Equilibrium analysis and route guidance in large-scale networks with MFD dynamics

Mehmet Yildirimoglu, Mohsen Ramezani, Nikolas Geroliminis *

Ecole Polytechnique Fédérale de Lausanne (EPFL), School of Architecture, Civil and Environmental Engineering (ENAC), Urban Transport Systems Laboratory (LUTS), Lausanne, Switzerland

ABSTRACT

Recent studies have demonstrated that Macroscopic Fundamental Diagram (MFD), which provides an aggregated model of urban traffic dynamics linking network production and density, offers a new generation of real-time traffic management strategies to improve the network performance. However, the effect of route choice behavior on MFD modeling in case of heterogeneous urban networks is still unexplored. The paper advances in this direction by firstly extending two MFD-based traffic models with different granularity of vehicle accumulation state and route choice behavior aggregation. This configuration enables us to address limited traffic state observability and to scrutinize implications of drivers' route choice in MFD modeling. We consider a city that is partitioned in a small number of large-size regions (aggregated model) where each region consists of medium-size sub-regions (more detailed model) exhibiting a well-defined MFD. This paper proposes a route guidance advisory control system based on the aggregated model as a large-scale traffic management strategy that utilizes aggregated traffic states while sub-regional information is partially known. In addition, we investigate the effect of equilibrium conditions (i.e. user equilibrium and system optimum) on the overall network performance, in particular MFD functions.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Research on congestion propagation in large urban networks has been mainly based on micro-simulations where traffic dynamics are defined at the link level. However, because of unpredictability of travelers' behavior and high complexity of traffic physical modeling, microscopic simulation models are time consuming and case dependent. In addition, traffic congestion management in large-scale urban systems is currently fragmented and uncoordinated with respect to optimizing the goals of travel efficiency and equity for multiple regions of a city. An alternative, which is investigated in this article, is a regional route guidance strategy, where a network is partitioned into homogeneous regions and drivers are given a regional path to follow (e.g. going through the city center or using the longer route at the periphery). On the other hand, the effect of driver adaptation models on the aggregate network performance is not fully explored. Microscopic simulation models depend on enroute decision mechanisms, where drivers update their routes based on instantaneous travel times at the network. This behavioral assumption may cause uneven distribution of congestion in the network and reduce the network flow especially in the unloading phase where certain parts of the network are already more congested than others. However,

* Corresponding author at: GC C2 389, Station 18, 1015 Lausanne, Switzerland. Tel.: +41 21 6932481; fax: +41 21 69 35060.
   E-mail addresses: mehmet.yildirimoglu@epfl.ch (M. Yildirimoglu), mohsen.ramezani@epfl.ch (M. Ramezani), nikolas.geroliminis@epfl.ch (N. Geroliminis).

under equilibrium conditions drivers are expected to make decisions based on experienced travel costs, which may lead to a more homogenous network in both loading and unloading phases. Understanding these interactions for heterogeneously congested cities is a big challenge, which will allow revisiting, redesigning, and integrating smarter traffic management approaches to alleviate congestion with a "system of systems" approach.

The traffic modeling in this paper is based on the network macroscopic fundamental diagram (MFD) that provides a unimodal, low-scatter, and demand-insensitive relationship between network vehicle density and space-mean flow in homogeneous urban areas (with small spatial link density heterogeneity) (Geroliminis and Daganzo, 2008). The idea of an MFD with an optimum accumulation belongs to Godfrey (1969) and similar approaches were introduced later by Herman and Prigogine (1979), Mahmassani et al. (1984) and Daganzo (2007). However, urban transportation networks exhibit uneven distribution of congestion which leads to a scattered flow-density relationship. Heterogeneity in congestion distribution can affect the shape/scatter or even the existence of MFD (Buisson and Ladier, 2009; Geroliminis and Sun, 2011). By using a grid network and considering variance of link density (over space) as an independent variable, Mazloumian et al. (2010) shows that MFD remains well-defined in subregions of the urban network. These results are very critical, because MFD concept can be useful for heterogeneously loaded cities, if the network can be partitioned into smaller homogenous areas. The effect of heterogeneity has been recently studied by many researchers with similar conclusions, see for example (Daganzo et al., 2011; Mahmassani et al., 2013b; Geroliminis and Sun, 2011; Doig et al., 2013) and others. Ji and Geroliminis (2012) develops a static partitioning mechanism to minimize the variance of link densities while maintaining a spatially compact shape. Resulting subregions can be used to develop macroscopic traffic control strategies. In case of hierarchical networks with respect to topology, e.g. mixed freeway/arterial networks, hybrid models might be utilized. For example, Haddad et al. (2013) models the urban part of a city with multi-region MFDs, while the freeway network is represented by the Cell Transmission Model (Daganzo, 1994). Transfer flows between the two models guarantee consistency and conservation of flows. More complicated network structures with strong directional flows, limited connectivity, and high variability in trip lengths and connection with MFD modeling and clustering should be further investigated. This is work in progress.

Mahmassani et al. (2013b) and Gayah and Daganzo (2011) investigate the effect of driver adaptation on the shape of MFD through microscopic simulation for different network sizes. They both identify that an increase in driver adaptivity through enroute decision models can influence and shrink the size of hysteresis loops in the network MFD. In addition, Knoop et al. (2012) has developed myopic local re-routing strategies considering the aggregated information from multiple sub-networks. This strategy does not consider the interaction of the vehicles in the following time periods. Leclercq and Geroliminis (2013) investigates approximations of user equilibrium (UE) and system optimum (SO) conditions of a simple network with MFD dynamics and two parallel routes with instantaneous travel times. Their findings reveal that SO improves system performance compared to UE mainly when initial network conditions are not in the congested regime of the MFDs. Nevertheless, the effect of equilibrium flows in real-sized networks has not been tested with MFD dynamics. In addition, stability analysis for MFD dynamics and control has been investigated in Haddad and Geroliminis (2012) and Haddad and Shraiber (2014).

Real-time large-scale traffic management strategies, e.g. multi-region perimeter control (Geroliminis et al., 2013; Haddad et al., 2013; 1), gating (Daganzo, 2007; Keyvan-Ekbatani et al., 2012; Keyvan-Ekbatani et al., 2015)) that benefit from parsimonious models with aggregated network dynamics, provide promising results towards a new generation of smart hierarchical strategies. On the other hand, the estimation of network traffic states for MFD analysis with different types of sensors identifies the applicability of MFD in large scale networks even if limited data exist, see for example (Ortigosa et al., 2013; Gayah and Dixit, 2013; Nagle and Gayah, 2014; Leclercq et al., 2014; Ji et al., 2014). Furthermore, a connection of travel time reliability with network heterogeneity based on MFD concepts have been investigated by Gayah et al. (2013), Mahmassani et al. (2013a), Yildirimoglu et al. (2015).

The primary motivation of the paper is to develop a network-level traffic management scheme to mitigate congestion in urban areas by considering the effect of route choice at an aggregated level. The management scheme is developed based on MFD and consists of a route guidance system that advises drivers a sequence of subregions to assist them in reaching their destination. This study extends the work in Yildirimoglu and Geroliminis (2014) to a route guidance system based on SO conditions. It is worth mentioning that we aim at network-level management strategies, thus the route guidance operates on subregional basis as opposed to conventional link-based route guidance, see for example (Papageorgiou, 1990; Ben-Akiva et al., 1997; Zhou et al., 2008; Ben-Elia et al., 2013; Zhu and Ukkusuri, 2013). The route guidance system can impact the travelers' route decision by providing them with useful information regarding the traffic states of the urban regions. Therefore, drivers can follow a series of subregions that has lower cost (in terms of travel time, fuel consumption, etc.), which might lead to a better overall system performance. The second motivation is to investigate the impact of driver adaptivity on the overall network performance, in particular MFDs. Most of the previous MFD estimations in the literature are based on one-shot simulations where driver adaptivity is incorporated by en-route decision mechanisms (e.g. current best or myopic local re-routing). This study tests the effect of dynamic UE (or DUE) and dynamic DSO (or DSO) flows in the network on the observed MFD functions and the existence of hysteresis loops. This article is the first attempt to integrate equilibrium flow conditions in the network MFD analysis, while different approaches have been utilized in Mahmassani et al. (2013b) (micro-simulation and instantaneous traffic conditions as opposed to equilibrium conditions), Gayah and Daganzo (2011) (a toy network with two rings), and Leclercq and Geroliminis (2013) (two parallel routes).

We investigate the problem where a heterogeneous city, in terms of link density, consists of several smaller and more homogeneous subregions, see Fig. 1, where each subregion (1–19) can be represented by a well-defined low scatter MFD.
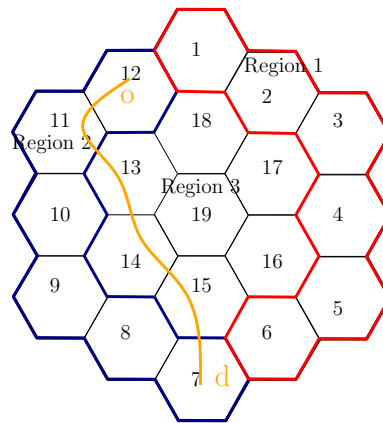
Fig. 1. The schematic of a multi-region urban network and a path example. The network consists of 19 sub-regions that are aggregated in 3 larger regions (Region 1 contains subregions 1–6, Region 2 contains 7–12, and Region 3 contains 13–19).

In result, the MFD of (larger) regions might exhibit hysteresis (due to heterogeneity) and have significant level of scatter in the congested regime. The partitioning of the city network to homogenous subregions enables us to model and study the effect of link density heterogeneity on the MFD characteristics, and also results in a configuration that obliges us to integrate a route choice model into the MFD traffic flow modeling. (Larger) regions should have a comparable size with the average trip length of the vehicles to allow alternative (subregional) route options. The modeling contribution of the paper is to address the dynamic UE and SO route choice within the MFD framework, which consists of a detailed modeling of traffic flow for each subregion. We introduce traffic dynamics at the subregional (detailed) and regional (parsimonious) level. While the detailed representation provides more information on network traffic state, some parameters of this model (e.g. route choice characteristics) might not be fully observable. Thus, the parsimonious model is essential for management (i.e. route guidance) purposes. Compared with previous works, the introduced regional dynamics is the generalized multi-region model (instead of 2-region model), incorporated with heterogeneity modeling (instead of assuming homogenous conditions), and with integrated route choice modeling (instead of no regional route choice). Additionally, this integration of 2 different scales of modeling unveils interesting regularities with respect to the trip length of drivers under different strategies. Note that the intension of this work is not to develop traffic management strategies for networks partitioned in too many regions. In this case, dynamic origin–destination matrices and other detailed information would be very difficult to be estimated with small errors. The subregional model can be considered as a more detailed (but not well-observed with data) representation (in control theory this is called the "plant"), while the regional model (3 regions in Fig. 1) is analyzed for management purposes ("model"). We apply the current analytical plant instead of a detailed micro-simulation, as this paper chooses a more methodological path, which allows us to create further insights of the dynamics of heterogeneity and route guidance strategy.

The remainder of the paper is organized as follows; in Section 2, we introduce the dynamics of the region-based model (optimization model) and the subregion-based model (plant). In Section 3, the methodological frameworks for dynamic traffic assignment (DTA) and route guidance (RG) system are elaborated. Section 4 presents and discusses results of a case study with different demand levels. Finally, Section 5 concludes the paper with future work directions.

2. Traffic models

In this study, we introduce two macroscopic traffic models: (i) a region-based model considering that the urban network is partitioned into a small number of regions, and (ii) a subregion-based model deploying a more detailed approach where the above regions are divided into smaller subregions. Two levels of modeling enables us to account for different layers of traffic state measurement and control and to incorporate heterogeneity effect in the urban network dynamics. The region-based model serves as the operation model that the management strategy (e.g. route guidance) is based on, while subregion-based model serves as the plant (reality) dynamics, which represents the ground truth traffic conditions. Region-based model integrates heterogeneity dynamics in two ways: (a) a time-varying average regional trip length and (b) an MFD depending on regional accumulation and the heterogeneity of the spatial distribution of congestion. On the other hand, the subregion-based model describes the evolution of traffic based on the MFD considering constant subregional average trip length, subregional paths, and the boundary flow capacity between subregions. We show in the article that the two models are consistent and that the less detailed model can integrate variable trip lengths, hysteresis loops and spatial heterogeneity in the distribution of congestion.

This configuration with two levels of traffic aggregation, i.e. regional and subregional, allows us to develop more realistic control schemes such that the partial observability of traffic states and control tractability can be investigated. While

time-dependent trip lengths have been considered in Ramezani et al. (2015) and Yildirimoglu and Geroliminis (2014), there are two significant modeling extensions in this work: (i) vehicles are allowed to cross the boundary between the regions more than one time (e.g. a trip with the sequence 11-13-10-14-8-7 is allowed) and (ii) equilibrium conditions have been satisfied in both models.

2.1. Region-based model

Let us assume that the urban network is partitioned into several subregions with well-defined MFDs. We consider a region as a set of multiple adjacent subregions, see Fig. 1, forming R regions, $\mathcal{R} = \{1, 2, \ldots, R\}$. Let $Q_{IJ}(t)$ (veh/s) denote the exogenous traffic flow demand generated in region I with *final* destination region J, $N_{IJ}(t)$ (veh) be the vehicle accumulation in region I with final destination region J, and $N_I(t)$ (veh) be the total accumulation in region I at time t; $I, J \in \mathcal{R}$; $N_I(t) = \sum_{J \in \mathcal{R}} N_{IJ}(t)$.

The traffic flow conservation equations of an R-region MFDs system are as follows:

$$\frac{dN_{II}(t)}{dt} = Q_{II}(t) - M_{II}^I(t) - \sum_{H \in \mathcal{V}_I} \hat{M}_{II}^H(t) + \sum_{H \in \mathcal{V}_I} \hat{M}_{HI}^I(t) \tag{1}$$

$$\frac{dN_{IJ}(t)}{dt} = Q_{IJ}(t) - \sum_{H \in \mathcal{V}_I} \hat{M}_{IJ}^H(t) + \sum_{H \in \mathcal{V}_I} \hat{M}_{HJ}^I(t) \quad I \neq J \tag{2}$$

where $\mathcal{V}_I$ is the set of regions that are directly reachable from region I, i.e. adjacent regions to region I. $M_{II}^I(t)$ (veh/s) denotes the *internal trip completion rate* for accumulation in I with destination I (without going through another region), while the *transfer flow* for accumulation in I with final destination J through the next immediate region H is denoted by $\hat{M}_{IJ}^H(t)$ (veh/s) that considers the receiving flow capacity on the transfer flow from region I to region H. Note that, Eq. (1) and (2) permit the vehicle paths to include more than one crossing over the boundaries between the regions. For example a trip through subregions 11-13-19-15-7 (see Fig. 1), starts and ends in region 2 while crossing region 3.

Internal trip completion rates and transfer flows are estimated corresponding to the ratio between accumulations as:

$$M_{II}^I(t) = \theta_{II}^I(t) \cdot \frac{N_{II}(t)}{N_I(t)} \cdot \frac{F_I(N_I(t), \sigma(N_I(t)))}{L_{II}(t)}, \tag{3}$$

$$M_{IJ}^H(t) = \theta_{IJ}^H(t) \cdot \frac{N_{IJ}(t)}{N_I(t)} \cdot \frac{F_I(N_I(t), \sigma(N_I(t)))}{L_{IH}(t)}, \tag{4}$$

$$\widehat{M}_{IJ}^H(t) = \min \left[M_{IJ}^H(t), C_{IH}(N_H(t)) \cdot \frac{N_{IJ}(t) \cdot \theta_{IJ}^H(t)}{\sum_{K \in \mathcal{R}} \left(N_{IK}(t) \cdot \theta_{IK}^H(t) \right)} \right], \tag{5}$$

where $F_I(\cdot)$ (veh m/s) is the production MFD of region I (i.e. the total distance travelled per unit time in the region) that is a function of the region accumulation, $N_I(t)$, and the link density *heterogeneity* across all region I links, $\sigma(N_I(t))$. Moreover, $L_{II}(t)$ (m) is the average trip length corresponding to internal trips in region I, $L_{IH}(t)$ (m) is the average trip length corresponding to transfer trips from region I to its neighbor region $H \in \mathcal{V}_I$, and $\theta_{IJ}^H(t)$ is the percentage of outflow from region I to final destination J through the next immediate region H; hence $\sum_{H \in \mathcal{V}_I} \theta_{IJ}^H(t) = 1$.[1] Modeling the region link density heterogeneity, i.e. $\sigma(N_I(t))$, is investigated in Ramezani et al. (2015); while estimation of traffic states $L_{IJ}^H(t)$ and $\theta_{IJ}^H(t)$ is discussed in the next subsection.[2]

Eq. (5) estimates the transfer flow from region I to the next region H that is the minimum of two terms: (i) the sending flow from region I that solely depends on the region I accumulation and (ii) the product of the boundary capacity between regions I and H that is a function of region H accumulation, i.e. $C_{IH}(N_H(t))$, and the proportion of vehicles in I with final destination J through the next region H, i.e. $N_{IJ}(t) \cdot \theta_{IJ}^H(t)$, among all the vehicles that will cross the same boundary, i.e. $\sum_{K \in \mathcal{R}} \left(N_{IK}(t) \cdot \theta_{IK}^H(t) \right)$.

Eqs. (1)–(5) represent the traffic dynamics in an R-region urban network considering the heterogeneity effect and integrating an aggregated route choice model, which is further exploited in the route guidance system. The model assumes that drivers can choose any arbitrary sequence of regions as their path and cross region boundaries without any limitation. While in previous publications related to urban traffic control (e.g. Geroliminis et al., 2013; Haddad et al., 2013) region receiving capacity was omitted as perimeter control was operating in the boundary, in this modeling formulation the region receiving capacity is present in Eq. (1) and (2). The reason is that no control situations might significantly increase the accumulation of receiving regions and in this way the inter-transfer flows from sending regions are constrained. Note that the perimeter

---

[1] Note that $L_{IH}(t)$ is not the full trip length of vehicles, but the part of the trip in region I. Thus, trip 11-13-19-5-7 in Fig. 1, will first contribute to $L_{23}(t)$ and then to the $L_{32}(t)$ and finally to $L_{22}(t)$. In a similar manner this trip will be part of $N_{23}(t)$ accumulation while in region 2, part of $N_{32}(t)$ accumulation while in region 3 and finally part of $N_{22}(t)$ accumulation while returning to region 2 and until it finishes its trip.

[2] If vehicles are only allowed to cross the boundary between the regions at most one time (in the region-based model) then Eq. (1) and (2) are simplified to $dN_{II}(t)/dt = Q_{II}(t) - M_{II}(t) + \sum_{H \in \mathcal{V}_I} \widehat{M}_{HI}^I(t)$ and $dN_{IJ}(t)/dt = Q_{IJ}(t) - \sum_{H \in \mathcal{V}_I} \widehat{M}_{IJ}^H(t)$, $I \neq J, J = H$ as described in Ramezani et al. (2015), while $\theta_{II}^I(t)$, $\forall I \in \mathcal{R}$ and $\theta_{IJ}^H(t)$, $J = H$ in Eqs. 3 and 4 are equal to 1.

control variables are not included in the model as the control decisions here are $L_{IJ}^H(t)$ and $\theta_{IJ}^H(t)$, $I, J, H \in \mathcal{R}$, computed by the advisory route guidance system.

2.2. Subregion-based model

Consider subregion r with homogeneous distribution of congestion whose traffic performance is well described by MFD, $f_r(n_r(t))$, representing the subregion production (veh m/s) corresponding to the accumulation $n_r(t)$ (veh) at time t (s). The average subregion r speed is $v_r(t) = f_r(n_r(t))/n_r(t)$ (m/s), and trip completion rate is $m_r(t) = f_r(n_r(t))/l_r$ (veh/s), considering a constant subregional average trip length l_r (m) independent of destination or next region.

Let $n_{o,d}^{p,r}(t)$ denote the number of vehicles in subregion r at time t with origin subregion o, destination subregion d, and path p, i.e. the set containing the sequence of subregions to reach d starting from o; note that r belongs to p. Consequently $\sum_o \sum_d \sum_p n_{o,d}^{p,r}(t) = n_r(t)$.[3] Trip completion rate $m_{o,d}^{p,r}(t)$ for the vehicles in region r at time t with origin o, destination d, and path p reads:

$$m_{o,d}^{p,r}(t) = \frac{n_{o,d}^{p,r}(t)}{n_r(t)} \cdot m_r(t) = \frac{f_r(n_r(t))}{n_r(t)} \cdot \frac{n_{o,d}^{p,r}(t)}{l_r} = v_r(t) \cdot \frac{n_{o,d}^{p,r}(t)}{l_r}. \tag{6}$$

Let $q_{o,d}(t)$ denote the exogenous demand generated at time t, from origin o to destination d, and $q_{o,d}^p(t)$ be the assigned demand to path p; $\sum_p q_{o,d}^p(t) = q_{o,d}(t)$. Let $\hat{m}_{o,d}^{r \to p^+(r)}(t)$ be the transferring flow from subregion r to subregion $p^+(r)$, which is the next subregion in the sequence described by path p. Similarly, $p^-(r)$ is the previous subregion before r in path p. The subregion traffic dynamics are as follows. Note time t is omitted from the equations for the sake of notation simplicity.

$$\frac{dn_{o,d}^{p,r}}{dt} = \begin{cases} q_{o,d}^p - m_{o,d}^{p,r} & \text{(i) if } r = o \& r = d, \\ q_{o,d}^p - \hat{m}_{o,d}^{r \to p^+(r)} & \text{(ii) if } r = o \& r \neq d, \\ \hat{m}_{o,d}^{p^-(r) \to r} - m_{o,d}^{p,r} & \text{(iii) if } r \neq o \& r = d, \\ \hat{m}_{o,d}^{p^-(r) \to r} - \hat{m}_{o,d}^{r \to p^+(r)} & \text{(iv) otherwise.} \end{cases} \tag{7}$$

where

$$\hat{m}_{o,d}^{r \to p^+(r)} = \min[m_{o,d}^{p,r}, c_r^{p^+(r)}(n_{p^+(r)}) \cdot a_{o,d}^{r \to p^+(r)}]. \quad \forall r \neq d \tag{8}$$

Eq. (7) defines the rate of change in accumulation $n_{o,d}^{p,r}$ such that in case of (i) internal demand within the same subregion, the rate is simply the exogenous demand minus the trip completion rate which is not bounded by any capacity function. Note that the subregion-based model assumes that internal subregional demand never leaves the subregion; therefore, in this case the subregional path p consists of only one subregion. In case (ii) current subregion r is the origin and not the destination, then the rate is the exogenous demand minus the transfer flow to the next subregion in path p. In case (iii) current subregion r is destination and not the origin, the rate is defined as the transfer flow from the previous subregion minus the trip completion rate which is again not bounded by any capacity function. In (iv) other cases, the rate is equal to the transfer flow from the previous subregion minus the transfer flow to the next subregion.

Eq. (8) estimates transfer flow from subregion r to the next subregion $p^+(r)$ in path p for all subregions except destination subregion d. It is the minimum of two terms: (i) the sending flow from subregion r, which solely depends on the accumulation of subregion r, and (ii) the receiving capacity fraction of subregion $p^+(r)$ that is a function of two terms; $c_r^{p^+(r)}(n_{p^+(r)}(t))$ and $a_{o,d}^{r \to p^+(r)}$. Capacity at boundary between r and $p^+(r)$, i.e. $c_r^{p^+(r)}(n_{p^+(r)}(t))$, is a function of accumulation in the receiving region $n_{p^+(r)}$, while $a_{o,d}^{r \to p^+(r)}$ is the fraction of boundary capacity that is assigned to $n_{o,d}^{p,r}$, which can be estimated by Eq. (9). Conceptually speaking, Eq. (9) states that $a_{o,d}^{r \to p^+(r)}$ corresponding to (o, d, p, r) quartet depends on its relative accumulation among all traveler groups that cross the same boundary between subregion r and $p^+(r)$. This equation can be derived by Little's formula (Little, 1961).

$$a_{o,d}^{r \to p^+(r)} = \frac{n_{o,d}^{p,r}}{\sum_i \sum_j \sum_w \mathbf{1}_{p^+(r)}(w^+(r)) \cdot n_{i,j}^{w,r}} \tag{9}$$

where $\mathbf{1}_{p^+(r)}(w^+(r))$ is an indicator function with value equal to 1 if the next subregions in the paths p and w are the same, otherwise zero. Note that traffic modeling presented through Eqs. (6)–(9) is in compliance with the traffic model introduced in Yildirimoglu and Geroliminis (2014) except the constant average trip length assumption that is preserved here.

---

[3] Note that $n_r(t)$ can be estimated with good precision from different type of sensors in real time, while $n_{o,d}^{p,r}$ would require more data and is expected to exhibit larger estimation errors.

2.3. Transfer of variables from subregion-based to region-based model

This study aims at providing route guidance information in the subregion-based model (i.e. $q_{o,d}^p$; assigned flows on path p between o and d) by calculating equilibrium flows in the region-based model where actual paths are replaced with θ_{IJ}^H. This procedure requires the transfer of model parameters in certain steps. In the beginning of receding horizon, traffic states from subregion-based model should be converted to region-based model equivalents (i.e. N_{IJ}, θ_{IJ}^H, L_{IH}) and transferred to the route guidance model in order to correct the modeling error. In fact, in a real application, this step corresponds to traffic data acquisition from the network through loop detectors and probe vehicles. Similarly, equilibrium state in the region-based model has to be converted to $q_{o,d}^p$ values in order to apply them in the subregion-based model (plant). The following paragraph describes the estimation of region-based model states from subregion-based model states, while the transfer of equilibrium state from the region-based model to the subregion-based model will be further discussed in Section 3.2.

Let us consider region $I \in \mathcal{R}$ that consists of several subregions and that is heterogeneous in space with respect to link density. We use capital letters for variables related to regions and lower case letters for variables related to subregions. We denote $\mathcal{SR}$ the set of all subregions in the urban network, while $\mathcal{SR}_I$ is the set of subregions that belongs to region I. Let $\mathcal{P}^{od}$ be the set of all available paths connecting subregion o and d.

In order to define a regional path equivalent, P, of a subregional path, p, we define function $\mathbb{P}(\cdot)$. For example, if p^* is a sequence of subregions $\{12, 11, 13, 14, 15, 7\}$, then $\mathbb{P}(p^*) = P^* = \{2, 3, 2, 2\}$ (see Fig. 1, where subregions $\{12, 11\} \in \mathcal{SR}_2, \{13, 14, 15\} \in \mathcal{SR}_3$, and $\{7\} \in \mathcal{SR}_2$.) Note that the destination region along the path is always added to the end of the sequence P in order to signal the destination region. Let $p_r^{|p|}$ denote the subregion r and all the following subregions in path p. Eq. 10 represents function $\mathbb{N}(\cdot)$ that helps define the next region along the path and regroup the variables accordingly. This function checks the next region in path p given the current subregion $r \in \mathcal{SR}_I$ while ignoring the remaining part of the path.

$$\mathbb{N}\begin{pmatrix} I, H \\ p, r \end{pmatrix} = \begin{cases} 1 & \text{if } \mathbb{P}\left(p_r^{|p|}\right) = \{I, H, \ldots\}, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Eqs. (11)–(13) present the estimation of variables N_{IJ}, L_{II} and L_{IH}, respectively. Note time t is omitted for the sake of simplicity.

$$N_{IJ} = \sum_{o \in \mathcal{SR}} \sum_{d \in \mathcal{SR}_J} \sum_{p \in \mathcal{P}^{od}} \sum_{\substack{r \in \\ (p \cap \mathcal{SR}_I)}} n_{o,d}^{p,r} \tag{11}$$

$$L_{II} = \frac{\sum_{o \in \mathcal{SR}} \sum_{d \in \mathcal{SR}_I} \sum_{p \in \mathcal{P}^{od}} \sum_{r \in p} \left(\mathbb{N}\begin{pmatrix} I, I \\ p, r \end{pmatrix} \cdot n_{o,d}^{p,r} \right)}{\sum_{r \in \mathcal{SR}_I} n_r} \cdot \frac{\sum_{r \in \mathcal{SR}_I} f_r(n_r)}{\sum_{o \in \mathcal{SR}} \sum_{d \in \mathcal{SR}_I} \sum_{p \in \mathcal{P}^{od}} m_{o,d}^{p,d}} \tag{12}$$

$$L_{IH} = \frac{\sum_{o \in \mathcal{SR}} \sum_{d \in \mathcal{SR}} \sum_{p \in \mathcal{P}^{od}} \sum_{r \in p} \left(\mathbb{N}\begin{pmatrix} I, H \\ p, r \end{pmatrix} \cdot n_{o,d}^{p,r} \right)}{\sum_{r \in \mathcal{SR}_I} n_r} \cdot \frac{\sum_{r \in \mathcal{SR}_I} f_r(n_r)}{\sum_{o \in \mathcal{SR}} \sum_{d \in \mathcal{SR}} \sum_{p \in \mathcal{P}^{od}} \sum_{\substack{r \in \\ (p \cap \mathcal{SR}_I)}} \left(\mathbf{1}_{\mathcal{SR}_H}(p^+(r)) \cdot \hat{m}_{o,d}^{r \to p^+(r)} \right)} \tag{13}$$

where $\mathbf{1}_{\mathcal{SR}_H}(p^+(r))$ is the indicator function with value equal to 1 if $p^+(r) \in \mathcal{SR}_H$. In fact, Eqs. (12) and (13) are rewritten versions of Eqs. (3) and (4) with the variables at the subregion-based model. For example Eq. 12 could also be formulated as $L_{II} = \left(\theta_{II}^I(t) \cdot N_{II}\right) / N_I \cdot F_I(N_I, \sigma(N_I)) / M_{II}^I$.

Eq. 14 estimates the aggregated route choice parameter θ_{IJ}^H, which represents the ratio of vehicles with next region H among the vehicles currently in region I and with destination region J. The denominator in Eq. 14 represents the accumulation in region I and with destination region J, while the numerator depicts part of the same accumulation with next region H.

$$\theta_{IJ}^H = \frac{\sum_{o \in \mathcal{SR}} \sum_{d \in \mathcal{SR}_J} \sum_{p \in \mathcal{P}^{od}} \sum_{r \in p} \left(\mathbb{N}\begin{pmatrix} I, H \\ p, r \end{pmatrix} \cdot n_{o,d}^{p,r} \right)}{\sum_{o \in \mathcal{SR}} \sum_{d \in \mathcal{SR}_J} \sum_{p \in \mathcal{P}^{od}} \sum_{\substack{r \in \\ (p \cap \mathcal{SR}_I)}} n_{o,d}^{p,r}} \tag{14}$$

Following the same example in Fig. 1 $o^* = 12$, $d^* = 7$ and $p^* = \{12, 11, 13, 14, 15, 7\}$, let us elaborate on Eqs. (11)–(14). Accumulation values $n_{12,7}^{p^*,12}$ and $n_{12,7}^{p^*,11}$ contribute to N_{23}, L_{23} and θ_{22}^3, as the next region along the path is 3. Note that $\mathbb{N}\begin{pmatrix} 2, 3 \\ p^*, 12 \end{pmatrix}$ and $\mathbb{N}\begin{pmatrix} 2, 3 \\ p^*, 11 \end{pmatrix}$ have the value of 1. Similarly, $n_{12,7}^{p^*,13}, n_{12,7}^{p^*,14}$ and $n_{12,7}^{p^*,15}$ contribute to N_{32}, L_{32} and θ_{32}^2. Finally, $n_{12,7}^{p^*,7}$ contributes to N_{22}, L_{22} and θ_{22}^2. In addition, the flow values at the boundary/destination $\hat{m}_{12,7}^{11 \to 13}, \hat{m}_{12,7}^{15 \to 7}$ and $m_{12,7}^{p^*,7}$ are incorporated in the calculation of L_{23}, L_{32} and L_{22}, respectively.

Through Eqs. (11)–(14), we assume full observability of the state at the subregion-based model. Unfortunately, these estimations are subject to measurement errors in a real implementation. However, a decent loop detector infrastructure and sufficient amount of probe data would make these estimations possible within a tolerable error level.

3. Methodological framework

The purpose of this section is threefold: (i) to establish DUE conditions in the subregion-based model, which is considered as the base scenario, (ii) to establish DSO conditions in the subregion-based model, which enforces the users to fully cooperate at the subregion level that is the upper envelope of network performance improvement, and (iii) to provide travelers with route guidance information based on DSO conditions in the region-based model. In this study, we do not use a detailed microscopic or mesoscopic simulation model to reach equilibrium conditions. However, the framework that we employ is capable of reproducing variable trip lengths, hysteresis loops, and spatial heterogeneity. On the other hand, the aim of route guidance system is to minimize total network delay in the region-based model. However, given the heterogeneity effect in the network and different traffic characteristics in the region-based and subregion-based models, it becomes challenging to extract route guidance information, i.e. assigning vehicles to subregional paths based on the conditions at the region-based model without detailed information about the subregions. This study also establishes DSO conditions in the subregion-based model to emphasize the cost of using a more aggregated model to develop route guidance information. Note that the model where we estimate operation decisions (e.g. route guidance) and the plant where we test the decisions should be different. Therefore, a route guidance system based on microscopic or mesoscopic simulation models, as well, is exposed to the same modeling error. Yildirimoglu and Geroliminis (2014) showed through micro-simulation that subregion dynamics can represent quite accurately the evolution of congestion even if the link dynamics of the urban network are not known.

Note that the current paper extends the work of Yildirimoglu and Geroliminis (2014) to a route guidance scheme that is based on DSO state. This reference makes the path-level route decisions on a conventional link-based graph (i.e. actual roads and intersections), produces aggregate route choice parameters, and establishes DUE conditions in the aggregated traffic model. However, it does not consider the interaction between the operation model (i.e. the region-based model) and the plant (i.e. the subregion-based model), and does not employ a rolling horizon procedure for the route guidance. In addition, the traffic model in Yildirimoglu and Geroliminis (2014) accounts for the route choice behavior by using explicit regional paths, while the region-based model in this study employs θ_{IJ}^{H} values.

This section consists of two subsections. The next subsection presents the methodology to establish DUE and DSO conditions in the subregion-based model, while the following introduces the route guidance strategy based on DSO state in the region-based model.

3.1. Dynamic traffic assignment (DTA) in the subregion-based model

This subsection tackles the problem of establishing DUE or DSO conditions at the subregional level, as expressed by time dependent subregional route choice estimation. The urban network is divided into subregions with low-scatter MFDs and constant average trip lengths, i.e. l_r. We assume that the subregional O-D demand is given.

Traffic equilibrium (DUE or DSO) can be formulated as a fixed-point problem, where an additional cycle of assignment and network loading steps yield the same traffic conditions. A well-known heuristic solution, method of successive averages (MSA), is a suitable method in our study considering the characteristics of the problem in hand. MSA has been used in both static and dynamic network equilibrium problems as an incremental assignment type heuristic (Daganzo and Sheffi, 1977; Mahmassani and Peeta, 1993). The method is based on predetermined step sizes along the descent direction. In other words, step size is not determined with respect to the characteristics of the current solution, which requires derivative information. Instead, it is determined a priori. Therefore, MSA stands as one of the most effective solution heuristics in case the derivative information is difficult to be acquired.

As DUE implementation requires equal and minimal experienced travel times on alternative paths at the same departure time, $T_{od}^{p}(t)$, the iterative traffic assignment procedure is performed based on time-dependent fastest path (Chabini, 1998). On the other hand, DSO formulation yields equal and minimal *marginal* travel times, $\tau_{od}^{p}(t)$, on the alternative paths. Time-dependent path marginal travel time represents the effect of one additional vehicle using the same path on the system travel time. Hence, DSO conditions can be established by assigning vehicles to the paths with minimum time-dependent marginal costs (Peeta and Mahmassani, 1995; Peeta and Ziliaskopoulos, 2001). To identify these paths, subregion marginal travel times, $\tau_r(t)$, are computed and a time-dependent minimum cost algorithm (Chabini, 1998) is deployed. Marginal travel time in subregion r is estimated as follows (time t is omitted for the sake of simplicity):

$$\tau_r = T_r + n_r \cdot \frac{dT_r}{dn_r} = \frac{l_r \cdot n_r}{f_r(n_r)} + n_r \cdot \frac{d}{dn_r} \frac{l_r \cdot n_r}{f_r(n_r)} = l_r \cdot n_r \left(\frac{2}{f_r(n_r)} - n_r \cdot \frac{df_r(n_r)/dn_r}{[f_r(n_r)]^2} \right) \tag{15}$$
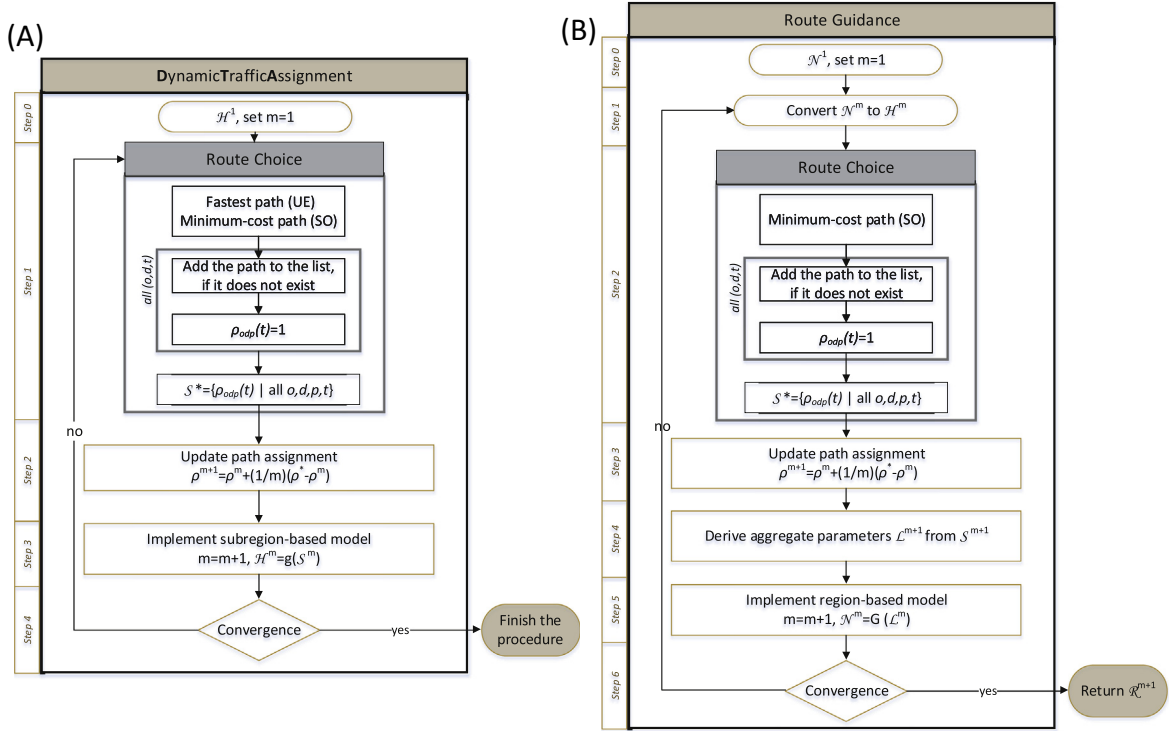
where T_r is the actual travel time of subregion r.

Fig. 2. (a) DTA in the subregion-based model, (b) route guidance system in the region-based model.

Let $\mathcal{H} = \{n_r(t)|\forall r, t\}$ and $\mathcal{S} = \{\rho_{odp}(t)| \forall o, d, p, t\}$, where $\rho_{odp}(t)$ denotes the ratio of the demand from origin o to destination d that chooses path p at time t (i.e. $q^p_{o,d}(t)/q_{o,d}(t)$). The flowchart to establish DUE/DSO state is presented in Fig. 2a, while the corresponding algorithm is summarized as follows:

Step 0: *Initialization*
- Set $m = 1$.
- Set $\mathcal{P}^{od} = \{\}$ $\forall o, d$.
- Initialize subregion accumulations $\mathcal{H}^1$. The initialization step can reflect an empty network.

Step 1: *Direction finding*
- For each destination subregion d, apply all-to-one time-dependent fastest path or time-dependent minimum cost path algorithm (Chabini, 1998) depending on DUE or DSO implementation, respectively.
- To find the auxiliary paths in DUE and DSO formulations, employ actual travel times $\{T_r(t)| \ \forall r, t\}$ and marginal travel times $\{\tau_r(t)|\forall r, t\}$, respectively. Note that they both result from the accumulation set in the current iteration; i.e. $\mathcal{H}^m$.
- Perform an all-or-nothing assignment for each (o, d, t) triplet, and assign $(\rho_{odp}(t))^* = 1$ for the chosen auxiliary path p.
- If the set $\mathcal{P}^{od}$ does not include the path p, add it to the set.

Step 2: *Calculating path assignment ratios*
- Set $(\rho_{odp}(t))^{m+1} = (\rho_{odp}(t))^m + (1/m) \cdot \left((\rho_{odp}(t))^* - (\rho_{odp}(t))^m \right)$ $\forall o, d, p, t$.

Step 3: *Update*
- Set $m = m + 1$.
- Create the set of path assignment ratios $\mathcal{S}^m = \{(\rho_{odp}(t))^m| \ \forall o, d, p, t\}$.
- Implement the subregion-based model, i.e. Eqs. (6)–(9), and get the new accumulations: $\mathcal{H}^m = g(\mathcal{S}^m)$. Note that $g(\cdot)$ represents the subregion-based model dynamics.

Step 4: *Stopping criteria*
- Evaluate: $M = \sum_r \sum_t \left[(n_r(t))^m - (n_r(t))^{m-1} \right]^2$.
- If $M \geqslant \epsilon$, where ϵ is a pre-defined threshold, go to Step 1. Otherwise, terminate the procedure.

3.2. Route guidance

The problem in this subsection is to establish DSO state in the region-based model and to provide travelers with route guidance information that satisfies the same conditions at the subregion-based model. This study produces the route guidance information through a rolling horizon framework, meaning that at each time step the method establishes DSO state in the operation model, i.e. the region-based model, over a rolling horizon and identifies a sequence of route choice decisions. However, only the first step of route choice decisions is applied in the plant, i.e. the subregion-based model. The procedure is then repeated with a shifted horizon in the next time step. This closed-loop system allows the route guidance strategy, through a feedback loop, to correct disturbances due to variations in the predicted demand and the errors between the plant and the operation model.

The region-based model employs time-dependent average trip lengths, i.e. $L_{II}(t)$ and $L_{IH}(t)$, aggregated route choice parameters, i.e. $\theta_{IJ}^H(t)$, and an MFD affected by the heterogeneity level in the network, while the subregion-based model utilizes constant average trip lengths, i.e. l_r, and path assignment ratios, i.e. $\rho_{odp}(t)$. Therefore, a route guidance strategy based on the region-based model requires the conversion of parameters to the subregion-based model. This constitutes a significant part of the route guidance system.

Let $\mathcal{N} = \{N_I(t) | \forall I, t\}$ and $\mathcal{L} = \{L_{IH}(t), \theta_{IJ}^H(t) | \quad \forall I, J, H, t\}$. Note that, in the rolling horizon framework, range of t is limited by the length of the horizon. The route guidance flowchart is presented in Fig. 2b, while the corresponding algorithm is summarized as follows:

Step 0: *Initialization*
- Set $m = 1$.
- Apply Eqs. (11)–(14) to provide the operation model with required variables (i.e. $N_{IJ}(t_0)$, $L_{IH}(t_0)$, $\theta_{IJ}^H(t_0)$). Note that Eqs. (11)–(14) return the variables that reflect the state at the plant only in the beginning of rolling horizon; i.e. t_0. In a real implementation, this step corresponds to traffic data acquisition from the network.
- Assume $N_{IJ}(t) = N_{IJ}(t_0), L_{IH}(t) = L_{IH}(t_0)$ and $\theta_{IJ}^H(t) = \theta_{IJ}^H(t_0)$ for $t > t_0$ and $\forall I, J, H$.
- At $t = t_0$, observe subregion accumulations in the plant, and calculate the relative accumulations for the subregions that belong to the same region, $\phi_r^I = n_r(t_0)/\sum_{i \in \mathcal{SR}_I} n_i(t_0); \quad \forall r \in \mathcal{SR}_I$.

Step 1: *Conversion of region accumulations*
- Assume relative accumulation set, ϕ_r^I, is constant over the rolling horizon, and convert regional accumulation set, $\mathcal{N}^m$, to subregional accumulation set, $\mathcal{H}^m$, i.e. $n_r(t) = N_I(t) \cdot \phi_r^I$ for $\forall t > t_0$ & $r \in \mathcal{SR}_I$. The region-based model in hand is able to evaluate regional accumulations. However, routing in the subregion-based model requires also the information on how the congestion is distributed within the region. This assumption allows us to account for the accumulation heterogeneity within the urban regions.

Step 2: *Direction finding*
- Apply Step 1 from DTA algorithm. Note that subregion accumulations, $\mathcal{H}^m$, deployed in this step are not exact as in DTA algorithm which is based on subregion-based model output. They are subject to the modeling error due to the granularity difference between the region-based and the subregion-based models and to the constant ratio assumption described above.

Step 3: *Calculating path assignment ratios*
- Apply Step 2 from DTA algorithm.

Step 4: *Producing aggregate route choice parameters*
- In order to update regional accumulations, one needs to implement region-based model dynamics with the updated route choice variables (i.e. $\mathcal{L}^m$). However, the previous step here provides the route choice decisions at the subregional level. In fact, there are two ways to estimate these aggregate route choice variables. The first one is done through observing and regrouping the state at the plant, as in Eqs. (12)–(14). This approach is exploited in Step 0 to complete the feedback loop and to correct the disturbances. However, subregion-based model is not available while we establish the equilibrium flows in the region-based model or while we produce route choice decisions based on predicted traffic state at the region-based model. The second approach, which we employ here, is a trajectory-based estimation of route choice variables, and it allows us to account for the impact of new path decisions. Assuming properties of static traffic models (e.g. departing flow immediately reaches its destination, and it exists simultaneously in all the links along its assigned path) apply here, this approach provides a practical estimation of the aggregate route choice variables.
- Denote $\widetilde{L}_{II}$, $\widetilde{L}_{IH}$, $\tilde{\theta}_{IJ}^H$ the aggregate route choice parameters only for the departing flow in the corresponding time period. In fact, they represent marginal effects of the newly assigned demand on L_{II}, L_{IH}, and θ_{IJ}^H. Therefore, a weighted average of them is later taken to compute the new aggregated route choice parameters. In compliance with regional path P, let us denote P_D the trip distance to cross in each region. Following the path example presented in Fig. 1, $p^* = \{12, 11, 13, 14, 15, 7\}$, regional path $P^* = \{2, 3, 2, 2\}$ and assuming same constant average

trip length (i.e. l_r) in each subregion, P_D^* is $\{2 \cdot l_r, 3 \cdot l_r, l_r\}$. (16)–(18) provide trajectory-based estimation of aggregate route choice variables only for the departing flow. Note time t is omitted, P^{end} represents the last/destination region in sequence P, and $P^+(R)$ is the next region after R in path sequence P.

$$\widetilde{L}_{II} = \frac{\sum_{o \in SR}\sum_{d \in SR_I}\sum_{p \in \mathcal{P}^{od}}\sum_{\substack{R \in (I \cap P \backslash P^{end}), \\ P = \mathbb{P}(p)}}\left(\mathbf{1}_I(P^+(R)) \cdot q_{o,d} \cdot \rho_{odp} \cdot P_D(R)\right)}{\sum_{o \in SR}\sum_{d \in SR_I}\sum_{p \in \mathcal{P}^{od}}\sum_{\substack{R \in (I \cap P \backslash P^{end}), \\ P = \mathbb{P}(p)}}\left(\mathbf{1}_I(P^+(R)) \cdot q_{o,d} \cdot \rho_{odp}\right)} \tag{16}$$

$$\widetilde{L}_{IH} = \frac{\sum_{o \in SR}\sum_{d \in SR}\sum_{p \in \mathcal{P}^{od}}\sum_{\substack{R \in (I \cap P \backslash P^{end}), \\ P = \mathbb{P}(p)}}\left(\mathbf{1}_H(P^+(R)) \cdot q_{o,d} \cdot \rho_{odp} \cdot P_D(R)\right)}{\sum_{o \in SR}\sum_{d \in SR}\sum_{p \in \mathcal{P}^{od}}\sum_{\substack{R \in (I \cap P \backslash P^{end}), \\ P = \mathbb{P}(p)}}\left(\mathbf{1}_H(P^+(R)) \cdot q_{o,d} \cdot \rho_{odp}\right)} \tag{17}$$

$$\tilde{\theta}_{IJ}^H = \frac{\sum_{o \in SR}\sum_{d \in SR_J}\sum_{p \in \mathcal{P}^{od}}\sum_{\substack{R \in (I \cap P \backslash P^{end}), \\ P = \mathbb{P}(p)}}\left(\mathbf{1}_H(P^+(R)) \cdot q_{o,d} \cdot \rho_{odp}\right)}{\sum_{o \in SR}\sum_{d \in SR_J}\sum_{p \in \mathcal{P}^{od}}\sum_{\substack{R \in (I \cap P), \\ P = \mathbb{P}(p)}}\left(q_{o,d} \cdot \rho_{odp}\right)} \tag{18}$$

- As opposed to Eqs. (12)–(14), Eqs. (16)–(18) make use of converted regional path P in order to produce corresponding regional parameters. In addition, while Eqs. (12)–(14) employ accumulation and outflow values observed through the plant, Eqs. (16)–(18) incorporates only the departing flow on various paths (i.e. $q_{o,d} \cdot \rho_{odp}$) and the distance on the assigned trajectories (i.e. $P_D(R)$). Again using the same example $p^* = \{12, 11, 13, 14, 15, 7\}$, assigned path flow at a departure time $q_{12,7} \cdot \rho_{12,7,p^*}$ simultaneously (i.e. at the very same departure time) contributes to the calculation of $\{\widetilde{L}_{23}, \tilde{\theta}_{22}^3\}$, $\{\widetilde{L}_{32}, \tilde{\theta}_{32}^2\}$, and $\{\widetilde{L}_{22}, \tilde{\theta}_{22}^2\}$ considering the regions $R \in \{2, 3, 2\}$, respectively, on path $P^* \backslash P^{*end}$. Note that the contribution to the trip length in each region is weighted with respect to the distance crossed inside the same region (i.e. P_D^*).

- Denote $\widetilde{N}_{IJ}$ the *instantaneous* accumulation in region I with destination region J only due to the departing flow in the corresponding time period. Note that it also appears in the denominator of Eq. (18).

$$\widetilde{N}_{IJ} = \sum_{o \in SR}\sum_{d \in SR_J}\sum_{p \in \mathcal{P}^{od}}\sum_{\substack{R \in (I \cap P), \\ P = \mathbb{P}(p)}}\left(q_{o,d} \cdot \rho_{odp}\right) \tag{19}$$

- As $\widetilde{L}_{II}$, $\widetilde{L}_{IH}$, and $\tilde{\theta}_{IJ}^H$ represent the marginal route choice parameters only for the departing flow, compute the weighted average of accumulation in the network and departing flow. This allows us to combine the effect of departing flow (i.e. $\widetilde{N}_{IJ}$) and existing accumulation (i.e. N_{IJ}) in the network.

$$(L_{II})^{m+1} = \frac{(N_{II})^m \cdot (L_{II})^m + \widetilde{N}_{II} \cdot \widetilde{L}_{II}}{(N_{II})^m + \widetilde{N}_{II}} \tag{20}$$

$$(L_{IH})^{m+1} = \frac{\sum_J\left((N_{IJ})^m \cdot (\theta_{IJ}^H)^m \cdot (L_{IH})^m\right) + \sum_J\left(\tilde{N}_{IJ} \cdot \tilde{\theta}_{IJ}^H \cdot \tilde{L}_{IH}\right)}{\sum_J\left((N_{IJ})^m \cdot (\theta_{IJ}^H)^m\right) + \sum_J\left(\tilde{N}_{IJ} \cdot \tilde{\theta}_{IJ}^H\right)} \tag{21}$$

$$(\theta_{IJ}^H)^{m+1} = \frac{(N_{IJ})^m \cdot (\theta_{IJ}^H)^m + \widetilde{N}_{IJ} \cdot \tilde{\theta}_{IJ}^H}{(N_{IJ})^m + \widetilde{N}_{IJ}} \tag{22}$$

Step 5: *Update*
- Set $m = m + 1$.
- Create $\mathcal{N}^m$.
- Implement region-based model, Eqs. (1)–(5), and get the new accumulations: $\mathcal{N}^m = \mathcal{G}(\mathcal{L}^m)$. Note that $\mathcal{G}(\cdot)$ represents the region-based model dynamics.

Step 6: *Stopping test*
- Evaluate: $M = \sum_I\sum_t\left[(N_I(t))^m - (N_I(t))^{m-1}\right]^2$.
- If $M \geqslant \epsilon$, where ϵ is a pre-defined threshold, go to Step 1. Otherwise, terminate the procedure.

Note that the convergence of MSA is not monotonic. This is because of random search direction (auxiliary values produced by stochastic network loading may sometimes point in a direction where objective function increases) and the fixed step size (predetermined step size, $\alpha_m = 1/m$, may overshoot the reduction in the objective function, as it incorporates no information related to the optimal solution neighborhood). In addition, one can claim that convergence criterion used in

MSA is forced to converge due to the nature of step size sequence $\{\alpha_m\}$. However, practical experience indicates reasonable convergence speed and existence of stable solution, before it is forced by the step size sequence.

4. Results

This section presents several numerical studies with various demand levels spanning from uncongested scenarios to gridlock conditions. We test the proposed route guidance strategy under these different traffic conditions. In addition, we investigate the consistency between two traffic models and discuss the impact of DUE and DSO flows on the network MFD.

The case study network consists of three regions, where region 1 and 2 designate the suburb of the urban network, each comprising of 6 subregions, and region 3 designates the city center of the urban network comprising of 7 subregions, as schematically shown in Fig. 1. Without loss of generality, we assume every subregion has the same production MFD and subregional average trip length. Note that the regional average trip lengths are time-varying, see Eqs. (12) and (13). The exogenous time-varying demand simulates one hour of morning peak followed by another hour of very low demand to fully clear the network. Region 1 and 2 generate most of the demand towards region 3 that, as the central business district, attracts most of the trips.

One of the main contributions of the paper is introducing two models with different levels of aggregation both in terms of vehicle accumulation state and route choice behavior. It is important that, despite their different granularity, the models can reproduce consistent description of aggregated traffic variables, that is necessary for real time traffic management implementation. To scrutinize the consistency of regional and subregional models, we test the models with similar exogenous demands and initial accumulations, i.e. $Q_{IJ}(t)$ and $q_{o,d}(t)$, and $N_{IJ}(0)$ and $n_{o,d}^{p,r}(0)$ are consistent. In addition, the only traffic state measurements needed for the region-based model are the regional average trip lengths, i.e. L_{II} and L_{IH}, and the aggregated route choice parameters, θ_{IJ}^H, that are available using Eqs. (12)–(14). Fig. 3(a) and (b) illustrate the evolution of $N_{IJ}(t)$ and $N_I(t) = \sum_J N_{IJ}(t)$ as estimated by the region-based ((1)–(5)) and subregion-based models ((6)–(8)), which indicate that the two models are consistent. In addition, Fig. 3(c) illustrates the results of the same test on a region-based model (as described in Ramezani et al. (2015) and FootNote 2 of the current paper) where vehicle routes include at most one boundary crossing. The inconsistency here exposes the significant effect of such modeling consideration when vehicles are allowed to cross more than once a boundary between two regions (in the aforementioned publication route choice was constrained to be consistent with this modeling assumption).

Fig. 4 demonstrates the existence of DUE and DSO traffic conditions through the heuristic method described in Section 3.1. Fig. 4(a) shows experienced travel times of two paths (4–16–19 and 4–17–19), along with the associated path
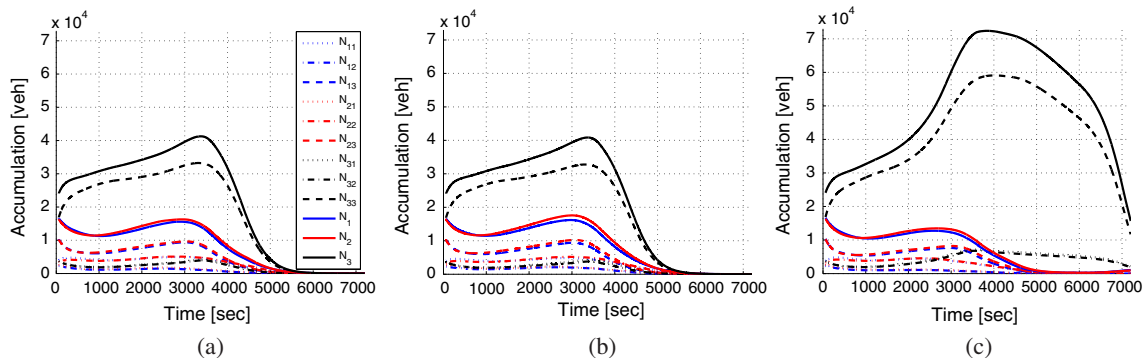

Fig. 3. Models consistency validation: (a) region-based model (operation model), (b) subregion-based model (plant), and (c) region-based model with assumption that there is at most one boundary crossing.
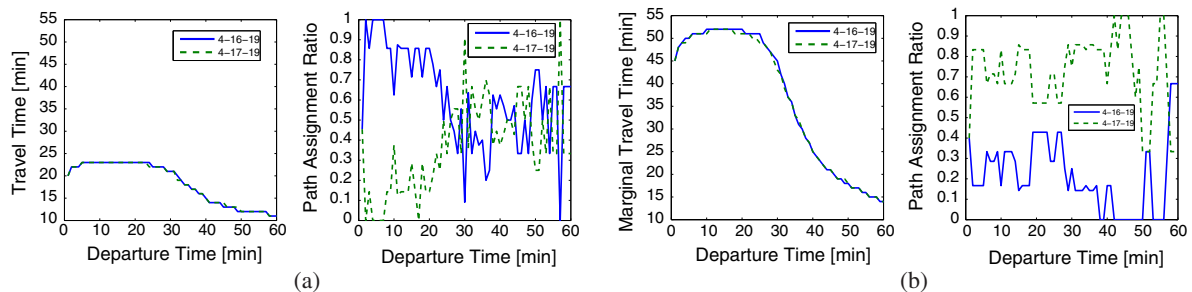

Fig. 4. Travel times and path assignment ratios for path 4–16–19 and path 4–17–19: (a) actual (experienced) travel time in case of DUE, (b) marginal (experienced) travel time in case of DSO.

Table 1
Total network delay (10^8 (veh s)).

	Demand Level			
	95%	100%	105%	110%
Dynamic user equilibrium (DUE)	2.31	2.77	2.95	4.01
Dynamic system optimum (DSO)	2.07 (10.4%)	2.39 (13.7%)	2.55 (13.6%)	3.31(17.5%)
Route guidance (RG)	2.07 (10.4%)	2.42 (12.6%)	2.65 (10.2%)	4.28(−6.7%)

assignment ratios. This specific case presents the traffic conditions corresponding to the DUE traffic assignment procedure (at the subregional level). Under DUE conditions, travelers tend to use the paths that provide lowest travel time among the alternatives. This results in minimal and equal travel times on the used alternative paths, apparent in Fig. 4(a). On the other hand, Fig. 4(b) depicts marginal travel times of the two paths and the associated path assignment ratios under DSO conditions. Similarly, the marginal travel times among the alternative paths are minimal and equal. Notably, the path assignment ratios in case of DUE and DSO are different. Note that Fig. 4 spans only first one hour of the simulation where the exogenous demand exists in the scenario, as the route choice decision is being made at the departure of the trip.

Table 1 lists the total network delay in three assignment scenarios for various demand levels. In all demand levels, as expected, DSO performs better than DUE. The benefit from this cooperative assignment strategy ranges from 10.4% to 17.5%. RG strategy produces results fairly close to the DSO conditions except 110% demand scenario. This is an important finding as the developed strategy proves its applicability in field by giving information to drivers what parts of the city to avoid. This strategy can also be integrated in a hierarchical traffic management framework as the first layer, where a more detailed second layer module could assign link-based routes to the drivers. Given the fast computation of RG strategy in this paper, such an approach could also ease the (sometimes) infeasible computational effort of detailed RG algorithms. The performance difference between DSO and RG scenarios can be represented as the cost of using a more aggregated model (i.e. the region-based model) to produce routing information for a relatively detailed model (i.e. the subregion-based model). The reason for the failure in 110% demand level might be the inconsistency between subregion-based and region-based model for the accumulation values close to the jam accumulation. Although subregion-based model does not allow values beyond jam accumulation, such a constraint does not exist in region-based model. Over the iterations in the route guidance model, inaccurate route choice estimations are done probable due to this inconsistency. In cases of very high demand (many sub-regions reach gridlock), a combination of route guidance with perimeter control is expected to produce better results.

Fig. 5 illustrates the MFD functions (production vs. accumulation) for different demand levels under three assignment scenarios. For all demand levels and DUE scenarios, region 3 experiences congestion (with states in the decreasing part of the MFD). 95%, 100% and 105% demand levels produce no significant hysteresis in any region with any of the assignment scenarios. On the other hand, 110% demand level creates extremely congested and almost gridlock conditions in many of the subregions in DUE and RG scenarios, which destroys the homogeneity of the regions, and causes hysteresis in the network unloading phase (see Fig. 5(j) and (l)). Even with 110% demand level, DSO is able to prevent gridlock in subregions and avoid hysteresis in region 3. Despite the fact that regions 1 and 2 are much less congested than region 3, they exhibit a significant hysteresis in this scenario (see Fig. 5(k)). This issue will be further investigated in the paper. Note that gridlock conditions are expected to be observed only during rare events (e.g. evacuation), and the region-based model is able to reproduce accurate MFD patterns and follow hysteresis loops (when exist) despite the lack of information for detailed route choices of drivers. Further research is required to investigate with real data the route choices of drivers at the sub-region level that will shed more light in proper assignment tools for simulation and modeling.

Fig. 6 illustrates subregional and regional accumulations for the three traffic assignment scenarios, i.e. DUE, DSO and RG strategy for 100% demand level. Notably, with DUE and DSO conditions, subregion accumulations are distributed in a harmonious way; they follow the same trend in the network loading and unloading phases. Although RG strategy is able to follow the same trend in region accumulations (see Fig. 6(e) and (f)), the evolution of subregion accumulations shows differences due to limitations between region-based and subregion-based models. Despite this limitation, RG strategy produces fairly close results to DSO (see Table 1). Note that odd-numbered subregions at the periphery (see Fig. 1) are connected to one single subregion from the inner core, while even-numbered subregions are connected with two of them. This largely affects the inflow and outflow characteristics of subregions, and causes inhomogeneity within the regions. The mild hysteresis in regions 1 and 2 presented in Fig. 5(d)–(i) is due to this hexagonal network structure and difference in accumulations between odd- and even-numbered regions (see Fig. 6(a)–(c)). This result indicates that equilibrium type assignments cannot establish homogeneity within the regions by itself; physical network structure and connectivity between different parts of the network play a significant role as well. This should be an interesting research priority.

Fig. 7 shows the subregion accumulations to visualize the spatial correlation structure and propagation of congestion in the network for demand level 100% and the DUE, DSO, and RG assignments. In this case study, traffic congestion mainly starts from the central subregion 19 and propagates in the remaining subregions of region 3, before it partially captures the subregions in the periphery (with less magnitude). In DUE conditions during the peak period, $t = 60$ (min), most of the sub-regions in region 3 are highly congested while regions 1 and 2 are uncongested, whereas in DSO conditions some drivers choose longer routes avoiding the center of congestion and thus, accumulations spread more homogeneously in
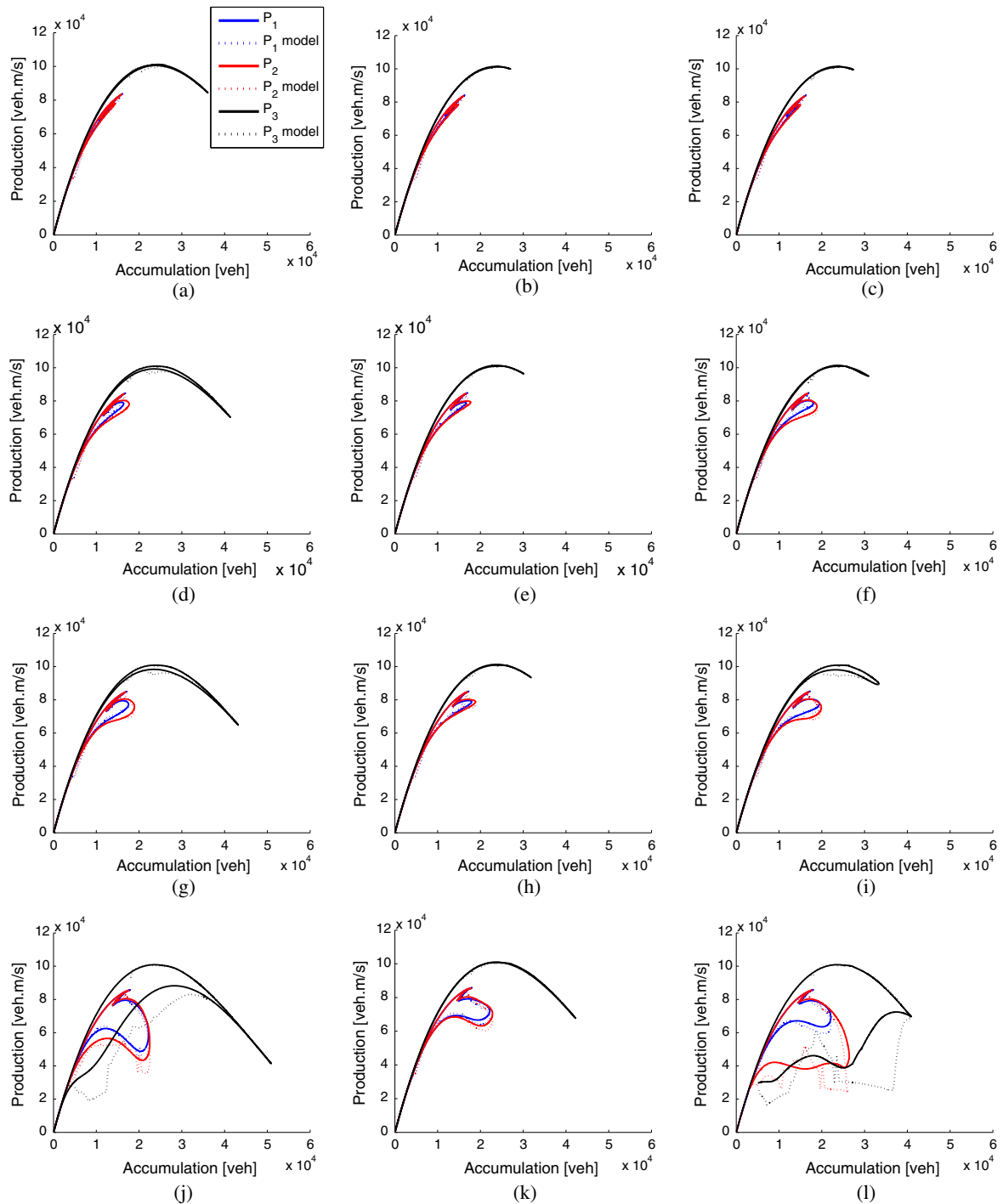
Fig. 5. MFDs for the 3 regions of Fig. 1, estimated with the regional and sub-regional models for different demand levels (rows) and different routing strategies (columns). (a) DUE 95%, (b) DSO 95%, (c) RG 95%, (d) DUE 100%, (e) DSO 100%, (f) RG 100%, (g) DUE 105%, (h) DSO 105%, (i) RG 105%, (j) DUE 110%, (k) DSO 110%, (l) RG 110%.

all the 3 regions. Note that small changes in regional route choices can drastically alter the level of congestion for the whole city. Such an implementation of route guidance could be applied without great effort and driver compliance might not be a significant issue as the deviation from DUE conditions is small.

Fig. 8 depicts the results from the "current best" or "en-route" assignment strategy with 100% demand level. In this strategy, drivers update their routes based on instantaneous travel times in every time interval. We notice that hysteresis
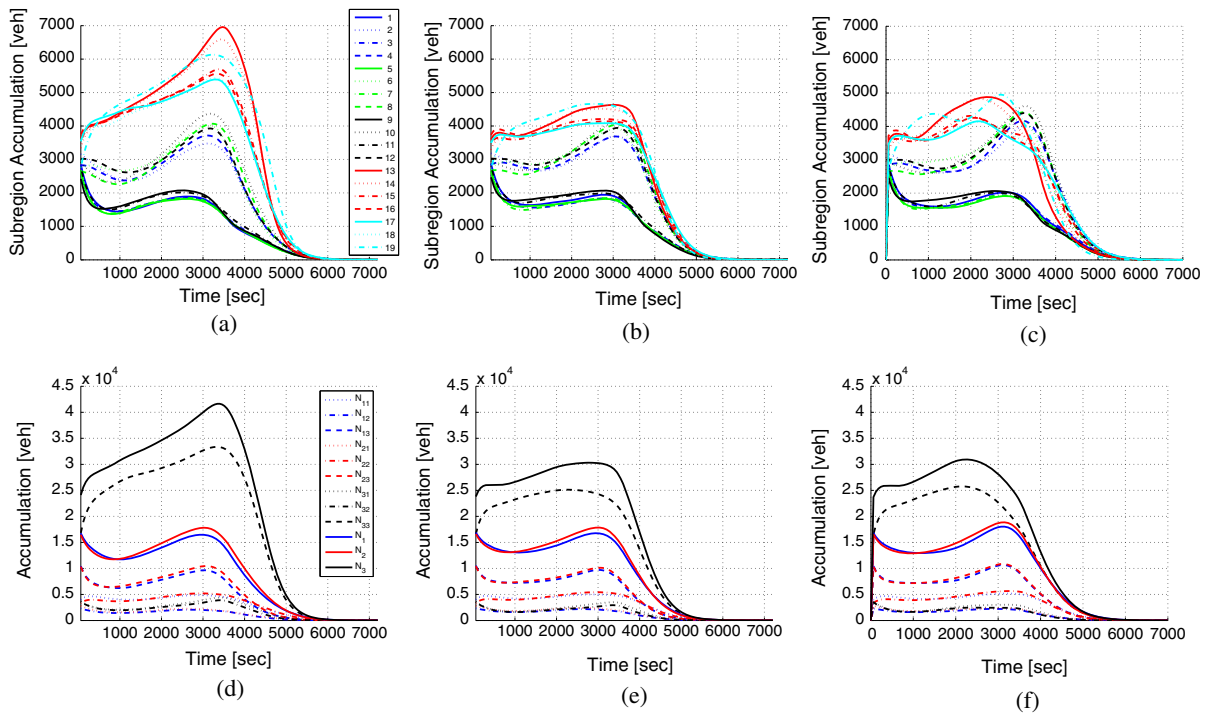
Fig. 6. 100% demand, subregional accumulations: (a) DUE, (b) DSO, and (c) RG. Regional accumulations: (d) DUE, (e) DSO, and (f) RG.
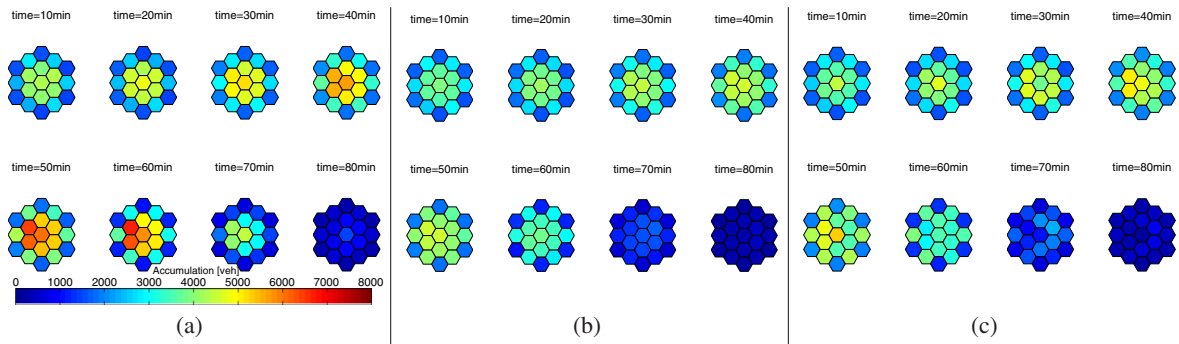

Fig. 7. 100% demand, subregion accumulations evolution over time. (a) DUE, (b) DSO, (c) RG.

phenomenon in MFD is quite significant here with respect to equilibrium scenarios (e.g. DUE, DSO or RG) investigated before (see Fig. 5(d)–(f)). Although there is no subregion that suffers from gridlock conditions (see Fig. 8(b)), especially the accumulation in region 3 is not properly distributed among subregions. This increases the inhomogeneity within region 3 and causes a significant hysteresis in the MFD shape (see Fig. 8(a)). MFDs for region 1 and 2 exhibit hysteresis too, but this is mainly due to the physical network structure discussed above. Note that all previous simulation-based MFD estimations rely on enroute assignment mechanisms and instantaneous travel times. Although en-route decision mechanisms are able to properly distribute the congestion in the network, it hardly represents the real driver behavior. Drivers might switch to the alternative routes in response to unexpected traffic conditions. However, in a "normal" day they are expected to rely on their past experiences. These results clearly indicate that the way route choice behavior is modeled has significant impacts on network-wide properties such as MFD functions. This is in accordance with micro-simulation findings in Mahmassani et al. (2013b) that more informed and adaptive drivers can improve the network performance.

DSO conditions, in principle, are difficult to be put into practice, because some drivers might experience higher travel costs than they do under DUE conditions. Nevertheless, as we show now, the developed regional route guidance strategy does not penalize a significant portion of travelers and has some strong potential for a field implementation. Fig. 9 depicts the distribution of travel time benefit across users for DSO and RG scenarios with respect to DUE conditions. Note that ~89%
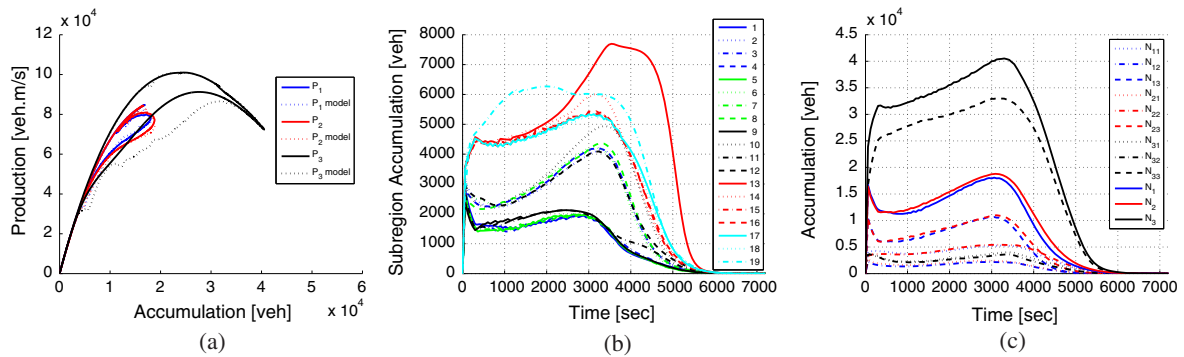
Fig. 8. En-route assignment with 100% demand: (a) MFD, (b) subregion accumulations, and (c) region accumulations.
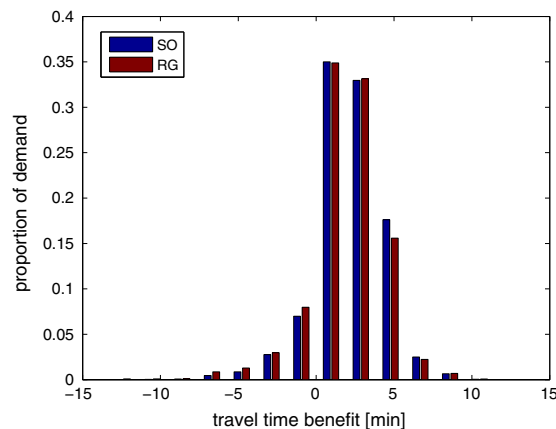

Fig. 9. Travel time benefit of users in DSO and RG scenarios with respect to DUE conditions.

and ∼87% of users benefit from DSO and RG, respectively. In addition, only ∼0.6% of users suffer from more than 5 min delay in both DSO and RG scenarios. These results indicate that a RG scenario based on DSO conditions can bring an overall benefit to the system in the cost of causing little delay to a low percentage of drivers. In fact, Jahn et al. (2005) reaches near-DSO state by integrating explicit user constraints that guarantee certain level of fairness across users. A similar approach could be used here to prevent delays more than a certain amount. This should be a future research priority.

5. Conclusion and future research

This paper has presented two MFD-based traffic models with different levels of vehicle accumulation state and route choice behavior aggregation. This configuration enables us to incorporate the effect of spatial congestion heterogeneity and route choice dynamics in MFD modeling and to investigate the consequence of limited traffic state measurements measurements in performance of RG strategy. We have integrated DUE and DSO flows in the analysis. The results clearly demonstrate that equilibrium state has a significant impact on the existence of hysteresis in MFD. Previous MFD estimations based on microscopic and mesoscopic simulation models where route choice behavior is modeled with en-route decision mechanisms suffer from hysteresis even in mildly congested scenarios. However, in this study, we observe that equilibrium flows create a non-hysteresis MFD except in very high demand scenarios where gridlock is not avoidable. In addition, a route guidance control system based on the parsimonious region-based model is developed, where its route guidance commands are applied in the subregion-based model. The results demonstrate that the advisory route guidance system produces conditions fairly close to DSO state, except very high demand scenarios where the consistency between the model and the plant should be further investigated.

For future research, the route guidance system can be integrated with the perimeter control strategy, which is expected to further improve homogeneity and network performance. This is challenging because the perimeter control decisions are optimized based on the region-based model, while the control decisions should be applied in the subregion-based model, and designing a realistic and accurate information feedback from the plant to the optimization or operation model needs further investigations. The interaction between the adaptive control and equilibrium state is an important issue to be addressed in future works. This problem could be overcome in a day-to-day assignment framework, where people adapt

to unexpected traffic conditions (created by the new control strategy) by taking different route decisions over days in addition to en-route decision mechanisms. Furthermore, a field test would provide more insights about the applicability and implications of the proposed control strategy.

Acknowledgements

References

Aboudolas, K., Geroliminis, N., . Perimeter and boundary flow control in multi-reservoir heterogeneous networks. Transport. Res. Part B: Methodol. 55, 265–281.

Ben-Akiva, M., Bierlaire, M., Bottom, J., Koutsopoulos, H., Mishalani, R., 1997. Development of a route guidance generation system for real-time application. In: Proceedings of the IFAC Transportation Systems Conference, Chania.

Ben-Elia, E., Di Pace, R., Bifulco, G.N., Shiftan, Y., 2013. The impact of travel information?s accuracy on route-choice. Transport. Res. Part C: Emerg. Technol. 26, 146–159.

Buisson, C., Ladier, C., 2009. Exploring the impact of homogeneity of traffic measurements on the existence of macroscopic fundamental diagrams. Transport. Res. Rec.: J. Transport. Res. Board 2124 (−1), 127–136.

Chabini, I., 1998. Discrete dynamic shortest path problems in transportation applications: complexity and algorithms with optimal run time. Transport. Res. Rec.: J. Transport. Res. Board 1 (1), 170–175.

Daganzo, C.F., 1994. The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Transport. Res. Part B: Methodol. 28 (4), 269–287.

Daganzo, C.F., 2007. Urban gridlock: macroscopic modeling and mitigation approaches. Transport. Res. Part B: Methodol. 41 (1), 49–62.

Daganzo, C.F., Gayah, V.V., Gonzales, E.J., 2011. Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability. Transport. Res. Part B: Methodol. 45 (1), 278–288.

Daganzo, C.F., Sheffi, Y., 1977. On stochastic models of traffic assignment. Transport. Sci. 11 (3), 253–274.

Doig, J.C., Gayah, V.V., Cassidy, M.J., 2013. Inhomogeneous flow patterns in undersaturated road networks. Transport. Res. Rec.: J. Transport. Res. Board 2390 (1), 68–75.

Gayah, V., Dixit, V., Guler, S., 2013. Relationship between mean and day-to-day variation in travel time in urban networks. EURO J. Transport. Log., 1–17

Gayah, V.V., Daganzo, C.F., 2011. Clockwise hysteresis loops in the macroscopic fundamental diagram: an effect of network instability. Transport. Res. Part B: Methodol. 45 (4), 643–655.

Gayah, V.V., Dixit, V.V., 2013. Using mobile probe data and the macroscopic fundamental diagram to estimate network densities. Transport. Res. Rec.: J. Transport. Res. Board 2390 (1), 76–86.

Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transport. Res. Part B: Methodol. 42 (9), 759–770.

Geroliminis, N., Haddad, J., Ramezani, M., 2013. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: a model predictive approach. IEEE Trans. Intell. Transport. Syst. 14 (1), 348–359.

Geroliminis, N., Sun, J., 2011. Properties of a well-defined macroscopic fundamental diagram for urban traffic. Transport. Res. Part B: Methodol. 45 (3), 605–617.

Godfrey, J., 1969. The mechanism of a road network. Traffic Eng. Control 11 (7), 323–327.

Haddad, J., Geroliminis, N., 2012. On the stability of traffic perimeter control in two-region urban cities. Transport. Res. Part B: Methodol. 46 (9), 1159–1176.

Haddad, J., Ramezani, M., Geroliminis, N., 2013. Cooperative traffic control of a mixed network with two urban regions and a freeway. Transport. Res. Part B: Methodol. 54, 17–36.

Haddad, J., Shraiber, A., 2014. Robust perimeter control design for an urban region. Transport. Res. Part B: Methodol. 68, 315–332.

Herman, R., Prigogine, I., 1979. A two-fluid approach to town traffic. Science 204 (4389), 148–151.

Jahn, O., Möhring, R.H., Schulz, A.S., Stier-Moses, N.E., 2005. System-optimal routing of traffic flows with user constraints in networks with congestion. Oper. Res. 53 (4), 600–616.

Ji, Y., Geroliminis, N., 2012. On the spatial partitioning of urban transportation networks. Transport. Res. Part B: Methodol. 46 (10), 1639–1656.

Ji, Y., Luo, J., Geroliminis, N., 2014. Empirical observations of congestion propagation and dynamic partitioning with probe data for large-scale systems. Transport. Res. Rec.: J. Transport. Res. Board 2422 (1), 1–11.

Keyvan-Ekbatani, M., Kouvelas, A., Papamichail, I., Papageorgiou, M., 2012. Exploiting the fundamental diagram of urban networks for feedback-based gating. Transport. Res. Part B: Methodol. 46 (10), 1393–1403.

Keyvan-Ekbatani, M., Yildirimoglu, M., Geroliminis, N., Papageorgiou, M., 2015. Multiple concentric gating traffic control in large-scale urban networks. IEEE Trans. Intell. Transport. Syst. (99), 1–14

Knoop, V., Hoogendoorn, S., VanLint, J.W., 2012. Routing strategies based on macroscopic fundamental diagram. Transport. Res. Rec.: J. Transport. Res. Board 2315 (1), 1–10.

Leclercq, L., Chiabaut, N., Trinquier, B., 2014. Macroscopic fundamental diagrams: a cross-comparison of estimation methods. Transport. Res. Part B: Methodol. 62 (0), 1–12.

Leclercq, L., Geroliminis, N., 2013. Estimating mfds in simple networks with route choice. Transport. Res. Part B: Methodol. 57, 468–484.

Little, J.D., 1961. A proof for the queuing formula: $L = \lambda w$. Oper. Res. 9 (3), 383–387.

Mahmassani, H., Williams, J., Herman, R., 1984. Investigation of network-level traffic flow relationships: some simulation results. Transport. Res. Rec. 971, 121–130.

Mahmassani, H.S., Hou, T., Saberi, M., 2013a. Connecting networkwide travel time reliability and the network fundamental diagram of traffic flow. Transport. Res. Rec.: J. Transport. Res. Board 2391 (1), 80–91.

Mahmassani, H.S., Peeta, S., 1993. Network performance under system optimal and user equilibrium dynamic assignments: implications for advanced traveler information systems. Transport. Res. Rec.: J. Transport. Res. Board 1408, 83–93.

Mahmassani, H.S., Saberi, M., Zockaie, A., 2013b. Urban network gridlock: theory, characteristics, and dynamics. Transport. Res. Part C: Emerg. Technol. 36 (0), 480–497.

Mazloumian, A., Geroliminis, N., Helbing, D., 2010. The spatial variability of vehicle densities as determinant of urban network capacity. Philos. Trans. Roy. Soc. A: Math. Phys. Eng. Sci. 368 (1928), 4627–4647.

Nagle, A.S., Gayah, V.V., 2014. Accuracy of networkwide traffic states estimated from mobile probe data. Transport. Res. Rec. 2421, 1–11.

Ortigosa, J., Menendez, M., Tapia, H., 2013. Study on the number and location of measurement points for an mfd perimeter control scheme: a case study of zurich. EURO J. Transport. Log., 1–22

Papageorgiou, M., 1990. Dynamic modeling, assignment, and route guidance in traffic networks. Transport. Res. Part B: Methodol. 24 (6), 471–495.

Peeta, S., Mahmassani, H.S., 1995. System optimal and user equilibrium time-dependent traffic assignment in congested networks. Ann. Oper. Res. 60 (1), 81–113.

Peeta, S., Ziliaskopoulos, A.K., 2001. Foundations of dynamic traffic assignment: the past, the present and the future. Networks Spat. Econ. 1 (3), 233–265.

Ramezani, M., Haddad, J., Geroliminis, N., 2015. Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control. Transport. Res. Part B: Methodol. 74, 1–19.

Yildirimoglu, M., Geroliminis, N., 2014. Approximating dynamic equilibrium conditions with macroscopic fundamental diagrams. Transport. Res. Part B: Methodol. 70, 186–200.

Yildirimoglu, M., Limniati, Y., Geroliminis, N., 2015. Investigating empirical implications of hysteresis in day-to-day travel time variability. Transport. Res. Part C: Emerg. Technol. http://dx.doi.org/10.1016/j.trc.2015.03.012.

Zhou, X., Mahmassani, H.S., Zhang, K., 2008. Dynamic micro-assignment modeling approach for integrated multimodal urban corridor management. Transport. Res. Part C: Emerg. Technol. 16 (2), 167–186.

Zhu, F., Ukkusuri, S.V., 2013. A cell based dynamic system optimum model with non-holding back flows. Transport. Res. Part C: Emerg. Technol. 36, 367–380.