Transportation Research Part B 74 (2015) 1–19

Contents lists available at ScienceDirect

Transportation Research Part B

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

Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control

Mohsen Ramezani [a], Jack Haddad [b,1], Nikolas Geroliminis [a,*]

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

[b] Technion – Israel Institute of Technology, Faculty of Civil and Environmental Engineering, Technion Sustainable Mobility and Robust Transportation (T-SMART) Laboratory, Technion City, Rabin Building, Haifa 32000, Israel

ARTICLE INFO

Article history:
Received 15 April 2014
Received in revised form 30 December 2014
Accepted 31 December 2014

Keywords:
Macroscopic Fundamental Diagram (MFD)
Heterogeneity
Perimeter control
Hierarchical control
Traffic hysteresis

ABSTRACT

Real traffic data and simulation analysis reveal that for some urban networks a well-defined Macroscopic Fundamental Diagram (MFD) exists, which provides a unimodal and low-scatter relationship between the network vehicle density and outflow. Recent studies demonstrate that link density heterogeneity plays a significant role in the shape and scatter level of MFD and can cause hysteresis loops that influence the network performance. Evidently, a more homogeneous network in terms of link density can result in higher network outflow, which implies a network performance improvement. In this article, we introduce two aggregated models, region- and subregion-based MFDs, to study the dynamics of heterogeneity and how they can affect the accuracy scatter and hysteresis of a multi-subregion MFD model. We also introduce a hierarchical perimeter flow control problem by integrating the MFD heterogeneous modeling. The perimeter flow controllers operate on the border between urban regions, and manipulate the percentages of flows that transfer between the regions such that the network delay is minimized and the distribution of congestion is more homogeneous. The first level of the hierarchical control problem can be solved by a model predictive control approach, where the prediction model is the aggregated parsimonious region-based MFD and the plant (reality) is formulated by the subregion-based MFDs, which is a more detailed model. At the lower level, a feedback controller of the hierarchical structure, tries to maximize the outflow of critical regions, by increasing their homogeneity. With inputs that can be observed with existing monitoring techniques and without the need for detailed traffic state information, the proposed framework succeeds to increase network flows and decrease the hysteresis loop of the MFD. Comparison with existing perimeter controllers without considering the more advanced heterogeneity modeling of MFD highlights the importance of such approach for traffic modeling and control.

1. Introduction

Efficient traffic control and management of large-scale transportation networks still remain a challenge both for traffic researchers and practitioners. Unlike microscopic approaches that usually utilize disaggregate traffic flow models, as

* Corresponding author at: GC C2 389, Station 18, 1015 Lausanne, Switzerland. Tel.: +41 21 6932481; fax: +41 21 69 35060.
E-mail addresses: mohsen.ramezani@epfl.ch (M. Ramezani), jh@technion.ac.il (J. Haddad), nikolas.geroliminis@epfl.ch (N. Geroliminis).
[1] Tel.: +972 778871742.

http://dx.doi.org/10.1016/j.trb.2014.12.010
0191-2615/

behavior of each vehicle is modeled in detail, e.g. car following and lane changing models, in this paper, we follow the macroscopic (network level) approach utilizing the Macroscopic Fundamental Diagram (MFD). The MFD aims at simplifying the micro-modeling task of the urban network, where the collective traffic flow dynamics of subnetworks capture the main characteristics of traffic congestion, such as the evolution of space-mean flows and densities in different regions of a city. Nevertheless, the MFD is not a universal law for all network structures and demands and if it is applied in a non-careful way it can hide critical patterns of congestion and result in inefficient control methods.

The MFD provides a unimodal, low-scatter relationship between network vehicle density (veh/km) and network space-mean flow or outflow (trip completion rate) (veh/h) for different network regions, if congestion is roughly homogeneous in the region. Recently, the macroscopic (network) traffic modeling has intensively attracted the traffic flow community. The physical model of MFD was initially proposed by Godfrey (1969) and observed with dynamic features in congested urban network in Yokohama by Geroliminis and Daganzo (2008), and investigated using empirical or simulated data by Buisson and Ladier (2009), Ji et al. (2010), Mazloumian et al. (2010), Zhang et al. (2013) and others. Earlier works had looked for MFD patterns in data from lightly congested real-world networks or in data from simulations with artificial routing rules and static demands (e.g. Mahmassani et al., 1987; Olszewski et al., 1995 and others), but did not demonstrate that an invariant MFD with dynamic features can arise. The observability of the MFD with different sensing techniques have been studied by Leclercq et al. (2014) and Ortigosa et al. (2014).

Studies Mazloumian et al. (2010), Geroliminis and Sun (2011b), Gayah and Daganzo (2011a), Mahmassani et al. (2013), and Knoop et al. (2013) have shown that networks with heterogeneous distribution of density exhibit network flows smaller than those that approximately meet homogeneity conditions (low spatial variance of link density), especially for high network densities. Networks with small variance of link densities have *a well-defined* MFD, i.e. low scatter of flows for the same densities. Heterogeneously congested networks might exhibit points below the upper envelope of an MFD or strong hysteresis loops (see for example freeway networks in Geroliminis and Sun (2011a) and Saberi and Mahmassani (2013)). Recently, in agreement with previous publications in heterogeneity, Mahmassani et al. (2013) proposed and calibrated with simulated data an MFD where the heterogeneity decreases the MFD output with a functional relationship. Following these findings the concept of an MFD can be applied for heterogeneously loaded cities with multiple pockets of congestion, if these cities can be partitioned into a small number of homogeneous clusters. Recent work Ji and Geroliminis (2012) created clustering algorithms for heterogeneous transportation networks with an objective to obtain small variance of link densities within a cluster. Understanding and modeling the dynamics of heterogeneity is a crucial and challenging question that can shed more light on how to develop smarter hierarchical traffic control schemes for large-scale urban networks. This paper moves towards this direction.

The MFD can be utilized to introduce elegant real-time control strategies to improve mobility and decrease delays in large urban networks, that local ones are unable to succeed, see for example Daganzo (2007), Haddad and Geroliminis (2012) and Geroliminis et al. (2013). Perimeter flow control strategies, i.e. manipulating the transfer flows at the perimeter border of the urban region, utilizing the concept of the MFD have been introduced for single-region cities in Daganzo (2007), Keyvan-Ekbatani et al. (2012), Gayah et al. (2014) and Haddad and Shraiber (2014), and for multi-region cities in Haddad and Geroliminis (2012), Geroliminis et al. (2013), and Aboudolas and Geroliminis (2013). Moreover, route guidance strategies with the utilization of MFD have been studied in Knoop et al. (2012) for grid networks without traffic lights. Gayah and Daganzo (2011b) and Leclercq and Geroliminis (2013) also studied simple routing strategies for two-bin or two-route network abstractions.

In Haddad et al. (2013) different control strategies with different levels of coordination have been introduced for metropolitan transportation networks that have a hierarchical structure which consists of freeways and urban roads. Previous works Daganzo and Geroliminis (2008), Geroliminis and Boyacı (2012) and Zhang et al. (2013) have shown that traffic-responsive signal control strategies and different signal settings can change the shape of the MFD and the critical accumulations. While in both works Geroliminis et al. (2013) and Haddad et al. (2013), do not explicitly model the effect of link heterogeneity, in this paper we aim at studying the effect of heterogeneity by introducing a new modeling approach that considers MFD of regions and smaller subregions. This work also models a simple route choice process between paths through subregions, and considers the effect of subregion flow receiving capacity.

The control problems in previous works, e.g. Geroliminis et al. (2013) and Haddad et al. (2013), have been solved by the model predictive control (MPC) approach. It was shown that this control approach can handle different levels of error in traffic demand and noise in MFDs shape. Nevertheless, the model and the plant in the MPC framework were inherently similar, though the errors in demand and the MFD distinguish between the two. A stronger level of dissimilarity between the model and the plant can provide a more convincing evidence for the applicability of such approaches in more realistic and complex networks.

The objectives of this paper are twofold, in modeling and control aspects. First, we further investigate the relation between the heterogeneity and the MFD. With respect to modeling, we investigate the dynamics of heterogeneity and how it can affect accuracy and scatter of a multi-region MFD model, which consists of variables that can be obtained with existing sensor technology. While there is some work how heterogeneity influences the shape of the MFD, there is no theoretical work to investigate how an asymmetric demand pattern can affect the distribution of congestion over time and space and its dynamic behavior. Existing MFD dynamic models as expressed in various publications are hysteresis-free and as a result the developed control frameworks based on such models cannot be trusted when hysteresis appears. With respect to control, our objective is to integrate the dynamics of heterogeneity in the optimization framework and design

perimeter control strategies that can decrease congestion heterogeneity and increase the network performance. As our analysis shows, this is a crucial step for efficient control strategies, where strong hysteresis loops appear. As we will show, considering an MFD without hysteresis and ignoring the dynamics of heterogeneity can result in situations where perimeter control is not beneficial for the system. Such an advanced model also allows to develop a two-level hierarchical control framework (e.g. see Doan et al., 2014) that decreases system delays and hysteresis loops, which are directly related with capacity loss at the network level.

The remainder of the paper is organized as follows: in Section 2 we describe the dynamics of the region-based model which integrates the effect of heterogeneity in the MFD. Then, a more detailed model of sub-regions, which can describe the dynamics of heterogeneity, is developed. In Section 3, a closed form expression of MFD as a function of mean and variance of link accumulation is obtained based on real data, while in Section 4 a hierarchical control framework based on MPC tries to optimize system performance based on the developed models. Decreasing the level of heterogeneity with control has a positive effect in the system delays and the avoidance of large hysteresis loops. Results of a case study and future work conclude the paper.

2. Modeling the dynamics of spatial density heterogeneity in urban regions with perimeter control

In this paper, we introduce two aggregated models with an objective to integrate the dynamics of heterogeneity in a network: (i) a *region-based model* considers networks partitioned into a small number of regions that might be split by perimeter controllers, and (ii) a *subregion-based model*, where each region of the above model is partitioned into subregions, see Fig. 1. Existing region-based dynamic models for single or multi-region networks (e.g. Daganzo, 2007; Geroliminis et al., 2013; Keyvan-Ekbatani et al., 2012) consider an MFD without hysteresis. Hysteresis creates multivalueness in the network flow for the same value of network accumulations. Given that these values are strongly influenced by the distribution of

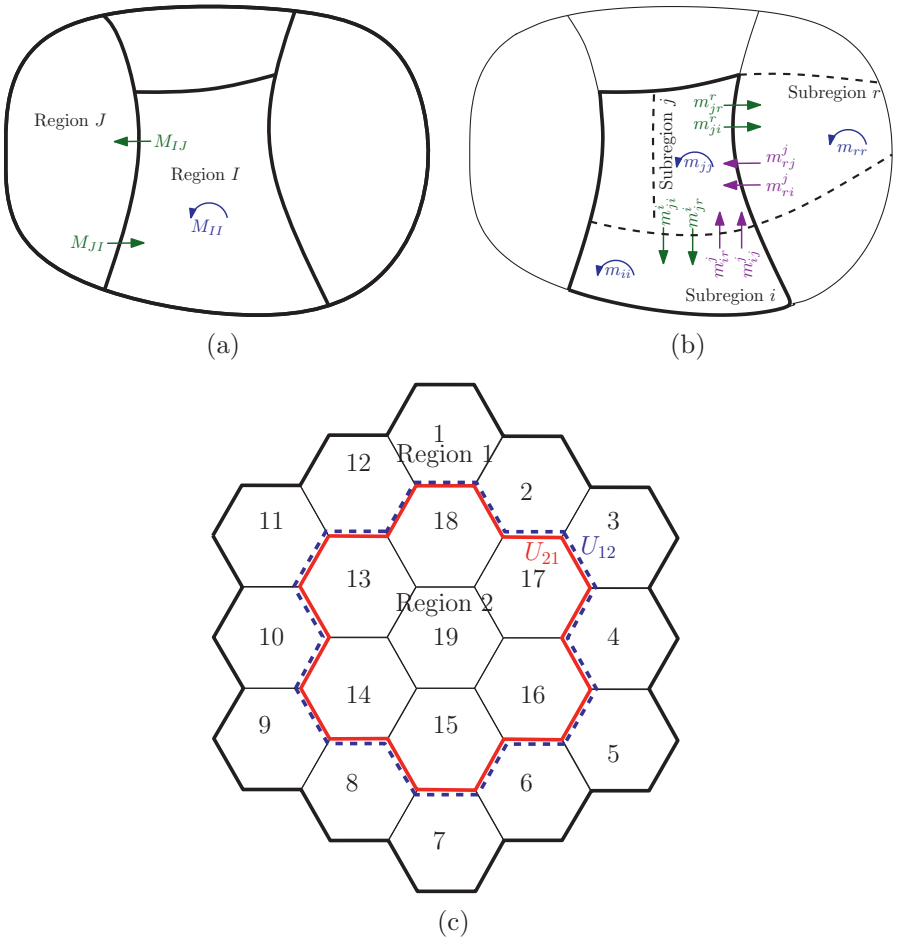

Fig. 1. A schematic urban network with (part of) internal and transfer flows for region I and subregions i, j, r in the (a) region- and (b) subregion-based models, respectively. (c) A case study network consists of two regions – Region 1 (the periphery) and Region 2 (city center) partitioned respectively into 12 and 7 subregions.

congestion, such a hysteresis cannot be an external input to the model, but it has to be integrated within the model and be influenced by the dynamics of heterogeneity. This is a challenging methodological step that requires the interconnection of a region-based and a subregion-based model. At the region-based model, the heterogeneity dynamics are integrated in the regional MFDs in two directions: (i) variant regional trip lengths and (ii) an MFD depending on regional accumulation and the heterogeneity of the spatial distribution of congestion. To integrate the dynamics of heterogeneity in the region-based model, a subregion-based model is needed that (a) describes the evolution of subregion accumulations, (b) integrates the heterogeneity dynamics in the subregion MFDs, (c) integrates a route choice model, and (d) models the effect of receiving (or boundary) capacity of subregion destination.

The constraint of receiving (or boundary) capacity has not been considered during the optimization process in previous control oriented publications (see for example Keyvan-Ekbatani et al., 2012; Geroliminis et al., 2013; Aboudolas and Geroliminis, 2013). The physical reasoning behind this assumption is that (i) boundary capacity decreases for accumulations much larger than the critical accumulation (see Geroliminis and Daganzo, 2007) and (ii) the control inputs will not allow the system to get close to gridlock. Nevertheless, perimeter controllers are acting only in the boundaries between regions and not all subregions are integrated in a perimeter control logic. Thus, such a constraint cannot be fully ignored now. Fig. 1(a) and (b) depict a schematic urban network with (part of) internal and transfer flows for region I and subregions i,j,r in the (i) region- and (ii) subregion-based models, respectively. All the related variables are introduced later in details.

2.1. Region-based model

Let us assume that an urban network is partitioned into R regions, $\mathcal{R}=\{1,2,\ldots,R\}$. Let $Q_{IJ}(t)$ (veh/s) be the traffic demand flow generated in region I with destination to region J, $N_{IJ}(t)$ (veh) be the accumulation in region I with region destination J; $I,J\in\mathcal{R}$, and $N_I(t)$ (veh) be the total accumulation in region I.

The total production $P_I(N_I(t),\sigma(N_I(t)))$ (veh · distance traveled per unit time) in region I is a function of the regional accumulation $N_I(t)$ and its variance across all links in the region, $\sigma(N_I(t))$, as has been reported in Mazloumian et al. (2010), Mahmassani et al. (2013), Geroliminis and Sun (2011b), and Knoop et al. (2013). The trip completion flow for region I is the sum of transfer flows, i.e. trips from I with direct destination $J, J\in\mathcal{H}_I$, where $\mathcal{H}_I$ is the set of regions that are directly reachable from (adjacent to) region I, plus the internal flow, i.e. trips from I with direct destination I. The *transfer flow* from I with destination to J is denoted by $M_{IJ}(t)$ (veh/s), while $M_{II}(t)$ denotes the *internal flow* from I with destination to I. They are calculated corresponding to the ratio between accumulations as follows

$$M_{II}(t)=\frac{N_{II}(t)}{N_I}\cdot\frac{P_I(N_I(t),\sigma(N_I(t)))}{L_{II}(t)}, \tag{1a}$$

$$M_{IJ}(t)=\frac{N_{IJ}(t)}{N_I}\cdot\frac{P_I(N_I(t),\sigma(N_I(t)))}{L_{IJ}(t)}, \tag{1b}$$

where $P_I(\cdot)$ (veh/s m) is the MFD production for region I at $N_I(t)$ with heterogeneity variance $\sigma(N_I(t))$, $L_{II}(t)$ (m) is the average trip length (space mean) for trips in region I, and $L_{IJ}(t)$ (m) is the average trip length for trips from region I to J. Note that the variable $\sigma(N_I(t))$ captures the link density spatial *heterogeneity* for an urban region (see later Eq. (11), where function $P_I(N_I(t),\sigma(N_I(t)))$ is described). Note that for flows $Q_{IJ}(t)$ where I and J are not adjacent, a sequence of regions should be known to develop the transfer flows between I and J, with J belongs to $\mathcal{H}_I$.

One of the main objectives of this work is to integrate the developed modeling in a control framework and investigate strategies that will decrease heterogeneity and network flow hysteresis loops. To this end, a semi-analytical approximative model of $\sigma(N_I(t))$ is required. To model $\sigma(N_I(t))$, we need first to develop an analytical model for subregion link density heterogeneity then aggregate the subregional heterogeneity into the regional one. Thus, in Section 3 we investigate the heterogeneity dynamics for a subregion based on a field dataset and introduce a method to aggregate and scale up the subregional heterogeneities to present link density heterogeneity in regions.

Perimeter controllers, $U_{IJ}(t)$ and $U_{JI}(t)$ (–), $0\leqslant U_{IJ}(t),U_{JI}(t)\leqslant 1$, might exist between each two regions I and J, $J\in\mathcal{H}_I$, that can constrain the transfer flows from I to J and from J to I, respectively. Consequently, the mass conservation equations of an R-region MFDs system are as follows:

$$\frac{dN_{II}(t)}{dt}=Q_{II}(t)-M_{II}(t)+\sum_{J\in\mathcal{H}_I}U_{JI}(t)\cdot M_{JI}(t), \tag{2}$$

$$\frac{dN_{IJ}(t)}{dt}=Q_{IJ}(t)-\sum_{J\in\mathcal{H}_I}U_{IJ}(t)\cdot M_{IJ}(t), \tag{3}$$

for $I=1,2,\ldots,R$ and $\forall J\in\mathcal{H}_I$. Note that $N_I(t)=N_{II}(t)+\sum_{J\in\mathcal{H}_I}N_{IJ}(t)$. These equations are a generalized (R regions instead of two) equations presented in Geroliminis et al. (2013) with integrated heterogeneity. Note that route choice modeling is not integrated in the *region-based* dynamic equations and this model is not aware that travellers make route choice decisions when conditions change. This is done on purpose, since traveller behavior might be difficult to be predicted in real time. It is also assumed that drivers are not allowed to cross a boundary more than once, e.g. a trip from region I to I by crossing region J is not considered. This will change the dynamic Eqs. (2) and (3) and more complicated accumulation states have

to be developed, which is beyond the scope of this work (e.g. number of vehicles in region I with destination I that will cross to region J and return in region I). This is under ongoing investigation. Nevertheless, this constraint does not apply to the detailed subregion-based model which is developed in the next section. The hierarchical control framework that is developed later is not influenced by such constraint.

2.2. Subregion-based model

The subregion-based model is a more detailed model since the urban region is considered as a collection of several smaller urban areas, called *subregions*, which still contain a significant number of links to be described by a low-scatter MFD. Each subregion accumulation evolves differently in time which allows to integrate the heterogeneity of the spatial distribution of congestion in the urban region (see terms related to MFD heterogeneity in (1a) and (1b)). This modeling approach will give us the opportunity to investigate more rigorously several assumptions in the MFD literature that have been empirically observed, e.g. trip length in a region is about constant, if and how route choice, perimeter control, and O-D affect the heterogeneity and the distribution of congestion. These are challenging research questions that have been raised by many researchers and it is not clear yet under what network conditions an MFD provides a decent representation of network performance. The purpose of the following formulation is to express in a consistent manner the variables in (1a)–(3), which represent region-based MFDs, accumulations, and trip lengths as a function of a more detailed model at the subregional level. While microscopic simulation might be an alternative instead of the described model, this paper chooses a more methodological path, which allows to create further insights of the dynamics of heterogeneity and the hierarchical control.

Let us consider region $I \in \mathcal{R}$ which is heterogeneous in space link density and consists of subregions, as schematically shown in Fig. 1(b). We use capital letters for variables related to regions and lower case letters for variables related to subregions. We denote $\mathcal{SR}$ as the set of all subregions in the urban network, while $\mathcal{SR}_I$ is the set of subregions that belongs to region I. Let $q_{ij}(t)$ (veh/s) be the demand from subregion i to subregion $j, n_{ij}(t)$ (veh) be the accumulation in subregion i with final subregion destination $j, \{i,j\} \in \mathcal{SR}$, and $n_i(t)$ (veh) be the total accumulation in subregion i, i.e. $n_i(t) = \sum_{j \in \mathcal{SR}} n_{ij}(t)$. The MFD production for subregion i, denoted by $p_i(t)$ (veh/s m), is the total distance traveled in subregion i by all vehicles $n_i(t)$, which is equal to the sum of the transfer and internal flows multiplied by the average trip length in subregion $i, l_i(t)$ (m).

Let $m^h_{ij}(t)$ (veh/s) be the *transfer flow* from subregion i with *final* subregion destination $j, i \neq j$, through the *immediate* next subregion $h \in \mathcal{H}_i$, where $\mathcal{H}_i$ is the set of subregions that are directly reachable from subregion i. The transfer flow is related to the ratio between subregion accumulations and corresponding trip length, i.e. $m^h_{ij}(t) = \theta^h_{ij}(t) \cdot n_{ij}(t)/n_i(t) \cdot p_i(n_i(t))/l_i(t)$, where $\theta^h_{ij}(t)$ $(-)$ is the flow percentage of the total transfer flows from subregion i to destination j that passes immediately through subregion h, thus $\sum_{h \in \mathcal{H}_i} \theta^h_{ij}(t) = 1$. Note that a simple route choice model is integrated in the *subregion-based* model, where $\theta^h_{ij}(t)$ are calculated by a logit model according to the travel times from i to j through the k (current best) shortest paths (sequence of subregions), which are calculated using Dijkstra's algorithm. The travel time for each path is calculated by summing travel times through subregions, where each subregion travel time is calculated as the fraction between the distance traveled inside the subregion (through its center) and its average speed $v_i(t)$ (m/s) calculated from the subregion MFD at the beginning of the trip, i.e. $v_i(t) = p_i(n_i(t))/n_i(t)$. Trip length within subregion i is assumed to be independent of origin, destination, and route choice, which is consistent with the field data in Geroliminis and Daganzo (2008) and the assumptions made for the region-based models of previous publications Geroliminis et al. (2013) and Haddad et al. (2013).

The *internal flow* from subregion i with destination to subregion i, denoted by $m_{ii}(t)$ (veh/s), is calculated by $m_{ii}(t) = n_{ii}(t)/n_i(t) \cdot p_i(n_i(t))/l_i(t)$. For instance, Fig. 1(b) illustrates part of the transfer and internal flows for a network considering only three subregions i, j, and r. The transfer flows between subregions j and r are $m^r_{ji}, m^r_{jr}, m^j_{rj}, m^j_{ri}$, and $m^j_{ji}, m^j_{ij}, m^j_{ir}, m^i_{jr}$ are the transfer flows between subregions j and i, while the internal flows for subregions i, j, r are m_{ii}, m_{jj}, m_{rr}, respectively. Note that the region-based model implicitly assumes that the internal regional trips never leave the region and also external trips cross the boundary between the regions only once. The route choice of subregion-based model meets these assumptions.

The subregion-based model also integrates the effect of flow receiving capacity of the destination subregion. In other words, flow transferring into a subregion might be restricted since accumulation at subregion destination is such high that there is not enough space to fully receive the incoming transfer flows. Receiving capacity is not integrated in the region-based model as the controllers at the boundary are expected to avoid these situations. Eq. (4) expresses the transfer flow as the minimum of two terms, (i) the sending flow upstream of the boundary (from region i) which depends on the accumulations of region i and (ii) the receiving flow which depends on the accumulation of region h. Such an approach has been integrated in mass conservation equations for 1st and 2nd order models of traffic flow, e.g. the Cell Transmission Model (Daganzo, 1994). The difference is that the 2nd term is an analogy of the remaining storage capacity of the receiving subregion. Therefore, we introduce a receiving capacity term into the transfer flow dynamic equations as follow

$$\hat{m}^h_{ij}(t) = \min\left(m^h_{ij}(t), \frac{m^h_{ij}(t)}{\sum_k m^h_{ik}(t)} \cdot r_{ih}(n_h(t))\right), \tag{4}$$

where $k \in \mathcal{SR}, k \neq i$, and $r_{ih}(\cdot)$ (veh/s) is the receiving flow capacity of subregion $h, h \in \mathcal{H}_i$, from subregion i. We consider that the receiving capacity is a piecewise function of $n_h(t)$ with two pieces, a constant value and a decreasing function, as follows

$$r_{ih}(n_h(t)) = \begin{cases} r_{ih}^{\max} & \text{if } 0 \leqslant n_h(t) < \alpha \cdot n_h^{\text{jam}}, \\ -\frac{r_{ih}^{\max}}{(1-\alpha)\cdot n_h^{\text{jam}}} \cdot n_h(t) + \frac{r_{ih}^{\max}}{1-\alpha} & \text{if } \alpha \cdot n_h^{\text{jam}} \leqslant n_h(t) \leqslant n_h^{\text{jam}}, \end{cases} \tag{5}$$

where $r_{ih}^{\max}$ (veh/s) is the maximum value of the receiving capacity and boundary capacity, n_h^{jam} (veh) is the jammed accumulation of subregion h, and $0 < \alpha < 1$ is a parameter that defines the critical accumulation when the receiving capacity starts to decrease and can be estimated if real data from sensors are readily available.

The transfer flows might be controlled by subregion perimeter controllers on the border between subregions, e.g. $0 \leqslant u_{ih}(t)\ (-) \leqslant 1$ denotes the perimeter control input between subregions i and h. The mass conservation equations for the subregions are as follows

$$\frac{\mathrm{d}n_{ii}(t)}{\mathrm{d}t} = q_{ii}(t) - m_{ii}(t) + \sum_{h\in\mathcal{H}_i} u_{hi}(t)\cdot \hat{m}_{hi}^{i}(t), \tag{6}$$

$$\frac{\mathrm{d}n_{ij}(t)}{\mathrm{d}t} = q_{ij}(t) - \sum_{h\in\mathcal{H}_i} u_{ih}(t)\cdot \hat{m}_{ij}^{h}(t) + \sum_{h\in\mathcal{H}_i; h\neq j} u_{hi}(t)\cdot \hat{m}_{hj}^{i}(t) \quad \forall j \in \mathcal{H}_i, \tag{7}$$

$$\frac{\mathrm{d}n_{ir}(t)}{\mathrm{d}t} = q_{ir}(t) - \sum_{h\in\mathcal{H}_i} u_{ih}(t)\cdot \hat{m}_{ir}^{h}(t) + \sum_{h\in\mathcal{H}_i} u_{hi}(t)\cdot \hat{m}_{hr}^{i}(t) \quad i \neq r;\ \forall r \notin \mathcal{H}_i, \tag{8}$$

$\forall i \in \mathcal{SR}$. Eqs. (6)–(8) assume that perimeter controllers exist between each neighbor subregions, however, they still hold if the assumption is relaxed by setting the control inputs to be equal to 1. Note that our intension is not to control inter transfers between any two subregions, but only in the boundaries of the region-based model, see Fig. 1(c). In this way we will keep the computational effort small and we will not rely on information which is difficult to be obtained with real data, e.g. n_{ij} for each subregion. Nevertheless, as stated before the more detailed model will shed light on the dynamics of heterogeneity and how it can affect the performance of an MFD region-based model, which consists of variables that can be obtained with existing sensors more accurately.

Finally, the region internal and external average trip length described in (1a) and (1b), L_{II} and L_{IJ}, respectively, are estimated as follows (considering a steady state law as the ratio of travel production over outflow)

$$L_{II}(t) = \frac{\sum_{i\in\mathcal{SR}_I}\sum_{j\in\mathcal{SR}_I} n_{ij}(t)}{\sum_{i\in\mathcal{SR}_I} n_i(t)} \cdot \frac{\sum_{i\in\mathcal{SR}_I} p_i(n_i(t))}{\sum_{i\in\mathcal{SR}_I} m_{ii}(t)}, \tag{9a}$$

$$L_{IJ}(t) = \frac{\sum_{i\in\mathcal{SR}_I}\sum_{j\in\mathcal{SR}_J} n_{ij}(t)}{\sum_{i\in\mathcal{SR}_I} n_i(t)} \cdot \frac{\sum_{i\in\mathcal{SR}_I} p_i(n_i(t))}{\sum_{i\in\mathcal{SR}_I}\sum_{h\in\mathcal{SR}_J} m_{ij}^{h}(t)}. \tag{9b}$$

The estimation of L_{II} and L_{IJ} is based on the assumption that the region- and subregion-based models should be consistent and have the same internal and external region outflows in case of perfect information. Thus, (9a) and (9b) have respectively similar logic to (1a) and (1b), while the right hand sides are expressed in terms of detailed variables of subregion-based model. I.e., the internal outflow M_{II} in the region-based model is equivalent to the sum of all $m_{ii}, i \in \mathcal{SR}_I$, and the external outflow M_{IJ} in the region-based model is equivalent to the sum of all $m_{ij}^{h}(t), i \in \mathcal{SR}_I$ and $h \in \mathcal{SR}_J$.

3. A functional form for the effect of heterogeneity on MFDs: Field data analysis

The region-based MFD dynamic model of (1a) and (1b) requires a functional form of the regional production depending on accumulation of the region and standard deviation of the spatial distribution of link accumulation. While various studies have investigated the effect of heterogeneity in the MFD e.g. Geroliminis and Sun (2011a), Mazloumian et al. (2010), Knoop et al. (2013), and Mahmassani et al. (2013) a functional form is necessary as these MFD dynamics have to be integrated in a control framework. While Mahmassani et al. (2013) provide a functional form based on simulated data, in this section we investigate such a relation with field data. More specifically we re-scrutinize the Yokohama field data, investigated in Geroliminis and Daganzo (2008) and Geroliminis and Sun (2011b), to obtain further insights into the dynamics of link occupancy heterogeneity and its effect on the MFD. Our objective is to propose an analytical distribution that models the first two statistical moments of individual link occupancy distribution. This approach is motivated by previous publications towards this direction, which are described in more detail in Section 1.

The developed models of this work in Section 2 can also be implemented for different functional forms and the reader can skip this section without loss of continuity. Nevertheless, this analysis provides useful empirical analysis for heterogeneity. An interesting finding is that the spatial distribution of congestion has similar functional form with other physical systems that experience spatial correlation.

The data consist of the occupancy (%) of 540 links every 5 min from early morning to the end of the day. We are interested in mean and standard deviation (STD) of link occupancy, since mean occupancy is an indicator of network congestion level and STD of link occupancies can be regarded as the heterogeneity indicator of the network. Analyses demonstrate that the negative binomial (NB) distribution can provide accurate estimations for mean and STD of occupancies for the Yokohama network data. NB distributions can describe well the spread of different phenomena with spatial correlations, such as

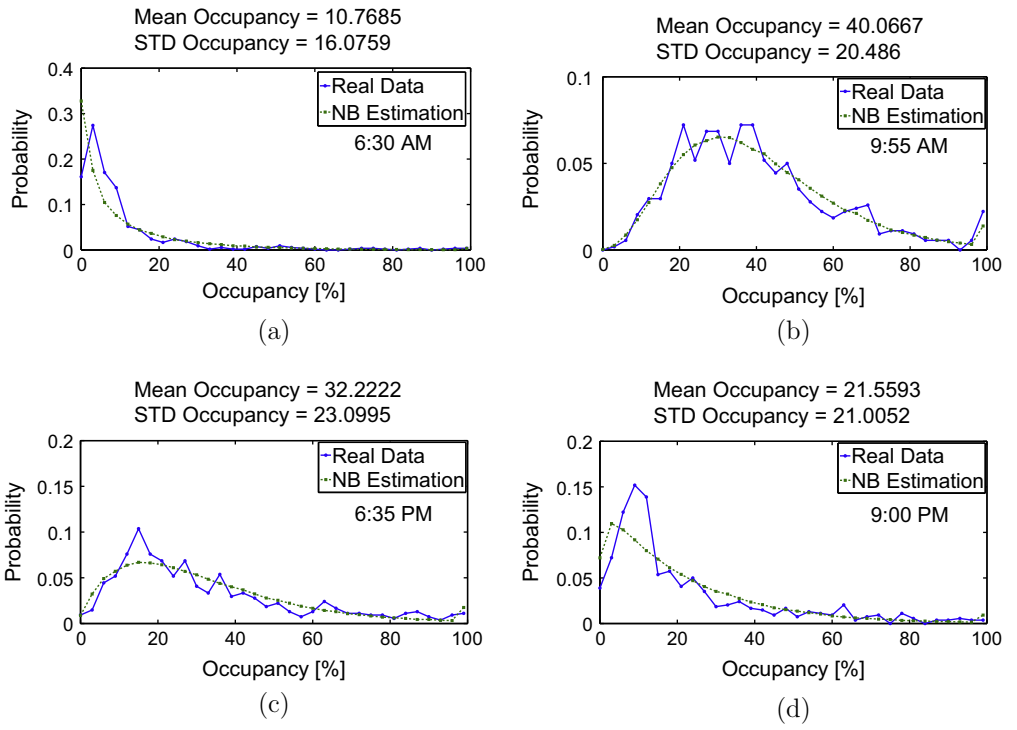

Fig. 2. Field data and the best NB estimated link occupancy distribution at four different times.

infectious diseases Lloyd-Smith et al. (2005), tree growth Clark (1998), and others. Note that Geroliminis and Sun (2011b) have derived a semi-analytical model of estimating this distribution based on spatial correlations between links, which might be challenging to be integrated in a control scheme. We choose to utilize the NB distribution, due to the ease of numerical calculations. NB is a discrete probability distribution of the number of successes in a sequence of binomial trials with probability of success, p, before a pre-specified number of failures, r, occurs. NB distribution is useful in modeling count data similar to the Poisson distribution, however, NB is more general and accurate to capture dispersion with spatial correlations than the Poisson distribution because its variance is greater than its mean. The NB probability mass function is

$$\Pr(X = x) = NB(x, r, p) = \binom{x + r - 1}{x} (1 - p)^r p^x. \tag{10}$$

Note that r can be interpreted as the number of congested links (failures), while p can be related to the occupancy that indicates congested state (probability of success or failure, if occupancy is normalized between zero and one).

Fig. 2 depicts the field data and the best NB fit (in maximum likelihood sense) representing link occupancy distribution at four different times during a day covering a wide range of traffic conditions from early uncongested to mid-day congested and evening mild-congested. The four cases have different mean, STD, and distribution of occupancies, while the NB estimation accurately models the link occupancy distribution. Moreover, it has been observed in Geroliminis and Sun (2011b) that there is a well-defined relationship between the average network occupancy and the STD of individual detector occupancy for Yokohama data. The NB estimation can reproduce a well-defined MFD similar to the MFD based on the field data even if link FD has significant scatter.

3.1. Effect of link occupancy heterogeneity on subregion MFD

Aforementioned observations confirm that the NB distribution can be regarded as a proper estimator of link occupancy distribution in a homogeneous subregion. Nevertheless, given the low scatter MFD of Yokohama, it is not possible to investigate the effect of heterogeneity for a large range of STD for a given subregion mean occupancy. To succeed this objective, we draw NB distributions with a range of sensible STD, o_{std}, for various average subregion occupancies, i.e. $o_{\text{u}} \in [5\text{—}75\%]$, and then estimate the subregion average flow with a low scatter link FD. The outcomes for $o_{\text{u}} = [5, 10, 15, \ldots, 75\%]$ are depicted in Fig. 3. To obtain a closed-form expression relating subregion average flow, q_{u}, to mean occupancy, o_{u}, and occupancy STD, o_{std}, we fit an exponential function to the data, i.e. $q_{\text{u}}(o_{\text{u}}, o_{\text{std}}) = (d_3 \cdot o_{\text{u}}^3 + d_2 \cdot o_{\text{u}}^2 + d_1 \cdot o_{\text{u}}) \cdot (a \cdot e^{b \cdot o_{\text{std}}} + c)$, where a, b, c, d_1, d_2, and d_3 are estimated parameters. The results reveal that the function that is product of a 3-degree polynomial, representative of a low-scatter MFD, and a exponential function, representative of heterogeneity effect, models accurately the MFD. Hence, the same decomposition of MFD function to two terms (i) an upper bound (low-scatter) MFD and (ii) the heterogeneity degradation is used in modeling the effect of link heterogeneity on region MFD, as we describe in the next subsection.

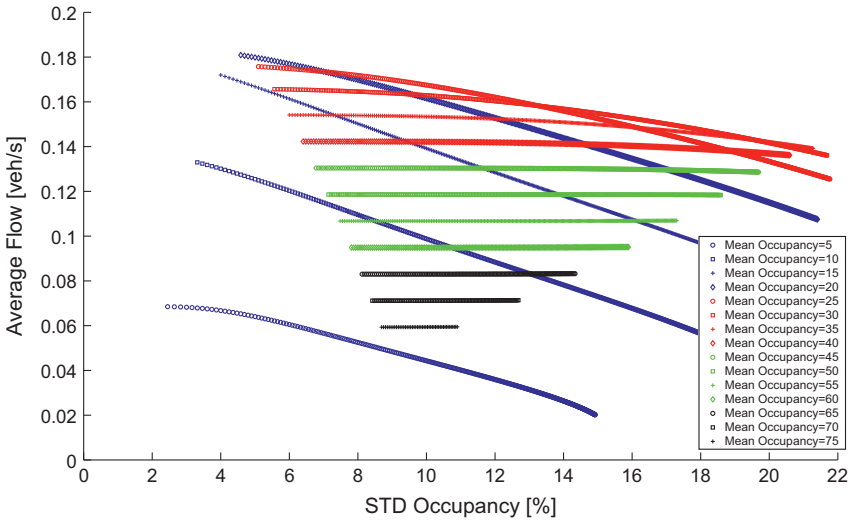

Fig. 3. Subregion average flow for different mean and STD occupancies.

The results show that in case of light conditions that the average subregion occupancy is low, increase in STD, i.e. the subregion becomes more heterogeneous, decreases the subregion average flow. As the subregion occupancy increases, the degrading effect of heterogeneity on the subregion average flow becomes less severe. About the subregion occupancy 50%, the best fit becomes almost a line with slope zero revealing that the subregion average flow is independent of the link occupancy heterogeneity.

Note that the validity of theses observations is based on the Yokohama field dataset which does not comprise very congested situations or particular cases which are unlikely in reality, e.g. a case with low network occupancy and high STD. This limitation prevents us to make a general statement, however with more field data specifically for congested situations a better understanding of heterogeneity effect on the MFD can be concluded. For example for cases close to gridlock, simulations from Mazloumian et al. (2010) showed that the distribution of congestion is bimodal with a fraction of links being around jam occupancy and another fraction close to zero occupancy, which might not be described well by an NB distribution.

3.2. Effect of link occupancy heterogeneity on region MFD

To obtain link density heterogeneity for urban region I, we assume that a well-defined relationship between the mean occupancy and the STD of link occupancies for each subregion i exists, $i \in \mathcal{SR}_I$, where $\mathcal{SR}_I$ denotes the set of subregions in region I. Thus, the STD of link occupancies for every subregion $i, i \in \mathcal{SR}_I$, and consequently, the link occupancy distribution, based on the NB distribution assumption, can be estimated given the mean occupancy of subregion i. Afterwards, NB distributions for all $i \in \mathcal{SR}_I$ are summed to capture the link occupancy distribution in region I and the STD of summation of NB distributions is an approximation of $\sigma(N_I(t))$. Fig. 4 illustrates the region MFD based on the

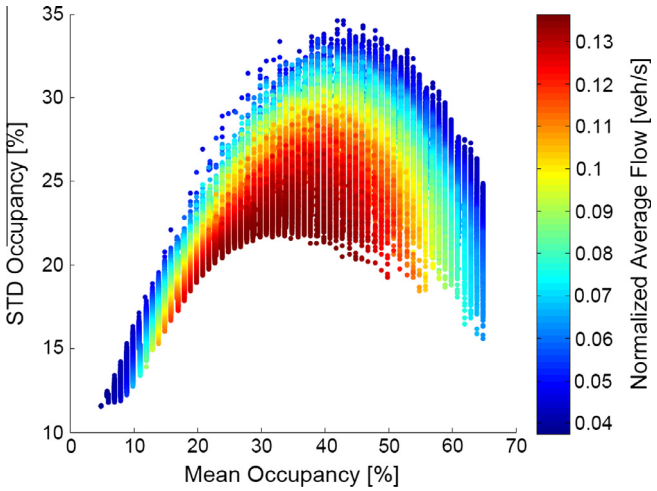

Fig. 4. The region production MFD as a function of mean and STD of occupancy.

aforementioned procedure. (Note that Fig. 3 depicts 2D presentation of subregion MFD whereas Fig. 4 is a 3D representation of region MFD.)

In a direct analogy with the formula presented in subsection 3.1, the production MFD of region I considering the link heterogeneity in region I is

$$P_I(N_I(t),\sigma(N_I(t))) = |\mathcal{SR}_I| \cdot \left[D_3 \cdot \left(\frac{N_I(t)}{|\mathcal{SR}_I|}\right)^3 + D_2 \cdot \left(\frac{N_I(t)}{|\mathcal{SR}_I|}\right)^2 + D_1 \cdot \frac{N_I(t)}{|\mathcal{SR}_I|}\right] \cdot \left(A \cdot e^{B\cdot(\sigma(N_I(t))-\sigma_h)} + (1-A)\right), \tag{11}$$

where $|\mathcal{SR}_I|$ denotes the number of subregions in region I, σ_h is the STD of summation of $|\mathcal{SR}_I|$ NB distributions with mean occupancy $N_I(t)/|\mathcal{SR}_I|$, and A, B, D_1, D_2, and D_3 are the estimated parameters that regulate the extent of link density heterogeneity effect on the region production. Note that (11) assumes that the region I production can be regarded as the product of two terms, the exponential term considering the heterogeneity and the production term which assumes homogeneous conditions corresponding to the upper bound (low-scatter) MFD. The functional form can approximate well the effect of heterogeneity of spatial distribution of congestion on region MFD and it can be integrated in the remaining of the paper to develop more advanced perimeter flow control strategies that can treat explicitly this effect.

4. Hierarchical control for heterogeneous networks

The previous section has provided, based on field-data analysis, a functional form related to the effect of heterogeneity on MFDs, which results in integrating heterogeneity dynamics in urban network modeling. This section aims at utilizing the integration of heterogeneity dynamics for control purposes.

The optimal perimeter control problem formulation and solution have been introduced for homogeneous networks, showing that applying a perimeter control strategy can improve the network performance of urban regions. Physically speaking a perimeter control strategy that assumes that all sub-regions in the one or the other boundary of the control have equal accumulations, might provide erroneous control actions if this is not the case. Applying similar control restrictions (e.g. same amount of transfer flows) in sub-regions with significantly different levels of congestion, might further increase congestion in some of them. Thus, by ignoring the effect of heterogeneity in the development of control it might lead in non-optimal results as we will show later. In Geroliminis et al. (2013) and Haddad et al. (2013), an MPC approach solution has been applied to minimize the total network delay, without considering the effect of the regions heterogeneity. While this might be successful when congestion is uniformly distributed, not all cities have such a property. Clustering algorithms have been proved efficient in decreasing the spatial heterogeneity, but in principle this is a feature of mobility that cannot totally disappear due to complex demand characteristics. Similar to homogeneous networks, in the current study the aim of perimeter control for heterogeneous networks is also to minimize the total network delay. However, giving the negative effect of heterogeneity in the network flow, one can introduce a new control scheme that is crafted for heterogeneous networks. In the following, we introduce *a hierarchical perimeter control framework* for heterogeneous networks, having two levels of control: a high-level controller aims at minimizing total delay with the help of MPC approach, and a low-level (feedback) controller aims at minimizing the region accumulation heterogeneity.

4.1. The high-level (MPC) controller

The aim of optimal perimeter control for heterogeneous networks is to minimize the network delay, defined as the integral of the network accumulation with respect to time, by manipulating the perimeter controllers. We utilize the MPC approach to solve the optimal control problem. The reader can refer to Geroliminis et al. (2013) and Haddad et al. (2013) for the application of MPC in perimeter control for homogeneous networks. Nevertheless, in the aforementioned publications the model and the plant were very similar with the only difference being some unknown stochastic term in demand and MFD. Developing an MPC framework with a very different model than plant is challenging and will shed more light in the possible application of such approaches in real life, where many network characteristics are unknown (e.g. route choice, sub-regional O-D tables). While microscopic simulation might be an alternative choice for the plant, this paper chooses the direction of two different types of aggregation (sub-regional and regional models), which allows to create further insights of the dynamics of heterogeneity and the methodological framework of traffic flow analysis. Other works (see for example, Keyvan-Ekbatani et al., 2012; Aboudolas and Geroliminis, 2013) have shown in a microsimulation environment that perimeter control strategies can significantly decrease network delays.

Both models, the subregion- and region-based models, are utilized in the MPC framework. The subregion-based model describes the traffic flow dynamics in detail (MPC-plant), while the region-based model is utilized to calculate the optimal control inputs in the optimization loop (MPC-model). Recall that the subregion-based model describes in more details the mass conservation dynamics based on subregional MFDs, which also integrates the constraints on the transfer flows by the receiving capacity, while the region-based model is the MPC-model that is suitable for performing tractable optimization. Note that the region-based model considers the effect of link heterogeneity, while this information is provided by the subregion-based model. Some of the variables of the sub-regional model might require significant estimation efforts and high density of sensors, which make the real-time implementation challenging. Nevertheless, the regional model, which

is utilized in the optimization framework, is based on information that can be estimated readily with standard monitoring and sensing techniques.

The MPC controller determines the optimal control inputs in a receding horizon manner, meaning that at each time step an objective function is optimized over a prediction horizon of K_p steps and a sequence of optimal control inputs are derived. Then the first sample of the control inputs is applied to the system and the procedure is carried out again with a shifted horizon. The closed-loop optimal control scheme in the MPC framework takes into account not only the errors between the plant and the model, but also the disturbances, e.g. variations in the expected demands, that might affect the system.

The optimal control problem is directly formulated as an MPC problem. Let k_c (−) and T_c (s) be the control time step and the control sample time, respectively. Then, the overall optimization problem is formulated as follows:

$$\min_{\tilde{U}_{IJ}(k_c),\tilde{U}_{JI}(k_c)} \quad T_c \cdot \sum_{I\in\mathcal{R},J\in\mathcal{H}_I} \sum_{0}^{K_p-1} N_{II}(k_c) + N_{IJ}(k_c) \tag{12}$$

subject to

$$N_{II}(k_c+1) = N_{II}(k_c) + T_c \cdot \left(Q_{II}(k_c) - M_{II}(k_c) + \sum_{J\in\mathcal{H}_I} U_{JI}(k_c)\cdot M_{JI}(k_c) \right), \tag{13}$$

$$N_{IJ}(k_c+1) = N_{IJ}(k_c) + T_c \cdot \left(Q_{IJ}(k_c) - \sum_{J\in\mathcal{H}_I} U_{IJ}(k_c)\cdot M_{IJ}(k_c) \right), \tag{14}$$

$$U_{IJ,\min} \leqslant U_{IJ}(k_c) \leqslant U_{IJ,\max}, \qquad \text{for } I = 1,2,\ldots,R \text{ and } \forall J \in \mathcal{H}_I. \tag{15}$$

The problem (12)–(15) is a nonlinear optimization problem and it can be solved using nonlinear optimization algorithms. $U_{IJ,\min}$ and $U_{IJ,\max}$ (−) are respectively the lower and upper bounds for the perimeter control inputs between regions I and J. The optimization variables defined over the prediction horizon K_p are $\widetilde{U}_{IJ}(k_c) = [U_{IJ}(k_c),\ldots,U_{IJ}(k_c+K_p-1)]^T$, where $U_{IJ}(k_c+l)$ for $l = 0,\ldots,K_p-1$ are the perimeter control inputs obtained by the MPC framework at every control time step k_c. The following subsection further elaborates the low-level feedback controller, which utilizes the high-level control inputs such that the regions become more homogeneous.

4.2. The low-level feedback homogeneity controller (FHC)

The goal of the high-level (MPC) controller of the hierarchical control framework is to minimize the total network delay. However, there is no explicit consideration of regional accumulation heterogeneity in the high-level control. We also aim at minimizing the regional heterogeneity, and to achieve this goal, a low-level feedback homogeneity controller (FHC) is introduced, where the high-level MPC controller defines the set values for the control inputs. The FHC determines the subregional perimeter controls, u_{ij}, to control the subregional accumulations and minimize the accumulation heterogeneity. Note that subregions, which are not attached to the boundary between the regions cannot be directly controlled, e.g. subregion 19 in Fig. 1(c).

The high- and low-level controllers are not conflicting, but are complementary. The FHC will try to homogenize the region so that circulating flow increases for trips within the region, independently if the MPC controller increases or decreases the transfer flow. This hierarchical scheme can have significant benefits in real-life applications, where the perimeter controller might create strong local heterogeneities and spillbacks for the intersections in the proximity of the border. More homogeneous networks can improve both the average travel time (because of higher network outflow), and the travel time reliability, as shown by Saberi and Mahmassani (2013).

The FHC is a state feedback controller, which its control law is based on the feeded subregional accumulations n_i. While n_{ij} variables are more difficult to estimate, subregional accumulations without information for the final destination are easier to be monitored. The FHC aims at manipulating the subregional controllers u_{ij} to bring subregional accumulations as close as possible to desired accumulations. Defining appropriate desired accumulations should achieve our control goal to homogenize the whole region. It is clear that defining desired accumulations is not a trivial task, since they are not a priori known and may change over time. However, utilizing information from the high-level controller helps us in this task. Given that the high-level controller at each time step k_c predicts the regional accumulation states for K_p step ahead, these predicted accumulations can be considered, after dividing them by the number of subregions within the region, as the subregional accumulation set points for the FHC. The FHC control law for u_{ij} should consider both subregional accumulations n_i and n_j, because u_{ij} affects both subregional accumulations. Hence, the following control law, which is a multivariable integral discrete controller, is proposed:

$$u_{ij}(k_c) = u_{ij}(k_c-1) + K_1 \cdot \left(\frac{N_J(k_c+K_p-1)}{|\mathcal{SR}_J|} - n_j(k_c) \right) - K_2 \cdot \left(\frac{N_I(k_c+K_p-1)}{|\mathcal{SR}_I|} - n_i(k_c) \right), \tag{16}$$

where K_1 and K_2 are positive designed parameters. Note that $N_I(k_c+K_p-1)/|\mathcal{SR}_I|$ and $N_J(k_c+K_p-1)/|\mathcal{SR}_J|$ are the controller set points that change over time. Following the regulating problem, the control gains K_1 and K_2 are designed in this paper assuming that the set points are known constant. When we apply the designed controller, it might be unstable as the set points change over the time, but in this problem since the regional accumulations change smoothly with time (as we see

also later in the result section) and do not experience strong fluctuations, this allows the controller to be effective in tracking the time-varying set points.

The MPC output, U_{IJ}, can be applied in two different options: (i) where each $u_{ij}, i \in I$ and $j \in J$, is equal (or be very close) to U_{IJ} without considering that each sub-region can be treated differently, and (ii) where the collective effect of $u_{ij}, i \in I$ and $j \in J$, is almost equal to U_{IJ}, i.e.

$$\left| \frac{\sum_{i \in I, j \in J} u_{ij}}{|u_{IJ}|} - U_{IJ} \right| < \delta, \tag{17}$$

where $|u_{IJ}|$ denotes the number of subregional controllers between region I and J, and δ is a prescribed positive value, e.g. 0.2. The second option empowers the controller to manipulate each u_{ij} differently and individually in order to minimize the regional heterogeneity. We utilize the second option in the hierarchical framework and compare numerical results, in the following section, with and without considering the low-level controller to highlight its importance. It is known that feedback regulators of type (16) cannot handle directly constraints in the optimization. The FHC first determines the u_{ij}s based on (16), and then truncates the control outputs to satisfy (15) and then if necessary add or subtract a value from all u_{ij}s to also satisfy (17).

5. Comparison of control strategies

In this section, we present a case study example to explore the characteristics of the proposed region-based and subregion-based models along with the hierarchical control scheme. Moreover, we thoroughly investigate the effect of heterogeneity controller, the low-level FHC, on the control strategy performances. Note that the main modeling contribution of this

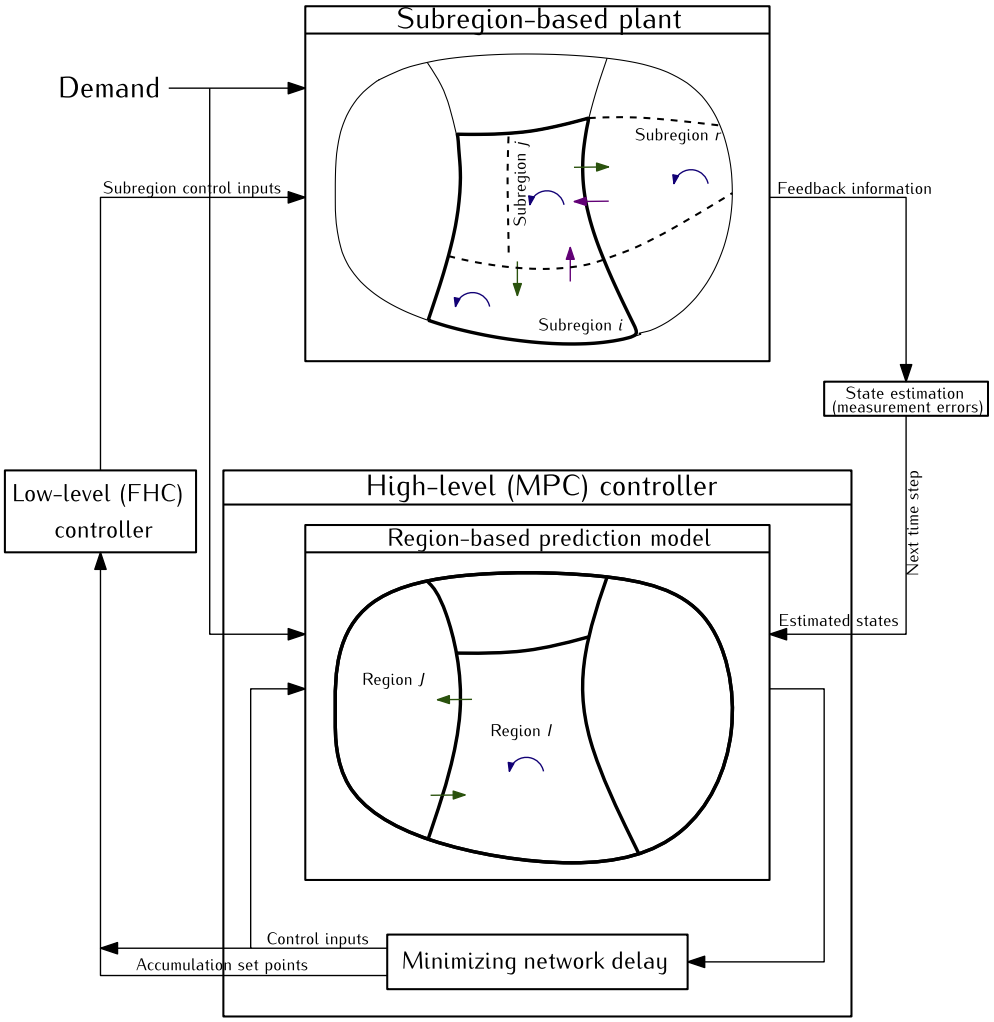

Fig. 5. The hierarchical perimeter control framework.

paper is developing two different models with different scales of aggregation and utilize them in the MPC framework as the prediction model and the plant, in contrast to Geroliminis et al. (2013) and Haddad et al. (2013) in which the dynamics of model and plant in the MPC frameworks were inherently similar, but the demand prediction errors and the MFD noisy scatter distinguish between them. With respect to control, we highlight the importance of hierarchical multi-layer control enhanced with advanced traffic modeling. This section emphasizes that if some parts of the modeling and/or the control are not fully implemented, the network traffic states are worse, especially under congested scenarios.

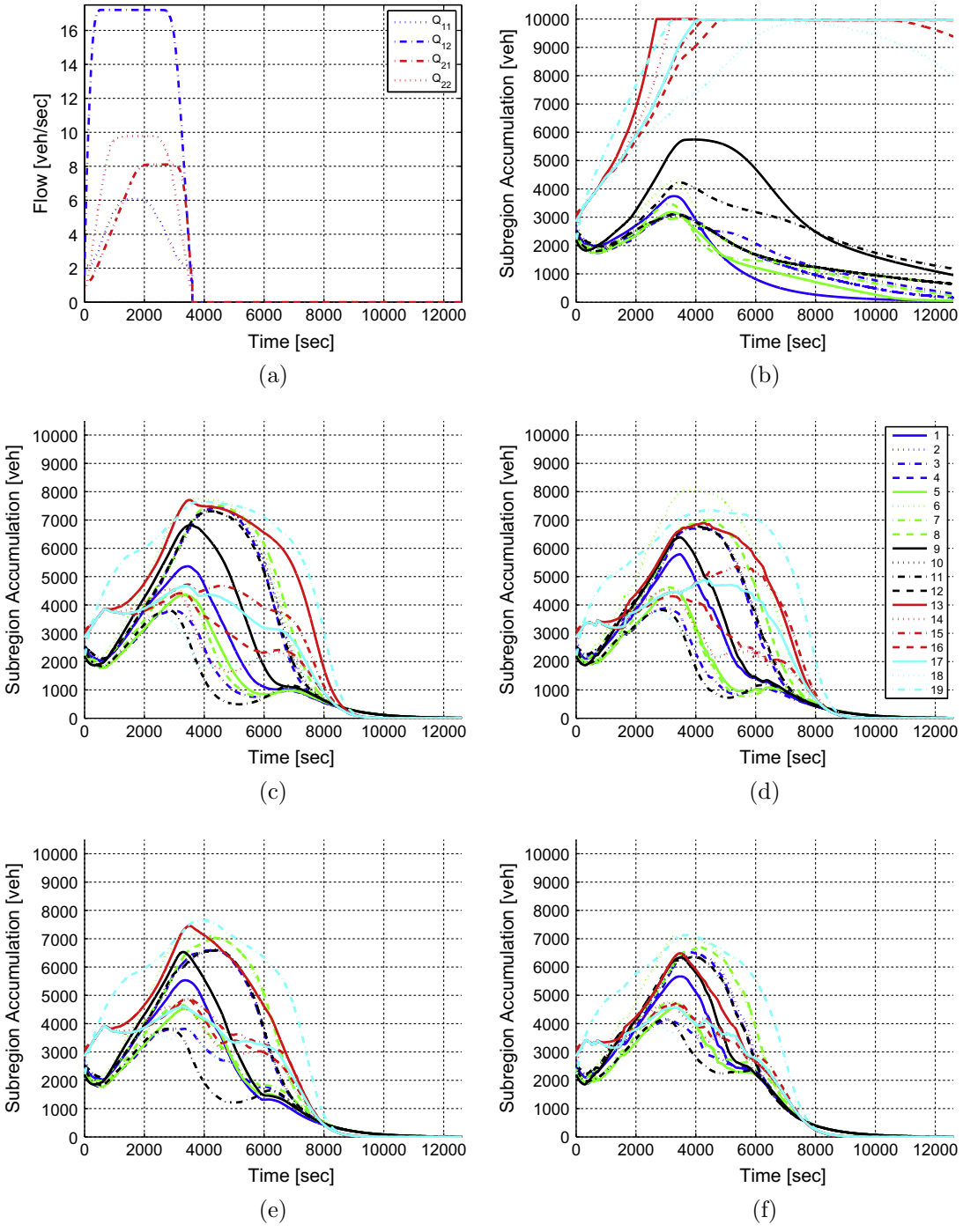

Fig. 6. The case study example with medium demand: (a) the time-varying demand profile, and subregional accumulations with (b) no control, (c) MPC, (d) MPC + HET, (e) MPC + FHC, and (f) MPC + HET + FHC. Note that (b)–(f) have similar legends.

The case study network consists of two regions, designating the periphery and city center of an urban network, each comprises of 12 and 7 subregions, respectively, as schematically shown in Fig. 1(c). Without loss of generality, we assume every subregion has the same (production) MFD consistent with the (production) MFD observed in Yokohama, and consequently, the well-defined relationship between mean and STD of subregion link occupancy exists, and the subregional average trip length is constant. Note that the *region* average trip lengths are varying as the model evolves, see Eqs. (9a) and (9b).

In all numerical runs, every subregion accumulation is initially identical and uncongested that results in initial regional accumulations as $N_1(0) = 29,000$ (veh) (71% of N_1^{crit}) and $N_2(0) = 19,000$ (veh) (80% of N_2^{crit}), where N_I^{crit} denotes the region I accumulation that maximizes the production MFD, P_I. To model the observation errors, a normal random component is added to the measurements from the plant (subregion-based model), i.e. the errors are introduced in $N_{II}(t), N_{IJ}(t), L_{II}(t)$, and $L_{IJ}(t)$, see Fig. 5. Also, a uniform random component is added to the perimeter control sequence to model the flow stochasticity of supply and demand at the boundary of regions, see the small fluctuations in Fig. 10(a)–(d).

The exogenous time varying demand, shown in Fig. 6(a), simulates one hour of morning peak followed by two and half hours of low demand to fully clear the network, while region 1 generates most of the demand towards region 2 that as the central business district attracts trips. Moreover, we consider medium and high demands, where in the high demand the exogenous demand from region 1 to region 2 is 10% higher than the medium demand shown in Fig. 6(a). The selected MPC controller parameters are as follows: the prediction horizon $K_p = 20$, the control lower bound $U_{\min} = 0.1$, and the upper bound $U_{\max} = 0.9$.

We compare five control strategies that are essential to investigate the importance of proper heterogeneity modeling and hierarchical control: (i) no control where there is no restriction on the perimeter transfer flows, (ii) MPC, only the high-level control, ignoring the modeling of heterogeneity and assuming an MFD without any hysteresis (similar to Geroliminis et al., 2013), (iii) MPC + HET, only the high-level control that utilizes the heterogeneity modeling of Section 2, (iv) MPC + FHC, that is the hierarchical control ignoring the heterogeneity modeling by assuming an MFD with no hysteresis, and (v) MPC + HET + FHC, that is the hierarchical control structure that applies MPC at the upper level and FHC at the lower level considering the heterogeneity modeling. These extensive tests enables to highlight the importance of heterogeneity modeling and control on the performance of proposed traffic control for heterogeneous urban networks.

Fig. 6(b)–(f) depict the evolution of subregional accumulations $n_i(t)$ over the simulation duration for the five control strategies in case of medium demand. The gridlock is apparent for no control case as shown in Fig. 6(b), while the rest of strategies manage to clear the network, see Fig. 6(c)–(f). Fig. 6(b) also demonstrates that the accumulation of each subregion cannot exceed the jam accumulation, i.e. 10,000 (veh), which is captured by considering the receiving capacity of subregions in the subregion-based model. A more careful investigation is necessary to compare the different control strategies. Table 1 lists the total network delay (averaged over 5 runs) for the control strategies with the two different levels of demand where the values in parenthesis designate the improvement over the MPC strategy without heterogeneity modeling and without lower level control. It is apparent that utilizing the heterogeneity modeling without integrating the FHC controller to decrease the level of heterogeneity, will improve the system delays, but obviously is worse than the two-level hierarchical control. In addition, the hierarchical control without heterogeneity modeling (i.e. MPC + FHC) provides similar performance with MPC + HET. Thus a careful consideration of heterogeneity in the modeling and control frameworks decreases the total network delay.

To understand why the strategies provide different performance, we initially investigate the MPC strategy. The MPC control strategy ignores the heterogeneity effect in the regional MFD, i.e. the exponential term in (11) is equal to 1. Fig. 7 depicts the control sequences and MFD for medium demand with the MPC control strategy, where strong hysteresis loops are apparent and notably the model MFD is identical in the loading and unloading phases. The Table 1 results indicate 10% and 30% increase in control performance with the MPC + HET + FHC strategy over the MPC strategy respectively in case of medium and high demand. Moreover, assuming the lower envelop of MFD as the optimization model, i.e. the exponential term in (11) is equal to its minimum value, results in worse outcomes. Note that Keyvan-Ekbatani et al. (2012, 2013) are able to keep accumulation at the critical values but cannot avoid hysteresis loops in the MFD with control. Avoiding hysteresis loops can be very beneficial for the overall network performance as we see in the analysis.

Figs. 8–10 highlight in details the importance of the proposed hierarchical control scheme. Note that the following illustrative comparisons are between the MPC + HET control strategy and the two-level hierarchical control strategy (MPC + HET + FHC). MPC + HET is superior to the MPC strategy which has been shown to be superior to standard simple control strategies, such as the "bang bang" control approach (Geroliminis et al., 2013), thus this is a strict test. Fig. 8 presents time-series of accumulations, Fig. 9 shows the corresponding MFDs, and Fig. 10 demonstrates the control actions. This is described in more details in the following paragraphs.

Table 1
Total network delay (veh·sec·10^6).

Demand	No control	MPC	MPC + HET	MPC + FHC	MPC + HET + FHC
Medium	1069.4	573.8 (–)	546.2 (4.8%)	541.9 (5.6%)	518.0 (9.7%)
High	1204.4	930.5 (–)	881.5 (5.3%)	851.0 (8.5%)	636.8 (31.6%)

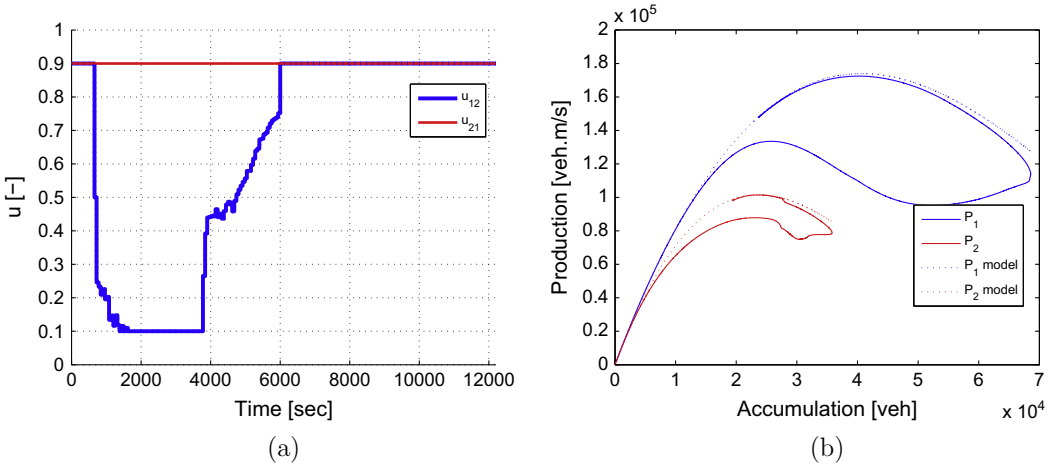

Fig. 7. The case study example for medium demand with the MPC control strategy (without heterogeneity consideration in modeling): (a) control inputs, and (b) MFD productions.

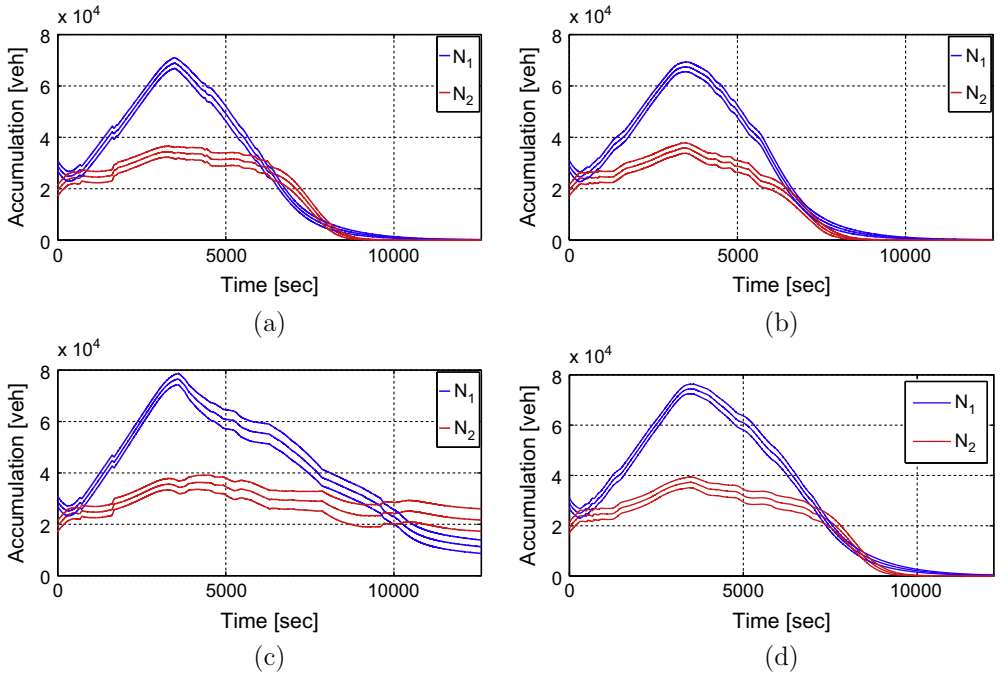

Fig. 8. Accumulation results obtained with MPC + HET control strategy and MPC + HET + FHC for the medium demand in (a) and (b), and for the high demand in (c) and (d), respectively.

The corresponding regional accumulations are illustrated in Fig. 8 for region 1 (center) and region 2 (periphery), while the lines above (or below) of the accumulation curve represent plus (or minus) one STD of the accumulation, a heterogeneity index. The accumulation results obtained with the MPC + HET and MPC + HET + FHC control strategies for the medium demand are depicted in Fig. 8(a) and (b), while the results for the high demand are depicted in Fig. 8(c) and (d), respectively. Accumulations look almost identical in the onset of congestion when heterogeneity index is small (remember that initial conditions have small spatial heterogeneity), however the mean accumulation and the accumulation heterogeneity are different in the offset of congestion. For region 2, the duration of congestion period is shorter and this influences also region 1 in the offset of congestion. Note that region 2 attracts more trips than region 1, even if it has smaller size. In case of high demand scenario with MPC + HET strategy, some subregions face gridlock and the network is not fully cleared, note the residual accumulation at the end of simulation. As MPC + HET + FHC can avoid high level of congestion in subregions, the improvement over MPC + HET is 28%. If initial conditions are also more heterogeneous an even better performance is

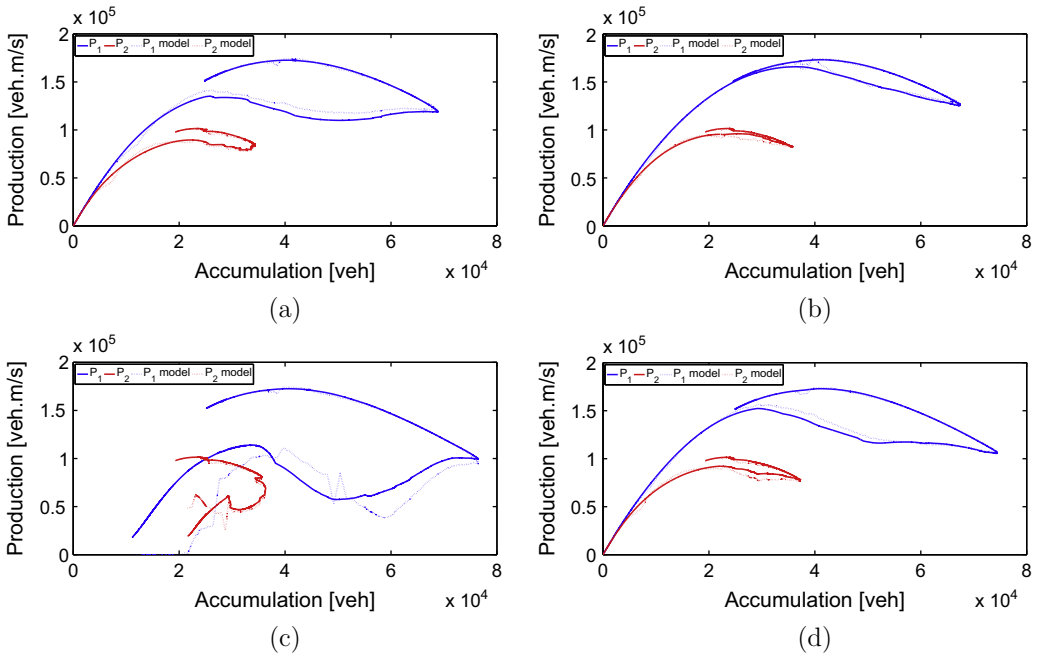

Fig. 9. Production MFD for MPC + HET and MPC + HET + FHC for the medium demand in (a) and (b), and for the high demand in (c) and (d), respectively.

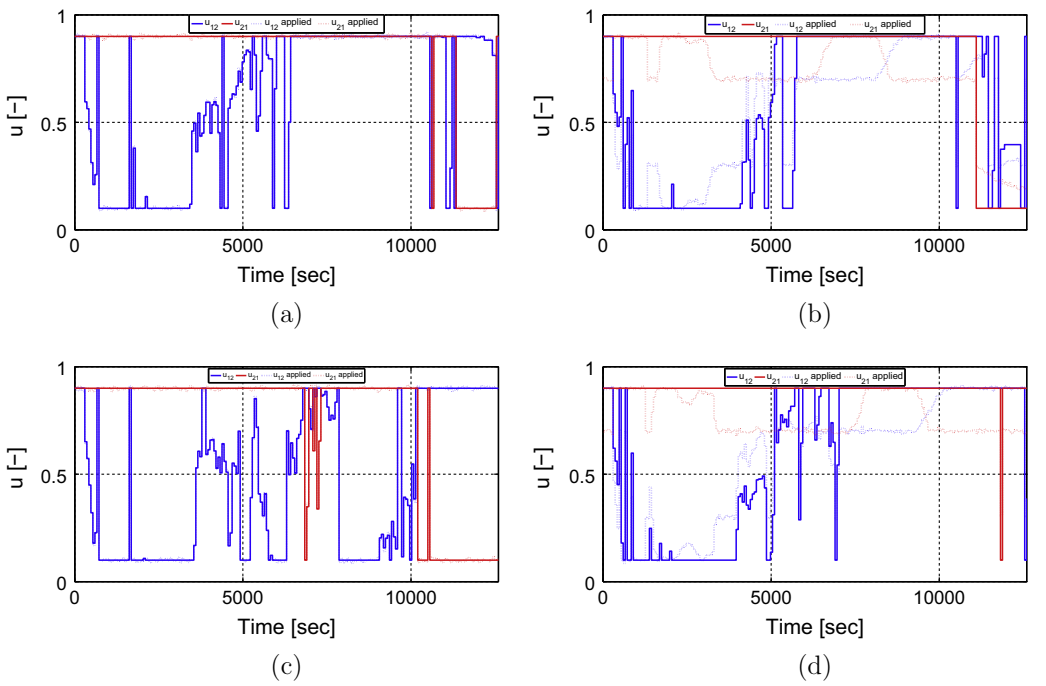

Fig. 10. Control input results obtained with MPC + HET and MPC + HET + FHC for the medium demand in (a) and (b), and for the high demand in (c) and (d), respectively.

expected for the advanced controller. While accumulations do not provide a full picture on how the hierarchical control framework improves the mobility levels, this is clear once MFDs are described for the regions.

The production MFDs obtained with the MPC + HET and MPC + HET + FHC strategies for the medium demand are shown in Fig. 9(a) and (b), and for the high demand in Fig. 9(c) and (d), respectively. Note that the figures show the MFDs obtained by the subregion model (noted as P_1 and P_2) and the MFDs of the region model (noted as P_1 model and P_2 model) as esti-

mated by (11). It is clear that (11) can capture well the effect of heterogeneity in the production MFD of the regions. Evidently, FHC improves the performance of the urban network by minimizing the extent of the hysteresis in both region MFDs during the unloading of the network, as shown in Fig. 9. Hysteresis can be considered as a strong inefficiency of the system as for the same level of vehicles in the network, the performance is worse and vehicles have to spend longer times in the network. Note that while the controller tries to protect region 2, it also succeeds to improve the production (and also the trip completion) for region 1, by decreasing the level of hysteresis to small values. This is succeeded by distributing accumulations in a more uniform pattern by controlling the subregional inter-transfers. This is an important finding as many previous

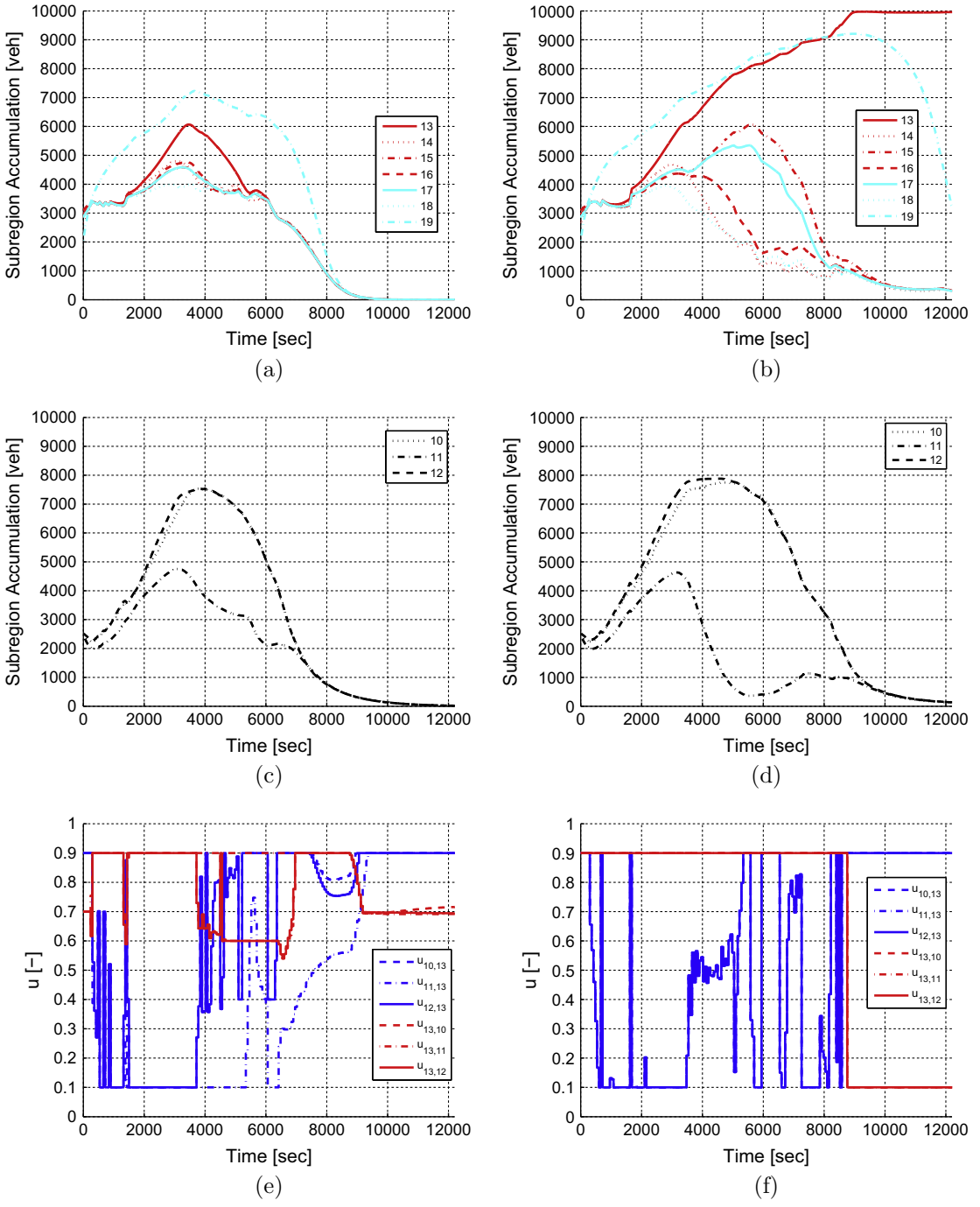

Fig. 11. The case study example with high demand: (a) subregion accumulations in region 2 with FHC, (b) subregion accumulations in region 2 without FHC, (c) subregion 10, 11, 12 accumulations with FHC, (d) subregion 10, 11, 12 accumulations without FHC, (e) control inputs between subregion 13 and 10, 11, 12 with FHC, and (f) control inputs between subregion 13 and 10, 11, 12 without FHC.

investigations in the shape of the MFD with real data and simulations observed hysteresis loops in many cases that result in a decrease in the network performance. It is worth to mention that, the production and outflow MFDs are related with the time-varying region average trip length. This variable affects the dynamics of MFDs, notably the hysteresis loop shape, i.e. clockwise vs. counterclockwise. Average trip length has a significant role in our understanding of urban traffic dynamics and reveals substantial information on the human mobility pattern in urban area. Further research on this direction could shed more light on properties of urban average trip length distribution.

Fig. 10 depicts the control sequences of MPC + HET and MPC + HET + FHC strategies at the regional level (i.e. U_{12} and U_{21}) for medium (a and b) and high demand (c and d). Each figure contains the estimated values of U_{12} and U_{21} by the optimization model (12)–(15), while U_{12} applied and U_{21} applied show the ones implemented utilizing the FHC (16) and (17). Fig. 10(a) and (c) do not contain the FHC, thus the two lines are almost identical (except some small random error during implementation). Nevertheless, in the FHC case (Fig. 10(b) and (d)) differences are substantial as the FHC tries to equalize subregion accumulations and it is allowed to deviate by the factor δ (see (17)) from the estimated values. The control sequences show similar trends. For instance, at the very beginning of the control process, the controllers do not restrict inter flow transfers since both regions are uncongested. While afterwards, as region 2 becomes more congested and attracts more trips, the controllers attempt to protect region 2 accumulation by changing U_{12} from $U_{\max}$ to $U_{\min}$ in a smooth manner, since without any restriction region 2 will face gridlock, see the subregional accumulation with no control in Fig. 6(b). The overall situation remains invariant till the end of morning demand $t = 3600$ (s), then because of decrease in the demand (unloading phase), regions shift towards the uncongested state. Thus, the controllers gradually permit more vehicles to enter to the city center, region 2, by altering U_{12} to $U_{\max}$. Though the similar trend, the MPC + HET + FHC applied control sequence (dashed lines in Fig. 10(b) and (d)) deviate around the MPC control value, see (17). The FHC modifies the MPC control sequence to manipulate transfer flows between subregions. This offers a flexible framework to directly control the heterogeneity. The dashed lines are the applied perimeter control inputs that achieve the twofold objectives of the hierarchical (MPC + HET + FHC) control strategy, to minimize the total network delay by simultaneously maximizing the network outflow and minimizing the regional accumulation heterogeneity. We test the same experiment study with δ equal to infinity, see (17), which means no restriction on FHC to follow the MPC control values. This results in less heterogeneity in regions however the total network delay is worse than the case with $\delta = 0.2$. The outcome demonstrates that selecting the goal of the controller solely to make regions more uncongested or solely more homogeneous in a network and ignore other aspects (e.g. overall network state, future prediction of traffic state, etc.) is not beneficial.

While results in Figs. 8–10 focus on the regional characteristics of the system, to further investigate the effect of FHC, the accumulation of subregions in region 2 for high demand case with MPC + HET + FHC (left column) and MPC + HET (right column) are shown in Fig. 11. Apparently, MPC + HET + FHC brings the accumulation of all subregions to zero by the end of the experiment whereas without FHC, subregion 13 goes to gridlock. To investigate the dynamics of subregion 13, Fig. 11(c) and (d) depict subregion 10, 11, 12 (the neighbors of subregion 13 in region 1) accumulations and Fig. 11(e) and (f) depict control inputs between subregion 13 and subregion 10, 11, 12, respectively for MPC + HET with and without FHC. As expected, control inputs between region 1 and region 2, i.e. $u_{10,13}, u_{11,13}, u_{12,13}$, are identical in case of MPC without FHC (similarly for control inputs between region 2 and region 1). However, FHC provides a traffic-responsive perimeter flow control strategy that is based on subregion traffic states to distribute the traffic congestion more efficiently among subregions. Consequently, FHC homogenizes region 2 accumulation and stabilizes subregion 13 accumulation by manipulating control inputs (see Fig. 11(e)), such that the dispersion of subregion accumulations over time is smaller compared to Fig. 11(b). It is evident that subregion 19 accumulation is less consistent with other subregions as it is uncontrollable with perimeter controllers. Note that in Fig. 11(e) and (f) the noise in control input is eliminated for illustration purposes.

6. Conclusion and future work

This paper has presented two urban traffic models based on the MFD at different levels of spatial aggregations to model the dynamics of density heterogeneity. A heterogeneous urban region can be partitioned into homogeneous subregions as the detailed model aims at modeling the accumulation dynamics of subregions, while the dynamics of urban regions are modeled in an aggregated manner.

We utilize the subregion- and region-based model as the plant and the optimization model in the MPC framework to formulate the optimal perimeter control for urban regions. We integrate variable perimeter control inputs for each subregion in the region boundary to actively control the density heterogeneity. The results in this paper can be utilized to develop efficient hierarchical control strategies for heterogeneously congested cities. Another research direction is related to model route choice with experienced travel time estimation and identification of equilibrium conditions with the MFD concept. (A recent attempt can be seen in Yildirimoglu and Geroliminis (2014).)

A challenging modeling direction is how to describe the aggregated modeling dynamics of regions, i.e. (2) and (3), when routes pass through the same regions more than once (e.g. a trip sequence of subregions in Fig. 1). This is currently under investigation and the dynamic framework will be considered in a future publication. All recent efforts related to control and MFD for multi-region networks assume a stationary boundary in time and space. In cases of congestion propagation in time and space and formation of congestion in different regions of a city (see e.g. Ji et al., 2014), a dynamic partitioning associated with a dynamic boundary control framework should be studied. For example, in our analysis subregion 19 was

uncontrollable as it was not close to the boundary between the two regions, hence high accumulations where observed during the evolution of congestion. A control strategy with dynamic boundaries is expected to further improve homogeneity and network performance. This is a research priority. Another challenging direction is to further investigate the effect of regional route choice in the modeling and control of MFDs (see for example Gayah and Daganzo (2011b) and Leclercq and Geroliminis (2013) for simple networks). A field test will shed more light in the applicability of the above methodologies and will create additional modeling and control insights. Such a test is under preparation in two large scale experiments.

Acknowledgement

This research was partially supported by ERC Starting Grant "METAFERW: Modeling and controlling traffic congestion and propagation in large-scale urban multimodal networks".

References

Aboudolas, K., Geroliminis, N., 2013. Perimeter and boundary flow control in multi-reservoir heterogeneous networks. Transportation Research Part B 55, 265–281.
Buisson, C., Ladier, C., 2009. Exploring the impact of homogeneity of traffic measurements on the existence of macroscopic fundamental diagrams. Transportation Research Record 2124, 127–136.
Clark, J.S., 1998. Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. The American Naturalist 152 (2), 204–224.
Daganzo, C.F., 1994. The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Transportation Research Part B 28 (4), 269–287.
Daganzo, C.F., 2007. Urban gridlock: macroscopic modeling and mitigation approaches. Transportation Research Part B 41 (1), 49–62.
Daganzo, C.F., Geroliminis, N., 2008. An analytical approximation for the macroscopic fundamental diagram of urban traffic. Transportation Research Part B 42 (9), 771–781.
Doan, M., Keviczky, T., Schutter, B., 2014. A hierarchical MPC approach with guaranteed feasibility for dynamically coupled linear systems. In: Maestre, J.M., Negenborn, R.R. (Eds.), Distributed Model Predictive Control Made Easy, Intelligent Systems, Control and Automation: Science and Engineering, vol. 69. Springer, Netherlands, pp. 393–406.
Gayah, V.V., Daganzo, C.F., 2011a. Clockwise hysteresis loops in the macroscopic fundamental diagram: an effect of network instability. Transportation Research Part B 45 (4), 643–655.
Gayah, V.V., Daganzo, C.F., 2011b. Exploring the effect of turning maneuvers and route choice on a simple network. Transportation Research Record 2249, 15–19.
Gayah, V.V., Gao, X.S., Nagle, A.S., 2014. On the impacts of locally adaptive signal control on urban network stability and the macroscopic fundamental diagram. Transportation Research Part B 70, 255–268.
Geroliminis, N., Boyacı, B., 2012. The effect of variability of urban systems characteristics in the network capacity. Transportation Research Part B 46 (10), 1607–1623.
Geroliminis, N., Daganzo, C.F., 2007. Macroscopic modeling of traffic in cities. In: Transportation Research Board 86th Annual Meeting. Washington, DC, Paper No. 07-0413.
Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transportation Research Part B 42 (9), 759–770.
Geroliminis, N., Sun, J., 2011a. Hysteresis phenomena of a macroscopic fundamental diagram in freeway networks. Transportation Research Part A 45 (9), 966–979.
Geroliminis, N., Sun, J., 2011b. Properties of a well-defined macroscopic fundamental diagram for urban traffic. Transportation Research Part B 45 (3), 605–617.
Geroliminis, N., Haddad, J., Ramezani, M., 2013. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: a model predictive approach. IEEE Transactions on Intelligent Transportation Systems 14 (1), 348–359.
Godfrey, J.W., 1969. The mechanism of a road network. Traffic Engineering and Control 11 (7), 323–327.
Haddad, J., Geroliminis, N., 2012. On the stability of traffic perimeter control in two-region urban cities. Transportation Research Part B 46 (1), 1159–1176.
Haddad, J., Shraiber, A., 2014. Robust perimeter control design for an urban region. Transportation Research Part B 68, 315–332.
Haddad, J., Ramezani, M., Geroliminis, N., 2013. Cooperative traffic control of a mixed network with two urban regions and a freeway. Transportation Research Part B 54, 17–36.
Ji, Y., Geroliminis, N., 2012. On the spatial partitioning of urban transportation networks. Transportation Research Part B 46 (10), 1639–1656.
Ji, Y., Daamen, W., Hoogendoorn, S., Hoogendoorn-Lanser, S., Qian, X., 2010. Macroscopic fundamental diagram: investigating its shape using simulation data. Transportation Research Record 2161, 42–48.
Ji, Y., Luo, J., Geroliminis, N., 2014. Empirical observations of congestion propagation and dynamic partitioning with probe data for large scale systems. Transportation Research Record 2422 (2), 1–11.
Keyvan-Ekbatani, M., Kouvelas, A., Papamichail, I., Papageorgiou, M., 2012. Exploiting the fundamental diagram of urban networks for feedback-based gating. Transportation Research Part B 46 (10), 1393–1403.
Keyvan-Ekbatani, M., Yildirimoglu, M., Geroliminis, N., Papageorgiou, M., Oct 2013. Traffic signal perimeter control with multiple boundaries for large urban networks. In: 16th International IEEE Conference on Intelligent Transportation Systems – (ITSC), 2013, pp. 1004–1009.
Knoop, V.L., Hoogendoorn, S.P., Van Lint, J.W.C., 2012. Routing strategies based on the macroscopic fundamental diagram. Transportation Research Record 2315, 1–10.
Knoop, V., Hoogendoorn, S., van Lint, H., 2013. The impact of traffic dynamics on the macroscopic fundamental diagram. In: 92nd Annual Meeting of Transportation Research Board. Washington, DC, USA.
Leclercq, L., Geroliminis, N., 2013. Estimating MFDs in simple networks with route choice. Transportation Research Part B 57, 468–484.
Leclercq, L., Chiabaut, N., Trinquier, B., 2014. Macroscopic fundamental diagrams: a cross-comparison of estimation methods. Transportation Research Part B, 1–12.
Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E., Getz, W.M., 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438, 355–359.
Mahmassani, H., Williams, J., Herman, R., 1987. Performance of urban traffic networks. In: Gartner, N., Wilson, N. (Eds.), Proceedings of the 10th International Symposium on Transportation and Traffic Theory. Elsevier, Amsterdam, The Netherlands.
Mahmassani, H.S., Saberi, M., Zockaie, A., 2013. Urban network gridlock: theory, characteristics, and dynamics. Transportation Research Part C: Emerging Technologies 36, 480–497.
Mazloumian, A., Geroliminis, N., Helbing, D., 2010. The spatial variability of vehicle densities as determinant of urban network capacity. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368 (1928), 4627–4647.
Olszewski, P., Fan, H.S.L., Tan, Y.-W., 1995. Area-wide traffic speed-flow model for the Singapore CBD. Transportation Research Part A 29A (4), 273–281.

Ortigosa, J., Menendez, M., Tapia, H., 2014. Study on the number and location of measurement points for an MFD perimeter control scheme: a case study of Zurich. EURO Journal on Transportation and Logistics 3 (3–4), 245–266.
Saberi, M., Mahmassani, H.S., 2013. Empirical characterization and interpretation of hysteresis and capacity drop phenomena in freeway networks. Transportation Research Record (2391), 44–55.
Yildirimoglu, M., Geroliminis, N., 2014. Approximating dynamic equilibrium conditions with macroscopic fundamental diagrams. Transportation Research Part B: Methodological 70, 186–200.
Zhang, L., Garoni, T., de Gier, J., 2013. A comparative study of macroscopic fundamental diagrams of arterial road networks governed by adaptive traffic signal systems. Transportation Research Part B 49, 1–23.