t_l'' - t_l'$. The vehicles arriving between t_l' and t_l^A experience the additional travel time

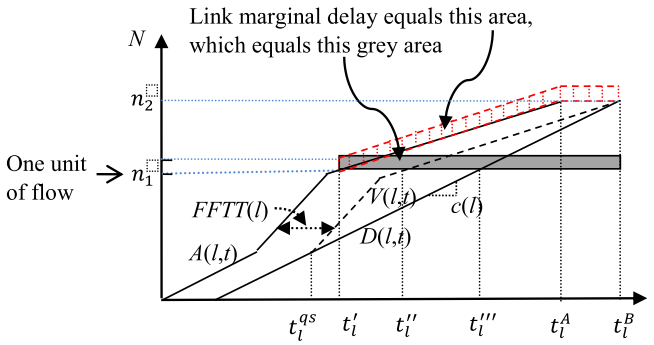


Fig. 6. Illustration of link marginal travel time on a congested link.

$1/c(l)$, because it takes $1/c(l)$ to discharge this perturbation vehicle. In our notation, this means that the change in the travel time $\Delta tt = 1/c(l)$. The impact on the travel times of the impacted vehicles arriving between t_l' and t_l^A can be derived from Fig. 6 as follows.

$$mTT\left(l, t_l'\right) = t_l^B - t_l''' = (n_2 - n_1)\frac{1}{c(l)}. \tag{30}$$

where n_1 denotes the additional (perturbation) vehicle arriving at link l at time t_l', while n_2 is that last vehicle impacted by the additional vehicle n_1.

Similarly, the link marginal emission can be divided into two parts: the emission of the additional vehicle and its impact on the emissions of the following vehicles (or the change in emission for the impacted vehicles). Let $\Delta EC(l, t)$ be the change in emission for one vehicle entering the link behind the unit perturbation vehicle. The following proposition derives the change in emission for the impacted vehicles, $mEC(l, t)$, and its relationship with the change in travel time for the impacted vehicles, $mTT(l, t)$.

Proposition 2. *$mEC(l, t) = (n2 - n1)\Delta EC(l, t)$, which is a multiple of the change in travel time for the impacted vehicles: $mEC(l, t) = \gamma(l) \times mTT(l, t)$.*

Proof. Since the change in travel time is the sum of the change in the time spent in queue and the change in the actual free-flow travel time (i.e., $\Delta t_w(l, t) = \Delta t_Q(l, t) + \Delta t_F(l, t)$) and $\Delta t_w(l, t) = 1/c(l)$,

$$\Delta t_F(l, t) = \frac{1}{c(l)} - \Delta t_Q(l, t), \tag{31}$$

According to Eq. (9), the change in the time spent in queue, $\Delta t_Q(l, t)$, is a fixed multiple of the change in the travel time, $\Delta t_w(l, t)$:

$$\Delta t_Q(l, t) = \frac{\Delta t_w(l, t)}{1 - \frac{v_{Q(l)}}{v_F(l)}} = \frac{1/c(l)}{1 - \frac{v_{Q(l)}}{v_F(l)}}. \tag{32}$$

Then, **according** to Eq. (11), the change in emission for one vehicle entering the link behind the unit perturbation vehicle is as follows.

$$\begin{aligned}\Delta EC(l, t) &= \Delta t_Q(l, t) EC(v_Q(l)) + \Delta t_F(l, t) EC(v_F(l)) \\ &= \frac{1}{c(l)}\left[\frac{1}{1 - \frac{v_{Q(l)}}{v_F(l)}} EC(v_Q(l)) + \left(1 - \frac{1}{1 - \frac{v_{Q(l)}}{v_F(l)}}\right) EC(v_F(l))\right].\end{aligned} \tag{33}$$

Based on Eq. (30), the change in emission for the impacted vehicles can be derived as follows.

$$mEC(l, t) = (n2 - n1)\Delta EC(l, t). \tag{34}$$

Let $\gamma(l) = \left[\frac{1}{1 - \frac{v_{Q(l)}}{v_F(l)}} EC(v_Q(l)) + (1 - \frac{1}{1 - \frac{v_{Q(l)}}{v_F(l)}}) EC(v_F(l))\right]$. Then, $mEC(l, t) = (n2 - n1)\frac{\gamma(l)}{c(l)} = \gamma(l) \times mTT(l, t)$.
This completes the proof.

The link marginal emission η_l^t is the sum of the emission of the additional vehicle and $mEC(l, t)$. Note that in Eq. (34) $\gamma(l)$ is a constant multiplier, because $v_F(l)$ and $v_Q(l)$ are given constants, and $EC(v_F(l))$ and $EC(v_Q(l))$ can be determined according to the emission function, such as Eq. (7). The queueing speed $v_Q(l)$ of link l is determined mainly based on the bottleneck discharge capacity. For instance, Lu et al. (2013) discussed a method for the outflow capacity at merge or diverge junctions.

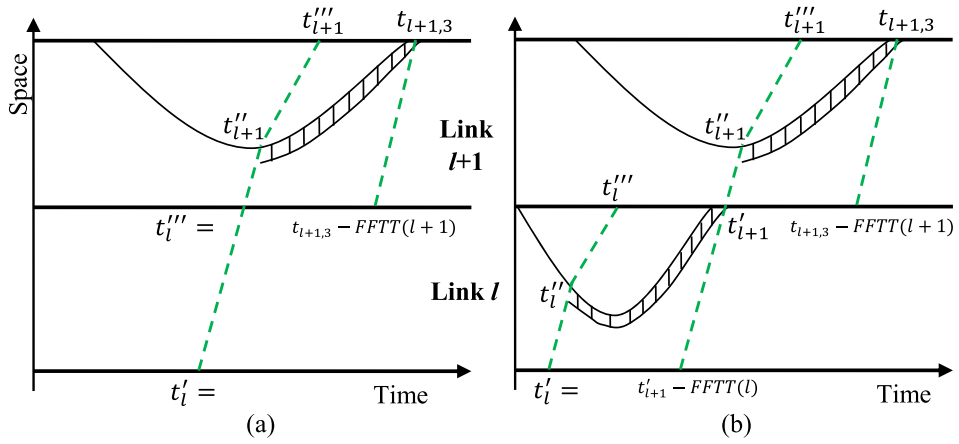


Fig. 7. Path marginal emission analysis without queue spillback.

The above derivation of the link marginal emission is based on the assumption of a constant bottleneck discharge capacity, $c(l)$. However, bottleneck capacities are time-dependent in the presence of signal controls at intersections or incidents on links, which result in time-dependent queueing speeds. Moreover, if a downstream link is so congested that the queue spills back to the current link, then there is no space for entering traffic. When link or bottleneck capacities and queuing speeds are time-dependent, the vehicular emission estimation will be much more complex, compared to the case with constant discharge capacities.

A few previous studies had proposed approaches to deal with the case of time-dependent bottleneck capacity along the line of Newell's simplified kinematic wave model (Newell, 1993a; Newell, 1993b). Lawson et al. (1997) had extended the SQM to a special case where the bottleneck capacity changes once at a known time due to an under-saturated traffic signal. Erera et al. (1998) further extended the approach to allow multiple changes in bottleneck capacity and piecewise linear concave flow-density relationship, and Daganzo (2001) relaxed the requirement of a concave flow-density relationship. Recently, Cetin (2012) developed models based on the shockwave theory (i.e., the Lighthill-Whitham-Richards theory) for estimating time-dependent queue lengths at signalized intersections based on the data from probe vehicles. However, practical network applications of these approaches remain very challenging, as it is very difficult to construct input-output diagrams for the more general case of time-dependent bottleneck capacity.

In this paper, we adopt the simulation-based DNLE(**d**), which combines Newell's simplified kinematic wave model and Newell's linear car-following model (Zhou and Taylor, 2014; Zhou et al., 2015a) to generate second-by-second vehicle speeds in response to time-dependent bottleneck capacities. In the proposed simulation-based ESODTA model (see Fig. 5), after the simulation-based network loading at Step 6, we check whether or not the discharge capacity of a link changes (i.e., if a signal is present on a link or queue spills back to the link under consideration). If the discharge capacity of a link does not change for passing vehicles, we can directly apply Proposition 2 to evaluate the link marginal emission. Otherwise, if the discharge capacity changes due to the signal or the queue spills back from the downstream, in order to apply Proposition 2 to obtain approximation of link marginal emissions, we need to fetch the corresponding bottleneck capacity value from the simulation results for each vehicle being analyzed, then use a single value of the discharge capacity $c(l)$ as an approximation across a certain time period from the entrance timestamp of the perturbation vehicle to the end timestamp of the queue. Note that this approximation scheme could lead to potential approximation errors when the values of $c(l)$ are highly dynamic, but it is important to recognize that the marginal emission analysis based on simulation results are essentially numerical approximation for location conditions in nature, not to mention that path marginal emission calculation could bring another degree of approximation errors as illustrated in Section 4.3 as well as identified by previous studies, such as Shen et al. (2007a).

4.3. Path marginal emission evaluation

Evaluating path marginal emissions in dynamic and congested traffic networks requires explicitly tracing the perturbation propagation of an additional unit of inflow along a path. This issue has also been recognized by Shen et al. (2007a), Qian and Zhang (2011) and Lu et al. (2013) on evaluating path marginal delays (or travel times).

Consider a freeway or an arterial segment with two sequential links without merges and diverges, link l and link $l+1$. Under congested conditions, there are three basic cases of interest, when the additional unit of vehicle arrives at this segment at time $t_l^{'}$ ($=\tau$, the departure time).

(i) There is a bottleneck on the downstream link $l+1$ and the queue on link $l+1$ does not spill back to link l; link l is in free-flow condition while link $l+1$ is partially congested (Fig. 7(a)).

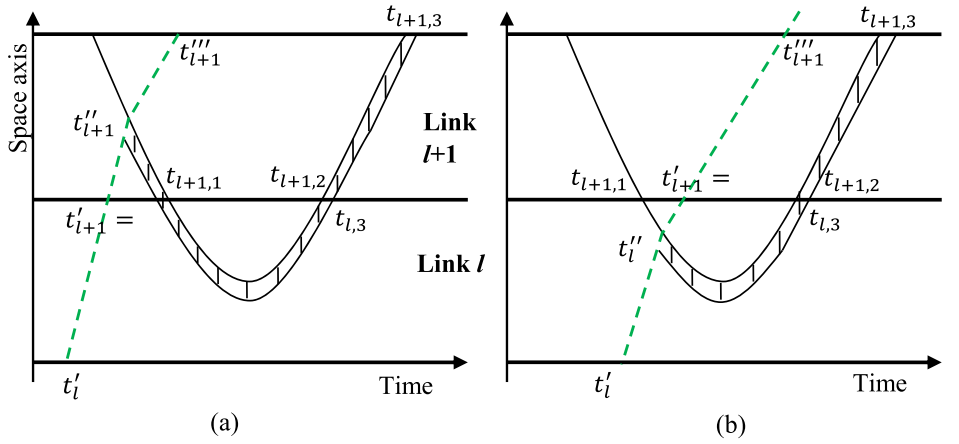


Fig. 8. Path marginal emission analysis with queue spillback from downstream link.

(ii) There is a bottleneck on each of the two links, and the two bottlenecks are independent (e.g., link $l+1$ is 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 are partially congested (Fig. 7(b)).

(iii) There is a bottleneck on the downstream link $l+1$ and the queue on link $l+1$ spills back to link l; that is, link l is partially congested while link l is fully congested (Fig. 8).

In case (i), the first link is not impacted by that additional vehicle, while the second link's impacted regime is from t'_{l+1} to $t_{l+1,3}-FFTT(l+1)$. In case (ii), the first link has an impacted regime spanning a time period from t'_l to $t'_{l+1}-FFTT(l)$, and the second link's impacted regime is from t'_{l+1} to $t_{l+1,3}-FFTT(l+1)$. Note that in case (ii), $t'_{l+1}=t_{l,3}$; that is, the perturbation propagates to link $l+1$ when the queue on link l vanishes at time $t_{l,3}$.

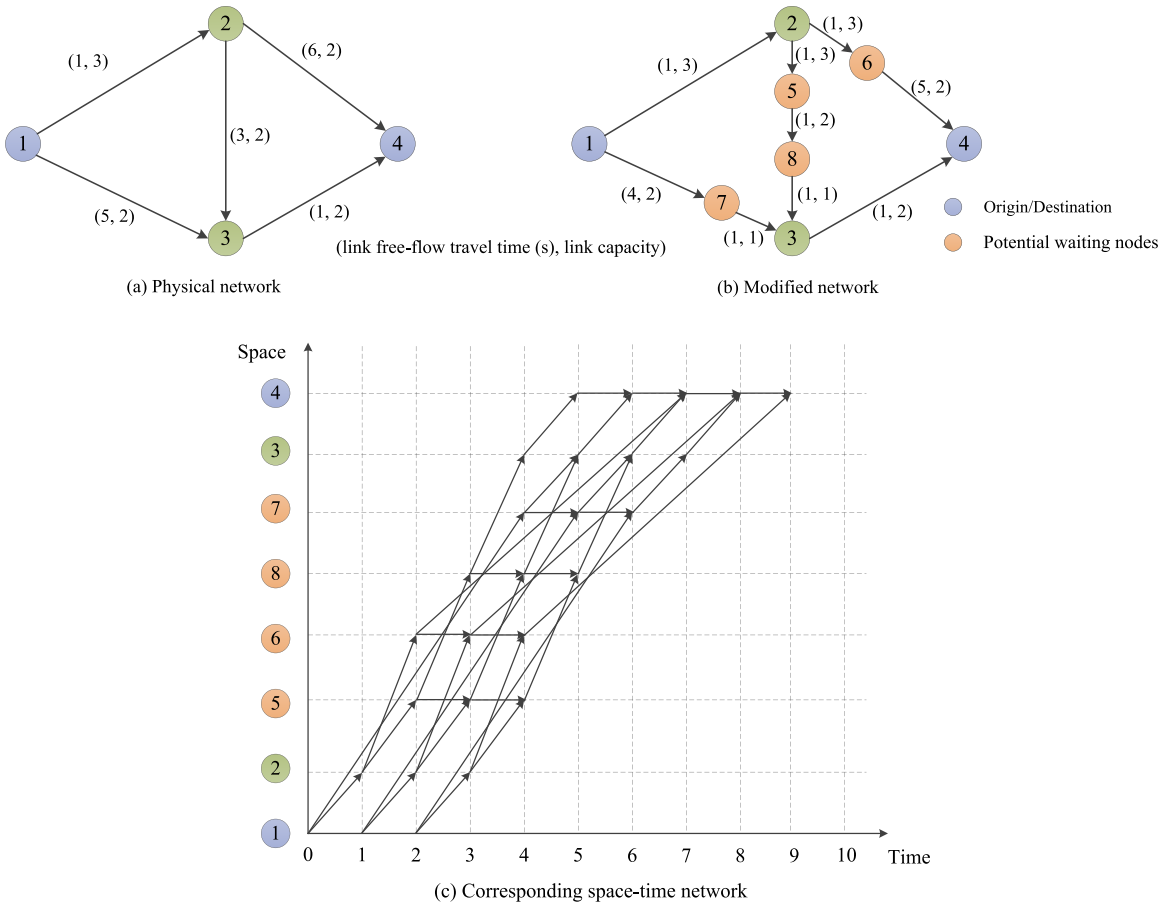
In case (iii), if the additional vehicle does not encounter the queue on link l (Fig. 8(a)), then the link marginal emission on this link is zero. The perturbation then moves to link $l+1$ with an impacted period from t'_{l+1} to $t_{l+1,1}$. After detecting the next nearest event timestamp, $t_{ne}=t_{l+1,1}$, which corresponds to a queue spillback event, we need to trace back to link l to take time period $[t_{l+1,1}-FFTT(l),\ t_{l+1,2}-FFTT(l)]$ into consideration. Finally, we move to link $l+1$ to cover the last impacted regime from $t_{l+1,2}$ to $t_{l+1,3}-FFTT(l+1)$. Thus, the proposed algorithm for evaluating path marginal emissions needs to incorporate a backtracking mechanism, in order to explicitly consider the difference pieces of the impacted regime over multiple links.

If the additional vehicle encounters the queue on link l (Fig. 8(b)), the first link's impacted regime spans a time period from t'_l to $t_{l,3}-FFTT(l)$, and the second link's impacted regime is from $t_{l,3}$ to $t_{l+1,3}-FFTT(l+1)$. Again, the perturbation can enter link $l+1$ at time $t_{l+1,2}$ only when the queue on link l vanishes at time $t_{l,3}$.

Based on the above analysis for the impacted regimes on two consecutive links due to an additional vehicle, this research proposes a method for evaluating the path marginal emission of a path $p \in P(w,\tau)$ with multiple links $l=1,\ldots,L$. The path marginal emission due to an additional vehicle consists of the path emission of that vehicle and additional emissions generated by impacted vehicles. Starting from the first link $l=1$ and given departure time τ, the proposed algorithm, presented in Algorithm 1, keeps accumulating path marginal emission by adding the link marginal emission (see Section 4.2) for each impacted period and traces the perturbation propagation based on the next nearest event timestamp which is used to guide the evaluation procedure advancing to next link (or impacted period) or returning to last link in the queue spillback case.

Algorithm 1
Evaluation of the path marginal emission for a triplet (p, w, τ).

Initialize $t'=\tau$, $l=1$, and $\eta^{\tau}_{wp}=ec^{\tau}_{wp}$, where ec^{τ}_{wp} denotes the emission of path $p \in P(w,\tau)$.
Do while link index $l \le L$ (L is the number of links)
Step 1: Obtain the next (nearest) event timestamp, t_{ne}. The next event may correspond to the end time of congestion (i.e., t_3), the beginning of queue spillback (i.e., t_1) or $t_{ne}=t'+FFTT(l)$ (under uncongested conditions).
Step 2: Evaluate link marginal emission, $\eta_l^{t'}$, using the methods described in Section 4.2.
Step 3: Accumulate the path marginal emission $\eta^{\tau}_{wp}=\eta^{\tau}_{wp}+\eta_l^{t'}$.
Step 4: Move to the next link, and update the starting time, t', of next analysis period. If next event timestamp t_{ne} corresponds to the end time of congestion or is in a uncongested time period, then $t'=t_{ne}+FFTT(l)$ and $l=l+1$; otherwise (t_{ne} corresponds to a queue spillback case) $l=l-1$ and $t'=t_{ne}-FFTT(l)$.
End



(a) Physical network

(b) Modified network

(c) Corresponding space-time network

Fig. 9. Physical traffic network, modified traffic network, and its corresponding space-time network.

5. Numerical experiments

This section presents two numerical examples. The first one demonstrates the analytical ESODTA model presented in Section 3, and the second one illustrates the solution algorithm for the simulation-based ESODTA presented in Section 4. All of the numerical tests are conducted on a Dell Precision T7610 Workstation with 20 CPU cores (40 computational threads) and 192 G RAM.

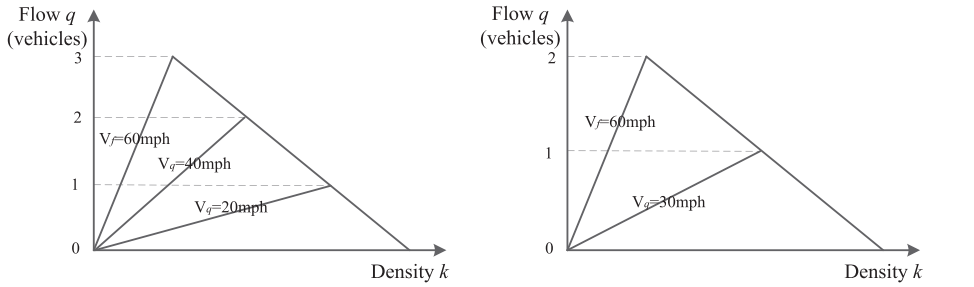
5.1. Numerical example for the analytical ESODTA in a small network

For illustrative purposes, one simple network is created with 4 nodes, 5 links, and one OD pair (1, 4), as shown in Fig. 9(a). The link free-flow travel time and link capacity of each link are displayed in the parenthesis near that link. Based on the description in Section 3.2, the traffic network is modified and shown in Fig. 9(b) by considering the merge and diverge junctions to better represent travel delay for emission calculation. The corresponding expanded space-time network is depicted in Fig. 9(c). All of the links are assumed to have the same free-flow speed value of 60 mph. The fundamental q-k diagrams under different link capacities are shown in Fig. 10. Based on the v_f and v_q on each link and the emission cost function in Eq. (7), we can calculate the emission cost parameter at free-flow travelling arcs and virtual waiting arcs using Eq. (20).

The ESODTA model formulated based on the expanded space-time network (Section 3.3) is solved using the commercial solver, GAMS. The Lagrangian relaxation approach, presented in Section 3.4, is also implemented to obtain a lower bound. The related source code can be downloaded at the ResearchGate website: https://www.researchgate.net/publication/295856604_Eco-system_optimal_time-dependent_flow_assignment_in_a_congested_network#share

5.1.1. Sensitivity analysis on travel demand

The first set of sensitivity analysis is conducted to examine the impact of different travel demand levels on the solution of ESODTA. Three levels of travel demand are tested and the specific values are defined in Table 2. The gap between the



(a) the fundamental q-k diagram for link capacity = 3 (b) the fundamental q-k diagram for link capacity = 2

Fig. 10. Fundamental q-k diagrams under different link capacities.

Table 2
Time-dependent travel demand under three different levels.

Demand Level (OD pair 1→4)	Departure Time $t=0$	Departure Time $t=1$	Departure Time $t=2$
Level 1	5	4	1
Level 2	5	4	2
Level 3	5	4	3

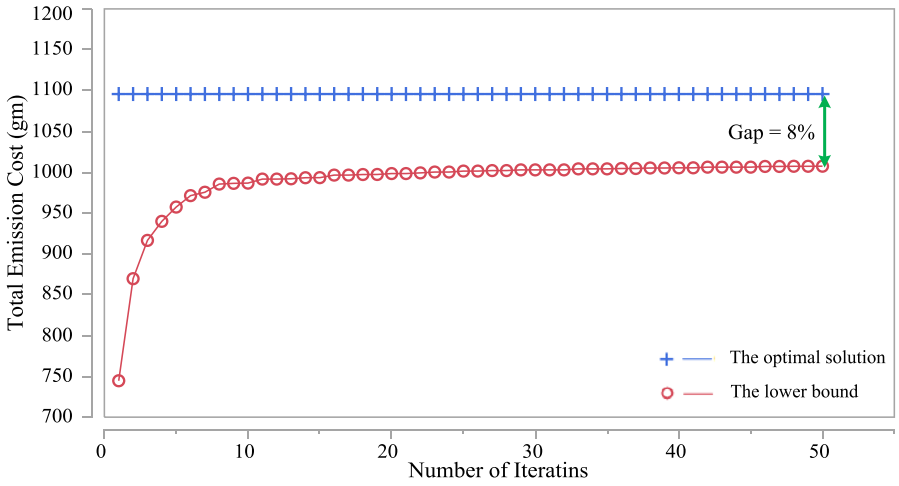


Fig. 11. The gap between the optimal solution and the lower bound under demand level 1.

Table 3
The comparison of ESODTA and TSODTA under three different levels of travel demand.

Demand Level		Level 1	Level 2	Level 3
ESODTA	Optimal Emission Cost (gm)	1095.72	1225.44	1355.16
	Corresponding Travel Time (min)	67	72	77
TSODTA	Corresponding Emission Cost (gm)	1104.84	1230.00	1355.16
	Optimal Travel Time (min)	63	70	77
Comparison between ESODTA and TSODTA	Emission Reduction by ESODTA (gm)	9.12	4.56	0
	Travel Time Reduction by TSODTA (min)	4	2	0

optimal solution and the lower bound under demand level 1 is shown in Fig. 11. The gap between the lower bound and the optimal solution after 50 iterations is about 8%. We also calculate the gap values under demand level 2 and level 3, which are 8.6% and 8.5%, respectively. The results show that the quality of the lower bound obtained using the Lagrangain relaxation approach is acceptable.

In addition, we solve the travel time-based SODTA (TSODTA) problem, in which the total emission cost in the objective function Eq. (16) is replaced with the total travel time. The comparison of ESODTA and TSODTA is presented in Table 3. As shown in Table 3, (i) the optimal total emission cost of ESODTA and the optimal total travel time of TSODTA increase with the increase of travel demand; (ii) The emission reduction in ESODTA is gained at the expense of an increased total travel

Table 4
The comparison of ESODTA and TSODTA under three different cases.

Cases		Case 1	Case 2	Case 3
ESODTA	Optimal Emission Cost (gm)	975.394	968.28	966.00
	Corresponding Travel Time (min)	59	61	62
TSODTA	Corresponding Emission Cost (gm)	975.394	977.40	977.40
	Optimal Travel Time (min)	59	57	57
Comparison between ESODTA and TSODTA	Emission Reduction by ESODTA (gm)	0	9.12	11.4
	Travel Time Reduction by TSODTA (min)	0	4	5

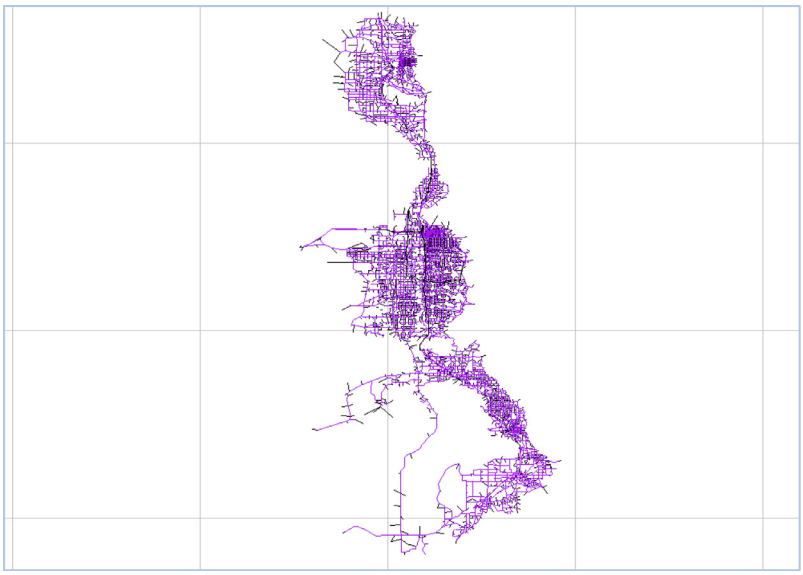


Fig. 12. Salt Lake City regional traffic network.

time, and vice versa in TSODTA. In practice, the decision maker chooses between the ESODTA and the travel time-based SODTA depending on his or her goal.

5.1.2. Sensitivity analysis on link capacities

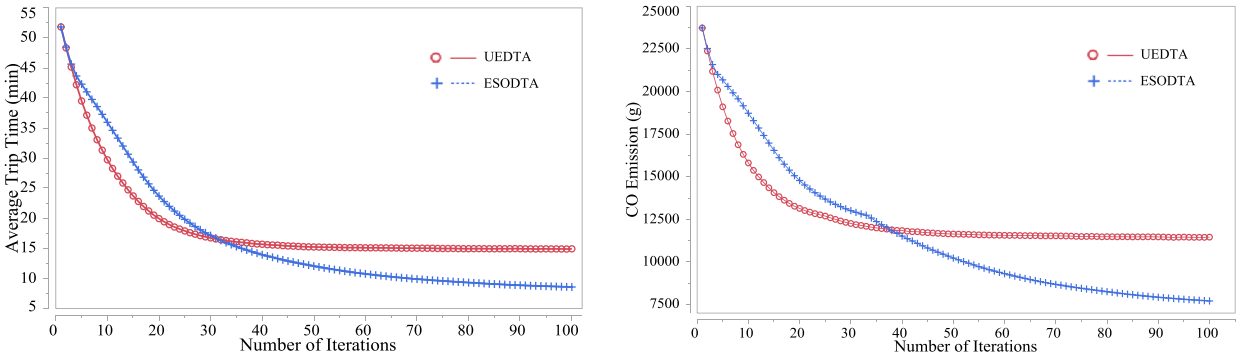
The second set of sensitivity analysis is conducted to examine the impact of different link capacities on the solution of ESODTA. We assume there are 5 agents departing at time 0 and 4 agents departing at time 1. The capacity of link (2, 4) is changed to be 1, 2 and 3 agents per min, respectively, in three different cases (case 1, 2, and 3). It is reminded that the queuing speed could change if the capacity is modified. After 50 iterations, the gap values of case 1, 2, and 3 between the lower bound and the optimal solution are 7.8%, 7.2% and 7.0%, respectively.

The travel-time based SODTA problem is also solved, and the comparison of ESODTA and TSODTA under three different cases is presented in Table 4. As shown in Table 4, (i) the optimal total emission cost of ESODTA and the optimal total travel time of TSODTA decrease with the increase of the capacity of link (2, 4). It indicates that the increased capacity on link (2, 4) could lead to a lower queueing time on that link and hence reducing the optimal total emission cost of ESODTA and the optimal total travel time of TSODTA. (ii) Similar to the sensitivity analysis on travel demand level, the emission reduction in ESODTA is gained at the expense of an increased total travel time, and vice versa in TSODTA.

5.2. Numerical example for the simulation-based ESODTA in a large-scale network

This numerical experiment is conducted on the Salt Lake City regional traffic network with 13,923 nodes, 26,768 links and 2302 zones shown in Fig. 12 (Zhou, et al., 2015b). The total number of vehicles is about 1.35 million for a 3-hour demand loading period, 15:00–18:00.

Although the primary focus of this experiment is on the simulation-based ESODTA model, presented in Section 4.1, we also solve the travel time-based user equilibrium DTA (UEDTA) problem (Lu et al., 2009) and compare the results of the two models. Average trip time and average CO emission of the two models for 100 iterations are depicted in Fig. 13. We can observe that: (i) in Fig. 13(a) although the objective of the ESODTA is to minimize the total emission, its final average trip time is also less than that of the UEDTA model; (ii) in Fig. 13(b) the final average CO emission of the ESODTA model is less than that of the UEDTA model, because travelers choose the least marginal emission routes in the ESODTA model.



(a) Average trip time of two models

(b) CO emission of two models

Fig. 13. Comparisons of average trip time and average CO emission of ESODTA and UEDTA models.

According to the final average trip time and average CO emission of the two models, one vehicle in average can save 3700 gs of CO emissions and 6 minutes of travel time if the routing policy of the ESODTA model is adopted. Considering the total demand of 1.35 million vehicles, significant savings in total emission and travel time can be achieved by adopting the routing policy of the ESODTA model. Meanwhile, in order to improve the computation efficiency, the two simulation-based algorithms are implemented using a parallel computing technique enhanced by an open-source mesoscopic traffic assignment simulator, DTALite by (Zhou and Taylor, 2014). The computational time for one iteration of the solution algorithms is about 2 minutes in the computing environment used. Note that, because (i) the approximation method is used to obtain approximate path marginal emission and path marginal travel time and (ii) the determination of the optimal step size in each iteration is computationally prohibitive, precise system optimal solutions are difficult to be achieved by the simulation-based approach for this large-scale network application. Therefore, we do not intend to test the travel time-based SODTA in this paper. The precise comparison between ESODTA and TSODTA using the analytical model can be found in Section 5.1.

6. Conclusions

This research generalizes the normative capability of DTA to sustainable transportation network modelling in terms of emission. The ESODTA model is proposed to obtain time-varying path flows in a congested network where travellers behave cooperatively in selecting paths to minimize total network emissions. The analytical ESODTA model with constant bottleneck capacities is formulated based on the expanded space-time network constructed using the simplified queuing model (Section 3). The column generation-based heuristic for the simulation-based ESODTA model consists of a mesoscopic DNL model and a gradient projection-based descent direction method for updating time-dependent path assignments (Section 4). Furthermore, link marginal emission is derived and its relationship with link marginal travel time is discussed. The numerical examples demonstrate the effectiveness of the model and the algorithm for obtaining the routing policy which minimizes total network emissions, and the emission reduction in ESODTA is gained at the expense of an increased total travel time. In addition, compared to the UEDTA, significant savings in total emission and travel time can be achieved by adopting the routing policy of the ESODTA model.

The solution to the ESODTA model provides a benchmark (or lower bound) in total system emission when drivers are unanimously guided by a central controller against other less environmentally-efficient routing policies, such as green user equilibrium (GUE) flow patterns. Moreover, the Eco-SO path flows provide a basis of generating green routing policies for advanced eco-friendly route guidance provision systems that consider different market penetration rates. Link and path marginal emissions, which are essential input to the ESODTA algorithm, are also valuable for deriving emission charges on road users.

It should be remarked that one thorny issue in the analytical SODTA model is a holding-back phenomenon; that is a vehicle is held on a link even though there is enough capacity for that vehicle to move onto its downstream link. A number of notable approaches have been proposed based on mixed integer models (e.g., Zhang et al., 2010a; Zheng and Chiu, 2011; Doan and Ukkusuri, 2012) or nonlinear programming models (Carey and Subrahmanian, 2000; Shen et al., 2007b) to address this unrealistic phenomenon in a SODTA solution. Recently, Zhu and Ukkusuri (2013) presented a penalty-based linear model that generates a non-holding-back solution based on the post-process of a prior solved SODTA solution.

This holding-back phenomenon could also exist in the solution (in terms of agent space-time trajectories) to the proposed ESODTA model, presented in Section 3.3. To resolve this holding-back issue, one could add side constraints that force vehicles to move forward if the downstream capacity is available. This type of if-then condition constraints is typically handled through the big-M method in integer programing, which could lead to extra computational burden. To enable the analytical SODTA model for large-scale networks, a future research direction could be on developing a space-time network processing tool that (i) automatically pre-builds potential waiting nodes and handles different capacity allocation rules as shown in Fig. 3, and (ii) prevents holding-back phenomena through adding valid if-then condition constraints into the proposed

integer programming model. Note that in the simulation-based approach presented in Section 4, the holding-back phenomenon could still be found in the time-dependent shortest path calculation results (in the extended space-time network), but generally it does not exist in the final simulated vehicle trajectories as the simulation model always moves vehicles forward along their assigned paths when the downstream capacity is available.

Acknowledgements

The research reported here was partially sponsored by U.S. Department of Transportation (DOT) University Transportation Centers, "Development and Demonstration of Advanced Methods for Quantifying Freight Truck Activity, Energy Use, and Emissions", partially funded by U.S. Department of Energy's (DOE) Advanced Research Projects Agency – Energy (ARPA-E), "Traveler Response Architecture using Advanced Novel Signaling for Network Efficiency in Transportation (TRANSNET)", partially funded by National Science Foundation – United States under Grant No. CMMI 1538105 "Collaborative Research: Improving Spatial Observability of Dynamic Traffic Systems through Active Mobile Sensor Networks and Crowdsourced Data", and partially funded by Utah Department of Transportation, "Simplified web-based decision support method for traffic management and work zone analysis". The work presented in this paper remains the sole responsibility of the authors.

References

Abdul Aziz, H.M., Ukkusuri, S.V., 2012. Integration of environmental objectives in a system optimal dynamic traffic assignment model. Comput. Aided Civil Infrastruct. Eng. 27 (7), 494–511.
Bai, S., Nie, Y., Niemeier, D.A., 2007. The impact of speed post-processing methods on regional mobile emissions estimation. Transp. Res. Part D 12 (5), 307–324.
Benedek, C.M., Rilett, L.R., 1998. Equitable traffic assignment with environmental cost function. J. Transp. Eng. 124, 16–22.
Boriboonsomsin, K., Barth, M., 2008. Impacts of freeway high-occupancy vehicle lane configuration on vehicle emissions. Transp. Res. Part D 13 (2), 112–125.
Carey, M., 1987. Optimal time-varying flows on congested networks. Oper. Res. 35 (1), 58–69.
Carey, M., 1992. Nonconvexity of the dynamic traffic assignment problem. Transp. Res. Part B 26 (2), 127–133.
Carey, M., Subrahmanian, E., 2000. An approach to modelling time-varying flows on congested networks. Transp. Res. Part B 34 (3), 157–183.
Carey, M., Watling, D., 2012. Dynamic traffic assignment approximating the kinematic wave model: system optimum, marginal costs, externalities and tolls. Transp. Res. Part B 46 (5), 634–648.
Cetin, M., 2012. Estimating queue dynamics at signalized intersections from probe vehicle data: methodology based on kinematic wave model. Transp. Res. Rec. 2315, 164–172.
Daganzo, C.F., 1994. The cell transmission model: a simple dynamic representation of highway traffic. Transp. Res. Part B 28 (4), 269–287.
Daganzo, C.F., 1995. The cell transmission model, part II: network traffic. Transp. Res. Part B 29 (2), 79–93.
Daganzo, C.F., 2001. A simple traffic analysis procedure. Netw. Spatial Econ. 1 (1-2), 77–101.
Daganzo, C.F., 2006. In traffic flow, cellular automata = kinematic waves. Transp. Res. Part B 40 (5), 396–403.
Doan, K., Ukkusuri, S.V., 2012. On the holding-back problem in the cell transmission based dynamic traffic assignment models. Transp. Res. Part B 46 (9), 1218–1238.
Doan, K., Ukkusuri, S.V., 2015. Dynamic system optimal model for multi-OD traffic networks with an advanced spatial queuing model. Transp. Res. Part C 51, 41–65.
Erera, A.L., Lawson, T.W., Daganzo, C.F., 1998. Simple, generalized method for analysis of traffic upstream of a bottleneck. Transp. Res. Rec. 1646, 132–140.
Frey, H.C., Liu, B., 2013. Development and evaluation of a simplified version of MOVES for coupling with a traffic simulation model. In: Proceedings, 91st Annual Meeting of the Transportation Research Board. Washington, DC.
Friesz, T.L., Luque, J., Tobin, R.L., Wie, B.-Y., 1989. Dynamic network traffic assignment considered as a continuous time optimal control problem. Oper. Res. 37 (6), 893–901.
Gaker, D., Zheng, Y., Walker, J., 2010. Experimental economics in transportation: focus on social influences and provision of information. Transp. Res. Rec. 2156, 47–55.
Gaker, D., Vautin, D., Vij, A., Walker, J.L., 2011. The power and value of green in promoting sustainable transport behavior. Environ. Res. Lett. 6 (3), 1–11.
Garcia, A., Reaume, D., Smith, R.L., 2000. Fictitious play for finding system optimal routings in dynamic traffic networks. Transp. Res. Part B 34 (2), 147–156.
Ghali, M.O., Smith, M.J., 1995. A model for the dynamic system optimum traffic assignment problem. Transp. Res. Part B 29 (3), 155–170.
Greene, D., Schafer, A., 2003. Reducing Greenhouse Gas Emissions from U.S. Transportation. Center for Climate and Energy Solutions http://www.c2es.org/docUploads/ustransp.pdf.
Huang, Y., Yang, L., Tang, T., Cao, F., Gao, Z., 2016. Saving energy and improving service quality: bicriteria train scheduling in urban rail transit systems. IEEE Trans. Intell. Transp. Syst. doi:10.1109/TITS.2016.2549282, in press.
Lafortune, S., Sengupta, R., Kaufman, D.E., Smith, R., 1993. Dynamic system-optimal traffic assignment using a state space model. Transp. Res. Part B 27 (6), 451–472.
Lawson, T.W., Lovell, D.J., Daganzo, C.F., 1997. Using input-output diagram to determine spatial and temporal extents of a queue upstream of a bottleneck. Transp. Res. Rec. 140–147 No. 1572.
Leclercq, L., 2007. Hybrid approaches to the solutions of the "Lighthill-Whitham-Richards" model. Transp. Res. Part B 41 (7), 701–709.
Li, P., Mirchandani, P., Zhou, X., 2015. Solving simultaneous route guidance and traffic signal optimization problem using space-phase-time hypernetwork. Transp. Res. Part B 81, 103–130.
Lighthill, M., Whitham, G.., 1955. On kinematic waves II: a theory of traffic flow on long crowded roads. Proc. R. Soc. London, Part 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. Transp. Res. Part B 43 (3), 345–364.
Lu, C.-C., Zhou, X., Zhang, K., 2013. Dynamic origin-destination demand flow estimation under congested traffic conditions. Transp. Res. Part C 34, 16–37.
Ma, R., Ban, X.J., Pang, J.S., 2014. Continuous-time dynamic system optimum for single-destination traffic networks with queue spillbacks. Transp. Res. Part B 68, 98–122.
Ma, R., Ban, X.J., Szeto, W.Y., 2015. Emission modeling and pricing in dynamic traffic networks. Transp. Res. Procedia 9, 106–129.
Magnanti, T., Perakis, G., 1997. Averaging schemes for variational inequalities and systems of equations. Math. Oper. Res. 22 (3), 568–587.
Mandavilli, S., Rys, M.J., Russell, E.R., 2008. Environmental impace of modern roundabouts. Int. J. Ind. Ergon. 38 (2), 135–142.
Merchant, D.K., Nemhauser, G.L., 1978. A model and an algorithm for the dynamic traffic assignment problems. Transp. Sci. 12 (3), 183–199.
Mounce, R., Smith, M., 2007. Uniqueness of Equilirbium in steady state and dynamic traffic networks. In: Allsop, R.E., Bell, M.G.H., Heydecker, B.G. (Eds.), Transportation and Traffic Theory. Elsevier, pp. 281–299.
Munoz, J.C., Laval, J.A., 2006. System optimum dynamic traffic assignment graphical solution method for a congested freeway and one destination. Transp. Res. Part B 40 (1), 1–15.

Nagurney, A., Ramanujam, P., Dhanda, K.K., 1998. Multimodal traffic network equilibrium model with emission pollution permits: compliance vs noncompliance. Transp. Res. Part D 35, 349–374.
Nagurney, A., Dong, J., Mokhtarian, P.L., 2002. Traffic network equilibrium and the environment: a multicriteria decision-making perspective. In: Kontoghiorges, E., Rustem, B., Siokos, S. (Eds.), Computational Methods in Decision-Making, Economics and Finance. Kluwer Academic Publishers, Dordrecht, pp. 501–523.
Newell, G.F., 1993a. A simplified theory on kinematic waves in highway traffic, part I: general theory. Transp. Res. Part B 27 (4), 281–287.
Newell, G.F., 1993b. A simplified theory on kinematic waves in highway traffic, part II: queueing at freeway bottlenecks. Transp. Res. Part B 27 (4), 289–303.
Newell, G.F., 1993c. A simplified theory on kinematic waves in highway traffic, part III: multi-destination flows. Transp. Res. Part B 27 (4), 305–313.
Newell, G., 2002. A simplified car-following theory: a lower order model. Transp. Res. B 36 (3), 195–205.
Panis, L., Broekx, S., Ronghui, L., 2006. Modeling instantaneous traffic emission and the influence of traffic speed limits. Sci. Total Environ. 371 (1-3), 270–285.
Peeta, S., Mahmassani, H.S., 1995a. System optimal and user equilibrium time-dependent traffic assignment in congested networks. Ann. Oper. Res. 60 (1), 81–113.
Peeta, S., Mahmassani, H.S., 1995b. Multiple user classes real-time traffic assignment for on-line operations: a rolling horizon solution framework. Transp. Res. Part C: Emerg. Technol. 3 (2), 83–98.
Peeta, S., Zhou, C., 2006. Stochastic quasi-gradient algorithm for the off-line stochastic dynamic traffic assignment problem. Transp. Res. Part B Methodol. 40 (3), 179–206.
Peeta, S., Ziliaskopoulos, A.K., 2001. Foundations of dynamic traffic assignment: the past, the present and the future. Netw. Spatial Econ. 1 (3/4), 233–266.
Qian, Z., Zhang, H.M., 2011. Computing individual path marginal cost in networks with queue spillbacks. Transp. Res. Rec. 2263, 9–18.
Qian, Z., Shen, W., Zhang, H.M.., 2012. System-optimal dynamic traffic assignment with and without queue spillback: its path-based formulation and solution via approximate path marginal cost. Transp. Research part B Methodol. 46 (7), 874–893.
Richards, P.I., 1956. Shock waves on the highway. Oper. Res. 4 (1), 42–51.
Sbayti, H., Lu, C.-C., Mahmassani, H.S., 2007. Efficient implementations of the method of successive averages in simulation-based DTA models for large-scale network applications. Transp. Res. Rec. 2029, 22–30.
Shen, W., Nie, Y., Zhang, H.M., 2007a. On path marginal cost analysis and its relation to dynamic system-optimal traffic assignment. In: Allsop, R.E., Bell, M.G.H., Heydecker, B.G. (Eds.), Transportation and Traffic Theory. Elsevier, pp. 319–352.
Shen, W., Nie, Y., Zhang, H., 2007b. Dynamic network simplex method for designing emergency evacuation plans. Transp. Res. Rec. 2022, 83–93.
Shen, W., Zhang, H.M., 2009. On the morning commute problem in a corridor network with multiple bottlenecks: its system-optimal traffic flow patterns and the realizing tolling scheme. Transp. Res. Part B 43 (3), 267–284.
Szeto, W.Y., Jaber, X., Wong, S.C., 2012. Road network equilibrium approaches to environmental sustainability. Transp. Rev. 32 (4), 491–518.
Tong, L., Zhou, X., Miller, H.J., 2015. Transportation network design for maximizing space–time accessibility. Transp. Res. Part B Methodol. 81, 555–576.
Tzeng, G.H., Chen, C.H., 1993. Multiobjective decision making for traffic assignment. IEEE Trans. Eng. Manage. 40, 180–187.
U.S. EPA, 2009. Draft Moto Vehicle Emission Simulator (MOVES) – Software Design and Reference Manual. U. S. Environmental Protection Agency Technical report EPA-420-B-09-007.
Vallamsundar, S., Lin, J., Konduri, K., Zhou, X., Pendyala, R.M., 2016. A comprehensive modeling framework for transportation-induced population exposure assessment. Transp. Res Part D Transp. Environ. 46, 94–113.
Wie, B.-W., Tobin, R.L., Friesz, T.L., 1994. The augmented Lagrangian method for solving dynamic network traffic assignment models in discrete time. Transp. Sci. 28 (3), 204–220.
Yang, L., Li, S., Gao, Y., Gao, Z., 2015. A coordinated routing model with optimized velocity for train scheduling on a single-track railway line. Int. J. Intell. Syst. 30, 3–22.
Yang, L., Zhang, Y., Li, S., Gao, Y., 2016. A two-stage stochastic optimization model for the transfer activity choice in metro networks. Transp. Res. Part B 83, 271–297.
Yin, J., Tang, T., Yang, L., Gao, Z., Ran, B., 2016. Energy-efficient metro train rescheduling with uncertain time-variant passenger demands: an approximated dynamic programming approach. Transp. Res. Part B 91, 178–210.
Zhang, H.M., Nie, Y., Qian, Z., 2013a. Modelling network flow with and without link interactions: the cases of point queue, spatial queue and cell transmission model. Transp. B Transp. Dyn. 1 (1), 33–51.
Zhang, K., Mahmassani, H.S., Lu, C.-C., 2013b. Dynamic pricing, heterogeneous users and perception error: probit-based bi-criterion dynamic stochastic user equilibrium assignment. Transp. Res. Part C 27, 189–204.
Zhang, Y., Lv, J., Ying, Q., 2010a. Traffic assignment considering air quality. Transp. Res. Part D 15, 497–502.
Zheng, H., Chiu, Y.C., 2011. A network flow algorithm for the cell-based single-destination system optimal dynamic traffic assignment problem. Transp. Sci. 45 (1), 121–137.
Zhou, X., Taylor, J., 2014. DTALite: a queue-based mesoscopic traffic simulator for fast model evaluation and calibration. Cogent Eng. 1 (1), 961345.
Zhou, X., Tanvir, S., Lei, H., Taylor, J., Liu, B., Rouphail, N.M., Frey, H.C., 2015a. Integrating a simplified emission estimation model and mesoscopic dynamic traffic simulator to efficiently evaluate emission impacts of traffic management strategies. Transp. Res. Part D 37, 123–136.
Zhou, X., Zlatkovic, M., Farhan, M., 2015b. Simplified Web-Based Decision Support Method for Traffic Management and Work Zone Analysis. For Utah Department of Transportation Research Division Report No. UT-15.09.
Zhang, L., Yin, Y., Lou, Y., 2010b. Robust signal timing for arterials under day-to-day demand variations. Transp. Res. Rec. 2192, 156–166.
Zhu, F., Lo, H.K., Lin, H.-Z., 2013. Delay and emissions modelling for signalized intersections. Transp. B Transp. Dyn. 1 (2), 111–135.
Zhu, F., Ukkusuri, S.V., 2013. A cell based dynamic system optimum model with non-holding back flows. Transp. Res. Part C 36, 367–380.
Ziliaskopoulos, A.K., 2000. A linear programming model for the single destination system optimum dynamic traffic assignment problem. Transp. Sci. 34 (1), 1–12.
Ziliaskopoulos, A.K., Mahmassani, H.S., 1993. Time dependent shortest-path algorithm for real-time Intelligent Vehicle Highway System applications. Transp. Res. Rec. 1408, 94–100.