Transportation Research Part C 94 (2018) 203–219

Contents lists available at ScienceDirect

Transportation Research Part C

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

Dynamic modeling and control of taxi services in large-scale urban networks: A macroscopic approach

Mohsen Ramezani [a,*], Mehdi Nourinejad [b]

[a] School of Civil Engineering, The University of Sydney, Australia
[b] Department of Civil Engineering, University of Toronto, Toronto, Canada

ARTICLE INFO

Article history:
Received 9 August 2017
Received in revised form 15 August 2017
Accepted 16 August 2017
Available online 1 September 2017

Keywords:
Meeting functions
Multi-user taxi market
Multimodal traffic
Network Fundamental Diagram (NFD)
Receding horizon

ABSTRACT

Taxis are increasingly becoming a prominent mobility mode in many major cities due to their accessibility and convenience. The growing number of taxi trips and the increasing contribution of taxis to traffic congestion are cause for concern when vacant taxis are not distributed optimally within the city and are unable to find unserved passengers effectively. A way of improving taxi operations is to deploy a taxi dispatch system that matches the vacant taxis and waiting passengers while considering the search friction dynamics. This paper presents a network-scale taxi dispatch model that takes into account the interrelated impact of normal traffic flows and taxi dynamics while optimizing for an effective dispatching system. The proposed model builds on the concept of the macroscopic fundamental diagram (MFD) to represent the dynamic evolution of traffic conditions. The model considers multiple taxi service firms operating in a heterogeneously congested city, where the city is assumed to be partitioned into multiple regions each represented with a well-defined MFD. A model predictive control approach is devised to control the taxi dispatch system. The results show that lack of the taxi dispatching system leads to severe accumulation of unserved taxi passengers and vacant taxis in different regions whereas the dispatch system improves the taxi service performance and reduces traffic congestion by regulating the network towards the undersaturated condition. The proposed framework demonstrates sound potential management schemes for emerging mobility solutions such as fleet of automated vehicles and demand-responsive transit services.

1. Introduction and motivation

Taxi service is a prominent mode of transportation in many major cities by providing a tailor-made and complimentary mobility solution to the public transport system. Despite their popularity, taxis contribute substantially to congestion as they tend to circulate in search of passengers in heavy-demand central business districts. The search process is initiated every time a taxi drops off a passenger and becomes vacant. The vacant taxi then cruises to find unserved passengers either in the same zone (where the former passenger was dropped off) or by driving to another zone. The heavy presence of vacant taxi movements is especially overwhelming because it impedes normal traffic and can lead to higher vehicle distance traveled in cities where taxis make up a significant ratio of the traffic mix. (Huo et al., 2012). In New York City, for instance, taxis make up 10% of the transportation mode share with approximately 450,000 taxi trips per day (Taxi and Limousine

* Corresponding author.
E-mail address: mohsen.ramezani@sydney.edu.au (M. Ramezani).

http://dx.doi.org/10.1016/j.trc.2017.08.011
0968-090X/

Nomenclature

Sets

R	set of homogeneous regions
U_i	set of regions in the vicinity of region i

Functions

$b_{ij}(t)$	taxi boarding rate from region i to region j at time t
G_i	trip completion rate in region i
P_i	total trip production in region i

Parameters

C	taxi fleet size
$l_i^{\mathrm{o}}(t)$	average trip length of occupied taxis in region i at time t
$l_i^{\mathrm{d}}(t)$	average trip length of dispatched taxis in region i at time t
$l_i^{\mathrm{n}}(t)$	average trip length of normal traffic in region i at time t
$q_{ij}^{\mathrm{c}}(t)$	taxi passenger demand from region i to region j at time t
$q_{ij}^{\mathrm{n}}(t)$	normal traffic demand from region i to region j at time t
γ	boarding function elasticity
α_k	weight of performance measure k in MPC objective function

Variables

$c_i(t)$	number of taxis in region i at time t
$c_i^{\mathrm{v}}(t)$	number of vacant taxis searching in region i at time t
c_{ij}^{d}	number of taxis dispatched from region i to j at time t
c_{ij}^{o}	number of occupied taxis traveling from region i to region j at time t
n_i	number of personal vehicles in region i at time t
n_{ij}	number of personal vehicles in region i with destination j at time t
p_i	number of waiting passengers in region i at time t
p_{ij}	number of waiting passengers in region i who wish to travel to region j at time t
$M_{ij}^{\mathrm{n}}(t)$	normal traffic transfer flow between region i and j at time t
$M_{ij}^{\mathrm{o}}(t)$	occupied taxi transfer flow between region i and j at time t
$M_{ij}^{\mathrm{d}}(t)$	dispatched taxi transfer flow between region i and j at time t
$M_{ii}^{\mathrm{o}}(t)$	taxi trip completion rate in region i at time t
$M_{ii}^{\mathrm{n}}(t)$	normal traffic trip completion rate in region i at time t
w_{ij}	rate of dispatching taxis from region i to region j at time t
$v_i(t)$	average speed in region i at time t

Commission, 2014). Other cities with moderate levels of taxi traffic are becoming increasingly aware of the future impacts of taxi traffic growth on normal traffic. The City of Toronto, Canada, for instance, estimates to have a total of 2.1 million additional taxi trips by 2022 (City of Toronto, 2016), which requires imposing suitable taxi regulation measures and developing intelligent taxi management systems.

An inefficient taxi service leads to longer vacant taxi travel times, longer passenger waiting times, lower taxi utilization, and traffic congestion. Hence, efficient taxi dispatching should be considered as an imperative part of taxi management systems, which uses real-time information of the fleet to enhance the overall performance of the taxi service. Zhan et al. (2015) showed that taxi dispatch systems, in an ideal case, can reduce the total cost of empty taxi trips by up to 90%. A holistic taxi dispatch system must take into account the interrelated effects of taxis on other traffic modes, and vice versa, while optimizing a network-wide objective criterion. On one hand, excessive and unwarranted taxi dispatching can further impede mixed traffic flow in cities where taxis make up a substantial ratio of the traffic. On the other hand, taxis (either occupied, vacant, or dispatched) are themselves affected by normal traffic conditions as they can incur a long travel time in presence of traffic congestion. This study proposes a dynamic bimodal traffic flow model and a taxi dispatching system that incorporates the interrelated dynamics of taxis and other transport modes (e.g. personal vehicles) to optimize a desirable network objective, which is assumed to be a weighted measure of several performance indicators such as passenger waiting time (for a taxi), taxi searching time (for an unserved passenger), and total network travel time of personal vehicles and occupied taxis. The presented taxi dispatching control method is robust, computationally tractable, and realistically applicable for taxi regulation that also advances the findings of taxi market analysis.

Taxi regulation has been the subject of many studies as early as the 1970s. The prevalent literature on taxi regulation focuses on finding the optimal fare price and taxi fleet size (Beesley, 1973), modeling the network-wide movement of taxis (Yang and Wong, 1998; Jung et al., 2014; Sayarshad and Chow, 2016), elastic taxi demand (Wong et al., 2001), competitive

taxi markets (Yang et al., 2002), stochastic taxi passenger demand (Zhang and Ukkusuri, 2016), taxi mode choice (Wong et al., 2008), taxi e-hailing (He and Shen, 2015; Wang et al., 2016), local taxi movements in a cell-based network (Wong et al., 2014), demand-responsive services (Amirgholy and Gonzales, 2016), and analyzing the friction in the taxi-passenger meeting process (Yang and Yang, 2011; Yang et al., 2014; Wang et al., 2016; Zha et al., 2016). A common ground among these studies is the equilibrium assumption in the vacant taxi movement decisions. Under equilibrium conditions, each vacant taxi travels to the closest and most profitable zone to search for customers. Consequently, at equilibrium, no vacant taxi can unilaterally change its destination, i.e. the next search zone, for a lower cost. From these studies, only a few consider the interrelated effects of taxis and normal traffic: Wong et al. (2001) developed a bi-level model which characterizes the simultaneous movements of vacant and occupied taxis as well as normal traffic flows, Wong et al. (2008) incorporated mode choice (between taking a taxi or a personal vehicle) in a traffic equilibrium model with taxi flows, and Yang et al. (2014) proposed a bilateral taxi-passenger meeting model while considering the congestion externalities of normal traffic and taxi flows. These studies and taxi equilibrium models in particular, however, fall short in representing the dynamic features of traffic congestion, do not capture the spatiotemporal accumulation of vacant taxis and passengers, and require intensive calibration and estimation of origin-destination trip matrices to derive an equilibrium pattern of link travel times. Moreover, the models of taxi equilibrium pose severe empirical limitations in realistic application as they employ link travel cost functions which cannot specify accurately the inter-day traffic dynamics (Tsekeris and Geroliminis, 2013).

With the recent proliferation of taxi hailing apps, such as Uber-taxi and Didi KuaiDi-taxi, optimization of taxi dispatching plays an important role to efficiently match unserved passengers with vacant taxis (Nie, 2017). Many of the studies on optimal taxi dispatching address the problem as either a variant of the vehicle routing problem (VRP) (Ghiani et al., 2003; Wong and Bell, 2006; Pillac et al., 2013; Hosni et al., 2014; Jung et al., 2016) or the bipartite graph matching problem (Agatz et al., 2012; Zhan et al., 2015; Nourinejad and Roorda, 2016). Under the VRP formulation, each taxi is assigned to sequentially pick up a number of passengers and under the bipartite graph formulation, each taxi is matched with the closest passenger in its vicinity. Both modeling frameworks can easily take into account the impact of traffic conditions (in terms of travel times) on the optimal taxi dispatch plan. However, neither of the two frameworks capture the impact of taxis on normal traffic flows which can be detrimental in major cities with a high taxi mode share. As an example, Lee et al. (2004), Seow et al. (2010), and Zhan et al. (2015) consider the impact of normal traffic on taxis by proposing a taxi dispatching system that uses real-time traffic conditions to match vacant taxis with the nearest (in terms of travel time) waiting passengers. Although these studies consider the impact of normal traffic flow on taxi dispatching, the reverse relationship (i.e., impact of taxis on normal traffic flow) is disregarded. Incorporating the effects of normal traffic flows on taxi dispatch models is a non-trivial task because it requires a modeling framework that captures the complex dynamics of traffic flow propagation and can be applied on large-scale networks.

The recent developments on urban network traffic flow modeling show promising outcomes to replicate the dynamics of traffic congestion and propagation in large-scale networks (Daganzo and Geroliminis, 2008; Geroliminis and Daganzo, 2008). The proposed modeling builds on the Macroscopic (Network) Fundamental Diagram (MFD) approach that provides a relationship between network-wide aggregated traffic states. The MFD captures the collective traffic flow dynamics of an urban region and relates the urban region outflow (production) to the accumulation (density) of vehicles while a homogeneous traffic state is prevalent (Buisson and Ladier, 2009; Geroliminis and Sun, 2011b; Saberi et al., 2014; Laval and Castrillón, 2015).

The analytically tractable nature of the MFD has lead to several studies in traffic modeling and management applications including route guidance management (Yildirimoglu et al., 2015; Knoop et al., 2012), traffic control in large-scale urban networks (Geroliminis et al., 2013; Keyvan-Ekbatani et al., 2012, 2015; Haddad et al., 2013, 2017), parking management (Zheng and Geroliminis, 2016), and traffic management of multi-modal networks (Zheng and Geroliminis, 2013; Geroliminis et al., 2014). Furthermore, several studies have investigated methods of estimating the MFD state (Gayah and Dixit, 2013; Leclercq et al., 2014; Ambühl and Menendez, 2016), the properties of a well-defined MFD (Geroliminis and Sun, 2011a,b), effect of routing on MFD dynamics (Leclercq and Geroliminis, 2013; Yildirimoglu et al., 2015), and properties of congestion heterogeneity (Daganzo et al., 2011; Ji and Geroliminis, 2012; Mahmassani et al., 2013; Ramezani et al., 2015).

Using the MFD as the basis of large-scale urban traffic modeling, this paper aims at developing a dynamic bimodal (cars and taxis) traffic modeling and control strategy, i.e. taxi dispatching, to improve urban mobility and mitigate congestion in cities. The proposed model incorporates two aggregate models, (i) a taxi-passenger meeting function that considers network conditions (i.e. network average speed) and (ii) MFD-based traffic dynamics of personal vehicles and taxis. The taxi-passenger meeting function models the interactions between the vacant taxis and unserved passengers without explicit consideration of (vacant taxi) trip length, while the personal vehicles and occupied and dispatched taxi dynamics are modeled based on MFD and given average trip lengths. In numerical experiments, we demonstrate that ignoring the effect of network average speed on the taxi-passenger meeting function leads to significant performance reduction in the dispatch control system. Furthermore, the model is general to capture the dynamics of multiple taxi firms operating in the same city. This study quantifies the benefits (such as reducing passenger waiting time and increasing taxi fleet utilization) of equipping one taxi firm with a taxi dispatch system. Moreover, we present several taxi dispatch strategies and assess the impact of each strategy on the taxi passengers waiting time and traffic flows of occupied and vacant taxis and normal vehicles.

The taxi dispatch scheme is developed based on the Model Predictive Control (MPC) approach that tackles the optimal taxi dispatch control problem. It was shown that the MPC approach is robust to traffic state (e.g. number of taxis, personal

cars, and passengers) measurement error (Geroliminis et al., 2013; Haddad et al., 2013) and modeling mismatches (Ramezani et al., 2015) in large-scale traffic control problems. The robustness of MPC approach is a direct consequence of a feedback loop within the MPC framework that, at every control decision step, prevents the accumulation of errors over time. The performance of the developed MPC dispatch system, with state measurement error, shows convincing evidence of the applicability of this approach in realistic case studies.

The remainder of the paper is organized as follows. Section 2 presents the taxi-passenger boarding function modeling and estimation followed by the incorporation of taxis and personal cars with MFD modeling. Section 3 elaborates the control procedure of the large-scale taxi dispatching system. Section 4 introduces the proposed taxis dispatching strategies. The results are discussed in Section 5. Conclusions and future works are drawn in Section 6.

2. Methodology

The modeling of the proposed bimodal traffic dynamics are based on two aggregate models, (i) a network-wide taxi-passenger search and boarding model that incorporates the prevailing network traffic states (Section 2.1) and (ii) a large-scale MFD-based traffic model that integrates the both modes, i.e. the normal vehicles and taxis (Section 2.3). The model considers an urban network that is heterogeneously congested, while assuming a partitioning procedure is applied on the network that results in a set of homogeneously congested regions, where the dynamics of normal traffic and taxis within each region are represented by a well-defined MFD. Furthermore, the model considers multiple taxi firms that might have different operational characteristics such as fleet size or fleet management scheme. For the sake of brevity, we introduce the R-region, 1 taxi firm model. The general case of R-region, F-taxi firm is introduced and elaborated in Appendix A.

2.1. Modeling taxi-passenger boarding

The majority of studies on taxi modeling have focused on the effects of taxi availability on customer waiting time and the resulting market equilibrium. Recently, however, more emphasis is directed towards the bilateral passenger-taxi searching and the emergent meeting (boarding) relationship. To model the meeting process of vacant taxis and unserved passengers, some studies have employed the concept of bilateral meeting functions (Yang et al., 2010, 2014; Yang and Yang, 2011). Meeting functions present the rate that two groups of agents (e.g. passengers and taxis) meet each other as a function of the group size of the two agents and other network characteristics such as average network speed.

The meeting function models the matching dynamics between agents that need to spend resources (e.g. moving in the network searching) to be matched with each other. For taxi services, a meeting function was first introduced by Schroeter (1983). Later, Lagos (2000) developed a theoretical model of the meeting frictions and dynamics that results in an aggregate meeting function. The searching and meeting phenomenon of taxis and passengers in urban networks exhibits frictions that is a state when both vacant taxis and unserved passengers are in the system. (Note that in single point systems, e.g. taxi stands in airports, and assuming zero taxi loading time, the queue of taxis and waiting passengers cannot coexist.) The bilateral searching and meeting phenomenon between agents has been studied in economics to model implications and outcomes of labor markets (Mortensen and Pissarides, 1994; Petrongolo and Pissarides, 2001).

The already investigated taxi-passenger meeting functions assume fixed travel speed. For a dynamic taxi dispatch system, however, the urban network speed is dynamically changing based on the level of traffic congestion captured with MFD modeling. Hence, it is important to understand how the meeting function relates to speed and other features of the network.

The meeting (boarding) rate function B is defined as:

$$b = B(c^v, p, v) \tag{1}$$

where $c^v$is the density of vacant taxis [taxi/m^2], p is the density of passengers [passenger/m^2], and v [m/s] is the network average speed. The meeting function has the following properties: $\partial B/\partial c^v > 0, \partial B/\partial p > 0$, and $\partial B/\partial v > 0$. In addition, $B \to 0$ as either $c^v \to 0, p \to 0$, or $v \to 0$. Furthermore, the meeting function elasticities with respect to c^v, p, and v are, respectively

$$\gamma_1 = \frac{\partial B}{\partial c^v}\frac{c^v}{B} \tag{2}$$

$$\gamma_2 = \frac{\partial B}{\partial p}\frac{p}{B} \tag{3}$$

$$\gamma_3 = \frac{\partial B}{\partial v}\frac{v}{B} \tag{4}$$

To employ the meeting function in the MFD based taxi dispatch model, the boarding rate $b_{ij}(t)$ [1] of passengers that are traveling from region i to region $j$at time t is estimated as:

[1] The unit of boarding (meeting) rate is the number of pick up of unserved passengers by vacant taxis per unit time, i.e. $(customer - taxi/s)$.

$$b_{ij}(t) = \frac{p_{ij}(t)}{p_i(t)} B(c_i^v(t), p_i(t), v_i(t)) \tag{5}$$

where $c_i^v(t)$ is the density of vacant taxis in region i, $p_i(t)$ is the density of unserved passengers in region i, $p_{ij}(t)$ is the density of unserved passengers who wish to travel from i to j, and $v_i(t)$ is the region i speed at time t estimated from the MFD. The ratio $p_{ij}(t)/p_i(t)$ in Eq. (5) is the probability that a randomly selected unserved passenger in region i travels to region j at time t. This is consistent with the assumption that traffic congestion and demand of taxi trips are homogeneous in the regions.

2.2. Estimation of the boarding function

We now investigate the properties of the meeting (boarding) function using a simulation model. The purpose of the simulation model is to assess the impact of the vacant taxi density c^v, unserved passenger density p, and the network average velocity v on the meeting function. The simulation model is set up as follows. Consider a Manhattan grid network with a common link speed v [m/s] which is an input of the model (in our study there are 121 nodes and 200 bi-directional links). Note that as the area of the region is constant and fixed, the density of taxis is directly proportional to the number of taxis in the region. In the network, there are c^v vacant taxis that cruise around in search of p unserved passengers. The simulation model is a discrete-event simulation where the taxi cruising behavior is assumed such that when a taxi reaches an intersection, it randomly chooses to enter one of the three adjacent links, i.e. the taxi either turns left or right or proceeds to drive through at the intersection. This is a valid assumption on the search pattern of taxis when passengers are homogeneously distributed in the network. Each vacant taxi continues to cruise until it finds a passenger. When a vacant taxi meets a passenger, they are matched and eliminated from the simulation. After each taxi-passenger meeting, a new vacant taxi and a new passenger are randomly generated in the network such that the number of vacant taxis and passengers remains constant throughout the simulation. The simulation lasts for a period of time to smooth the effect of transient conditions. The agents in the discrete-event simulation are (moving) vacant taxis and (not-moving) unserved passengers while the locations of taxis (on which link they are) are tracked. Hence this is a microscopic simulation, though the detailed driving behavior and dynamics of taxis are not modeled.

With these assumptions, a meeting rate is obtained for the triple (c^v, p, v). By varying the range of the triple (c^v, p, v), we fit the simulated meeting rates to following Cobb-Douglas type meeting function

$$B(c^v(t), p(t), v(t)) = A c^v(t)^{\gamma_1} p(t)^{\gamma_2} v(t)^{\gamma_3} \tag{6}$$

where A is a constant and γ_1, γ_2, and γ_3 are the time-invariant meeting function elasticities. The simulated and estimated meeting rates are presented in Fig. 1 for three cases of low, medium, and high network average velocity. As is evident in Fig. 1, the meeting rate increases non-linearly with respect to c^v and p. The estimated elasticities in the Cobb-Douglas function are all statistically significant and the model has a goodness-of-fit with *adjusted*-$R^2 = 0.82$.

The current research on the taxi-passenger meeting function commonly assumes that the network average speed v is constant and hence disregards the impact of the network average speed on the meeting function. While this assumption is justified for the existing static equilibrium models where velocity is constant (Yang and Yang, 2011; Yang et al., 2014), the assumption loses its validity in dynamic systems where v is continually changing with respect to the number of vehicles in the network. Thus, it is important to have an explicit representation of v in the meeting function as is done in Eq. (6). To assess the effect of disregarding the network average speed in the meeting function, we fit the simulated meeting rates to a second Cobb-Douglas type meeting function of the form $B(t) = A c^v(t)^{\gamma_1} p(t)^{\gamma_2}$ where A is a constant and γ_1 and γ_2 are the meeting function elasticities. This newly estimated model has a lower goodness-of-fit with *adjusted*-$R^2 = 0.51$ which demonstrates the importance of including the network average speed in the meeting function in dynamic traffic modeling where traffic congestion changes over time. Furthermore in the results Section, we investigate and compare the performance of a dispatch strategy that does not consider the effect of the network average speed on the meeting function.

The presented simulation model assumes that each vacant taxi cruises randomly until it finds a passenger. This assumption does not accurately represent cases where passengers and vacant taxis are matched through a communication channel such as a taxi hailing mobile application. To assess whether the bilateral meeting model can also capture this type of matching behavior, we develop a simulation model that assumes a ratio of the passengers and vacant taxis are equipped and have access to a matching device. Each equipped vacant taxi is automatically matched with the closest (in terms of distance) equipped passenger. Once matched, the taxi travels the shortest route to reach the location of the passenger. We calculate the meeting rate as a function of the number of passengers and taxis and we use the Cobb-Douglas function to estimate the two elasticities (γ_1 and γ_2 as shown in Eq. (6)). We find that the bilateral meeting function accurately captures, with a high goodness-of-fit ratio, the boarding process with equipped taxis and passengers. The estimated elasticities show that returns-to-scale (i.e., $\gamma_1 + \gamma_2$) increases with respect to the taxi hailing application penetration ratio. Future research can further reveal the effect of technological advancements on the meeting function properties.

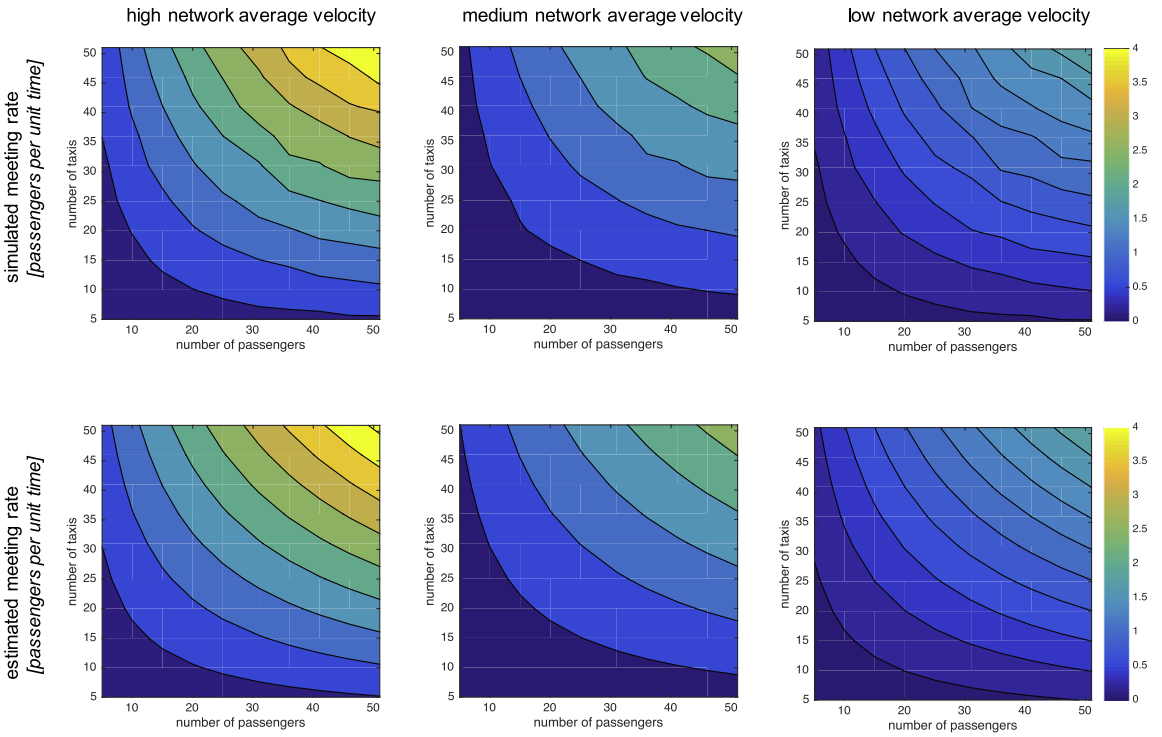

Fig. 1. Simulated (top) and estimated (bottom) boarding functions at high (left), medium (middle), and low (right) network average velocities.

2.3. Model formulation: integration of MFD within taxi dispatching

Consider an urban network that is partitioned into a set of homogeneous regions denoted by R as shown in Fig. 2. The fleet of taxis in every region i is divided into occupied, vacant, and dispatched taxis. The occupied taxis are assumed to move one passenger from her/his origin to her/his destination. Consequently, the occupied taxi movement pattern is similar to personal vehicles. On the contrary, the movement pattern of vacant taxis is a search cruising behavior with no explicit destination. That is, each vacant taxi is assumed to search in the same region until it finds an unserved passenger. This assumption is valid when the regions are large enough so that it is economically unjustified for vacant taxis to voluntarily travel to another region unless they are dispatched. The dispatched taxis are the vacant taxis that the dispatching system advises to move to other regions. Although the dispatched taxis do not carry any passenger, they can (i) reduce traffic congestion in case of

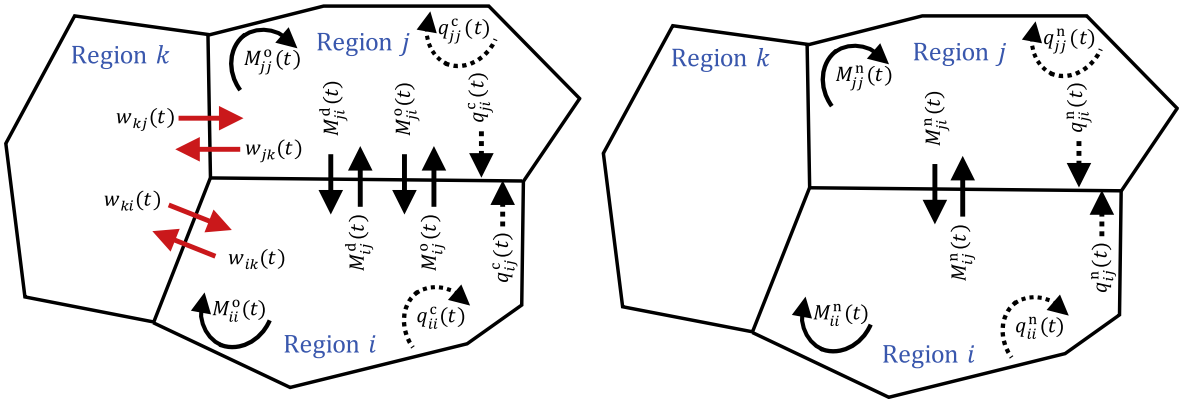

Fig. 2. Schematic of a heterogeneous urban network that is partitioned to several homogeneous regions with well-defined low scatter MFD representations: (left) internal and external taxi movements (solid lines) and exogenous taxi travel demands (dashed lines), (right) internal and external normal vehicle movements (solid lines) and exogenous normal vehicle travel demands (dashed lines). The model integrates both traffic flow dynamics. The red arrows indicate part of the taxi dispatch movements. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

excessive accumulation of cars in a region and (ii) increase the boarding rate of unserved passengers (i.e. equivalently increase the taxi fleet utilization) in case there is an imbalance between the vacant taxis and unserved passengers in a region. The following equations present conservation of taxis at each region:

$$c_i^{\mathrm{v}}(t) + c_i^{\mathrm{o}}(t) + c_i^{\mathrm{d}}(t) = c_i(t) \quad \forall i \in R \tag{7}$$

$$\sum_{i \in R} c_i(t) = C \tag{8}$$

where $c_i^{\mathrm{v}}(t), c_i^{\mathrm{o}}(t)$ and $c_i^{\mathrm{d}}(t)$ denote the number of vacant, occupied, and dispatched taxis in region i at time t, respectively, $c_i(t)$ denotes the total number of taxis in region i at time t, and C is the total taxi fleet size. Eq. (7) ensures that the total number of occupied, vacant, and dispatched taxis is the total number of taxis in each region, and Eq. (8) represents the total taxi fleet size. Note that the generalized model to consider multiple taxi firms is introduced in Appendix A.

Each region i is modeled with an MFD defined by the trip completion rate function $G_i(n_i(t) + c_i(t)) = P_i(n_i(t) + c_i(t))/l_i(t)$ [*vehicles per unit time*] where $P_i(.)$ [*vehicle distance traveled per unit time*] is the total production at region i, $l_i(t)$ is the average trip length in region i, and $n_i(t)$ is the total number of personal vehicles in region i at time t. The average speed of region i at time t is $v_i(t) = P_i(n_i(t) + c_i(t))/(n_i(t) + c_i(t))$.

By assuming that trip completion flows are proportional to accumulations, the transfer flows between region i and j for occupied taxis $M_{ij}^{\mathrm{o}}(t)$, personal vehicles $M_{ij}^{\mathrm{n}}(t)$, and dispatched taxis $M_{ij}^{\mathrm{d}}(t)$ at time t are respectively the following:

$$M_{ij}^{\mathrm{o}}(t) = \frac{c_{ij}^{\mathrm{o}}(t)}{n_i(t) + c_i(t)} \frac{P_i(n_i(t) + c_i(t))}{l_i^{\mathrm{o}}(t)} \quad \forall i \in R, j \in U_i \tag{9}$$

$$M_{ij}^{\mathrm{n}}(t) = \frac{n_{ij}(t)}{n_i(t) + c_i(t)} \frac{P_i(n_i(t) + c_i(t))}{l_i^{\mathrm{n}}(t)} \quad \forall i \in R, j \in U_i \tag{10}$$

$$M_{ij}^{\mathrm{d}}(t) = \frac{c_{ij}^{\mathrm{d}}(t)}{n_i(t) + c_i(t)} \frac{P_i(n_i(t) + c_i(t))}{l_i^{\mathrm{d}}(t)} \quad \forall i \in R, j \in U_i \tag{11}$$

where $c_{ij}^{\mathrm{d}}(t), c_{ij}^{\mathrm{o}}(t)$, and $n_{ij}(t)$ are the number of dispatched taxis, occupied taxis, and personal vehicles that travel from region i to region j at time t, respectively; and $l_i^{\mathrm{d}}(t), l_i^{\mathrm{o}}(t)$, and $l_i^{\mathrm{n}}(t)$ are the average travel distances of dispatched taxis, occupied taxis, and personal vehicles in region i at time t, respectively. U_i denotes the set of regions that are in the direct vicinity of region i. In addition, $\sum_{j \in U_i} c_{ij}^{\mathrm{d}}(t) = c_i^{\mathrm{d}}(t), \sum_{j \in (U_i \cup i)} c_{ij}^{\mathrm{o}}(t) = c_i^{\mathrm{o}}(t)$, and $\sum_{j \in (U_i \cup i)} n_{ij}(t) = n_i(t)$. Accordingly, the trip completion flows within the regions are

$$M_{ii}^{\mathrm{o}}(t) = \frac{c_{ii}^{\mathrm{o}}(t)}{n_i(t) + c_i(t)} \frac{P_i(n_i(t) + c_i(t))}{l_i^{\mathrm{o}}(t)} \quad \forall i \in R \tag{12}$$

$$M_{ii}^{\mathrm{n}}(t) = \frac{n_{ii}(t)}{n_i(t) + c_i(t)} \frac{P_i(n_i(t) + c_i(t))}{l_i^{\mathrm{n}}(t)} \quad \forall i \in R \tag{13}$$

where $M_{ii}^{\mathrm{o}}(t)$ is the rate of occupied taxi trip completion and $M_{ii}^{\mathrm{n}}(t)$ is the rate of personal vehicle trip completion within region i at time t. There is no internal trip completion rate for dispatched taxis because by definition the dispatched taxis are traveling from their current region to a neighbor region to avoid excessive search cruising for unserved passengers.

The evolution of occupied taxis based on their destination (i.e. internal or external) over time is formulated as:

$$\frac{\mathrm{d}c_{ii}^{\mathrm{o}}(t)}{\mathrm{d}t} = b_{ii}(t) + \sum_{j \in U_i} M_{ji}^{\mathrm{o}}(t) - M_{ii}^{\mathrm{o}}(t) \quad \forall i \in R \tag{14}$$

$$\frac{\mathrm{d}c_{ij}^{\mathrm{o}}(t)}{\mathrm{d}t} = b_{ij}(t) - M_{ij}^{\mathrm{o}}(t) \quad \forall i \in R, j \in U_i \tag{15}$$

In Eq. (14) and (15), the first terms of the RHS denote the meeting rate of vacant taxis and unserved passengers, which indicate the change rate of vacant taxis to occupied taxis (i.e. $b_{ij}(t)$ is the rate that vacant taxis board passengers and become occupied, hence the positive sign in (14) and (15), and accordingly, the negative sign in (20). Moreover, $b_{ij}(t)$ indicates the rate that unserved waiting passengers board taxis, hence the negative sign in (18) and (19).)

Conservation of the number of personal vehicles is ensured via the following equations:

$$\frac{\mathrm{d}n_{ii}(t)}{\mathrm{d}t} = q_{ii}^{\mathrm{n}}(t) + \sum_{j \in U_i} M_{ji}^{\mathrm{n}}(t) - M_{ii}^{\mathrm{n}}(t) \quad \forall i \in R \tag{16}$$

$$\frac{\mathrm{d}n_{ij}(t)}{\mathrm{d}t} = q_{ij}^{\mathrm{n}}(t) - M_{ij}^{\mathrm{n}}(t) \quad \forall i \in R, j \in U_i \tag{17}$$

where $q_{ij}^{\mathrm{n}}(t)$ is the exogenous demand of personal vehicles traveling from region i to region j ($j \in U_i$) at time t.

The evolution of the taxi passenger demand (unserved passengers) is tracked through the following equations:

$$\frac{dp_{ii}(t)}{dt} = q_{ii}^{c}(t) - b_{ii}(t) \quad \forall i \in R \tag{18}$$

$$\frac{dp_{ij}(t)}{dt} = q_{ij}^{c}(t) - b_{ij}(t) \quad \forall i \in R, j \in U_i \tag{19}$$

where $q_{ij}^{c}(t)$ is the exogenous passenger taxi demand from region i to region j ($j \in U_i$) at time t. We assume that the travel demand for personal vehicles and taxis are inputs of the model without any explicit elasticity on the network conditions. This can be relaxed by integrating a demand model that considers the network performance on the demand split between the two modes. This is a future research direction.

The conservation of the number of vacant taxis is ensured via the following equation:

$$\frac{dc_i^{v}(t)}{dt} = M_{ii}^{o}(t) + \sum_{j \in U_i} M_{ji}^{d}(t) - \sum_{j \in \{U_i \cup i\}} b_{ij}(t) - \sum_{j \in U_i} w_{ij}(t) \quad \forall i \in R \tag{20}$$

where $w_{ij}(t)$ is the rate of dispatching taxis from region i to region j at time t. The dispatching rate, i.e. $w_{ij}(t)$, is the control variable in the model that represents the rate that a part of vacant taxis in region i are dispatched to travel to region j to search for unserved passengers. Thus, $w_{ij}(t)$ is the change rate of vacant taxis to dispatched taxis, which is demonstrated as a negative term in Eq. (20) and a positive term in Eq. (21). In addition, it is assumed that the vacant taxis remain in their region and there is no systematic change of regions in the search behavior of vacant taxis, unless they are dispatched to another region. To account for the vacant taxis that voluntarily change their search region, we introduce measurement noise that captures fluctuations in number of vacant taxis.

Finally, the dynamics of dispatched taxis are modeled as:

$$\frac{dc_{ij}^{d}(t)}{dt} = w_{ij}(t) - M_{ij}^{d}(t) \quad \forall i \in R, j \in U_i \tag{21}$$

This equation assumes that once a vacant taxi is dispatched its flag is set as occupied (though without carrying a passenger) till it reaches to the other region and become a vacant taxi in the new region, see the positive sign of $M_{ij}^{d}(t)$ in (20). Thus, we expect that once the dispatch system determines $w_{ij}(t) > 0, c_{ij}^{d}(t)$ exhibits a sharp raise followed by a smooth decay that is the direct outcome of MFD dynamics.

3. Model predictive control for taxis dispatching

The aim of the taxi dispatch system is to regulate taxi dispatch rates (i.e., $w_{ij}(t)$) to minimize the total network delay that is defined as the integral of the weighted accumulations of personal vehicles, taxis (occupied, vacant, and dispatched), and passengers with respect to time:

$$\min_{w_{ij}(t)} J = \int_{t_0}^{t_f} \left(\alpha_1 \sum_{i \in R} n_i(t) + \alpha_2 \sum_{i \in R} c_i^{o}(t) + \alpha_3 \sum_{i \in R} c_i^{v}(t) + \alpha_4 \sum_{i \in R} c_i^{d}(t) + \alpha_5 \sum_{i \in R} p_i(t) \right) dt \tag{22}$$

where $t_f[s]$ is the final time and $t \in [t_0, t_f]$ is the control time. The first term of the objective function is the total travel time of normal traffic, the second term is the total travel time of occupied taxis that is a proxy of the taxi fleet utilization factor, the third term of the objective function is the total travel time of vacant taxis that is a proxy of taxi search time, the fourth term in the objective function is the total travel time of dispatched taxis, and the last term denotes the total waiting time of all passengers. The weights (α_1 to α_5) in the objective function J regulate the trade off between the rank of the five components. For instance, a taxi dispatch system devised by a taxi company, targeting for a higher taxi service quality, stipulates a high α_5 and a low α_1, whereas a city operator devising a holistic taxi dispatching system uses a more balanced weighting scheme. In the numerical case studies presented in Section 5, we consider different combination of the weights to scrutinize the effect of weights on the system performance.

The taxi dispatching control problem is solved by the model predictive control (MPC) approach. The MPC approach is suitable for performing tractable optimization which can be implemented for real-time applications. The MPC framework effectively handles different levels of error in traffic demand and noise in traffic state measurements (Geroliminis et al., 2013). In addition, MPC is efficient to address constraints on states and control variables. Specifically, the MPC solution, $w_{ij}(t)$, should satisfy the following:

$$0 \leqslant n_{ii}(t), n_{ij}(t), p_{ii}(t), p_{ij}(t), c_{ii}^{o}(t), c_{ij}^{o}(t), c_i^{v}(t), c_{ij}^{d}(t) \quad \forall i \in R, j \in U_i, t \in [t_0, t_f] \tag{23}$$

while

$$n_i(t) + c_i(t) \leqslant n_i^{jam} \quad \forall i \in R, t \in [t_0, t_f] \tag{24}$$

$$0 \leqslant w_{ij}(t) \quad \forall i \in R, j \in U_i, t \in [t_0, t_f] \tag{25}$$

where constraints (23) ensure that the personal vehicles, taxis, and passengers accumulations remain positive, constraints (24) set an upper-bound on each region total vehicle accumulations (i.e., n_i^{jam} for region i), and constraints (25) ensure non-negativity of taxi dispatch rates. In addition, it is straightforward to set an upper bound on the taxi dispatch rates if there is a policy in this regard enforced by the taxi firms.

The MPC framework is developed based on the receding horizon feature (or rolling time horizon). At each time step, the MPC optimizes the objective function J over a finite prediction horizon of N_p time steps to derive a sequence of optimal control inputs. Afterwards, only the first time duration of the control input is applied to the system and the procedure is carried out again with a shifted horizon, see Fig. 3. Note that the prediction horizon is shorter than the total control process time such that the control time $t \in [t_0, t_f]$ is covered by several overlapping prediction time horizons. Furthermore, to regulate the optimization computation resources, one can introduce a control horizon of N_c time steps such that within each optimization procedure only the first N_c time steps are optimized and the rest of the control inputs remain constant.

The proposed MPC framework in Fig. 3 includes a taxi and personal travel demand prediction module that is the input to the prediction model. In this study, we add a white noise to the actual travel demands to replicate prediction error of travel demands. The feedback loop from the plant to the prediction model provides an estimation of traffic states. To account for traffic state estimation errors, we assume an additive white noise to the traffic states. The taxi dispatch controller consists of the prediction model, i.e. Eqs. (5)–(21) and an optimization tool that minimizes (22) subject to (23)–(25). The optimization output is the taxi dispatch rate between the regions, i.e. $w_{ij}(t)$, which is applied to the plant for one interval. The main systematic difference between the plant and the optimization model is modeling errors that comprise errors in MFD and bilateral search function parameters.

The selection of the prediction horizon N_p and the control horizon N_c affects the performance of the MPC taxi dispatch controller. The prediction horizon should be long enough so the model accurately predicts the complex dynamics of the system traffic states. Thus, a longer prediction horizon improves the performances of the MPC controller but increases the optimization computation time that may create difficulties for real-time field implementation. Similar consideration is valid for the control horizon in regard to the trade off between computation complexity and controller performance. Fig. 4 depicts the results of tuning the MPC framework parameters N_p and N_c where the relative improvements of the objective function corresponding to the MPC controller compared with the no dispatch policy over a range of N_p and N_c are shown.

Fig. 4 illustrates that the controller improvement is less sensitive to the control horizon N_c, where $N_c \geqslant 2$ yields similar results, whereas $N_c = 1$ results in a low MPC performance. Note that, for prediction horizon $N_p \leqslant 22$, the MPC controller does

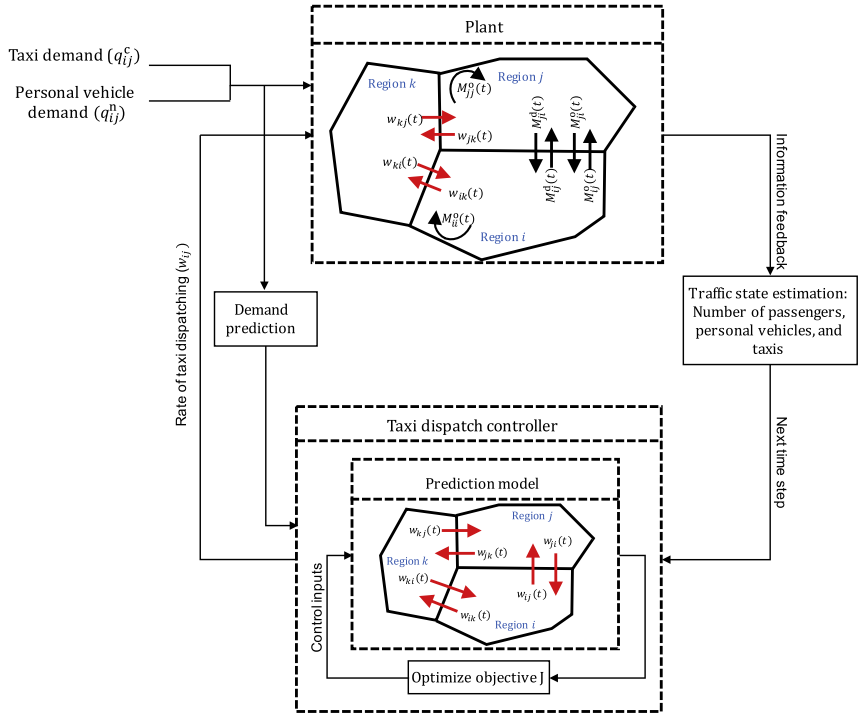

Fig. 3. Proposed MPC framework for taxi dispatching. The Demand prediction module provides a prediction of demand during the prediction horizon to the controller. The Traffic state estimation module provides the current state measurements to the controller. Both parts include prediction errors and estimation noise.

not perform better than the no dispatch policy. Accordingly, we choose the MPC parameters as $N_p = 40$ and $N_c = 4$ for the case study examples while the control step duration is set to 60 [s].

4. Taxi dispatch control strategies

Choosing a taxi dispatch control strategy depends on the objective of the dispatcher entity. As an example, a privately run taxi firm may dispatch its taxi fleet to improve the level-of-service by decreasing the average passenger waiting time and increasing the taxi fleet utilization, whereas a publicly run transport provider may consider the network-wide congestion effects of taxi dispatching as well. In this section, we introduce five taxi dispatch control strategies to investigate the properties of the proposed modeling and control framework. The strategies are designed to demonstrate the benefits of taxi dispatching using the MPC framework. The first controller, called the "no dispatch" strategy, assumes lack of any form of taxi dispatching. Analysis and comparison of this strategy to other strategies demonstrates the benefits of implementing a taxi dispatch system. The second strategy is the "MPC controller" which is designed to minimize the total travel time of all vehicles (i.e., taxis and normal vehicles) and the total waiting time of passengers. The MPC controller is established by setting all the weights of the objective function (22) equal to one, i.e., $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 1$. The third controller is the "MPC without velocity modeling". This controller allows us to assess the importance of including the network average velocity in the boarding function which was highlighted in Section 2.1. The only difference between the second and third controllers is the modeling of the boarding function in the prediction model, i.e. Eq. (6) in "MPC without velocity modeling" does not consider v.

The fourth controller is the "MPC for passenger" that is developed to prioritize the minimization of waiting time of passengers. This controller is designed by setting all the weights of the objective function (22) equal to one (i.e., $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$) except the final weight assigned to the passenger waiting time (i.e., $\alpha_5 = 10$). This weighting scheme prioritizes passenger waiting time over the other factors of the objective function. In practice, this controller is of interest in case of adverse weather where it is critical to minimize passenger waiting time instead of minimizing network travel time or the passenger travel time inside taxis. The final controller represents a myopic dispatch strategy where taxis are dispatched at each time step according to number of vacant taxis and passengers in each region at that time step. The myopic strategy does not take into account the impact of the current dispatch decision on the future states of the network. The main goal of the "Myopic dispatch" controller is to balance the ratio of demand (unserved passengers) to supply (vacant taxis) at every time step in each region to increase the meeting rate of unserved passengers and vacant taxis.

5. Numerical experiments

5.1. Two-region city with one taxi firm case study

Consider a two-region city where the vehicle and taxi demands, i.e., $q_{ij}^{n}(t)$ and $q_{ij}^{c}(t)$, represent a time-varying directional morning rush hour towards the city center. The initial condition of number of taxis is: $c_1^o(0) = 0; c_2^o(0) = 0; c_1^v(0) = 1000; c_2^v(0) = 1000; C = 2000$. The initial car accumulation values are set to $n_1(0) = 4750$ and $n_2(0) = 3500$. Region 1 is assumed to be more congested than region 2 at time $t = 0$. Without loss of generality, it is also assumed that both regions have a similar MFD (consistent with the observed filed data in Geroliminis and Daganzo

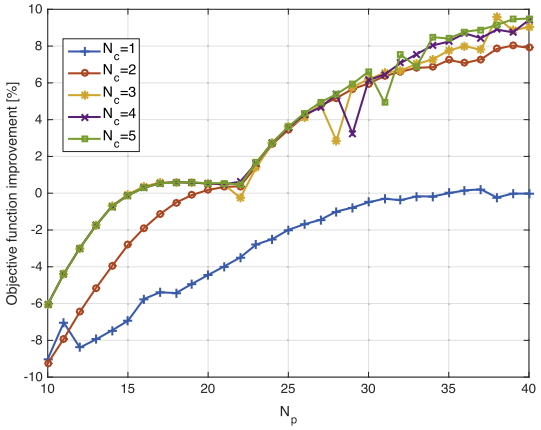

Fig. 4. The relative improvement of MPC performance against no dispatch policy for tuning of MPC approach parameters. We select $N_p = 40$ and $N_c = 4$ to conduct the numerical tests.

(2008)), boarding rate functions with similar $\gamma_1, \gamma_2, \gamma_3$, and average trip lengths l_i^n and l_i^o. The control time step is set to 60 [s], i.e. the dispatch rates are optimized every 60 [s] and remain fixed during the control time step interval.

The evolution of the personal vehicles, taxi passengers, occupied taxis, and vacant taxis is depicted in Fig. 5 where the left and right columns (of Fig. 5) demonstrate respectively the case of no taxi dispatching and the case of taxi dispatching with the MPC controller. For the no-dispatching case, it is evident that there exists a significant taxi passenger demand in region 1 peaking at time $t = 3200$, see Fig. 5(c), while the vacant taxis are predominantly searching in region 2, see Fig. 5(g). The inherent imbalance between the vacant taxis and taxis passengers necessitates the need of a taxi dispatch system. With taxi dispatching, vacant taxis are dispatched from region 2 to region 1 at time $t = 1680$ to serve the taxi passenger demand in region 1. Under this dispatch policy, Fig. 5(f) shows an increase in the number of occupied taxi trips in region 1 and Fig. 5(h) shows a decline in the the number of vacant taxis in region 2 at time $t = 1680$. Hence, the taxi passengers are effectively matched with the vacant taxis that would otherwise be searching in region 2 instead of region 1.

The effectiveness of taxis dispatching is also evident in extra personal vehicle accumulation in the no-dispatching case (Fig. 5(a)) compared to the dispatching case (Fig. 5(b)) with the latter demonstrating lower traffic congestion especially in region 1. The improved traffic conditions are a result of lowering vacant taxis flows with systematic dispatching. Furthermore, at the end of the rush hour, at time $t = 3600$, there are a large number of taxi passengers that have not yet found a taxi to board, see Fig. 5(c), and there are a large number of vacant taxis, see Fig. 5(g). With dispatching, however, Fig. 5(d) shows that there are no remaining taxi passengers at time $t = 3600$ and Fig. 5(h) shows that there are significantly less vacant taxis over the time with dispatching control method compared to the no dispatching case. Thus, the dispatch system while providing a more efficient taxi service also reduces the objective function of the model (i.e. the network total travel and waiting times) by 9% by pushing the network towards the undersaturated condition.

The rate of dispatching and the number of dispatched taxis in each region are illustrated in Fig. 6. The rate of dispatching taxis from region 2 to region 1 peaks at time 1680 to serve the accumulated taxi passenger demand in region 1, see Fig. 6(a). Consequently the number of taxis dispatched from region 2 to region 1 increases at time 1680 as well, see Fig. 6(b).

The MFDs of the no-dispatching and the dispatching cases are illustrated in Fig. 7. Under the no-dispatching case, Fig. 7(a) shows that the network is predominantly in the congested state because vacant taxis are not optimally distributed. Under the dispatching case, on the other hand, Fig. 7(b) depicts a higher tendency that the network is in the undersaturated condition.

The five controller strategies introduced in Section 4 are applied on the case study and the results are presented in Table 1 where the columns represent the average waiting time of passengers and the average travel time of taxis and personal vehicles. Table 1 shows that the MPC controller has the lowest total travel and passenger waiting time because of its property of predicting future states of congestion. The myopic dispatch strategy, on the other hand, has the highest total travel time and passenger waiting time because the decisions of this controller are made based on the instantaneous state of the network without any consideration of future states. The myopic controller performs even worse than the no-dispatch policy because the imprudent dispatching of taxis in the myopic controller leads to traffic congestion. This can be seen in Table 1 where the myopic dispatch strategy results in the highest dispatched taxi travel time, indicating an unnecessary dispatching because of the myopic nature of the controller. In addition, myopic dispatching has a severe congestion effect of 28% on normal traffic travel time as is evident in Table 1.

Table 1 also demonstrates that the MPC-without-velocity-modeling performs worse than the MPC because the bilateral meeting function is not accurately estimated when network velocity is disregarded. It is specifically crucial to consider the network average speed in taxi dispatching design when the traffic state is at a critical state where a few additional (taxis) vehicles shift the traffic state from the uncongested regime (the increasing part of MFD) to the congested regime (the decreasing part of MFD). Furthermore, the MPC-without-velocity has a high passenger waiting time and a high taxi dispatch time. This shows that the taxis are not optimally dispatched to the zones where they are needed.

Finally, Table 1 shows that the MPC-for-passengers controller substantially lowers passenger waiting time by 20% compared to the no-dispatch policy while MPC reduces passenger waiting time by 7%. This is expected as "MPC for passengers" control strategy prioritizes the minimization of passenger waiting time by increasing its corresponding weight in the objective function. This setting of weights has a minor effect of 0.4% increase on the normal traffic travel time compared to the no-dispatch policy. Accordingly, as the two-region network is more congested, the occupied taxi travel time increases by 8% since they have to spend more time in the urban network. However, the overall passenger waiting time and vehicles and taxis travel times is 0.3% lower than the no-dispatch strategy which is a network-level improvement. This result indicates that minimizing passenger waiting time is a practical dispatching strategy in cases of adverse weather conditions where it is critical that taxis provide an acceptable level-of-service to passengers.

5.2. Two-region city with two taxi firms case study

Consider the same two-region case study with two taxi firms: taxi firm 1 and taxi firm 2. Taxi firm 1 is equipped with the means to dispatch its taxis whereas taxi firm 2 does not benefit from dispatching. Each of the two firms start off with 500 taxis in each of the two regions, i.e. there are 2000 taxis in the network in total.

The results of the multi-firm case are presented in Fig. 8 where Fig. 8(a) and (b) show respectively the number of vacant taxis for taxi firm 1 and taxi firm 2. Taxi firm 1 has far fewer vacant taxis in both regions because it is able to dispatch its

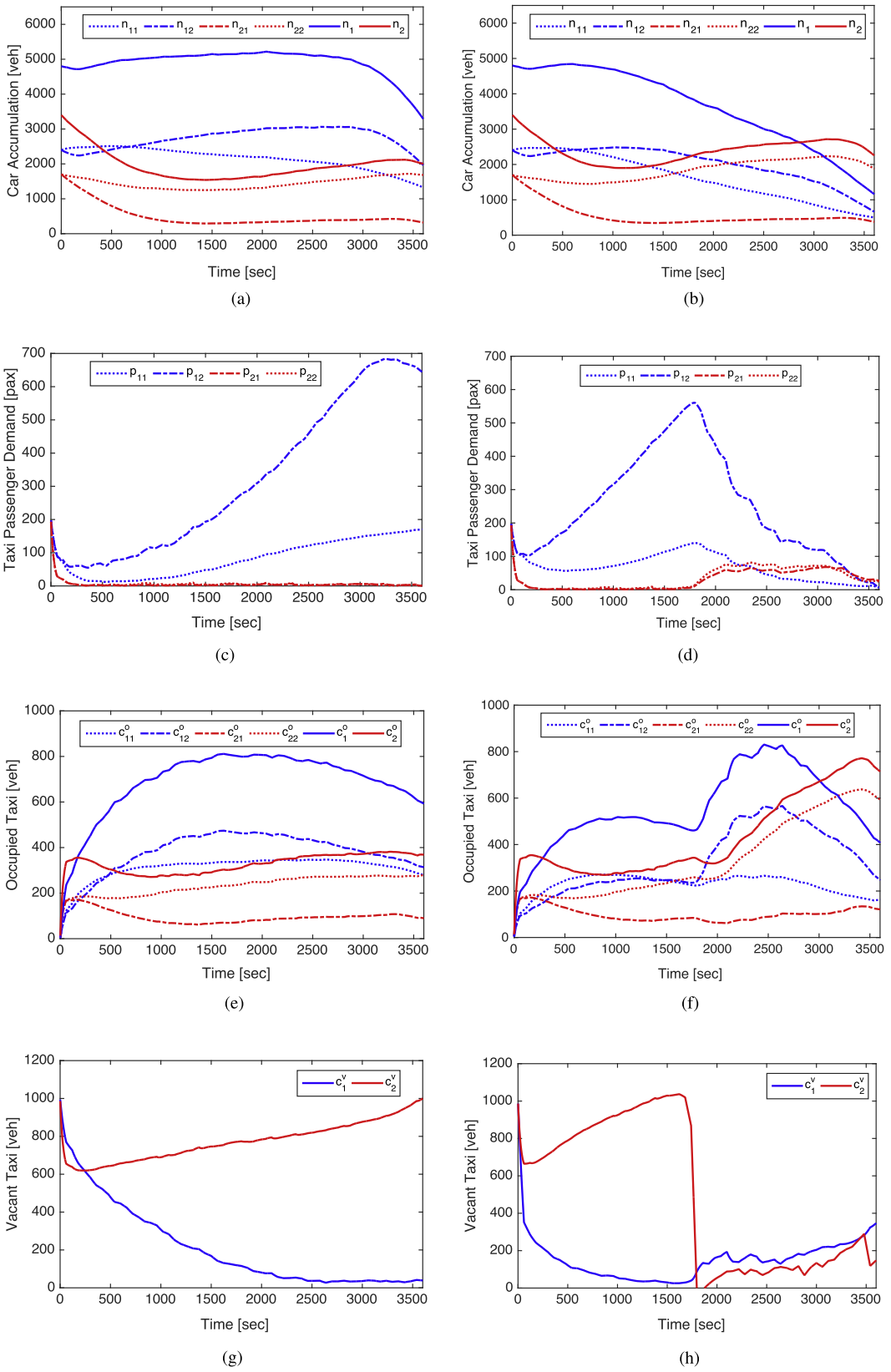

Fig. 5. Results of no dispatching case (left column); MPC taxi dispatching (right column).

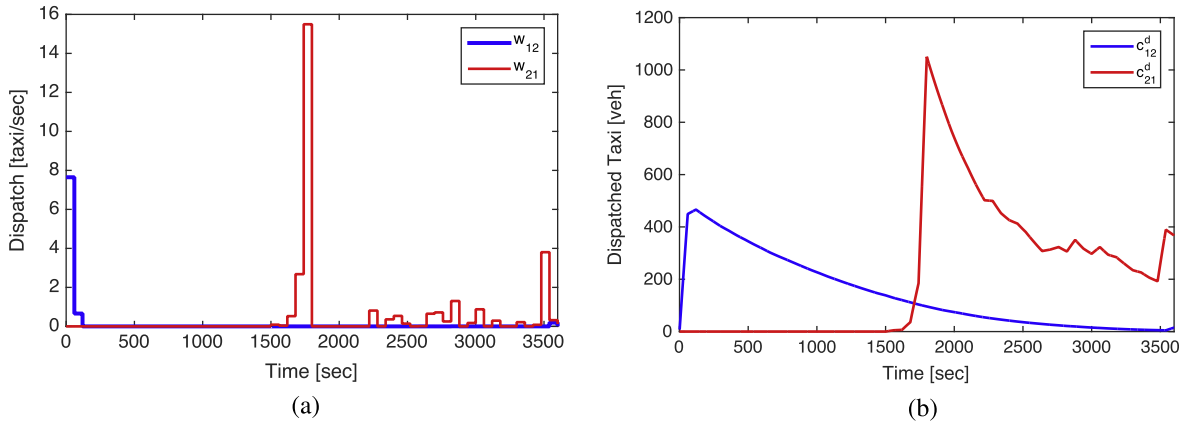

Fig. 6. (a) Rate of dispatching taxis. (b) Number of dispatched taxis.

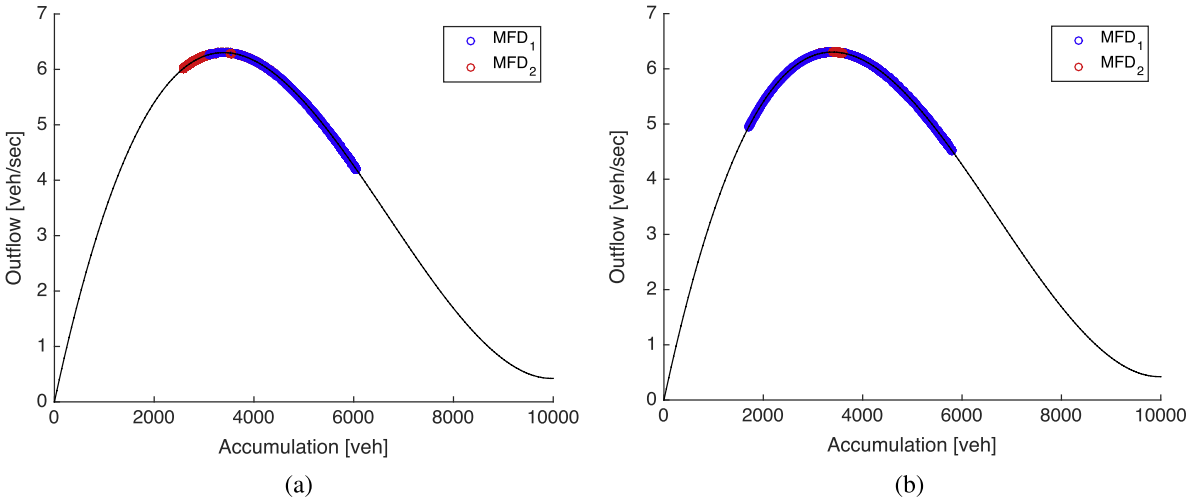

Fig. 7. (a) MFD without taxi dispatching. (b) MFD with taxi dispatching.

Table 1
Vehicle travel time and passenger waiting time ($\times 10^6$). The numbers in brackets represent the percentage change with respect to the no-dispatch policy.

Control strategy	Normal traffic travel time	Occupied taxi travel time	Vacant taxi travel time	Dispatched taxi travel time	Passenger waiting time	Total travel and waiting time
No dispatch	24.6	3.7	3.6	0	1.5	33.4
MPC	21.8 (−11%)	3.6 (−3%)	2.2 (−39%)	1.5	1.4 (−7%)	30.5 (−9%)
MPC without velocity modeling	21.9 (−11%)	2.9 (−21%)	1.8 (−50%)	2.6	2.7 (80%)	31.9 (−4%)
MPC for passengers	24.7 (0.4%)	4.0 (8%)	3.0 (−17%)	0.3	1.2 (−20%)	33.3 (−0.3%)
Myopic dispatch	31.5 (28%)	3.4 (−8%)	1.1 (−69%)	2.8	3.1 (107%)	41.9 (25%)

vacant taxis to the region with more unserved passengers. The Number of waiting passengers is presented in Fig. 8(c) where it is shown that the majority of passengers are accumulated in region 1. To serve the passengers in region 1, taxi firm 1 starts to dispatch its taxis from region 2 to region 1 at time $t = 880$ as shown in Fig. 8(d). The vacant taxis of taxi firm 2, on the other hand, are predominantly located in region 2 (see Fig. 8(b)) where there are very few unserved passengers. Hence, taxi firm 1, due to its taxi dispatch system, is able to make more efficient use of its fleet compared to taxi firm 2. A future research direction is to further investigate the effect of taxi market control (e.g. dispatching) in multi-firm situations where a dynamic game governs the system.

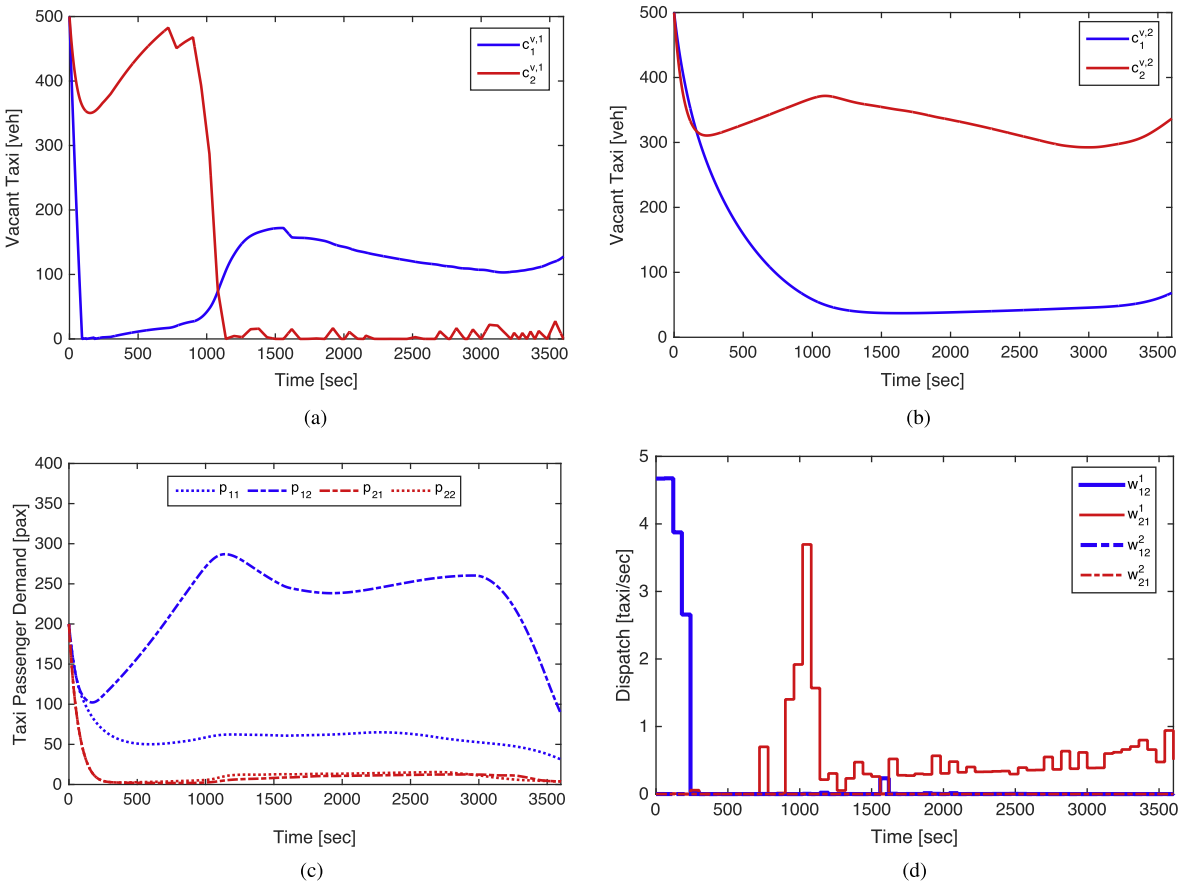

Fig. 8. Results of two-region two-taxi firm case study. Taxi firm 1 is deployed with predictive dispatching contrary to taxi firm 2. (a) Vacant taxis of firm 1; (b) Vacant taxis of firm 2; (c) Number of unserved passengers; (d) Taxi dispatch rates.

6. Summary and future research

This paper has introduced a parsimonious, dynamic, and network-scale taxi dispatch scheme in cities that considers the effect of traffic congestion on the overall performance of the system. The traffic modeling is based on the macroscopic network fundamental diagram (MFD), which incorporates the joint dynamics of personal vehicles and taxis. The vacant taxi dynamics are modeled as an aggregated two-agent search process with search friction that is a function of the density of the two agents, i.e. unserved passengers and vacant taxis, and the prevailing urban network condition, i.e. network average speed provided by MFD. The urban network is assumed to be partitioned to several homogeneous regions with well-defined MFDs while the taxi dispatch rate between the regions are optimized to achieve a network wide objective function. The proposed traffic flow modeling considers multiple regions and taxi firms. A model predictive control (MPC) approach is developed to find the optimal taxi dispatching rates. Numerical experiments reveal the properties of the MPC taxi dispatch scheme and demonstrate the proposed taxi dispatching system offers significant improvements on the performance of the urban system in terms of total delay of personal vehicles, waiting time of passengers, and taxi fleet utilization.

A future research direction is to integrate the taxi dispatching system with the perimeter traffic flow control strategy (Geroliminis et al., 2013; Ramezani et al., 2015), which replicates a quasi-system optimum traffic management strategy in case the city traffic operator is in charge of the taxi system. This is challenging because the mobility market dynamics and design, e.g. fare policy and entry regulations, and the interaction between multiple taxi agencies need further investigations. Scrutinizing the proposed taxi dispatch modeling and control in detail through microscopic traffic simulation can also shed light on the applicability and implications of the method. Another challenging research direction is to study the effect of recent technological advances in taxi e-hailing, booking, and ride-sharing on the modeling and dynamics of the taxi system. Furthermore, analysis of the effect of taxi dispatching on the heterogeneity of traffic congestion is an important issue to be addressed in future works.

Appendix A. Multi-firm taxi dispatching

Let us assume there are F taxi firms operating in the R-region urban network. The meeting (boarding) function of taxi-passenger matching process considering multiple taxi firms and different types of passengers according to their destination and assuming a homogeneous region reads:

$$b_{ij}^{f}(t) = \frac{c_i^{\mathrm{v},f}(t)}{c_i^{\mathrm{v}}(t)} \frac{p_{ij}(t)}{p_i(t)} B(c_i^{\mathrm{v}}(t), p_i(t), v_i(t)) \quad \forall i \in R, j \in U_i, f \in F \tag{A.1}$$

where $b_{ij}^{f}(t)$ is the meeting rate of taxis of firm f with unserved passengers in region i with destination j at time t and $c_i^{\mathrm{v},f}(t)$ denotes the number of vacant taxis of taxi firm f in region i at time t. The overall boarding function is defined as $b_{ij}(t) = \sum_{f\in F} b_{ij}^{f}(t) = \frac{p_{ij}(t)}{p_i(t)} B(c_i^{\mathrm{v}}(t), p_i(t), v_i(t))$.

The conservation of taxis is ensured through the following:

$$\sum_{f\in F} c_i^{\mathrm{v},f}(t) = c_i^{\mathrm{v}}(t) \quad \forall i \in R \tag{A.2}$$

$$\sum_{f\in F} c_{ij}^{\mathrm{o},f}(t) = c_{ij}^{\mathrm{o}}(t) \quad \forall i \in R, j \in U_i \tag{A.3}$$

$$\sum_{f\in F} c_{ij}^{\mathrm{d},f}(t) = c_{ij}^{\mathrm{d}}(t) \quad \forall i \in R, j \in U_i \tag{A.4}$$

where $c_{ij}^{\mathrm{o},f}(t)$ and $c_{ij}^{\mathrm{d},f}(t)$ are respectively, the number of occupied and dispatched taxis of taxi firm f in region i with destination j at time t.

The internal trip completion rate of occupied taxis of taxi firm f in region i at time t is:

$$M_{ii}^{\mathrm{o},f}(t) = \frac{c_{ii}^{\mathrm{o},f}(t)}{n_i(t) + c_i(t)} \frac{P_i(n_i(t) + c_i(t))}{l_i^{\mathrm{o}}(t)} \quad \forall i \in R, f \in F \tag{A.5}$$

Accordingly, the inter-regional transfer flows are:

$$M_{ij}^{\mathrm{o},f}(t) = \frac{c_{ij}^{\mathrm{o},f}(t)}{n_i(t) + c_i(t)} \frac{P_i(n_i(t) + c_i(t))}{l_i^{\mathrm{o}}(t)} \quad \forall i \in R, j \in U_i, f \in F \tag{A.6}$$

$$M_{ij}^{\mathrm{d},f}(t) = \frac{c_{ij}^{\mathrm{d},f}(t)}{n_i(t) + c_i(t)} \frac{P_i(n_i(t) + c_i(t))}{l_i^{\mathrm{d}}(t)} \quad \forall i \in R, j \in U_i, f \in F \tag{A.7}$$

where $M_{ij}^{\mathrm{o},f}(t)$ and $M_{ij}^{\mathrm{d},f}(t)$ denote the transfer flow of occupied and dispatched taxis of taxi frim f from region i to region j at time t respectively.

Using the above equations, the mass conservation equations of occupied taxis of taxi firm f are

$$\frac{\mathrm{d}c_{ii}^{\mathrm{o},f}(t)}{\mathrm{d}t} = b_{ii}^{f}(t) + \sum_{j\in U_i} M_{ji}^{\mathrm{o},f}(t) - M_{ii}^{\mathrm{o},f}(t) \quad \forall i \in R, f \in F \tag{A.8}$$

$$\frac{\mathrm{d}c_{ij}^{\mathrm{o},f}(t)}{\mathrm{d}t} = b_{ij}^{f}(t) - \sum_{j\in U_i} M_{ij}^{\mathrm{o},f}(t) \quad \forall i \in R, j \in U_i, f \in F \tag{A.9}$$

Accordingly, the evolution of the number of dispatched taxis is

$$\frac{\mathrm{d}c_{ij}^{\mathrm{d},f}(t)}{\mathrm{d}t} = w_{ij}^{f}(t) - \sum_{j\in U_i} M_{ij}^{\mathrm{d},f}(t) \quad \forall i \in R, j \in U_i, f \in F \tag{A.10}$$

where $w_{ij}^{f}(t)$ represents the dispatch rate implemented by taxi-firm f for its vacant taxis in region i to move to region j at time t. Ultimately, the following equation tracks the number of vacant taxis of taxi firm f:

$$\frac{\mathrm{d}c_i^{\mathrm{v},f}(t)}{\mathrm{d}t} = M_{ii}^{\mathrm{o},f}(t) + \sum_{j\in U_i} M_{ji}^{\mathrm{d},f}(t) - \sum_{j\in (U_i \cup i)} b_{ij}^{f}(t) - \sum_{j\in U_i} w_{ij}^{f}(t) \quad \forall i \in R, f \in F \tag{A.11}$$

Note that Eqs. (11), (13), and (16)–(19) remain unchanged and valid.

References

Agatz, N., Erera, A., Savelsbergh, M., Wang, X., 2012. Optimization for dynamic ride-sharing: a review. Eur. J. Oper. Res. 223 (2), 295–303.

Ambühl, L., Menendez, M., 2016. Data fusion algorithm for macroscopic fundamental diagram estimation. Transp. Res. Part C: Emerg. Technol. 71, 184–197.
Amirgholy, M., Gonzales, E.J., 2016. Demand responsive transit systems with time-dependent demand: user equilibrium, system optimum, and management strategy. Transp. Res. Part B: Methodol. 92, 234–252.
Beesley, M.E., 1973. Regulation of taxis. Econ. J. 83 (329), 150–172.
Buisson, C., Ladier, C., 2009. Exploring the impact of homogeneity of traffic measurements on the existence of macroscopic fundamental diagrams. Transp. Res. Rec.: J. Transp. Res. Board (2124), 127–136.
City of Toronto, 2016. City of Toronto taxi fact sheet.
Daganzo, C.F., Gayah, V.V., Gonzales, E.J., 2011. Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability. Transp. Res. Part B: Methodol. 45 (1), 278–288.
Daganzo, C.F., Geroliminis, N., 2008. An analytical approximation for the macroscopic fundamental diagram of urban traffic. Transp. Res. Part B: Methodol. 42 (9), 771–781.
Gayah, V., Dixit, V., 2013. Using mobile probe data and the macroscopic fundamental diagram to estimate network densities: tests using microsimulation. Transp. Res. Rec.: J. Transp. Res. Board (2390), 76–86.
Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transp. Res. Part B: Methodol. 42 (9), 759–770.
Geroliminis, N., Haddad, J., Ramezani, M., 2013. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: a model predictive approach. IEEE Trans. Intell. Transp. Syst. 14 (1), 348–359.
Geroliminis, N., Sun, J., 2011a. Hysteresis phenomena of a macroscopic fundamental diagram in freeway networks. Transp. Res. Part A: Policy Pract. 45 (9), 966–979.
Geroliminis, N., Sun, J., 2011b. Properties of a well-defined macroscopic fundamental diagram for urban traffic. Transp. Res. Part B: Methodol. 45 (3), 605–617.
Geroliminis, N., Zheng, N., Ampountolas, K., 2014. A three-dimensional macroscopic fundamental diagram for mixed bi-modal urban networks. Transp. Res. Part C: Emerg. Technol. 42, 168–181.
Ghiani, G., Guerriero, F., Laporte, G., Musmanno, R., 2003. Real-time vehicle routing: Solution concepts, algorithms and parallel computing strategies. Eur. J. Oper. Res. 151 (1), 1–11.
Haddad, J., 2017. Optimal perimeter control synthesis for two urban regions with aggregate boundary queue dynamics. Transp. Res. Part B: Methodol. 96, 1–25.
Haddad, J., Ramezani, M., Geroliminis, N., 2013. Cooperative traffic control of a mixed network with two urban regions and a freeway. Transp. Res. Part B: Methodol. 54, 17–36.
He, F., Shen, Z.-J.M., 2015. Modeling taxi services with smartphone-based e-hailing applications. Transp. Res. Part C: Emerg. Technol. 58, 93–106.
Hosni, H., Naoum-Sawaya, J., Artail, H., 2014. The shared-taxi problem: formulation and solution methods. Transp. Res. Part B: Methodol. 70, 303–318.
Huo, H., Zhang, Q., He, K., Yao, Z., Wang, M., 2012. Vehicle-use intensity in china: current status and future trend. Energy Policy 43, 6–16.
Ji, Y., Geroliminis, N., 2012. On the spatial partitioning of urban transportation networks. Transp. Res. Part B: Methodol. 46 (10), 1639–1656.
Jung, J., Chow, J.Y., Jayakrishnan, R., Park, J.Y., 2014. Stochastic dynamic itinerary interception refueling location problem with queue delay for electric taxi charging stations. Transp. Res. Part C: Emerg. Technol. 40, 123–142.
Jung, J., Jayakrishnan, R., Park, J.Y., 2016. Dynamic shared-taxi dispatch algorithm with hybrid-simulated annealing. Comput.-Aided Civil Infrastruct. Eng. 31 (4), 275–291.
Keyvan-Ekbatani, M., Kouvelas, A., Papamichail, I., Papageorgiou, M., 2012. Exploiting the fundamental diagram of urban networks for feedback-based gating. Transp. Res. Part B: Methodol. 46 (10), 1393–1403.
Keyvan-Ekbatani, M., Yildirimoglu, M., Geroliminis, N., Papageorgiou, M., 2015. Multiple concentric gating traffic control in large-scale urban networks. IEEE Trans. Intell. Transp. Syst. 16 (4), 2141–2154.
Knoop, V., Hoogendoorn, S., Van Lint, J., 2012. Routing strategies based on macroscopic fundamental diagram. Transp. Res. Rec.: J. Transp. Res. Board (2315), 1–10.
Lagos, R., 2000. An alternative approach to search frictions. J. Polit. Econ. 108 (5), 851–873.
Laval, J.A., Castrillón, F., 2015. Stochastic approximations for the macroscopic fundamental diagram of urban networks. Trans. Res. Part B: Methodol. 81, 904–916.
Leclercq, L., Chiabaut, N., Trinquier, B., 2014. Macroscopic fundamental diagrams: a cross-comparison of estimation methods. Transp. Res. Part B: Methodol. 62, 1–12.
Leclercq, L., Geroliminis, N., 2013. Estimating mfds in simple networks with route choice. Transp. Res. Part B: Methodol. 57, 468–484.
Lee, D.-H., Wang, H., Cheu, R., Teo, S., 2004. Taxi dispatch system based on current demands and real-time traffic conditions. Transp. Res. Rec.: J. Transp. Res. Board (1882), 193–200.
Mahmassani, H.S., Saberi, M., Zockaie, A., 2013. Urban network gridlock: theory, characteristics, and dynamics. Transp. Res. Part C: Emerg. Technol. 36, 480–497.
Mortensen, D.T., Pissarides, C.A., 1994. Job creation and job destruction in the theory of unemployment. Rev. Econ. Stud. 61 (3), 397–415.
Nie, Y.M., 2017. How can the taxi industry survive the tide of ridesourcing? evidence from shenzhen, china. Transportation Research Part C: Emerging Technologies 79, 242–256.
Nourinejad, M., Roorda, M.J., 2016. Agent based model for dynamic ridesharing. Transp. Res. Part C: Emerg. Technol. 64, 117–132.
Petrongolo, B., Pissarides, C.A., 2001. Looking into the black box: a survey of the matching function. J. Econ. Literat. 39 (2), 390–431.
Pillac, V., Gendreau, M., Guéret, C., Medaglia, A.L., 2013. A review of dynamic vehicle routing problems. Eur. J. Oper. Res. 225 (1), 1–11.
Ramezani, M., Haddad, J., Geroliminis, N., 2015. Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control. Transp. Res. Part B: Methodol. 74, 1–19.
Saberi, M., Mahmassani, H.S., Zockaie, A., 2014. Network capacity, traffic instability, and adaptive driving: findings from simulated urban network experiments. EURO J. Transp. Logist. 3 (3–4), 289–308.
Sayarshad, H.R., Chow, J.Y., 2016. Survey and empirical evaluation of nonhomogeneous arrival process models with taxi data. J. Adv. Transp. 50 (7), 1275–1294.
Schroeter, J.R., 1983. A model of taxi service under fare structure and fleet size regulation. Bell J. Econ., 81–96
Seow, K.T., Dang, N.H., Lee, D.-H., 2010. A collaborative multiagent taxi-dispatch system. IEEE Trans. Automat. Sci. Eng. 7 (3), 607–616.
Taxi and Limousine Commission, 2014. 2014 Taxicab FactBook.
Tsekeris, T., Geroliminis, N., 2013. City size, network structure and traffic congestion. J. Urban Econ. 76, 1–14.
Wang, X., He, F., Yang, H., Gao, H.O., 2016. Pricing strategies for a taxi-hailing platform. Transp. Res. Part E: Logist. Transp. Rev. 93, 212–231.
Wong, K., Bell, M.G., 2006. The optimal dispatching of taxis under congestion: a rolling horizon approach. J. Adv. Transp. 40 (2), 203–220.
Wong, K., Wong, S., Yang, H., 2001. Modeling urban taxi services in congested road networks with elastic demand. Transp. Res. Part B: Methodol. 35 (9), 819–842.
Wong, K., Wong, S., Yang, H., Wu, J., 2008. Modeling urban taxi services with multiple user classes and vehicle modes. Transp. Res. Part B: Methodol. 42 (10), 985–1007.
Wong, R., Szeto, W., Wong, S., 2014. A cell-based logit-opportunity taxi customer-search model. Transp. Res. Part C: Emerg. Technol. 48, 84–96.
Yang, H., Leung, C.W., Wong, S., Bell, M.G., 2010. Equilibria of bilateral taxi–customer searching and meeting on networks. Transp. Res. Part B: Methodol. 44 (8), 1067–1083.
Yang, H., Wong, S., 1998. A network model of urban taxi services. Transp. Res. Part B: Methodol. 32 (4), 235–246.

Yang, H., Wong, S.C., Wong, K., 2002. Demand–supply equilibrium of taxi services in a network under competition and regulation. Transp. Res. Part B: Methodol. 36 (9), 799–819.
Yang, H., Yang, T., 2011. Equilibrium properties of taxi markets with search frictions. Transp. Res. Part B: Methodol. 45 (4), 696–713.
Yang, T., Yang, H., Wong, S.C., 2014. Taxi services with search frictions and congestion externalities. J. Adv. Transp. 48 (6), 575–587.
Yildirimoglu, M., Ramezani, M., Geroliminis, N., 2015. Equilibrium analysis and route guidance in large-scale networks with mfd dynamics. Transp. Res. Part C: Emerg. Technol. 59, 404–420.
Zha, L., Yin, Y., Yang, H., 2016. Economic analysis of ride-sourcing markets. Transp. Res. Part C: Emerg. Technol. 71, 249–266.
Zhan, X., Qian, X., Ukkusuri, S.V., 2015. A graph-based approach to measuring the efficiency of an urban taxi service system.
Zhang, W., Ukkusuri, S.V., 2016. Optimal fleet size and fare setting in emerging taxi markets with stochastic demand. In: Computer-Aided Civil and Infrastructure Engineering.
Zheng, N., Geroliminis, N., 2013. On the distribution of urban road space for multimodal congested networks. Transp. Res. Part B: Methodol. 57, 326–341.
Zheng, N., Geroliminis, N., 2016. Modeling and optimization of multimodal urban networks with limited parking and dynamic pricing. Transp. Res. Part B: Methodol. 83, 36–58.