https://www.sciencedirect.com/science/article/pii/S0191261518301152

https://doi.org/10.1016/j.trb.2018.10.007

Hierarchical Control of Heterogeneous Large-scale Urban Road Networks via Path Assignment and Regional Route Guidance

Mehmet Yildirimoglu[a,*], Isik Ilber Sirmatel[b], Nikolas Geroliminis[b]

[a] *School of Civil Engineering*
The University of Queensland, Australia
[b] *Urban Transport Systems Laboratory*
School of Architecture, Civil and Environmental Engineering
École Polytechnique Fédérale de Lausanne, Switzerland

Abstract

High level of detail renders microscopic traffic models impractical for control purposes and local control schemes cannot coordinate actions over large scale heterogeneously congested urban networks. Developing efficient models and control methods for large-scale urban road networks is, therefore, an important research challenge. Alleviating congestion via manipulation of traffic flows or assignment of vehicles to specific paths has a great potential in achieving efficient network usage. Motivated by this fact, this paper proposes a hierarchical traffic management system. The upper-level route guidance scheme optimizes network performance based on actuation via regional split ratios, whereas the lower-level path assignment mechanism recommends subregional paths for vehicles to follow, satisfying the regional split ratios in order to achieve said performance. Simulation results from a 49-subregion or 7-region network shows a great potential of the proposed scheme in achieving coordination and efficient use of network capacity, leading to increased mobility.

Keywords: Hierarchical control, Route guidance, Model predictive control, Macroscopic fundamental diagram, Large-scale urban networks

*corresponding author

1. Introduction

Real-time management of urban road traffic is becoming increasingly important with accelerating urbanization leading to ever higher levels of congestion in cities. Urban traffic control literature usually focuses on link-level dynamics and local control laws, which present difficulties for analysis and synthesis of large-scale traffic management schemes. The high level of detail in microscopic models, although desirable for simulation studies, leads to highly complex and thus possibly intractable models ill-suited for real-time control purposes. Furthermore, local control schemes, although successful in managing local traffic in undersaturated conditions, fail to achieve coordination with other parts of the urban network. These shortcomings emphasize the need for developing aggregated models of large-scale urban networks and designing network-level control schemes that can efficiently manage major components of the city traffic, while at the same time providing a realistic coordination between the actions of the network-level controller and the commands sent to individual drivers for manipulating the macroscopic behavior.

The modeling approach adopted in this paper is based on the macroscopic fundamental diagram (MFD) of urban traffic. First proposed by Godfrey (1969), and later shown to exist under dynamic conditions in urban areas with homogeneous distribution of congestion in Geroliminis and Daganzo (2008), the MFD provides a unimodal, low-scatter, and demand-insensitive relationship between accumulation and production (and possibly trip completion flow) for an urban region. Despite having shortcomings related to heterogeneity (i.e., the region should have low heterogeneity for a well-defined MFD to exist) and hysteresis (the MFD may behave differently on the onset and offset of congestion), the MFD is a powerful modeling tool that enables the development of low-complexity dynamical models for large-scale urban road networks, leading to the design of network-level traffic management and control schemes.

Modeling and control of large-scale urban road networks via MFD receives increasing attention in the transportation community. Perimeter control strategies, i.e. restricting the flow at the boundary of urban regions, using the concept of MFD have been particularly attracting significant interest. Daganzo (2007) develops a bang-bang controller for a single region network and ensures that the average density remains in the uncongested part of the diagram. Keyvan-Ekbatani et al. (2012) develop a single region PI controller and implement the perimeter control in a microscopic simulation

environment. Keyvan-Ekbatani et al. (2015) consider a multiple concentric congestion scenario and develop a 2-level controller. Aboudolas and Geroliminis (2013) extend the problem to a multi-region framework and target critical accumulation values in all regions. However, this might not be optimal or feasible under heterogeneously congested conditions. Kouvelas et al. (2017) overcome this discrepancy using a data-driven optimization technique and estimate the optimal gain values and set points. Alternatively, Geroliminis et al. (2013) formulate the optimal perimeter control problem within a model predictive control framework and implement it on a 2-region network. Using the same framework, Ramezani et al. (2015) incorporate the heterogeneity aspects into the MPC framework and show that a 2-level controller can further improve conditions. In addition, Haddad and Shraiber (2014); Zhong et al. (2017) study the robustness and Haddad (2017a,b) investigate the optimality of the perimeter control strategies. Yang et al. (2017) considers the delay at the intersections along the perimeter and optimizes the network performance as a whole. These studies rely on the aggregated network dynamics expressed via MFD and herald the progress towards a new generation of network-level traffic control schemes.

Studies on network traffic state estimation via MFD using various sensors point out the advantages of an aggregated approach in face of limited data, see (Gayah and Dixit, 2013; Ortigosa et al., 2014; Nagle and Gayah, 2014; Leclercq et al., 2014; Ji et al., 2014; Saberi et al., 2014). And, clustering techniques to partition a heterogeneously congested city into homogeneous regions with well defined MFDs can be found in Saeedmanesh and Geroliminis (2016), Saeedmanesh and Geroliminis (2017), Lopez et al. (2017), An et al. (2017) and others. MFD-based models are also exploited to develop optimum pricing strategies, see Zheng et al. (2012, 2016).

Works relying on MFD models and exploring actuation via route guidance started appearing relatively recently in the literature. Simple routing strategies are studied in Gayah and Daganzo (2011) and Leclercq et al. (2013) for two-bin or two-route network abstractions. Management of grid networks without traffic lights is explored in Knoop et al. (2012) using MFD-based routing strategies. An optimal route guidance scheme based on model predictive control is developed in Hajiahmadi et al. (2013). A network-level MPC with integrated perimeter control and regional traffic distribution is proposed in Sirmatel and Geroliminis (2018) for alleviating congestion in a multi-region framework. Ramezani and Nourinejad (2017) develops an MFD-based taxi dispatch system to improve the taxi service performance

and to reduce traffic congestion. Menelaou et al. (2017) develop a heuristic solution to the route reservation problem that avoids traffic congestion and minimizes the travel time. These studies build on MFD's capability of representing large-scale traffic with few variables and aim for an improved routing configuration in the network. Nevertheless, as MFD represents the traffic at an aggregated scale, the resulting configuration at an upper-level (i.e. regional traffic distribution) needs to be translated to a lower-level signal (i.e. path) that can be followed by individual drivers. Yildirimoglu et al. (2015) build an iterative route guidance strategy (based on the assignment model in Yildirimoglu and Geroliminis (2014)) that addresses the hierarchical nature of the problem; however, this method presents computational difficulties and convergence issues especially in high congestion scenarios. This study, on the other hand, aims for an elegant optimization framework and moves toward an efficient hierarchical management scheme.

In this paper, we propose a hierarchical traffic management scheme for alleviating congestion and improving mobility in large-scale urban road networks. We develop an MFD-based strategy that considers optimization of routing variables at the upper level and manipulation of traffic flows via path assignment at the lower level. Previous works on MFD-based route guidance consider only a centralized aggregated approach. Their approach is important to test the feasibility of the developed strategies at the macro-level, but does not provide any insights about the implementation at a lower disaggregated level. This is a general criticism towards the MFD-based studies and there are recent efforts to apply some of these frameworks in more detailed case studies (see for example the work of Keyvan-Ekbatani et al. (2015) for equalizing queues at the boundary intersections associated with perimeter control, the work of Kouvelas et al. (2017) to apply a more complex MFD control framework in microsimulation, the work of DePrator et al. (2017) to see how the effect of left turns in a microscopic modeling environment influence the shape of the MFD). Nevertheless, there is no effort so far (to the best of our knowledge) to see how MFD-based route guidance can be implemented at a lower level. In our framework, the upper level employs an economic model predictive control (MPC) scheme (building on the work in Sirmatel and Geroliminis (2018)) with route guidance actuation at the upper layer considers dynamics with region MFDs and computes the regional split ratios (i.e., the percentages of vehicle flows exiting a region and entering a particular neighbouring region) that minimize the total time spent (TTS) in the network for a finite time horizon. An economic MPC scheme differs from

standard MPC formulations as in the former objective functions expressing economically optimal plant operation (such as minimizing time spent) are considered, whereas in the latter the objective function is related to standard control objectives such as stabilization or setpoint tracking. The lower level involves a decentralized subregional path assignment mechanism that computes the paths for vehicles appearing in each region such that the regional split ratios from the upper layer are achieved in the network. This mechanism is formulated as an optimization problem where the ordered control actuation from the upper layer is matched with the reconstructed or aggregated actions at the lower level. The study also considers a *plant* (i.e., the simulation model representing reality) where detailed paths and accumulation based subregion-level models are used (as proposed in Yildirimoglu et al. (2015)) and a *model* where control decisions are made with more aggregated, exclusively accumulation based region-level model of MFD dynamics. The *model* contains states that are easier to be estimated with multi-sensor data compared to the *plant* states. This region model is integrated with the heterogeneity variance term (as developed in Ramezani et al. (2015)) for modeling the effect of decreased outflow with increasing heterogeneity, and used as a prediction tool in the regional route guidance MPC.

The contributions of this paper are twofold: (1) we propose a decentralized integer linear programming (ILP) based path assignment mechanism that translates a set of desired regional split ratios (each specifying the percentage of flow exiting the region through a specific neighboring region) to the subregion-level paths and assigns vehicles appearing in the region to those paths, (2) we integrate the regional route guidance MPC with heterogeneity variance and dynamic average trip lengths so as to consider the effects of heterogeneity and heteroscedasticity in the MFD-based region model.

(1) While MFD-based control strategies lean on manipulation of aggregated or macroscopic flows, a route guidance strategy requires disaggregated or microscopic instructions to be provided to drivers. The ILP-based subregional path assignment mechanism we develop here is the first attempt to build the link between these two levels using an optimization framework. Note that the work in Sirmatel and Geroliminis (2018) does not consider the model-plant mismatch and works with the assumption that individual vehicle flows can be regrouped to execute the ordered split ratios in a region. Although the desired regional split ratios can be guaranteed by random sampling of vehicles and assigning them to the corresponding paths, the impact of such a strategy on the heterogeneity and the average trip distance in regions

might be detrimental to the performance of the network. Driver acceptability of these strategies should also be investigated as there is no guarantee that the proposed route has a better travel time at the lower level (MFD does not consider the exact routes of individuals). Therefore, there is need for a novel mechanism that can properly distribute the individual vehicle flows and guarantee the desired regional split values while maintaining the heterogeneity and average distance measures.

(2) The regional route guidance MPC proposed in Sirmatel and Geroliminis (2018) does not consider any effects related to heterogeneity and assumes a constant average trip length for each region. However, as the network considered in this paper is made of subregions, the region MFDs are adversely affected from the resulting heterogeneous distribution of congestion among them and are subject to dynamical variations of regional average trip lengths. To consider these two effects within the regional route guidance MPC framework, the formulation of Sirmatel and Geroliminis (2018) is extended in this paper to include the heterogeneity variance effects and dynamic average trip lengths (as proposed in Yildirimoglu et al. (2015) and Ramezani et al. (2015)). The joint operation of the ILP-based path assignment mechanism and the regional route guidance MPC as a hierarchical traffic management scheme is tested in simulations for heterogeneously congested conditions in a large-scale urban network with MFD dynamics, the results of which indicate the potential of the proposed scheme in alleviating congestion and improving mobility in urban networks.

The remainder of this paper is structured as follows. In Section 2, we present the two levels of traffic modeling; region and subregion model. In Section 3, we elaborate on the mechanism of the hierarchical traffic management system that consists of regional route guidance MPC and ILP-based subregional path assignment. In Section 4, we discuss the results from a case study. And, we conclude the paper in the last section with final remarks.

2. MFD-Based Modeling of Large-Scale Urban Networks

In this section, we introduce two traffic models: (i) a region model considering an urban network partitioned into a small number of regions, and (ii) a subregion model defining dynamics for a far more detailed network representation where the above regions are divided into smaller subregions. A network consisting of 7 regions and 49 subregions is schematically shown in Figure 1.

To build a hierarchical control scheme, we use the approach of Ramezani et al. (2015) with two modeling levels; subregion model representing the traffic reality and region model representing the operation or prediction model. In the subregional representation, the network is actually not partitioned into 49 subregions, but simply this collection of 49 subregions is the network itself. In other words, the subregions, for the purposes of this work, are the smallest particles and represent the network itself. The subregion dynamics, together with the presumed subregional MFD parameters, are considered to represent the reality of the urban network, where we assume that: (1) There is no internal routing inside a subregion (as details beyond the subregion are not modeled), (2) there are no interactions between the subregional average trip lengths and the path assignment decisions (the average distance the vehicles cross in a subregion is assumed constant and same for everyone), (3) the traffic performance is represented by a stable MFD (note that we also test the proposed model later with some scatter in subregion MFD). In a more detailed representation (e.g., microscopic simulation), subregions can be replaced with links (sections between intersections), where also there is no route choice (the only option is to cross the whole link), trip distance is the same for all users and fundamental diagram is stable. Therefore, one can make an analogy between subregions and links. In contrast to this, in the region model, we assume that the network (as represented by the 49 subregions) is partitioned into 7 regions for control purposes, and this is the actual partitioning considered in the paper. Region sizes are important in the sense that partitioning the network into a large number of regions could potentially lead to computational problems regarding the route guidance MPC. In Sirmatel and Geroliminis (2018), the computational efficiency results suggest that for a network partitioned into 7 regions, the MPC schemes retain real-time feasibility, whereas a number much more than 7 could be expected to yield intractability problems.

2.1. Region Model

Consider an urban network $\mathcal{R}$ with heterogeneous distribution of accumulation, with a given partition into R regions, i.e., $\mathcal{R} = \{1, 2, \dots, R\}$. Inflow demand generated in region I with destination J is $Q_{IJ}(t)$ (veh/s), whereas $N^H_{IJ}(t)$ (veh) is the accumulation in region I with destination J that is going to transfer from I to H, with $N_I(t)$ (veh) expressing the total regional accumulation in I, at continuous time t; $I, J \in \mathcal{R}$; $N_I(t) \triangleq \sum_{J\in\mathcal{R}} \sum_{H\in\mathcal{V}_I} N^H_{IJ}(t)$, where $\mathcal{V}_I$ is the set of regions adjacent to I. For each region we consider

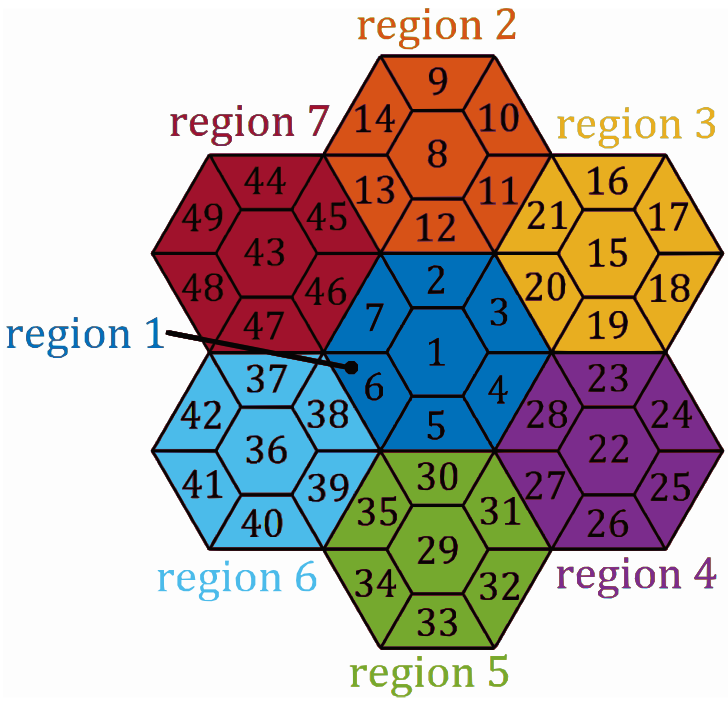

Figure 1: Schematic of a multi-region urban network, consisting of 7 regions each with 7 subregions (region 1 contains subregions 1 to 7, region 2 contains 8 to 14, etc.).

regional split ratios $\theta_{IJ}^{H}(t)$ (for $I, J \in \mathcal{R}$, $H \in \mathcal{V}_I$), that can distribute the transfer flows entering region I with destination J over neighboring regions $H \in \mathcal{V}_I$. Note here that in contrast to previous works (Yildirimoglu et al., 2015; Sirmatel and Geroliminis, 2018), where the regional split ratio $\theta_{IJ}^{H}(t)$ expresses a distribution of the flow exiting region I over its neighbors, in the present work it distributes the flow *entering* region I. This definition of $\theta_{IJ}^{H}(t)$ is consistent with the definition of the path assignment control inputs for the subregional model, which will be described in the next sections.

Mass conservation equations for an R-region MFDs network are:

$$\dot{N}_{II}^{H}(t) = \theta_{II}^{H}(t)\left(Q_{II}(t) + \sum_{G\in\mathcal{V}_I} M_{GI}^{I}(t)\right) - M_{II}^{H}(t) \quad H \in \mathcal{V}_I \cup I \tag{1a}$$

$$\dot{N}_{IJ}^{H}(t) = \theta_{IJ}^{H}(t)\left(Q_{IJ}(t) + \sum_{G\in\mathcal{V}_I} M_{GJ}^{I}(t)\right) - M_{IJ}^{H}(t) \quad H \in \mathcal{V}_I,\ I \neq J, \tag{1b}$$

for $I, J \in \mathcal{R}$. $M_{II}^{I}(t)$ (veh/s) is the exit (i.e., internal trip completion) flow from region I to destination I (exiting without leaving region I), whereas $M_{IJ}^{H}(t)$ (veh/s) is the flow transferring from I to H with destination J. The exit and transfer flows can be expressed as follows (following Ramezani et al. (2015)):

$$M_{II}^{I}(t) = \frac{N_{II}^{I}(t)}{N_I(t)} \frac{F_I(N_I(t))}{L_{II}(t)} \rho_I(N_I(t), \sigma(N_I(t))) \tag{2a}$$

$$M_{IJ}^{H}(t) = \frac{N_{IJ}^{H}(t)}{N_I(t)} \frac{F_I(N_I(t))}{L_{IH}(t)} \rho_I(N_I(t), \sigma(N_I(t))), \tag{2b}$$

where $F_I(\cdot)$ (veh.m/s) is the production MFD of region I as a function of regional accumulation $N_I(t)$ (for a 3rd degree polynomial approximation, the MFD is of the form $F_I(N_I(t)) = A\ N_I(t)^3 + B\ N_I(t)^2 + C\ N_I(t)$, where A, B, and C are the MFD parameters estimated from data), $L_{II}(t)$ and $L_{IH}(t)$ are the average trip lengths for internal trips inside region I and transferring trips from I to H, respectively, whereas $\rho_I(\cdot) \in [0, 1]$ is the heterogeneity coefficient of region I expressing the decrease in production due to heterogeneity ($\rho_I(\cdot) = 1$ if region I is perfectly homogeneous and it decreases with increasing heterogeneity), which can be formulated as follows (see Ramezani et al. (2015) for details):

$$\rho_I(N_I(t), \sigma(N_I(t))) = \beta \cdot (\mathrm{e}^{\gamma \cdot (\sigma(N_I(t)) - \sigma_{\mathrm{h}})} - 1) + 1 \quad \forall I \in \mathcal{R}, \tag{3}$$

where $\sigma(N_I(t))$ is the heterogeneity variance of region I, σ_{h} is the standard deviation of summation of negative binomial distributions of the subregions of region I with mean occupancy $N_I(t)/|\mathcal{I}|$ (with $\mathcal{I}$ the set of subregions in region I and $|\mathcal{I}|$ its size), whereas β and γ are estimated parameters describing the effect of heterogeneity in link density on the production of the region. Analyses based on real data demonstrate that the negative binomial distribution can provide accurate estimations for mean and standard deviation of occupancies for the Yokohama network (Ramezani et al., 2015). That means, one can accurately estimate the production in a region using two terms; (i) an upper bound (low-scatter) MFD (i.e. $F_I(N_I(t))$) and (ii) the heterogeneity degradation (i.e. ρ_I). While $F_I(N_I(t))$ can be represented with a 3$^{\mathrm{rd}}$-degree polynomial function, ρ_I is modeled with an exponential function. The parameters of these functions are network-specific values and might exhibit changes in different applications; however, it is important that

one uses a consistent set of parameters based on the same network and data set.

The regional split ratios $\theta_{IJ}^{H}(t)$ are control inputs which are to be computed by the network-level route guidance MPC, the design of which is studied in the next section. Note that recently there are efforts to address in heterogeneous trip lengths in more detail by considering a trip based formulation (see for example Leclercq et al. (2015), Mariotte et al. (2017) and Lamotte and Geroliminis (2017)). While these models might provide a better estimation of outflow, they are more tedious to be integrated in a control framework. This is an ongoing research direction.

2.2. Subregion Model

Subregion model presented in this section builds on Yildirimoglu and Geroliminis (2014), and it is necessary to develop the path assignment mechanism which will be introduced in the next section. Consider an urban network $\mathcal{SR}$ with heterogeneous distribution of accumulation and a given partition into SR subregions, i.e., $\mathcal{SR} = \{1, 2, \ldots, SR\}$. Now, consider a subregion $r \in \mathcal{SR}$ 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 continuous time t. 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 time, destination or next region.

Let $n_{o,d}^{p,r}(t)$ denote the number of vehicles in subregion r at time t with first subregion o in a given region I, destination subregion d and path p, i.e. the sequence of subregions from o to the last subregion $\bar{d}$ in I; $o, \bar{d}, r \in \mathcal{I}$, $d \in \mathcal{SR}$, $\sum_o \sum_d \sum_p n_{o,d}^{p,r}(t) = n_r(t)$. Note that r belongs to p, and all subregions along p are in region I. That means, $n_{o,d}^{p,r}(t)$ tracks the vehicles from the time they enter region I or start their trip until they leave it or reach their destination. Therefore, o is either the origin subregion where the demand is generated or the boundary subregion that receives flows from other regions. Similarly, $\bar{d}$ is either the destination subregion (i.e. $d = \bar{d}$) or the boundary subregion that sends flows to other regions. Trip completion rate or the transfer flow $m_{o,d}^{p,r}(t)$ for the same group of vehicles reads as follows:

$$m_{o,d}^{p,r}(t) = \frac{n_{o,d}^{p,r}(t)}{n_r(t)} \cdot m_r(t) = \frac{f_r(n_r(t))}{n_r(t)} \cdot \frac{n_{o,d}^{p,r}(t)}{l_r} = v_r(t) \cdot \frac{n_{o,d}^{p,r}(t)}{l_r}. \qquad (4)$$

If $d = r$, the above equation refers to flow leaving the network at subregion d. Otherwise, it represents the transfer flow from subregion r to subregion $p^+(r)$, i.e. the subregion following r in the path p. However, as the transfer flow is also subject to the boundary capacity between the subregions, we denote the actual transfer flow by $\hat{m}^{p,r}_{o,d}(t)$. Note that, in a similar way, $p^-(r)$ is the subregion preceding r in path p. Subregion traffic dynamics are then defined as follows:

$$\frac{\mathrm{d}n^{p,r}_{o,d}}{\mathrm{d}t} = \begin{cases} q^p_{o,d} - m^{p,r}_{o,d} & \text{(i) if } r = o \;\&\; r = d, \\ q^p_{o,d} - \hat{m}^{p,r}_{o,d} & \text{(ii) if } r = o \;\&\; r \neq d, \\ \hat{m}^{p,p^-(r)}_{o,d} - m^{p,r}_{o,d} & \text{(iii) if } r \neq o \;\&\; r = d, \\ \hat{m}^{p,p^-(r)}_{o,d} - \hat{m}^{p,r}_{o,d} & \text{(iv) otherwise.} \end{cases} \tag{5}$$

where

$$\hat{m}^{p,r}_{o,d} = \min[m^{p,r}_{o,d}, c^p_r(n_{p^+(r)}) \cdot a^{p,r}_{o,d}] \quad \forall\, r \neq d \tag{6}$$

$q_{o,d}$ denotes the sum of the exogenous demand generated in o or the flow transferred to o at time t with destination d, and $q^p_{o,d}$ represents the assigned flow to path p; $\sum_p q^p_{o,d} = q_{o,d}$. Note that time t is omitted from the above equations for the sake of notational simplicity. Additionally, $q^p_{o,d}$ is equal to $q_{o,d} \cdot \alpha^p_{o,d}$, where $\alpha^p_{o,d}$ is the path fraction and the decision variable to be computed by ILP-based path assignment scheme that will be described in the next section. Equation 5 defines the change in accumulation $n^{p,r}_{o,d}$ based on four cases. In (i), we deal with internal demand within the same subregion; therefore, the rate is simply the newly assigned flow minus the trip completion rate which is not bounded by any capacity function. 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 (ii), r is the first subregion in the region but not the destination. So, the rate is simply the assigned flow minus the transfer flow to the next subregion in path p. (iii) If r is destination but not the first subregion, then 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 other cases (iv), the rate is equal to the transfer flow from the previous subregion minus the transfer flow to the next subregion.

Equation 6 defines the actual transfer flow from 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 and (ii) the receiving capacity of subregion $p^+(r)$ that is a function of two terms; $c^p_r(n_{p^+(r)})$

and $a_{o,d}^{p,r}$. Capacity at boundary between r and $p^+(r)$, i.e. $c_r^p(n_{p^+(r)})$, is a decreasing function of accumulation, which represents the resistance of the subregion to absorb more traffic with increasing congestion. Additionally, $a_{o,d}^{p,r}$ is the fraction of boundary capacity that can be allocated to $\hat{m}_{o,d}^{p,r}$. Using $m_{o,d}^{p,r}$ values, we calculate the total number of vehicles heading for a particular boundary between two subregions. As not all the vehicles may be allowed to pass the boundary due to the finite capacity, we calculate the fraction of $m_{o,d}^{p,r}$ to the total flow heading for the same boundary and assign the corresponding fraction of the capacity to $\hat{m}_{o,d}^{p,r}$. Equation 6 suggests that only the minimum of the sending flow, i.e. $m_{o,d}^{p,r}$, and the allocated capacity, i.e. $c_r^p(n_{p^+(r)}) \cdot a_{o,d}^{p,r}$ can cross the boundary. This calculation follows the definition of Little (1961).

3. Hierarchical Traffic Management of Large-scale Urban Networks

In this section, we develop an integrated hierarchical route guidance system. The flowchart in Figure 2 summarizes the proposed hierarchical framework. On the upper level, the MPC scheme computes optimum regional split ratios based on the accumulation, average trip length and heterogeneity measurements taken from the subregion model. MPC assumes that average trip length and heterogeneity measures remain constant over the prediction horizon and minimizes the network delay by predicting the evolution of region accumulations. The optimum split ratios are then transferred to ILP-based path assignment scheme, which produces subregional path decisions in accordance with the aggregated split values within each region. This procedure assigns the transfer flow and the exogenous demand to the paths between the subregion they appear in (or the boundary subregion) and their destination (or the boundary subregion). The resulting path decisions calculated by the ILP-based path assignment scheme are then applied to the subregion model.

While the subregion and region models are defined based on the continuous time t, the hierarchical control framework operates on a discrete-time basis. The lower-level, consisting of the subregion model (i.e., the *plant*) and the ILP mechanism, is operated according to the simulation time step t_s (see the clock at lower right of Figure 2). The simulation time step t_s is an integer multiple of the simulation sampling time T_s (s), i.e., $t_s = m_s \cdot T_s$ with $m_s \in \mathbb{Z}_{\geq 0}$. In other words, the ILP commands $\alpha_{o,d}^p(t_s)$ are updated at each simulation time step t_s (i.e., each T_s seconds). The MPC scheme at

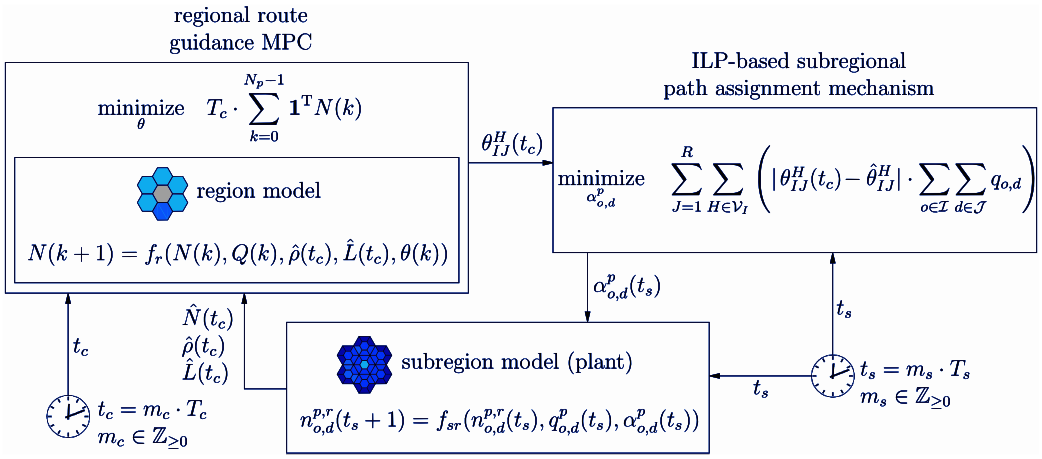

Figure 2: Structure of the proposed path assignment and regional route guidance based hierarchical traffic management scheme.

the upper-level, on the other hand, is operated according to the control time step t_c (see the clock at lower left of figure 2). The control time step t_c is an integer multiple of the control sampling time T_c (s), i.e., $t_c = m_c \cdot T_c$ with $m_c \in \mathbb{Z}_{\geq 0}$. Thus, the MPC scheme receives the traffic state information, and updates its decisions $\theta_{I,J}^H$ based on this information, at each control sampling time t_c (i.e., each T_c seconds). Note that control time step T_c is usually chosen as an integer multiple of T_s for convenience. Hence, the MPC command $\theta_{IJ}^H(t_c)$ is updated at each control time step t_c and kept constant between consecutive control time steps, while the ILP mechanism uses this constant value to compute the $\alpha_{o,d}^p(t_s)$ command until the control step ends (thus, using the same value for T_c/T_s times). The whole procedure is repeated in the next control time step in the receding horizon fashion.

3.1. ILP-based Subregional Path Assignment

In this section, we formulate an integer linear programming (ILP) problem in order to assign the flows in the subregion network so that they satisfy θ_{IJ}^H values ordered by route guidance MPC and produce L_{IH} values within a tolerance range. The formulation, which is repeated for each region I, reads

as follows:

$$\underset{\alpha^p_{o,d}}{\text{minimize}} \quad \sum_{J=1}^{R} \sum_{H \in \mathcal{V}_I} \left(|\theta^H_{IJ}(t_c) - \hat{\theta}^H_{IJ}| \cdot \sum_{o \in \mathcal{I}} \sum_{d \in \mathcal{J}} q_{o,d} \right) \tag{7}$$

$$\text{subject to} \quad \sum_{p \in \mathcal{P}^o} \alpha^p_{o,d} = 1, \quad \forall o \in \mathcal{I}, \forall d \tag{7a}$$

$$\alpha^p_{o,d} \in \{0, 1\} \quad \forall o \in \mathcal{I}, \forall d, \forall p \in \mathcal{P}^o \tag{7b}$$

$$\hat{\theta}^H_{IJ} = \frac{\sum_{o \in \mathcal{I}} \sum_{d \in \mathcal{J}} \sum_{p \in \mathcal{P}^o_H} \left(q_{o,d} \cdot \alpha^p_{o,d} \right)}{\sum_{o \in \mathcal{I}} \sum_{d \in \mathcal{J}} q_{o,d}} \quad \forall J, \forall H \in \mathcal{V}_I \tag{7c}$$

$$\hat{L}_{IH} = \frac{\sum_{o \in \mathcal{I}} \sum_{d} \sum_{p \in \mathcal{P}^o_H} \left(q_{o,d} \cdot \alpha^p_{o,d} \cdot l^p \right)}{\sum_{o \in \mathcal{I}} \sum_{d} \sum_{p \in \mathcal{P}^o_H} \left(q_{o,d} \cdot \alpha^p_{o,d} \right)} \quad \forall H \in \mathcal{V}_I \tag{7d}$$

$$(1 - \varepsilon) \cdot L_{IH}(t_c) \leq \hat{L}_{IH} \leq (1 + \varepsilon) \cdot L_{IH}(t_c) \quad \forall H \in \mathcal{V}_I, \tag{7e}$$

where $\theta^H_{IJ}(t_c)$ is the regional split ratio ordered by MPC at the control time step t_c and $L_{IH}(t_c)$ is the average trip length assumed to remain constant from the same control time step. $\hat{\theta}^H_{IJ}$ and $\hat{L}_{IH}$, on the other hand, are the corresponding variables that are reconstructed based on the trajectories of assigned flows. Also, denote $\mathcal{P}^o$ the set of all paths starting from o, $\mathcal{P}^o_H$ the subset of paths heading for neighboring region H, l^p the distance to be crossed along path p within region I and ε tolerance error between the observed and reconstructed trip lengths. Note that simulation time step t_s is omitted from the above equations for the sake of notational simplicity and the only static variables are the physical path distance l^p and tolerance error ε. The remaining variables, which are not followed by $.(t_c)$, should in fact be followed by $.(t_s)$.

Equation 7 minimizes the weighted average of the absolute difference between the ordered and reconstructed regional split ratios given the constraints presented from Equation 7a to 7e. The objective function considers the weighted average with respect to demand between regions (i.e. $\sum_{o \in \mathcal{I}} \sum_{d \in \mathcal{J}} q_{o,d}$) so as to attach more importance to high-volume pairs. Equation 7a guarantees that the demand between o and d is assigned to a path, while Equation 7b defines $\alpha^p_{o,d}$ as a binary variable and warrants an all-or-nothing assignment process. Considering the demand $q_{o,d}$ to be assigned between o and d (including the exogenous demand and the transfer flow),

Equation 7c defines the regional split ratios based on the assigned flows. The denominator in Equation 7c is the total flow to be assigned between I and J, while the numerator is the portion allocated with the routes targeting neighboring region H. Similarly, Equation 7d computes the average trip lengths based on the assigned paths. The denominator represents the assigned flow heading for the boundary between I and H, and the numerator defines the total distance traveled by them. Finally, Equation 7e defines the tolerance bounds that the reconstructed trip length should fall into. Note that the above formulation is repeated for each region I.

Since Equation 7d includes the decision variable $\alpha^p_{o,d}$ both in the numerator and the denominator, the resulting problem cannot be formulated as an ILP. Therefore, assuming that θ^H_{IJ} values ordered by MPC will be fully satisfied, we replace Equation 7d with the following:

$$\hat{L}_{IH} = \frac{\sum_{o\in\mathcal{I}}\sum_{d}\sum_{p\in\mathcal{P}^o_H}\left(q_{o,d}\cdot\alpha^p_{o,d}\cdot l^p\right)}{\sum_{o\in\mathcal{I}}\sum_{d}\left(q_{o,d}\cdot\delta_{dJ}\cdot\theta^H_{IJ}\right)} \quad \forall H\in\mathcal{V}_I \tag{8}$$

where δ_{dJ} is 1 if $d\in\mathcal{J}$, and 0 otherwise. We note that Equation 7d and Equation 8 are equal to each other if θ^H_{IJ} and $\hat{\theta}^H_{IJ}$ are the same. And, the above approximation does not affect the performance of the assignment scheme as long as θ^H_{IJ} and $\hat{\theta}^H_{IJ}$ are close to each other. Essentially, Equation 7d is replaced with Equation 8 relying on the assumption that the value of the objective function is zero. However, this does not force the objective function to be zero. The decision variables (i.e. $\alpha^p_{o,d}$) are still in the numerator of the formula presented in Equation 8, which allows the framework to test different values of $\hat{L}_{IH}$ and so $\hat{\theta}^H_{IJ}$. Obviously, the value of $\hat{L}_{IH}$ from Equation 7d will differ from that of Equation 8 in most cases where θ^H_{IJ} and $\hat{\theta}^H_{IJ}$ are not exactly the same. Nevertheless, this discrepancy should be minimal at the optimal solution where θ^H_{IJ} and $\hat{\theta}^H_{IJ}$ are expected to be similar. This replacement is crucial to keep the problem linear and tackle it with ILP solution methods. Nevertheless, we realize that Equation 8 does not account for dynamic conditions in the subregions; it only considers the distance to be crossed along the path, which is a static physical measure. In order to ensure homogeneity within the regions, we further modify Equation 8 and express the travel time from region I to region H with the below formula:

$$\frac{\hat{L}_{IH}}{V_I} = \frac{\sum_{o\in\mathcal{I}}\sum_{d}\sum_{p\in\mathcal{P}^o_H}\left(q_{o,d}\cdot\alpha^p_{o,d}\cdot\sum_{r\in p}\left(l_r/v_r\right)\right)}{\sum_{o\in\mathcal{I}}\sum_{d}\left(q_{o,d}\cdot\delta_{dJ}\cdot\theta^H_{IJ}\right)} \quad \forall H\in\mathcal{V}_I \tag{9}$$

where V_I denotes the average speed in region I (i.e. $F_I(N_I)/N_I$). Basically, we break the path distance l^p into subregions, and calculate the travel time in each subregion r using the static average distance l_r and the dynamic speed v_r changing with time t_s (the actual notation should be $v_r(t_s)$, but t_s is omitted for simplicity). The sum of them gives us the travel time on the subregional path p. If actual traffic conditions are ignored, certain subregions within a given region might be more (or less) congested depending on the routing and the subregional paths. To avoid such cases and improve homogeneity within regions, we replace Equation 8 with Equation 9 where we account for both distance and current traffic conditions in the network. Note that the target measure in Equation 9 is the average travel time from I to H, not the average distance which is a static network attribute. This allows us to react to uneven distribution of congestion across the subregions and ensure homogeneity within the regions. In other words, Equation 9 enables ILP to control the region-to-region travel times and avoids overloading of few subregions. We also note that we do not use Equation 9 the way it is presented; we take V_I to the right hand side of the equation to compute $\hat{L}_{IH}$ and substitute it into Equation 7e.

The path assignment mechanism could also be formulated as a linear programming problem, where $\alpha^p_{o,d}$ could be defined as a continuous variable between 0 and 1. Although this would significantly simplify the computation efforts, there may not always be enough demand to accommodate such split ratios. While we keep the formulation here as an ILP problem, we test the effects of this assumption later in Section 4.3.

The ILP problem is built with YALMIP (Löfberg, 2004), an efficient toolbox for modeling and optimization in MATLAB, and solved with Gurobi mixed integer linear programming solver. The optimization scheme is implemented using MATLAB 8.5.0 (R2015a), on a 64-bit Windows PC with 3.6-GHz Intel Core i7 processor and 16-GB RAM.

3.2. Regional Route Guidance MPC

We formulate the problem of finding the regional split ratio θ^H_{IJ} values that minimize total time spent (for a finite horizon) as the following discrete time economic nonlinear MPC problem (extending the work in Sirmatel and

Geroliminis (2018)):

$$
\begin{aligned}
\underset{\theta}{\text{minimize}} \quad & T_c \cdot \sum_{k=0}^{N_p-1} \mathbf{1}^{\mathrm{T}} N(k) \qquad (10)\\
\text{subject to} \quad & N(0) = \hat{N}(t_c)\\
& |\theta(0) - \hat{\theta}(t_c - 1)| \leq \Delta_\theta\\
& \text{for } k = 0, \ldots, N_p - 1 :\\
& \quad N(k+1) = f_r\left(N(k), Q(k), \rho(t_c), \hat{L}(t_c), \theta(k)\right)\\
& \quad 0 \leq N_I(k) \leq N_I^{\text{jam}},\ \forall I \in \mathcal{R}\\
& \quad 0 \leq \theta_{IJ}^H(k) \leq 1,\ \forall I, J \in \mathcal{R}, H \in \mathcal{V}_I \cup I\\
& \quad \sum_{H \in \mathcal{V}_I \cup I} \theta_{IJ}^H(k) = 1,\ \forall I, J \in \mathcal{R}\\
& \quad \text{if } k \geq N_c :\\
& \quad\quad \theta(k) = \theta(k-1),
\end{aligned}
$$

where T_c and t_c are the control sampling time and control time step, respectively (with $t_c = m_c \cdot T_c$ where $m_c \in \mathbb{Z}_{\geq 0}$), $N(k)$, $Q(k)$, $\hat{\rho}(t_c)$, $\hat{L}(t_c)$, and $\theta(k)$ are vectors containing all $N_{IJ}^H(k)$, $Q_{IJ}(k)$, $\hat{\rho}_I(t_c)$, $\hat{L}_{IH}(t_c)$ and $\theta_{IJ}^H(k)$ terms (with $\hat{\rho}_I(t_c) \triangleq \rho_I(\hat{N}_I(t_c), \sigma(\hat{N}_I(t_c)))$), respectively, with k being the control interval counter, f is the time discretized version of eq. (1)–(2), $\hat{N}(t_c)$, $\hat{\rho}(t_c)$, and $\hat{L}(t_c)$ are the measurements on accumulations, heterogeneity coefficient, and average trip lengths taken at the current control time step, respectively, $\hat{\theta}(t_c - 1)$ is the control input applied to the plant at the previous control time step, N_p and N_c are the prediction and control horizons, whereas Δ_θ is the rate limit on regional split ratios. Within the prediction horizon N_p, we assume that heterogeneity coefficients $\hat{\rho}(t_c)$ and average trip lengths $\hat{L}(t_c)$ remain constant. To relax this assumption one needs to either estimate a vector of accumulations and trip lengths valid for a finite horizon into the future (same length as N_p) or model the dynamics that define heterogeneity and average trip length as a function of route guidance control inputs. We also consider that as trip length is a difficult quantity even to measure and estimate, it will be even more difficult to predict it. Assuming a quantity constant, when it is difficult to predict, is a well-established approach in MPC literature (it is better not to predict when errors are expected to be

large). The prediction aspect is considered outside the scope of the paper, and could be considered for future work.

The optimization problem (10) is a nonconvex nonlinear program (NLP), which can be efficiently solved via, e.g., sequential quadratic programming or interior point solvers (see Diehl et al. (2009) for details). Software implementation of the MPC scheme is done using the CasADi (Andersson et al., 2018) toolbox in MATLAB 8.5.0 (R2015a), on a 64-bit Windows PC with 3.6-GHz Intel Core i7 processor and 16-GB RAM, using a direct collocation scheme (see, e.g., Diehl et al. (2006) for details) with the NLPs solved by the interior point solver IPOPT (Wächter and Biegler, 2006).

4. Case study

The simulation case study is based on a network with 49 subregions partitioned into 7 regions (see Figure 1). The path-based subregion-level model given in Equations (4)-(6) is used as the simulation model (i.e., the *plant* representing reality), whereas the region-level model given in Equations (1)-(2) is used as the prediction model of the route guidance MPC. Each region is assumed to have a production MFD with the parameters $A = 9.98 \cdot 10^{-8}$, $B = -0.002$, $C = 9.78$, jam accumulation $N^{\text{jam}} = 10^4$ (veh), critical accumulation $N^{\text{cr}} = N^{\text{jam}}/3$ (veh), which are consistent with the MFD observed in a part of downtown Yokohama (see Geroliminis and Daganzo (2008)). Subregions are assumed to have well-defined production MFDs, the parameters of which are scaled versions of the region MFDs, whereas the associated average trip lengths are constant and there is assumed to be no heterogeneity affecting the subregions. Region MFDs, on the other hand, are exposed to variations in the outflow as they are affected by the average trip lengths $L_{IH}(t)$ and heterogeneity coefficient $\rho_I(t)$, which are dynamically changing with the traffic conditions at the subregion level. Prediction and control horizons for the MPC are chosen as $N_p = 5$ and $N_c = 2$ (following the controller tuning results of Sirmatel and Geroliminis (2018) for a 7-region network as depicted in Figure 1). The traffic states in the plant are updated using a time discretized version of the subregion model given in (5), with a simulation sampling time of $T_s = 60$ s. Path assignment decisions of the ILP mechanism are also updated together with the traffic states each $T_s = 60$ s, whereas the MPC operates with a control sampling time of $T_c = 240$ s. Thus, the ILP mechanism uses the same route guidance decision coming from the MPC for the 4 simulation time steps belonging to the same control time step.

The subregional path assignment mechanism is essentially in charge of tracking two signal classes; split ratios (i.e. θ_{IJ}^{H}) and average trip lengths (i.e. L_{IH}). While the ILP formulation minimizes the difference between the ordered split ratios (i.e. θ_{IJ}^{H}) and reconstructed split ratios (i.e. $\hat{\theta}_{IJ}^{H}$) (Equation 7), it ensures that the reconstructed trip lengths (i.e. $\hat{L}_{IH}$) are close to the trip lengths observed from the plant (i.e. L_{IH}) and are within the bounds defined by ε (see Equation 7e). Regional route guidance MPC assumes that average trip length measures remain constant over the prediction horizon and minimizes the network delay by changing the split ratios. The subregional path assignment mechanism follows the same rationale and aims to match the reconstructed split ratios with the ordered ones (see Equation 7) while maintaining the reconstructed average trip lengths in the vicinity of observed ones (see Equation 7e). Accordingly, the objective function in the ILP is the sum of absolute difference between θ_{IJ}^{H} and $\hat{\theta}_{IJ}^{H}$ values, and the tolerance range of $\hat{L}_{IH}$ with respect to L_{IH} is added as a constraint in the optimization problem. The value of ε, in this study, is 0.05. And, it has been chosen such that the outflow from the regions is not largely affected by the changing trip length values, and yet the overall framework is flexible enough to follow the ordered split ratio signals. Note that the mismatch between ordered and reconstructed patterns can always be taken into account in the next time step by the feedback mechanism of the control framework, and we observe that the results are not very sensitive to the changes within the range $\varepsilon \in [0.025, 0.10]$.

Additionally, ILP formulation presented in Equation 7 includes the set of paths (i.e. $\mathcal{P}^{o}$) which requires the enumeration of alternatives between the subregion pairs (in the same region). As the enumeration of all paths would present computational difficulties at a large-scale region and include unrealistic routes, we limit our analysis with the first 5 physical (distance-based) shortest paths between the subregion pairs. Therefore, the resulting scheme does not offer a "perfectly optimal" solution where few agents are largely penalized to reach social equilibrium; instead, it considers limited willingness of travellers to switch to alternative routes and provides a constrained social equilibrium solution where no traveller is given an exceedingly long path (similar to Jahn et al. (2005)). While SO conditions might generate longer paths for a few users, with the above consideration, it is more likely that users follow and comply with the outcome of guidance strategy. To determine the shortest paths, we use the physical network properties (e.g., connectivity, distance) of the 49-subregion representation, just like one would

do with a link-level representation of a network. In other words, we build a graph where nodes are subregions and the neighboring nodes (as seen in Figure 1) are connected to each other with an edge whose value is equal to the subregional average trip length (i.e., l_r). We then identify the shortest paths between the subregions using this graph. An alternative to the static (distance-based) choice set we use here might be to consider dynamic traffic conditions and update the choice in every time step with the time-based shortest paths. Nevertheless, our current framework, which builds on the YALMIP toolbox, does not allow such changes because of the coding-related limitations. While the toolbox makes development of optimization problems very simple, it requires the ILP-based mechanism to be built in advance with a choice set. Therefore, such dynamic changes in the choice set are not possible.

Note that the subregion model that is used for testing the hierarchical control does not appear anywhere inside the two levels of the hierarchical control scheme. Hence, there is a significant difference between the models used for optimization and the model used for replicating reality; the "plant". The ILP-based path assignment mechanism does not use a dynamical model at all; it simply relies on the physical network properties (e.g., connectivity, distance, etc.) that are easy to be collated. On the other hand, the route guidance MPC uses the region model which is very different than the subregion model in the following aspects: a) the region model considers the aggregated representation of 7 subregions as a single region and scrutinizes traffic flows between regions not subregions, b) the subregions have well-defined MFDs, whereas the region MFDs are subject to heterogeneity effects according to the congestion distribution at the subregional level (as modeled by Equation 3).

4.1. Numerical results

To test the performance of the proposed hierarchical route guidance (RG) scheme (the structure of which is given as a block diagram in Figure 2) against a no control (NC) case, we conduct a simulation experiment based on a congested scenario for the two cases. In both NC and RG, drivers receive a path information when they enter the network or a new region, and the actuations are updated every control time step (i.e. 4 minutes) to have a fair comparison between the scenarios. Depending on trip to be made, the path consists of sequence of subregions from origin subregion or region boundary to the destination subregion or region boundary. In NC, the path decisions

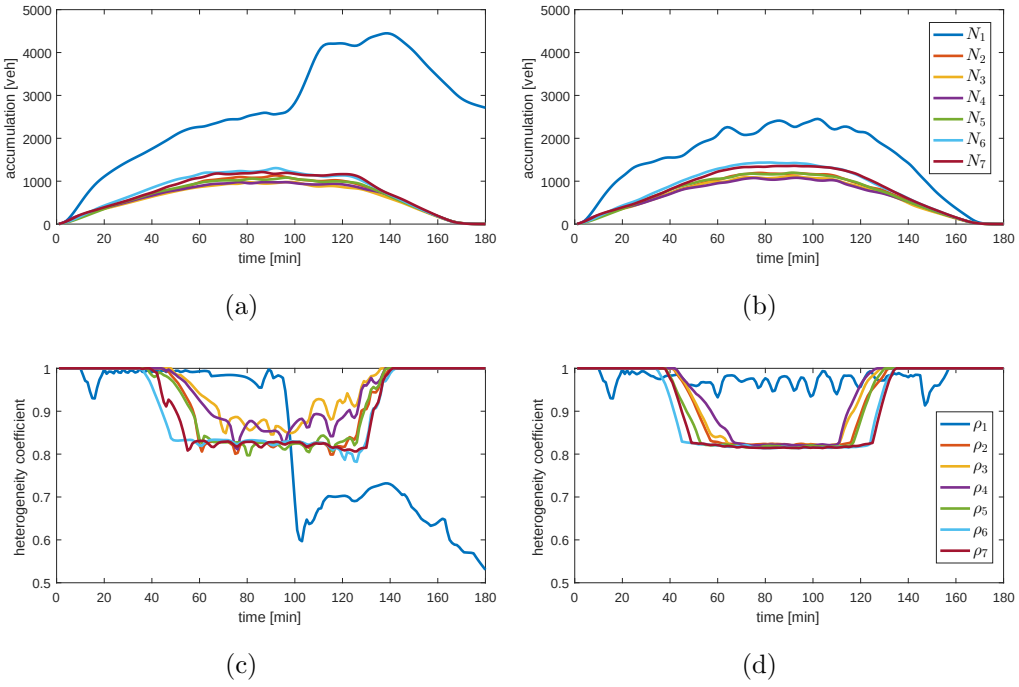

Figure 3: Results of the congested scenario, showing region accumulations $N_I(t)$ (a)-(b) and heterogeneity coefficients $\rho_I(t)$ (c)-(d) as functions of time, for the no control case (a)-(c) and the hierarchical route guidance scheme (b)-(d).

reflect the shortest path based on instantaneous traffic conditions, while RG assigns the vehicles to the paths produced by the hierarchical scheme. In both NC and RG scenarios, the network starts empty but is exposed to increased inflow demands as time progresses. As the vehicles are closer to their origin than to their destination in the beginning of the simulation, the first 20 minutes is considered the warm-up period and RG system is activated after that.

Figure 3 shows the regional accumulations (Figure 3a and 3b) and heterogeneity coefficients as described by Equation 3 (Figure 3c and 3d) for the two scenarios. The figure clearly indicates that the RG scheme is efficient in alleviating congestion, whereas in the NC case, congestion cannot be avoided in the city center (i.e. region 1). By comparing Figure 3a and 3b we observe that the peripheral regions (i.e., regions 2 to 7) carry slightly more traffic in RG scenario than in NC scenario, which in turn helps the central region stay uncongested. In NC scenario, the accumulation in the central region escalates

to very high values at around 100 min, and the network is not emptied at the end of the 3-hour simulation. To have a fair comparison between the two scenarios, we run the simulation as long as there are vehicles in the system and calculate the total time spent when all the vehicles reach their destination. This calculation results in $2.32 \cdot 10^4(veh.h)$ and $1.72 \cdot 10^4(veh.h)$ for NC and RG scenarios, respectively, which corresponds to around 27% improvement with RG. An important reason that RG performs significantly better is not only a subset of vehicles does not cross the center region, but also the level of homogeneity is higher in the central region that further increases the outflow. Note that, as presented in Eq. 2, low ρ_I values lead to low outflows from the region. In NC scenario, a sharp decrease in the heterogeneity coefficient (i.e. ρ_1) signals the deterioration in the congestion distribution in region 1, which is followed by high accumulation values in the same region. On the other hand, RG strategy achieves to produce coefficient values close to 1 and guarantees that congestion distribution stays consistent throughout the simulation in the central region. Additionally, in both NC and RG, we observe that the heterogeneity coefficient in the peripheral regions seems to converge to a value around 0.8 as the simulation progresses. This is mainly due to the physical structure of the network; the subregions at the outer layer of the network (e.g. subregions 10, 16 and 25) are rarely used by through traffic. The traffic load they carry is inherently low compared with other subregions in the region, which causes a lower heterogeneity coefficient for the region they belong to. While the homogenization of the regions is not considered as part of the objective function in the regional route guidance MPC, ILP-based subregional path assignment accounts for the subregion speeds at the current time step and constraints the change in region-to-region travel time through Equation 9. This enables ILP to react to uneven distribution of congestion and homogenize the traffic across subregions. Therefore, as Figure 3 depicts, the improvement comes from both regional route discipline and increased homogeneity within the regions. In the RG scenario, the accumulation in the central region is significantly lower and the heterogeneity coefficient is substantially higher.

Figure 4(a) and (b) depict the subregion accumulations of region 1 (i.e., the central region) in NC and RG scenarios, respectively. While accumulation values are comparable across the scenarios in the first half of the simulation, the central subregion (or subregion 1) starts attracting significant demand at around 90 (min) and reaches gridlock state few minutes later. On the other hand, although central subregion always carries a higher traffic, RG

strategy ensures a greater level of homogeneity in the region and does not cause gridlock state in any subregion. Figure 4(c) presents the production MFD of region 1 in two scenarios which results from the accumulations introduced in Figure 4(a) and (b). The city center suffers from a significant capacity drop in NC scenario, while RG scenario is able to keep region 1 in the uncongested regime and guarantee higher production values despite few scatter (see the red circles in Figure 4(c)). Note that the capacity drop we observe in NC scenario is due to the jump in the heterogeneity coefficient values presented in Figure 3(c). The difference in MFD patterns is also re-

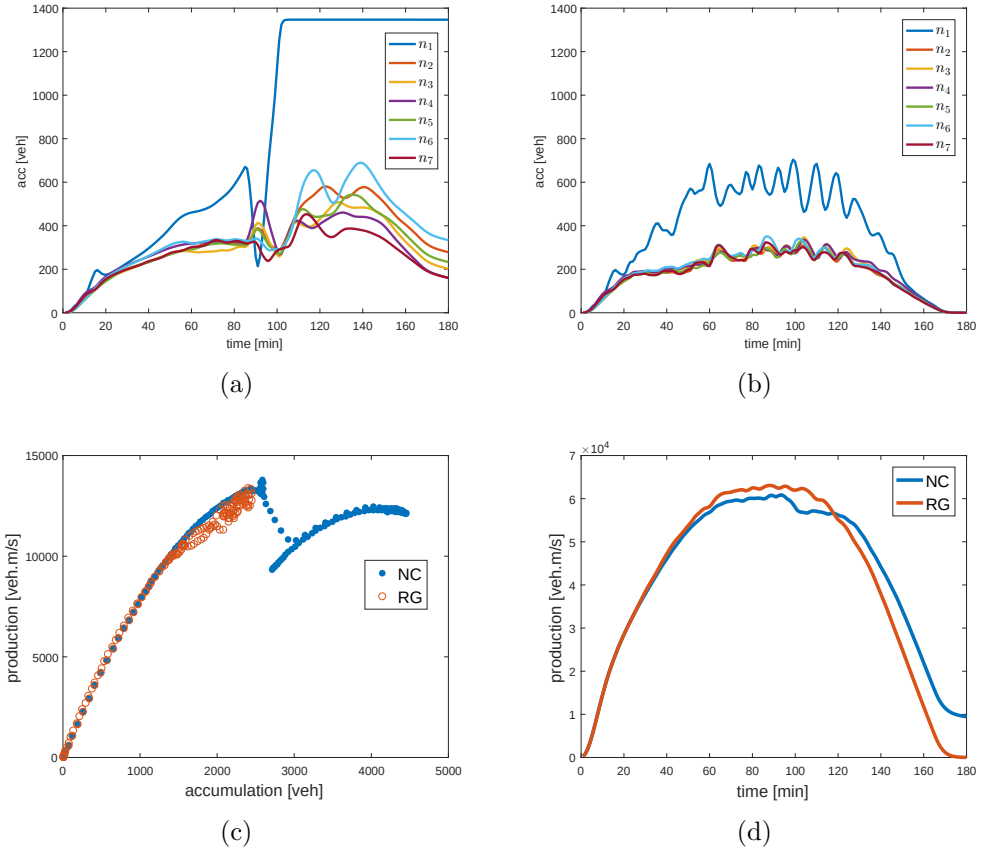

Figure 4: Results of the congested scenario comparing subregion accumulations of region 1 (a)-(b), production MFDs of region 1 (c), and the network production as a function of time (d), for the no control case (NC) (a) and the hierarchical route guidance scheme (RG) (b).

flected in the overall network production; Figure 4(d) shows that RG is able to produce higher production during the peak hour and empty the network earlier than NC scenario. Note the non-zero production values in NC at the end of the simulation.

Figure 5 presents the average trip length values that result from NC and RG scenarios in region 5. As can be seen from Figure 1, region 5 has 3 neighboring regions. Hence, including the internal trips, Figure 5 depicts 4 curves separately for NC and RG scenarios. At the start of the simulation, most vehicles are closer to their origin than to their destination; therefore, the outflow values are very small and the average trip length values are very high. However, we note that, in all scenarios, they quickly converge to stable values (like a warm up period). The average trip lengths to neighboring peripheral regions (i.e. L_{54} and L_{56}) slightly increase after the activation of RG, while the one to the central region (i.e. L_{51}) and to itself (i.e. L_{55}) remain more or less the same. This is due to the change in the assignment patterns; as there are more vehicles using the peripheral network with the implementation of RG, it is not possible to keep the average trip length values at the low level observed in NC. However, both L_{54} and L_{56} quickly converge to slightly higher values and remain approximately stable until the network unloading stage. The simulation ends with relatively low distance values in both scenarios as a result of the region being emptied and most vehicles being

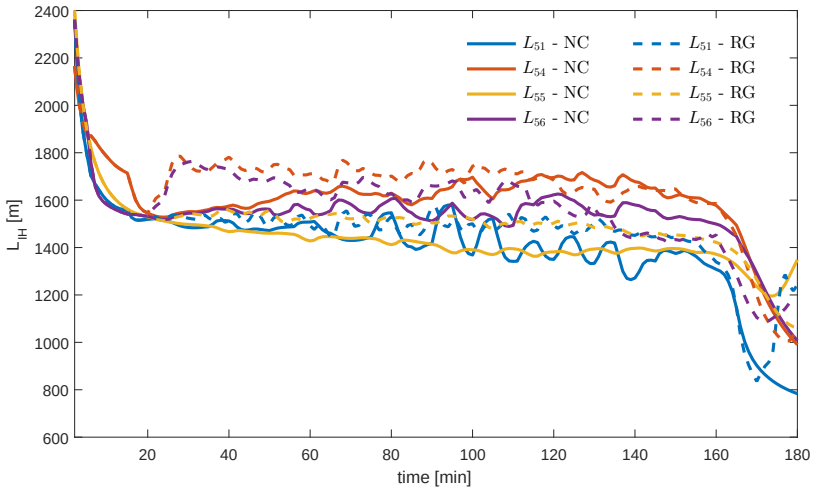

Figure 5: Average trip length values in region 5.

closer to their destination. Trip length graphs for other peripheral regions are omitted due to space limitation. Nevertheless, they represent similar patterns.

Figure 6 compares NC with RG regarding the proportion of accumulations (i.e. N_{IJ}^H/N_{IJ}) and presents the ordered and reconstructed split ratios (i.e. θ and $\hat{\theta}$) in RG. Note that split ratios apply to transfer flows (between regions) and the newly generated exogenous demand, while the proportion of accumulation defines the route choice patterns for all circulating vehicles. For the illustration purposes, we choose the vehicles traveling from region 5 to regions 2 and 7. As previously mentioned, the first 20 min of the simula-

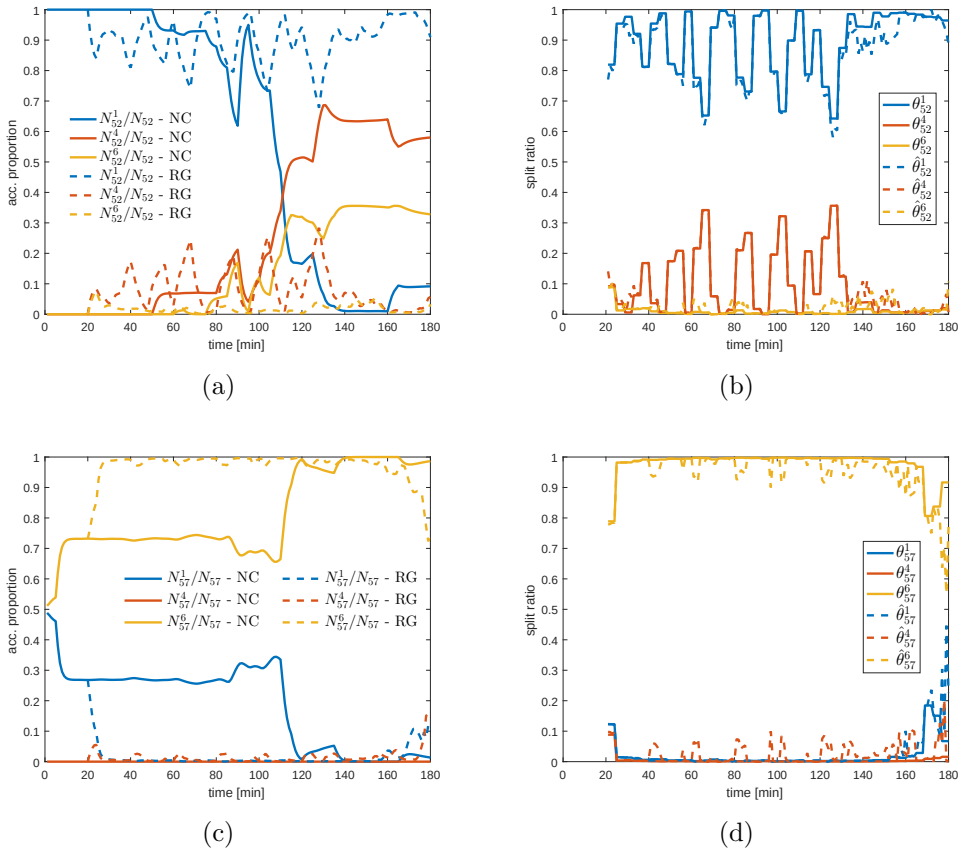

Figure 6: (a), (c) Resulting accumulation proportions in NC and RG; (b), (d) Ordered and reconstructed split ratios, θ and $\hat{\theta}$, in RG for vehicles going from region 5 to regions 2 and 7.

tion is considered the warm-up period during which RG is not active. That is why accumulation proportions are the same for NC and RG in the first 20 min (see Figure 6(a) and 6(c)).

Figure 6(a) and 6(b) introduce the accumulation proportions and split ratios, respectively, for the vehicles traveling from region 5 to 2. All vehicles start off by choosing region 1 in NC (see Figure 6(a)), as it provides the (physical) shortest alternative (see Figure 1). Due to changing traffic conditions, after $t = 50$(min), alternative paths that do not cross the central region become more appealing, and some vehicles start using the idle capacity at the periphery of the network. However, this does not save the central region from getting overly congested. On the contrary, RG, after being activated at $t = 20$(min), assigns 0-35% of the (newly entering) demand to the peripheral regions (see the red and yellow curves in Figure 6(b)) and ensures that the central region has a more balanced distribution of accumulation (see Figure 6(a)). Note that the ordered and reconstructed split ratios in Figure 6(b) are very close to each other for this particular demand pair throughout the simulation. Figure 6(c) and 6(d) present the split ratios and accumulation proportions, respectively, for the vehicles traveling from region 5 to 7. According to the regional representation of the network (see Figure 1), the central region 1 and the peripheral region 6 provide equally appealing alternatives (in terms of distance). In NC scenario, initially, approximately 30% of the accumulation crosses the central region; however, in response to hyper-congestion in the center, travelers switch to peripheral regions towards the end of the simulation. On the other hand, RG guides almost all the vehicles through the peripheral region 6 and protects the central region from congestion. Note that regional split ratios are rapidly changing in RG at the end of the simulation, which is due to the network being emptied. Turning RG off below a certain accumulation level could easily prevent this behavior, but is not expected to significantly change the results as there are very few vehicles in the network within this period.

Figure 7 and 8 provide a series of snapshots over time that depicts the evolution of regional and subregional accumulations in the network, respectively. In Figure 7, we see that congestion is rather evenly distributed across the regions in RG, while NC scheme is overloading the central region. Note that the central region has some residual accumulation at the end of the time horizon, while the network is completely emptied in RG. Figure 8 provides a zoom-in view of the accumulations in the network. We observe that the subregions at the outermost layer of the network are used at below their

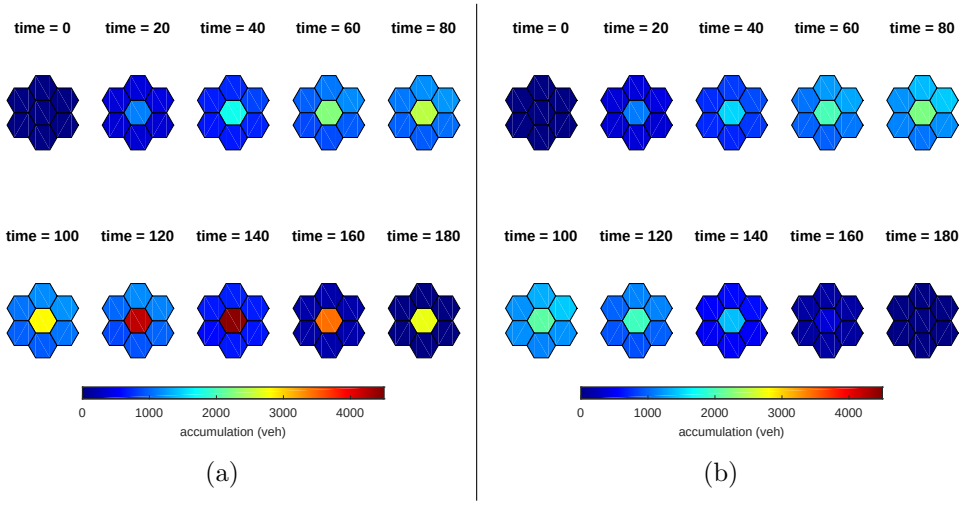

Figure 7: Evolution of region accumulations over time (min). (a) NC, (b) RG.

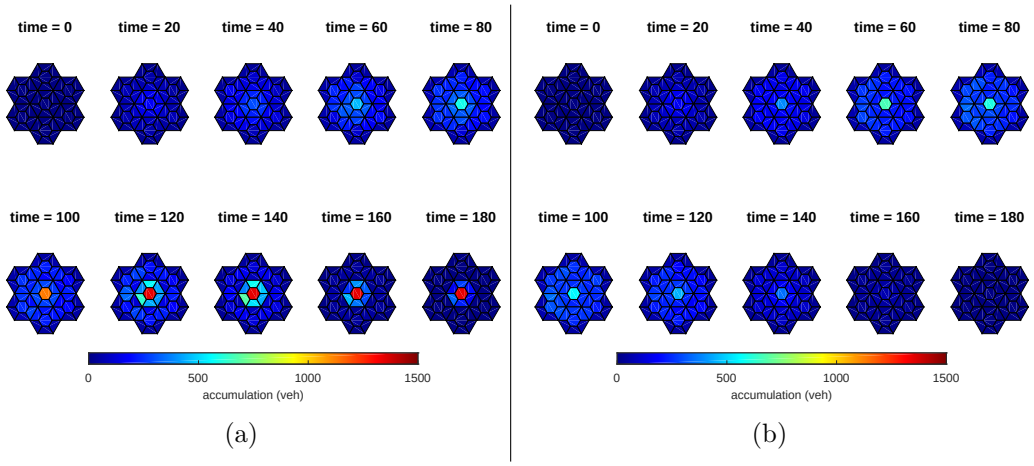

Figure 8: Evolution of subregion accumulations over time (min). (a) NC, (b) RG.

capacity throughout the simulation in both scenarios. As these subregions are not critical in terms of the network connectivity, they are mostly used by vehicles entering or leaving the network through them. In addition, we note that the subregion at the core of the network (subregion 1 in Figure 1) gets gridlocked (i.e. reaches jam accumulation) at around 120 min, and it cannot be rescued from that state until the simulation end. In overall, these results show that the RG scheme can distribute congestion evenly over the network using the authority over path assignment and route guidance, leading to an efficient use of network capacity and ultimately to increased mobility.

4.2. Comparison with Perimeter Control Actuated MPC

To evaluate the performance of the proposed hierarchical control scheme in comparison with a perimeter control case, an MPC with only perimeter control type actuation is tested using the simulation experiment with the congested scenario. The perimeter control MPC is constructed in a similar vein with the proposed regional route guidance MPC: the multi-region MFDs network based MPC formulation from Sirmatel and Geroliminis (2018) is extended with heterogeneity variance and dynamic trip length terms as in equation (10), while perimeter control inputs (i.e. $U_{IJ}(t)$) are introduced into the formulation as decision variables and the regional split ratios $\theta^H_{IJ}(t)$ are defined as measurements (for the MPC) that are assumed constant for the prediction horizon. On the subregion level the drivers are free to choose their own routes as in the no control case. As the perimeter control MPC could not cope with the gridlock-level congestion in subregion 1 due to the increasing inhomogeneity, it is supplemented with a simple bang-bang perimeter controller that protects this subregion from severe congestion. Note that while the perimeter control MPC controls the boundaries between regions, the bang-bang controller restricts the inflow from subregions 2-7 to subregion 1. In fact, incorporation of the bang-bang controller adds 6 new borders to be controlled in the network, and in a real-world context, it may not be always possible to control the intersections in the city center (represented by region 1). However, to conduct a fair comparison here, we design a perimeter control (PC) strategy that combines the MPC actuations at the region boundaries and the bang-bang actions at the border of subregion 1. A discussion on why the perimeter control MPC itself does not work properly will be provided in the following paragraph.

Figure 9(a) presents the accumulation of region 1 for NC, RG and PC (i.e. MPC+bang-bang) scenarios. We clearly see that although PC can improve

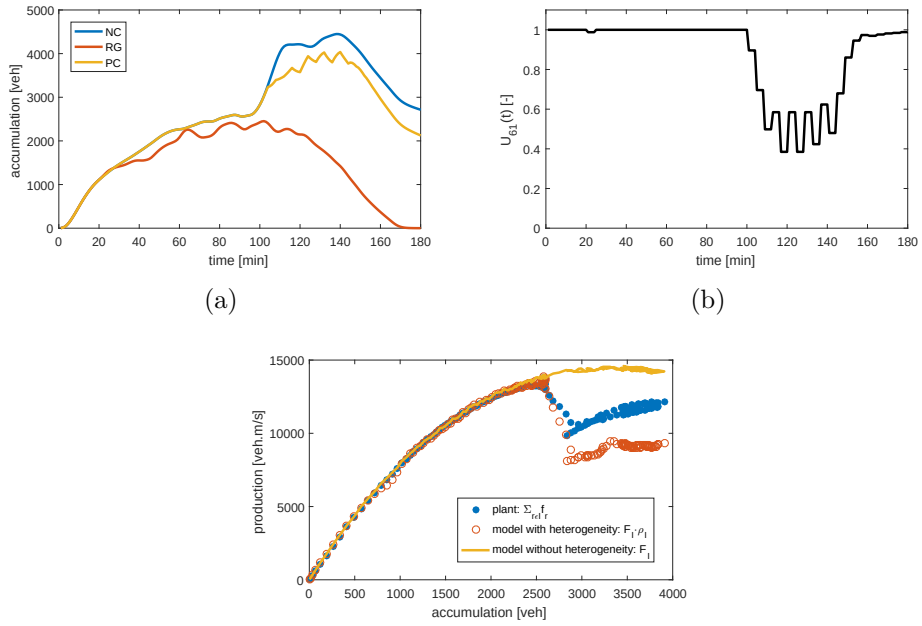

Figure 9: (a) Accumulation of region 1 in the no control case (NC), the proposed route guidance (RG), and the perimeter control (PC), (b) Perimeter control input $U_{61}(t)$ for the PC case, (c) Production MFDs in region 1.

over the NC scenario by protecting region 1, it is not as effective as RG in the alleviation of congestion. Figure 9(b) presents the perimeter control action at the boundary between regions 6 and 1, which corresponds to the border between subregions 38 and 6 at the plant. While the maximum outflow is maintained at the border until around 100 (min), the controller reacts to the increasing accumulation in region 1 in the following time steps and restricts the inflow to the region 1. However, PC is not able to clear the network until the end of the 3-hour simulation. As in NC scenario, we run the simulation until the network is empty in PC scenario and calculate the total time spent in the network. PC results in $2.12 \cdot 10^4(veh.h)$, which presents around 9% improvement over NC scenario, while RG yields 27% reduction in the total time spent. Finally, Figure 9(c) compares the production MFDs of region 1 in three cases: observations from the plant, production model with heterogeneity coefficient presented in Equation 3 and production model with full homogeneity assumption (upper envelope for the production values). Note

that this analysis represents a scenario where the perimeter control only consists of MPC actuations. We clearly see that the production estimation with the heterogeneity model does not provide an accurate approximation for the plant measurements. As only one subregion is very congested and the others are not, region 1 exhibits a highly imbalanced congestion scenario where the production model with heterogeneity coefficient fails to provide accurate estimations. We observe that the resulting heterogeneity coefficient values push the production estimations further down than the plant measurements. This is the main reason why perimeter control MPC itself cannot cope with the congested scenario in hand. On the other hand, RG strategy improves the homogeneity inside the regions, keeps the traffic states within the limit of trackable values and improves the traffic conditions in the network.

4.3. Sensitivity of the model

In this subsection, we test the sensitivity of our model with respect to certain design and scenario features; in particular (1) compliance rates of drivers to the guidance information, (2) path assignment characteristics (i.e., all-or-nothing vs. partial flows) and (3) noise in the plant characteristics (i.e., randomness in the subregion MFD).

First, we test our strategy with lower compliance rates. We keep the same formulation for the subregional path assignment and the regional route guidance, where we assume full compliance of users (note that the design of controller does not explicitly consider a given compliance rate). Nevertheless, at the testing stage, we provide the resulting $\alpha^p_{o,d}$ values only with the complying users and let the other users make route choice decisions in accordance with NC scenario. Figure 10(a) presents the accumulation values in region 1 (supposedly the most congested region) resulting from a number of compliance rates. Clearly, the higher the compliance rate is, the less congested region 1 becomes. Nevertheless, all the scenarios produce considerably better traffic conditions than NC scenario. Even with 25% compliance rate, traffic conditions are significantly improved and hyper-congestion (values higher than critical accumulation N^{cr}= 3333 veh) is avoided. That means, even a small percentage of users complying with the guidance information and avoiding potentially congested parts of the network, can bring major benefits to the system. The network performance with low compliance rates is similar to the full-compliance RG scenario, the resulting total time spent in the network is only 2-4% higher than that of 100% compliance scenario.

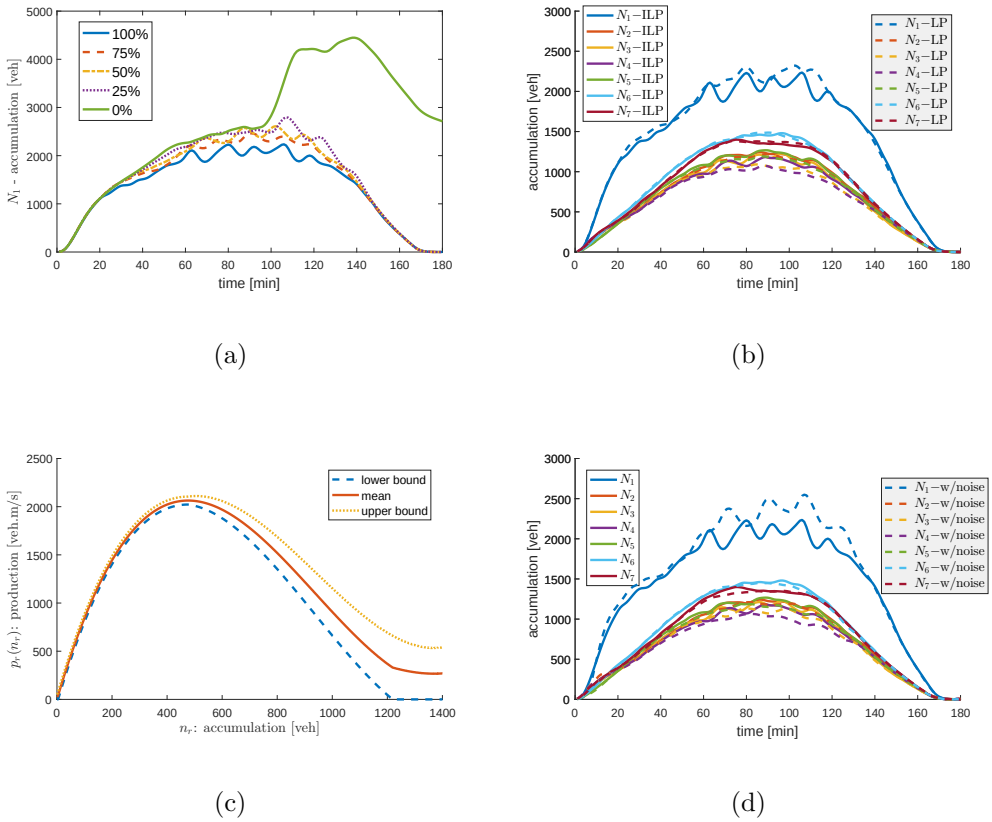

Figure 10: (a) Accumulation in region 1 with varying compliance rates, (b) accumulation in all regions with continuous and integer decision variables in the subregional path assignment formulation, (c) subregion MFD with uniformly distributed random noise between lower and upper bounds, (d) region accumulations resulting from the RG scenario with and without noise in subregion MFD.

Second, assuming that there may not be always enough demand to accommodate continuous split ratios (between the boundary nodes), we have applied all-or-nothing assignment and defined $\alpha_{o,d}^{p}$ as a binary variable. To test the effects of this assumption, we now describe it as a continuous variable between 0 and 1 and reformulate the subregion assignment mechanism as a linear programming (LP) problem. Figure 10(b) depicts the accumulation evolution in all regions for ILP and LP formulation. None of the regions exhibits a significant difference; however, we note minor changes particularly for region 1 between 60 min and 120 min. LP formulation leads to stable accumulation values within this period, while ILP formulation produces small up-and-downs. As LP can achieve a better tracking of signals (θ_{IJ}^{H} and L_{IH}) by adjusting continuous path flow distributions, it creates a more consistent accumulation profile even in the most congested period. Note that we do not observe a meaningful change in the performance measures across the two scenarios, which indicates that the model can as well be useful with all-or-nothing assumptions.

Third, we investigate the effect of additional noise in the plant characteristics. The purpose of the detailed 49-subregion simulator is to create a significant difference between what the model knows during optimization (i.e. 7-region model) and what influence comprehensive plant characteristics might have. The subregion model provides a simulation environment where many of the assumptions in the region model are released. For example, while routing is achieved with split ratios (θ_{IJ}^{H}) in the region model, subregional paths are incorporated into the plant in order to navigate the vehicles around the network. Nevertheless, we acknowledge that MFDs at the subregional level might experience scatter too (as link FDs do). Therefore, we release the assumption of deterministic MFDs at the subregional level, and we incorporate noise into the subregion MFD values. Figure 10(c) depicts the resulting subregion MFD curve along with its mean, upper and lower bounds. We assume that the production value corresponding to a certain accumulation level is uniformly distributed between upper and lower bounds. Note that the noise increases with accumulation in the region, which is consistent with the observations from real data and microscopic simulation models. We then incorporate the random MFD curves into our framework and apply the proposed model to evaluate the impact of noise on the overall performance. Figure 10(d) compares the regional accumulation values resulting from the scenario with noise to the base scenario where we assume deterministic MFD curves for the subregions. While the new scenario with noise leads to higher

accumulation values and greater variation in the central region (the most congested region), the framework is capable of mitigating the congestion. Total time spent in the two scenarios is very similar, which indicates the robustness of the proposed algorithm to the noise involved in estimations.

5. Conclusion

In this paper, we propose a hierarchical traffic management scheme based on path assignment and route guidance and report improved mobility in large-scale urban road networks. We describe region-level and subregion-level MFD-based dynamical traffic models and use them as a prediction model (for MPC) and as an evaluation model (i.e. *plant*), respectively. The contributions of the paper are twofold; (1) developing an ILP-based path assignment mechanism that can translate upper-level and aggregated control actuations into lower-level and disaggregated traffic decisions, (2) incorporating heterogeneity effect and variable trip lengths into the regional route guidance MPC framework. Efficiency of the proposed hierarchical scheme is tested in simulations in a 49 subregion network. The results indicate a great potential in making efficient use of network capacity via actuation over paths and achieving improved mobility. Such a hierarchical traffic management scheme can be implemented in real life applications, if data from GPS and loop detectors are combined to estimate the state variables described in the paper. Additionally, the proposed RG strategy is compared with the well studied perimeter control system and is proven to be more effective in congestion alleviation.

Future research should study the integration of the route guidance system with the perimeter control strategy, which is expected to further improve homogeneity and network performance. The perimeter control decisions that are optimized based on the region-based model could be further refined at the lower level (i.e. plant) through feedback controllers and limited traffic state information (as proposed by Ramezani et al. (2015)). Designing a realistic and accurate information feedback from the plant to the optimization or operation model needs further investigations. In addition, future research should look into replacing the subregional representation of the lower-level network with a more detailed link-level modeling of traffic.

References

Aboudolas, K., Geroliminis, N., 2013. Perimeter and boundary flow control in multi-reservoir heterogeneous networks. Transportation Research Part B: Methodological 55, 265–281.

An, K., Chiu, Y.C., Hu, X., Chen, X., 2017. A network partitioning algorithmic approach for macroscopic fundamental diagram-based hierarchical traffic network management. IEEE Transactions on Intelligent Transportation Systems PP, 1–10. URL: 10.1109/TITS.2017.2713808.

Andersson, J.A., Gillis, J., Horn, G., Rawlings, J.B., Diehl, M., 2018. CasADi: A software framework for nonlinear optimization and optimal control. Mathematical Programming Computation , 1–36.

Daganzo, C.F., 2007. Urban gridlock: Macroscopic modeling and mitigation approaches. Transportation Research Part B: Methodological 41, 49–62.

DePrator, A.J., Hitchcock, O., Gayah, V.V., 2017. Improving urban street network efficiency by prohibiting conflicting left turns at signalized intersections. Transportation Research Record: Journal of the Transportation Research Board , 58–69.

Diehl, M., Bock, H.G., Diedam, H., Wieber, P.B., 2006. Fast direct multiple shooting algorithms for optimal robot control, in: Fast motions in biomechanics and robotics. Springer, pp. 65–93.

Diehl, M., Ferreau, H.J., Haverbeke, N., 2009. Efficient numerical methods for nonlinear MPC and moving horizon estimation, in: Nonlinear model predictive control. Springer, pp. 391–417.

Gayah, V., Daganzo, C., 2011. Effects of turning maneuvers and route choice on a simple network. Transportation Research Record: Journal of the Transportation Research Board , 15–19.

Gayah, V., Dixit, V., 2013. Using mobile probe data and the macroscopic fundamental diagram to estimate network densities: Tests using microsimulation. Transportation Research Record: Journal of the Transportation Research Board , 76–86.

Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. Transportation Research Part B: Methodological 42, 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 Transactions on Intelligent Transportation Systems 14, 348–359.

Godfrey, J., 1969. The mechanism of a road network. Traffic Engineering and Control 11, 323–327.

Haddad, J., 2017a. Optimal coupled and decoupled perimeter control in one-region cities. Control Engineering Practice 61, 134–148.

Haddad, J., 2017b. Optimal perimeter control synthesis for two urban regions with aggregate boundary queue dynamics. Transportation Research Part B: Methodological 96, 1–25.

Haddad, J., Shraiber, A., 2014. Robust perimeter control design for an urban region. Transportation Research Part B: Methodological 68, 315–332.

Hajiahmadi, M., Knoop, V.L., De Schutter, B., Hellendoorn, H., 2013. Optimal dynamic route guidance: A model predictive approach using the macroscopic fundamental diagram, in: 16th International IEEE Conference on Intelligent Transportation Systems, IEEE. pp. 1022–1028.

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. Operations research 53, 600–616.

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: Journal of the Transportation Research Board , 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: Methodological 46, 1393–1403.

Keyvan-Ekbatani, M., Yildirimoglu, M., Geroliminis, N., Papageorgiou, M., 2015. Multiple concentric gating traffic control in large-scale urban networks. IEEE Transactions on Intelligent Transportation Systems 16, 2141–2154.

Knoop, V., Hoogendoorn, S., Van Lint, J., 2012. Routing strategies based on macroscopic fundamental diagram. Transportation Research Record: Journal of the Transportation Research Board , 1–10.

Kouvelas, A., Saeedmanesh, M., Geroliminis, N., 2017. Enhancing model-based feedback perimeter control with data-driven online adaptive optimization. Transportation Research Part B: Methodological 96, 26–45.

Lamotte, R., Geroliminis, N., 2017. The morning commute in urban areas with heterogeneous trip lengths. Transportation Research Part B: Methodological URL: https://doi.org/10.1016/j.trb.2017.08.023.

Leclercq, L., Chiabaut, N., Trinquier, B., 2014. Macroscopic fundamental diagrams: A cross-comparison of estimation methods. Transportation Research Part B: Methodological 62, 1–12.

Leclercq, L., Geroliminis, N., et al., 2013. Estimating mfds in simple networks with route choice. Transportation Research Part B: Methodological 57, 468–484.

Leclercq, L., Parzani, C., Knoop, V.L., Amourette, J., Hoogendoorn, S.P., 2015. Macroscopic traffic dynamics with heterogeneous route patterns. Transportation Research Procedia 7, 631–650.

Little, J.D., 1961. A proof for the queuing formula: L= λw. Operations research 9, 383–387.

Löfberg, J., 2004. YALMIP: A toolbox for modeling and optimization in MATLAB, in: IEEE International Symposium on Computer Aided Control Systems Design, IEEE. pp. 284–289.

Lopez, C., Leclercq, L., Krishnakumari, P., Chiabaut, N., Lint, H., 2017. Revealing the day-to-day regularity of urban congestion patterns with 3d speed maps. Scientific Reports 7, 14029.

Mariotte, G., Leclercq, L., Laval, J.A., 2017. Macroscopic urban dynamics: Analytical and numerical comparisons of existing models. Transportation Research Part B: Methodological 101, 245–267.

Menelaou, C., Kolios, P., Timotheou, S., Panayiotou, C., Polycarpou, M., 2017. Controlling road congestion via a low-complexity route reservation approach. Transportation research part C: emerging technologies 81, 118–136.

Nagle, A., Gayah, V., 2014. Accuracy of networkwide traffic states estimated from mobile probe data. Transportation Research Record: Journal of the Transportation Research Board , 1–11.

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, 245–266.

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

Ramezani, M., Nourinejad, M., 2017. Dynamic modeling and control of taxi services in large-scale urban networks: A macroscopic approach. Transportation Research Part C: Emerging Technologies .

Saberi, M., Mahmassani, H., Hou, T., Zockaie, A., 2014. Estimating network fundamental diagram using three-dimensional vehicle trajectories: Extending edie's definitions of traffic flow variables to networks. Transportation Research Record: Journal of the Transportation Research Board , 12–20.

Saeedmanesh, M., Geroliminis, N., 2016. Clustering of heterogeneous networks with directional flows based on snake similarities. Transportation Research Part B: Methodological 91, 250–269.

Saeedmanesh, M., Geroliminis, N., 2017. Dynamic clustering and propagation of congestion in heterogeneously congested urban traffic networks. Transportation Research Part B: Methodological 105, 193–211.

Sirmatel, I.I., Geroliminis, N., 2018. Economic model predictive control of large-scale urban road networks via perimeter control and regional route

guidance. IEEE Transactions on Intelligent Transportation Systems 19, 1112–1121. URL: https://doi.org/10.1109/TITS.2017.2716541.

Wächter, A., Biegler, L.T., 2006. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Mathematical Programming 106, 25–57.

Yang, K., Zheng, N., Menendez, M., 2017. Multi-scale perimeter control approach in a connected-vehicle environment. Transportation Research Part C: Emerging Technologies URL: https://doi.org/10.1016/j.trc.2017.08.014.

Yildirimoglu, M., Geroliminis, N., 2014. Approximating dynamic equilibrium conditions with macroscopic fundamental diagrams. Transportation Research Part B: Methodological 70, 186–200.

Yildirimoglu, M., Ramezani, M., Geroliminis, N., 2015. Equilibrium analysis and route guidance in large-scale networks with MFD dynamics. Transportation Research Part C: Emerging Technologies 59, 404–420.

Zheng, N., Rérat, G., Geroliminis, N., 2016. Time-dependent area-based pricing for multimodal systems with heterogeneous users in an agent-based environment. Transportation Research Part C: Emerging Technologies 62, 133–148.

Zheng, N., Waraich, R.A., Axhausen, K.W., Geroliminis, N., 2012. A dynamic cordon pricing scheme combining the macroscopic fundamental diagram and an agent-based traffic model. Transportation Research Part A: Policy and Practice 46, 1291–1303.

Zhong, R., Chen, C., Huang, Y., Sumalee, A., Lam, W., Xu, D., 2017. Robust perimeter control for two urban regions with macroscopic fundamental diagrams: A control-lyapunov function approach. Transportation Research Part B: Methodological URL: https://doi.org/10.1016/j.trb.2017.09.008.