20th International Symposium on Transportation and Traffic Theory

On the distribution of urban road space for multimodal congested networks

Nan Zheng[a] and Nikolas Geroliminis[a*]

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

Abstract

Transport systems in real cities are complex with many modes of transport sharing and competing for limited road space. This work intends to understand how space distributions for modes and interactions among modes affect network traffic performance. While the connection between performance of transport systems and general land allocation is the subject of extensive research, space allocation for interacting modes of transport is an open research question. Quantifying the impact of road space distribution on the performance of a congested multimodal transport system with a dynamic aggregated model remains a challenge. In this paper, a multimodal macroscopic fundamental diagram (MFD) is developed to represent the traffic dynamics of a multimodal transport system. Optimization is performed with the objective of minimizing the total passenger hours travelled (PHT) to serve the total demand by redistributing road space among modes. Pricing strategies are also investigated to provide a higher demand shift to more efficient modes. We find by an application to a bi-modal two-region city that (i) the proposed model captures the operational characteristics of each mode, and (ii) optimal dynamic space distribution strategies can be developed. In practice, the approach can serve as a physical dynamic model to inform space distribution strategies for policy makers with different goals of mobility.

1 INTRODUCTION

As cities around the world grow rapidly and more people through different modes compete for limited urban road infrastructure to travel, it is important to understand how this space can be managed to improve accessibility for travelers. The ultimate goal of research towards this direction is to develop modeling and optimization tools, which will contribute on how to redistribute city space to multiple transportation modes and to understand what level of mobility cities of different structures and topologies can achieve.

Management strategies can be implemented to partition a city so that road space is deliberately allocated between competing modes. Although the allocation of this space is a policy-oriented decision, it should be informed by the correct physics and dynamics of the multimodal flows. This system can be treated as an interconnected network of regions (sub-networks) with one or more modes moving. In this extension, different

* Corresponding author. Tel.: +41 21 69 3248 1; fax: +41 21 69 32479.
E-mail address: nikolas.geroliminis@epfl.ch

parts of a city can be subject to different management strategies (see for example Fig. 1). Perhaps bus-only streets are allocated only in the central region while other parts of the city allow vehicles to operate in mixed traffic. Changes in infrastructure, demand, or operations in one region have impact on the behavior of adjoining regions. While recent findings in the macroscopic modeling and dynamics of traffic in cities have provided knowledge of single-mode/ single-region cities and single-mode/multi-region cities, the understanding of multi-mode, multi-region cities is limited.
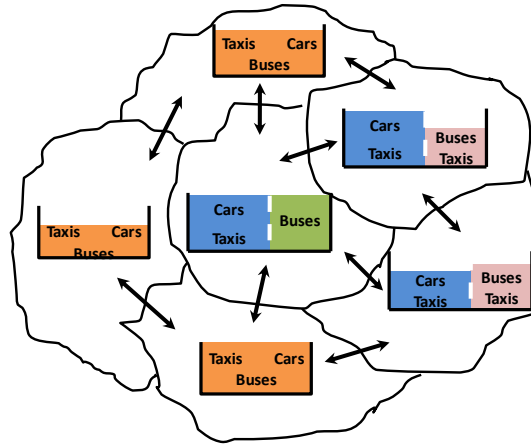

Fig. 1. A multi-region, multimodal system

Under a multimodal environment, space should be allocated taking into account spatiotemporal differences in the demand, the topology, and the control characteristics. These spatiotemporal decisions are important because if they are made incorrectly, space could be wasted. If this wasted space could be productively used by low-occupancy vehicles without affecting the more productive modes, mobility is being restricted. For example, recent studies in Californian freeways, have questioned the effectiveness of high-occupancy vehicle (HOV) lanes and have shown that HOV lanes are underutilized and the passenger capacity of freeways has decreased, resulting in heavier congestion levels [1].

Operational characteristics should be considered as well. Despite the different features of modes in terms of occupancy (number of passengers), driving behavior (speeds, acceleration and deceleration profiles, length), duration of travel, scheduled vs. non-scheduled service, a common characteristic is the following: all of these vehicles when moving to an urban environment make stops related to traffic congestion (e.g. red phases at traffic signals) and *other* stops, which also cause delays to the transportation system as a whole, e.g. buses stop at bus stops to board/alight passengers, taxis stop frequently and randomly when they search/pick up/deliver passengers, cars may stop/maneuver when search/find a parking spot. While there is a good understanding and vast literature of the dynamics and the modeling of congestion for congestion-related stops, the effect of *service or general purpose* stops in the overall performance of a transportation system still remains a challenge. It is intuitive that the effect of these stops during light demand conditions in the network capacity is almost negligible, but nowadays city centers are experiencing high level of congestion and the frequency in time and space of the service stops is significantly high.

In this paper, we present a macroscopic approach for optimizing road space allocation for multimodal transport systems with focus on two modes of transport, cars and buses and two regions, the city center and the periphery. Extensions to more complex city structures are also discussed. This approach should not only deal with the problems mentioned above, but also will switch the interest from the currently inefficient vehicle throughput based optimization to the more efficient for networks and society, passenger throughput optimization. Congestion pricing strategies for cars are also considered to further facilitate the demand shift when preferential treatment to buses is offered.

2 BACKGROUND

Existing literature on the physics of urban congestion can be divided generally into city-scale, street-scale, and car-to-car scale. City-scale investigations have thus far looked only at the behaviour of *one* mode and the involved dynamics of traffic congestion. Studies of multiple modes, on the other hand, have only been made at the street-level scale for simplistic time-independent scenarios. Planning studies have looked at public transport on a city scale, particularly buses on idealized road networks. Making road space allocation decisions, however, requires consideration of interactions and dynamics between multiple modes. To date, such considerations have been made only at the much finer street scale and still in a time-independent and unrealistic environment. Thus, the existing body of work leaves a gap to be filled—a physically realistic time-dependent, city-scale model including multiple modes that interact with each other.

There is a strong understanding and vast literature of congestion dynamics and spreading in street- or car-to-car-scale traffic systems with a single mode of traffic, e.g. a highway section with cars. Besides traffic scientists, mathematicians and physicists have also contributed to the field of traffic flow. Because of the numerous publications, we refer the reader to reference [2] for an overview. Briefly speaking, the main modeling approaches can be classified as follows: Car-following models deal with the non-linear interactions and dynamics of single vehicles [3]. To address computational burden, cellular automata describe the dynamics of vehicles in a coarse-grained way by discretizing space and time, (e.g. [4][5]). Gas-kinetic models formulate a partial differential equation for the spatio-temporal evolution of the vehicle density and the velocity distribution (e.g. [6]). First-order flow models are based on a partial differential equation for the density and a fundamental diagram relation (e.g. [7][8]). Second-order models contain an additional equation for non-steady state conditions. In continuum models, a network is approximated as a continuum in which users choose their routes in a two-dimensional space (e.g. [9]).

On the public transport side at city-scale, modelers have analyzed how systems should be designed, e.g. Wirasinghe et al. [10] considered how to systematically design a bus transit system for an idealized city with centralized demand. However, these models have been only applied to one mode, and only in the steady-state operation. Researches have also been done to look at how multiple modes can share the road, but only on the street-scale level. Sparks and May [11] developed a mathematical model to evaluate priority lanes for high occupancy vehicles on freeways. Later, Dahlgren [12] and Daganzo and Cassidy [13] studied how different modes use freeways, recognizing that if different modes serve different numbers of passengers, then analyses should not consider all vehicles with the same weight. But these works are limited to small scale systems. They looked at the effect on total passenger travel time if a lane on a specific road section were dedicated to multiple occupant vehicles. This consideration of different occupancies between vehicles is important because it recognizes that some modes carry more passengers than others. The importance of considering passengers rather than vehicles was earlier mentioned by Vuchic [14]. He criticized street-scale evaluations based only on vehicle flows, because multimodal systems should integrate the occupancies of each mode. A qualitative analysis of space allocation can be found in Gonzales et al [15]. The quantitative treatment of the transit process (e.g. network route design, scheduling) is reflected in a considerable amount of effort in numerous publications (e.g. [16] and others), and will not be addressed here.

Researchers have looked at allocating street space between more than one mode, through the dedication of a freeway lane to HOVs or a lane for buses on a city street ([17][18]). These methods have limited applicability, however, because they assume steady-state traffic flow and they also ignore the fluctuations and spill-over effects that typically characterize urban traffic congestion. Currie et al. [19] carried out impact analysis in planning studies of road space allocation. That analysis was based on a disaggregate micro-simulation which relies on intensive travel data inputs that are typically unreliable or unavailable. Researches in references [20], [21][22][23], focused on the connection between performance of transport systems and land use for transportation as a whole, without further consideration on allocation by usage.

More recently, Gonzales and Daganzo [24] examined system optimum solutions for a transport system with cars and public transit share spaces for the morning commute problem. The authors solve user and system optimum for the bottleneck model and system optimum for the network model (with an MFD representation). The proposed model recognizes that the network capacity for cars is reduced if transit operations receive dedicated space. However, quantifying the impact of this dedicated space on the performance of a transport system still remains an exclusive and challenging question. Mesbah et al. [25] proposed an optimization framework for optimizing transit priority and allocating dedicated bus lanes. The optimization aims at minimizing the total travel time of both car travelers and public users. However, travel time is estimated by BPR function which is not valid for they dynamic case. In both references, the impact of the operation of buses (e.g. dwell time) on the delay of travelers was not discussed.

This gap is due to the lack of a traffic model, which will represent the flow dynamics of a multimodal system as a result of road space allocation. While various theories have been proposed to macroscopically model urban networks ([26][27][28][29]), only until recently Geroliminis and Daganzo [30] demonstrated the existence of a fundamental model (the macroscopic fundamental diagram - MFD) on congestion dynamics of single-mode system with empirical data. This reference showed that (i) an MFD is a plot between network space-mean flow and density, (ii) the MFD is a property of the network itself (infrastructure and control) and not very sensitive to demand, i.e. the MFD should have a well-defined maximum and remain invariant when the demand changes and (iii) the space-mean flow is maximum for the same value of critical density of vehicles, for many origin-destination tables. Daganzo and Geroliminis [30] and Geroliminis and Boyaci [31] have derived analytical theories for the shape of the MFD as a function of network and intersection parameters, using variational theory (VT) for homogeneous and heterogeneous network topologies, respectively. Properties of well-defined MFDs, stability and scatter analysis and other simulation and experimental tests can be found in [32][33][34][35][36][37][38] and others. Recently, Gonzales et al. [39] observed through simulation an MFD for multimodal systems of cars and buses in the city center of Nairobi, Kenya.

Boyaci and Geroliminis [40] extended the VT theory for the multi-modal case and provided a semi-analytical approximation of the MFD, by considering that bus service related stops interact with traffic as hypothetical traffic signals with periodic characteristics. The operational characteristics of this hypothetical signal-bus stop depend on the dwell times and the frequencies of buses, while the capacity during service stops is *smaller* as buses might block traffic. In VT theory a number of different observers (forward and backward) move in the system of traffic signals and for each observer the average speed (because of delays in red phases and bus stops) and the maximum passing rate are estimated. These observers create associated "cuts" in a network flow-density diagram and the lower envelope of these cuts provides a good approximation of the MFD for homogeneously congested areas. While VT has been proved to provide a tight cut (see [30]) in case of corridors with uniform density (i.e. incoming turns are about equal to outcoming ones). Nevertheless, application of variational theory in simulated and real networks (with turns, heterogeneous topology, route choice etc) show that this approach is a good approximation if congestion is evenly distributed.

Building in the knowledge of the single-mode macroscopic modeling, developing the dynamics of multimodal systems is promising. To summarize, a macroscopic approach should be developed to estimate the dynamics and to optimize space allocation of multimodal urban road systems. We will show that (i) the proposed approach captures the operational characteristics of different modes, (ii) the resulting system performances are consistent with the physics of traffic given different road space strategies, e.g. with or without dedicated bus lanes, (iii) allocation of road space can be readily optimized in a static and dynamic way, and (iv) pricing strategies can further improve the efficiency of the system with less space dedicated for buses. Next section develops the methodological framework with respect to the multimodal dynamics and the optimization approach. Section 4 presents the results of a case study with a two-region city and two modes of transport, while further research is discussed in Section 5. Table 1 provides a nomenclature for the different variables and parameters utilized in the paper.

Table 1: List of main variables and parameters.

Variables	Description
ob_i^c	average number of passengers per car in region i
$Q_i^k(t)$	demand generated at time t in region i with next destination region k
$Q_i^{km}(t)$	demand generated at time t in region i with next destination region k, choosing mode m
$p_i^{kb}(t)$	Bus mode share for demand $Q_i^k(t)$
$O_i^m(n_i^b(t), n_i^c(t))$	flow exiting region i of mode m at time t as a function of accumulations
$r_i^{km}(t)$	fraction of O_i^m from region i to j with final destination k at time t
$P_i^m(t)$	total distance traveled by all vehicles of mode m in region i at time t
$O_{i \to j}^{km}(t)$	transfer flow of mode m from region i to j with final destination k
$\pi_i(t)$	space share allocation of region i at time t
$U_i^{km}(t)$	utility of using mode m at time t in region i with next destination region k
$n_i^{km}(t)$	accumulation of mode m in region i with next destination region k at time t
$n_i^m(t)$	accumulation of mode m currently in region i at time t
$ob_i^{kb}(t)$	average number of on-board passengers per bus in region i with destination k at time t
$\delta_{i \to j}^k(t)$	binary variable showing the sequence of the trip, with value equal to 1 if a trip from region i to k passes through j immediately after leaving i, and 0 otherwise
$OB_i^{kb}(t)$	total number of bus on-board passengers in region i with destination k at time t
$\alpha_i^b(t)$	parameter for the effect of dwell times on bus speed in region i at time t
$TT_i^b(t), V_i^b(t)$	average travel time and speed by using mode m in region i at time t
$TT_d^b(t)$	extra travel time spent due to bus dwell times in region i at time t
$\bar{L}'_{ib}$	average trip length of bus passengers in region i
$\bar{L}_{im}$	average trip length of mode m in region i
C_{bus}	storage capacity of buses (persons per vehicle)
$D_j^b(t)$	bus discomfort (crowdedness) in region i at time t
θ_i	percentage of on-board passengers reaching destination i
γ, β_1, β_2	parameters for mode choice and bus crowdness

3   METHODOLOGICAL FRAMEWORK

Consider a city divided into *N* regions as in Fig. 1, denoted by *i* =1,…, *N*. Criteria for partitioning a region (approximate size a few hundred links each) are: homogeneous distribution of congestion within each region to obtain a low scatter MFD (see for example [41]), similar topological characteristics and similar type of mode usage. Any region *i* is partitioned into sub-regions, each one containing a specific type of mode usage, e.g. it can be dedicated bus lanes, mixed traffic lanes, car-only lanes or any other special usage lanes. Demand and their aggregate origin-destination are considered known. We will later relax this assumption and investigate how uncertainty and errors in the demand influence our approach. The strategy of allocating fraction of space to each

sub-region in region i at time t is denoted by vector $\boldsymbol{\pi_i}(\boldsymbol{t})$ (which can be static e.g. 20% of space for bus-only lanes all the time, or dynamic e.g. 10% during off-peak while 20% during peak). The goal of road space optimization for a multi-modal transport system of a city is to minimize the total passenger hours travelled (PHT) over time for all modes of transport m that serve the total demand by redistributing the road space in areas with different usages of the city. Mathematically, the optimization problem is

$$\min_{\boldsymbol{\pi_i}(t)} Z = \sum_{t,i,m} \text{PHT}_{t,i,m}\left(\boldsymbol{\pi_i}(\boldsymbol{t})\right) \tag{1}$$

(1) is subject to the dynamics of the traffic system of the city. The dynamic interactions are illustrated in Fig. 2, where $\boldsymbol{n}(t)$ is the vector of accumulation of vehicle in different regions, $\boldsymbol{O}(t)$ the outflow, $\boldsymbol{Q}(t)$ the demand and $\boldsymbol{C}(t)$ the cost of travel for each of the regions. Given the initial state of the system at time t_0 and a space allocation strategy $\boldsymbol{\pi}(t)$, the system modeled by three parts that interact at every time step: the traffic state and travel cost of each mode at each region, the flow dynamics within and between regions and mode choice of the generated demand at each region. We develop the methodological framework in the following sections.
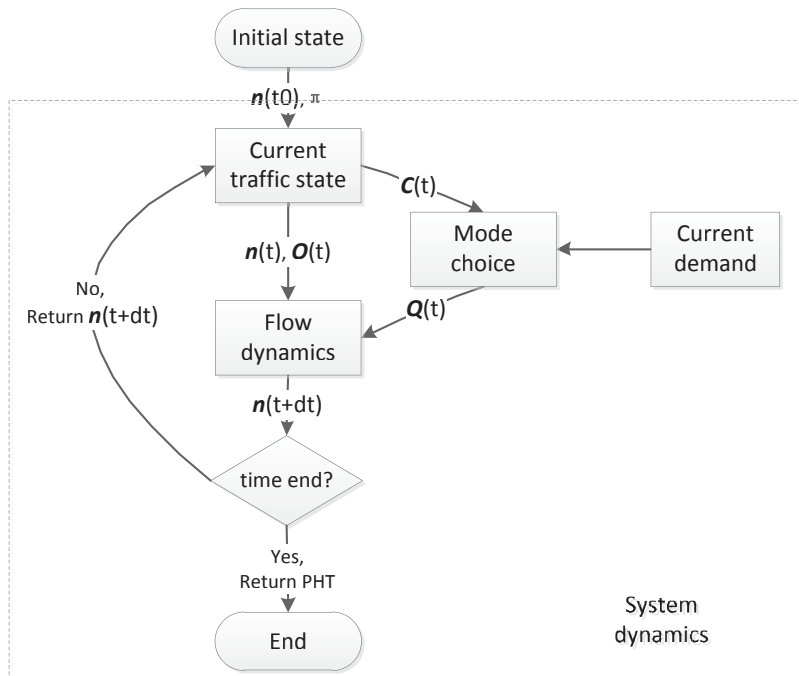

Fig. 2. System dynamics flow chart

3.1 Traffic flow dynamics

Let $Q_i^{km}(t)$ be the demand generated in region i with final destination k choosing to travel with mode m at time t. The total demand generated during this interval from i to k, $Q_i^k(t) = \sum_m Q_i^{km}(t)$. Route choice for each demand pair from i to k is defined exogenously with binary variables $\delta_{i \to j}^k(t)$, that show the sequence of the trip, with value equal to 1 if a trip from region i to k passes through j immediately after leaving i, and 0 otherwise. It is assumed that all trips generated from i with destination k follow the same sequence of regions in their route. For this paper it is assumed that the sequence of regions is not altered by traffic conditions. Let us now consider two modes bus and car, with indices c and b respectively $m = \{b, c\}$. Let $n_i^{km}(t)$ be the accumulation of mode m in region i with final destination k, $O_{i \to j}^{km}(t)$ be the transfer flow of mode m from region i to j with final destination k. Note that $n_i^m(t) = \sum_{k=1}^{N} n_i^{km}(t)$. Note also that for $i=k$, $O_{i \to j}^{km}$ is zero $\forall j \neq i$. $O_{i \to i}^{im}$ is the trip ending flow with final destination region i.

As long as flow $O_{i \to j}^{km}(t)$ enters j, it becomes part of accumulation $n_j^{km}(t)$. The dynamic equations between the state variables n_i^{km}, flow variables $O_{i \to j}^{km}$, and the demand Q_i^{km} of the system with N regions for the two modes are

$$n_i^{kc}(t+1) = n_i^{kc}(t) + \frac{Q_i^{kc}(t+1)}{ob_i^c} - \sum_{j=1}^{N} O_{i \to j}^{kc}(t) + \sum_{l=1}^{N} O_{l \to i}^{kc}(t), \qquad (2)$$

$$n_i^{kb}(t+1) = n_i^{kb}(t) - \sum_{j=1}^{N} O_{i \to j}^{kb}(t) + \sum_{l=1}^{N} O_{l \to i}^{kb}(t), \qquad (3)$$

where, ob_i^c is the occupancy of cars in region i, which is a constant (say 1-1.3 passenger per car) and will remain the same until the end of the trip and $\frac{Q_i^{kc}(t+1)}{ob_i^c}$ is the number of generated car trips. Both equations are discrete versions of a mass conservation differential equation, where t represents an interval. (2) states that the change of car accumulation n_i^{kc} over time equals to the internally generated demand from region i to k minus the outflow from region i with final destination k plus the inflow from all regions adjacent to i with final destination k, which is also the outflow for these regions. (3) is similar to (2), but without new generation of bus trips to the system (it is assumed that the amount of buses in service is usually constant during typical commuting hours). Note that the number of buses in each region at different times depends on the relative speeds and MFDs of the regions.

Flow $O_{i \to j}^{km}(t)$ is the minimum of two terms: the sending flow from region i, which depends on the accumulations of region i and the boundary capacity of region j, BC_j, which is a function of the receiving region's accumulations. Nevertheless, this constraint can be ignored during the optimization process. The physical reasoning besides this assumption is that (i) boundary capacity decreases for accumulations much larger than the critical accumulation that maximizes flow (see [42]), and (ii) optimized space allocation will not allow the system to get close to gridlock. Note also that $O_{i \to j}^{km}(t) = 0$, if $\delta_{i \to j}^{k}(t) = 0$.

The sending flow of $O_{i \to j}^{km}(t)$ is estimated by the macroscopic fundamental diagram O_i^m. The MFD of mode m of region i is denoted as $O_i^m(n_i^b(t), n_i^c(t))$. For our approach, the given condition is that the city is partitioned in homogeneously congested regions and that each region has a well-defined MFD. MFD O_i^m can be analytically estimated as a function of accumulation for different mode usages and space allocations using variational theory described in the previous section. Then,

$$O_{i \to j}^{km}(t) = \min\{O_i^m(n_i^b(t), n_i^c(t)) \cdot r_i^{km}(t), BC_j(n_i^b(t), n_i^c(t))\} \qquad (4)$$

where $r_i^{km}(t)$ is the proportion of the outgoing flow O_i^m from region i to j with final destination k. If the average distance traveled by travelers within region i is assumed to be independent of their destinations, then the proportion $r_i^{km}(t)$ can be estimated as a ratio of the related accumulations, according to Little's formula [43].

$$r_i^{km}(t) = \frac{n_i^{km}(t)}{n_i^m(t)} \qquad (5)$$

We should also estimate the dynamics of passengers for each mode and every region. With respect to the car, we just need to multiply both sides of (2) by ob_i^c. For simplicity, we assume $ob_i^c = 1$. With respect to the bus, we need to estimate how the occupancy of bus (in passengers) evolves over time according to demand and finished trips. The dynamics of bus passengers are very different than bus dynamics. A simple on-board passenger mass conservation equation is

$$OB_i^{kb}(t+1) = OB_i^{kb}(t) + Q_i^{kb}(t+1) - \sum_{j \neq i}^{N} O_{i \to j}^{kb}(t) \cdot ob_i^{kb}(t) + \tag{6}$$

$$+ \sum_{l=1}^{N} O_{l \to i}^{kb}(t) \cdot ob_l^{kb}(t) - b_i^k \cdot OB_i^{kb}(t) \cdot (1 - (1 - \theta_i)^z),$$

where

$$ob_i^{kb}(t) = \frac{OB_i^{kb}(t)}{n_i^{kb}(t)}, \tag{7}$$

$$\theta_i = \begin{cases} \left(\dfrac{\overline{L'}_{ib}}{s_i} \right)^{-1}, & if \ i = k \\ 0, & otherwise \end{cases} \tag{8}$$

In (6), OB_i^{kb} is the number of bus on-board passengers currently in region i with final destination k. (7) estimates ob_i^{kb}, the average number of on-board passengers per bus from region i to k. The RHS of (6) consists of the following terms: (i) the passenger inflow at time t+1 for the specific OD pair, (ii) the passengers that move outside region i while in the bus, (iii) the passengers that move inside region i while in the bus and (iv) the passengers that finish their trip inside region i. The last term is non-zero only for $i=k$ (binary $b_i^k = 1$ for $i=k$ and zero otherwise). This term approximates the passenger trip endings in region i as the mean of a Bernoulli trial which is repeated z times for each passenger with probability of success θ_i. At each stop in region i, every on-board passenger from the previous time step has probability θ_i of reaching its destination. $\overline{L'}_{ib}$ is the average trip length of bus passengers and s_i is spacing between bus stops for region i, i.e. passengers travel $1/\theta_i$ stops on average. Variable z is the number of stops that a bus travels during interval t, which is $V_i^b(t)T/s_i$ (T is the duration of the interval and $V_i^b(t)$ is the space-mean speed of buses in region i, which will be estimated in Section 3.2). Note that the trip length of a passenger, $\overline{L'}_{ib}$, is different than the trip length of a bus in the region, $\overline{L}_{ib}$. If more detailed data exists from real measurements, the generic model can be adjusted accordingly with a non-homogeneous Bernoulli process (or a Brownian motion). How the demand Q_i^k is divided between the two modes, Q_i^{kc} and Q_i^{kb} is discussed in Section 3.3.

3.2   Multimodal travel time estimation

The multimodal traffic flow dynamics of the previous section considered that the demand is known for each mode. This will be an input of a mode choice model, which is presented in the next section. To obtain this goal, it is important to identify the speed and travel time of each mode of transport. The estimation of space-mean speed $V_i^m(t)$ for each mode depends on the type of usage, i.e. mixed traffic of cars and buses or dedicated bus lanes. Keep in mind that variational theory estimation for shared-use of car and bus estimates the network flow and of cars for given car density given the conflicts and the operational characteristics of buses (frequency, dwell times etc). It does not estimate the speed of buses and extra effort is needed to account for the additional delay due to dwell times. This extra step is not required in separated bus lanes as there is only one mode of transport.

The space-mean speed in a time interval is by definition the ratio of the total distance travelled and the total hours traveled. Let $P_i^m(t)$ be the total distance travelled by vehicles of mode m (bus or car) in region i at time t and $\overline{L}_m$ the average trip length of mode m in region i. Then, for steady state queuing systems $P_i^m(t) = O_i^m(t) \cdot \overline{L}_{im}$ [43] and $V_i^m(t) \stackrel{\text{def}}{=} P_i^m(t) / n_i^m(t)$. Geroliminis and Daganzo [44] have shown that trip length of cars in downtown Yokohama was a time invariant variable. The average trip length of buses can be easily estimated as buses have fixed routes. By using the above equations and variational theory ([30][31]), we can estimate (i) O_i^c and V_i^c for car-only regions and mixed traffic car-bus regions and (ii) O_i^b for bus-only regions. For shared-use car-bus regions, the speed of buses further decreases because of dwell times for picking up and dropping off passengers by a parameter $0 < \alpha_i^b(t) \leq 1$ and $V_i^b(t) = V_i^c(t) \cdot \alpha_i^b(t)$. If we assume that buses have the same speed with cars when they are not stopped at bus stops then $\alpha^b(t)$ can be approximated as

$$\alpha_i^b(t) = \frac{TT_i^b(t)}{TT_i^b(t) + TT_d^b(t)} \qquad (9)$$

where $TT_d^b(t)$ is the total time spent at bus stops (total dwell time) in interval t, and $TT_i^b(t)$ is the average travel time of buses in region i excluding TT_d^b, which is given by $\bar{L}_{ib}/V_i^c(t)$. Average travel time of cars in region i, $TT_i^c(t)$, can be estimated as $\bar{L}_{ic}/V_i^c(t)$.

To validate all of the above assumptions, we performed a micro-simulation of San Francisco network taken from [42] which provides a low scatter MFD and can be estimated with variational theory with small error. While in the previous version this was a car-only simulation, we performed additional simulations for shared-use bus-car lanes and for dedicated bus lanes with different bus frequencies and dwell times. Figure 3 summarizes the MFDs in some of the scenarios for mixed traffic (Fig. 3a) and the dedicated bus lanes (Fig.3b), while Fig. 3c shows the average speed of cars, the speed of buses and estimated speed of buses according to the aforementioned methodology (time interval is 5min). A detailed analysis of these results will be reported in a future publication as it is beyond the scope of the current paper.
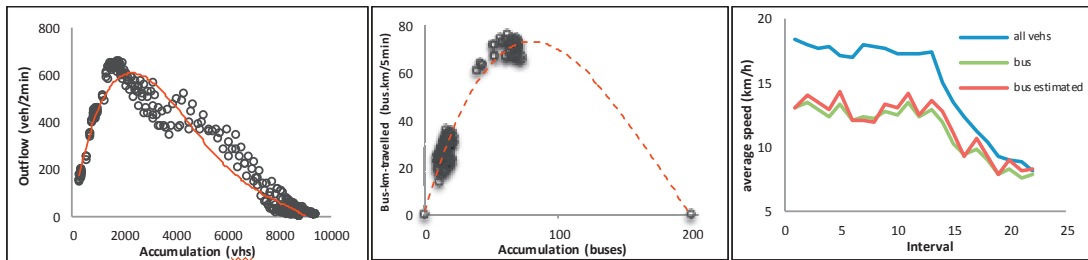

Fig. 3. (a) the MFD of mixed car-bus network; (b) The MFD of dedicated bus lanes; (c) Speed profile for mixed network

3.3   Mode choice

Suppose that the selection of the preferred mode is made by the travelers only once when they enter the network and start their trip. The total demand generated at each time interval $Q_i^k(t) = \sum_m Q_i^{km}(t)$ is assigned to the preferred modes through an endogenous approach. Mode choice of a traveler only happens at this moment and remains the same until her trip is finished. The choice of a mode is based on the utility of the mode, which is expressed as the travel cost of using that mode at the beginning of the trip. We consider that the travel cost of the whole route consists of travel time for cars and buses $TT_i^c(t)$ and $TT_i^b(t)$, on-board spatial discomfort $D_i^b(t)$ for buses and extra costs $C_i^c(t)$ for cars vs. buses (e.g. parking, tolls, fare) for each region i and interval t. Calculation of utility for car and bus for a specific origin destination pair (from i to k) is given by (10), where the summation term considers the set of all regions that a trip passes on its route from i to k, $\{S_i^k\}$, that can be estimated using binaries $\delta_{i \to j}^k(t)$.

$$U_i^{kc}(t) = - \sum_{j \in \{s_i^k\}} (TT_j^c(t) + C_j^c(t))$$
$$U_i^{kb}(t) = - \sum_{j \in \{S_i^k\}} (TT_j^b(t) + D_j^b(t)), \qquad (10)$$

where

$$D_j^b(t) = \gamma \cdot \left(\frac{ob_j^b(t)}{C_{bus}} \right)^2. \qquad (11)$$

The rationale of the spatial discomfort D_j^b is that (i) the utility of using bus decreases if number of on-board passengers (crowdedness) increases and (ii) it prevents the buses from reaching overcrowded conditions if buses are much faster than cars. Spatial discomfort of passengers D_j^b is a function of the average occupancy of a bus at

time t, which is the ratio between the average on-board passengers per bus in region j, $ob_j^b = \sum_{k=1}^{N} ob_j^{kb}$ and the storage passenger capacity of a bus, C_{bus}. As the unit of utility is time, the spatial discomfort term shall be transformed to time unit. Parameter γ should be calibrated with real data. Nevertheless, during the optimization process, which is described in details in the next section, we noticed that by changing the chosen value of γ by ±50%, objective function varies only by 2-3%. A different formulation with the same logic can be proposed for D_j^b, which should always be a monotonically increasing function of bus average occupancy.

Given the described utilities, we propose to model the mode choice of buses over time $p_i^{kb}(t)$ and $Q_i^{kb}(t)$ with a sequential approach, where the difference of utilities between the modes for a given origin destination pair is $\Delta U_i^k(t) = U_i^{kc}(t) - U_i^{kb}(t)$.

$$p_i^{kb}(t+1) = p_i^{kb}(t) + \beta_1 \cdot \Delta U_i^k(t) + \beta_2 \cdot \left(\Delta U_i^k(t) - \Delta U_i^k(t-1) \right)$$
$$Q_i^{kb}(t+1) = Q_i^k(t+1) \cdot p_i^{kb}(t+1)$$

(12)

(12) states that (i) if traffic conditions do not change, mode choice percentage remains the same as of the previous period and (ii) if traffic conditions change, mode choice percentage changes in proportion to the difference in utilities as expressed by ΔU_i^k, and to the evolution of ΔU_i^k. In such a way, we assume that all travelers make smart choices based on current and historic information. Note that travelers choose the mode of travel only once in the beginning of their trip based on the traffic conditions of this time. Parameters β_1 and β_2 need to be calibrated with real data. Alternatively, one can use the traditional logit-form model. Since we are trying to model a dynamic choice process and we consider that a dynamic estimation of a logit model might be a challenging task, we utilize the proposed approach.

(12) tries to succeed a set point, which is equal utilities of modes in our case (equilibrium in the choices). This type of controllers is a generic widely used control loop feedback mechanism. The equation calculates an error value as the difference between a measured process variable and a desired setpoint. It attempts to minimize the error ΔU_i^k by adjusting the process inputs, p_i^{kb}. The adjustment depends on the present error (term β_1) and the accumulation of past errors (term β_2). Sensitivity analysis of parameters γ, β_1 and β_2 is performed later.

3.4 Optimization framework

After modeling the system dynamics, we can now describe the optimization framework. Given values for the space allocation vector $\boldsymbol{\pi}$, which include temporal and spatial variables, the total passenger hours travelled (PHT) can be obtained as (where T is the duration of the interval t)

$$PHT(\pi) = \sum_t \sum_i \sum_k \left(n_i^{kc}(t) \cdot ob_j^{kc} + OB_i^{kb}(t) \right) \cdot T.$$

(13)

Optimization of (13) is highly non-linear. We solve this problem by a non-linear programming method, the sequential quadratic programming (SQP). SQP method solves a sequence of optimization sub-problems, each of which optimizes a quadratic model of the objective subject to a linearization of the constraints. Consider a non-linear programming problem of the form $\min_\pi PHT(\pi)$ subject to constraints $g_M(\pi) \geq 0$ and $h_M(\pi) = 0$. In our case, $g_M(\pi)$ will state the constraints of implementation of space allocation in practice, e.g. space should not be frequently reallocated. $h_M(\pi)$ will be system dynamics equations introduced in the previous sections. The general Lagrangian for this problem is defined as (λ and σ are Lagrange multipliers)

$$\mathcal{L}(\pi, \lambda, \sigma) = PHT(\pi) + \sum_M \lambda_M \cdot g_M(\pi) + \sum_M \sigma_M \cdot h_M(\pi).$$

(14)

At an iteration π_K, an approximation is made of the Hessian of the Lagrangian function, using a quasi-Newton updating method. This is then used to generate a quadratic programming sub-problem, and it can be solved using any QP algorithm. The solution is used to form a search direction d_K for searching a new iterate $\pi_{K+1} = \pi_K + \alpha_K d_K$, where the step length α_K is determined by an appropriate line search procedure. The iteration will

continue until stop criterion is achieved: the change of objective function is below a certain threshold. For detailed description of the SQP algorithm, please refer to reference [45]. We apply this algorithm for multiple initial values (around 1000) to avoid convergence to local minima, which might be the case for a non-smooth objective function.

4   RESULTS

An application of the developed methodological framework is performed with a case study of a two-region city. Consider an urban network with two concentric regions, as shown in Fig. 4. Mixed traffic of buses and cars occurs in the outside region (periphery), while a fraction of road space in the center region is dedicated to buses. The radius of the center region is $R_1 = 1.6km$ and of the periphery is $R_1 + R_2 = 3.2km$. The road networks of the two regions are well connected at the border. We simulate an urban road traffic system for 4-hours (80 time units), a typical morning or evening period. Demand has a symmetric trapezoidal shape with time and the length of peak period is equal to 1hr. A 70% fraction of the demand generated in the periphery will travel to the city center and 30% fraction of the demand generated in the center will travel to the periphery of the city. The dynamics of the system follow the equations defined in section 3.1, 3.2 and 3.3, while MFDs for different sub-areas are estimated with variational theory. Two modes of transport are considered available in the system, car and bus. A fraction 10% of the users are captive and do not have access to cars. The network topology, signal settings and traffic parameters are as follows: (i) two-lane one-way roads with block lengths $l_{b1} = 154m$ and $l_{b2} = 308m$, (ii) signal settings for all intersections, green, $g = 40sec$; cycle, $c = 90sec$ and offset, $o = 0sec$, (iii) a triangular fundamental diagram for each link with free flow speed for buses and cars, $u_f = 15m/sec$; jam density, $k_{jam}^c = 0.15veh/m$ for cars and $k_{jam}^b = 0.05veh/m$ and congested wave speed for buses and cars, $w = 5m/sec$.

Static (Section 4.1) and dynamic (Section 4.2) space allocations are investigated. We also investigate how an area-based (Section 4.3) pricing strategy can further facilitate demand shift from cars to public transport. Sensitivity analysis for the non-physical parameters of the model is performed in Section 4.4, while the effect of demand increase and uncertainty is also integrated in the analysis.
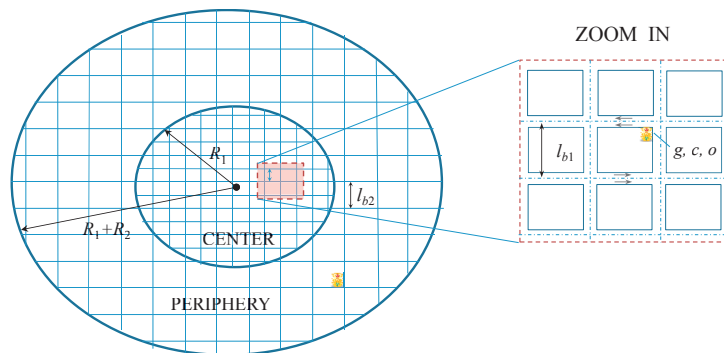

Fig. 4. An illustration of the case study

4.1   Static space allocation

We first investigate the performance of the traffic system, where a constant amount of space is allocated to dedicated bus lanes in the center region during all times. The results of the basic scenario where buses and cars share the whole network system without bus lanes, and of a scenario where 10% and 15% of the space in the center region is dedicated to bus lanes, are shown in Fig. 5. Note that the system performance of the latter scenarios significantly improves. The highly congested part disappears, the center of the city operates close to the maximum outflow of the car MFD, and total person hours traveled (PHT) during peak hours reduces by more than half.

Since we apply a static space allocation which has only one space variable to be optimized, we can simply enumerate all the possible values of space allocation and compare the results. Result of such an experiment is shown in Fig. 6a, where PHTs are plotted over percentage of space for bus lanes. The small "kink" for $\pi = 3\%$ is because of the behavior of captive users that choose bus even if car is a faster mode. The optimal space allocation can be found around 10%. The SQP optimization approach of the previous section provides identical results. From the figure, an interesting observation is that the resulting PHTs are similar for $\pi = 10\%$ to 15%. This means that small variations in the demand or errors in the parameters of the model will still produce close-to-optimality results and PHTs will remain low.
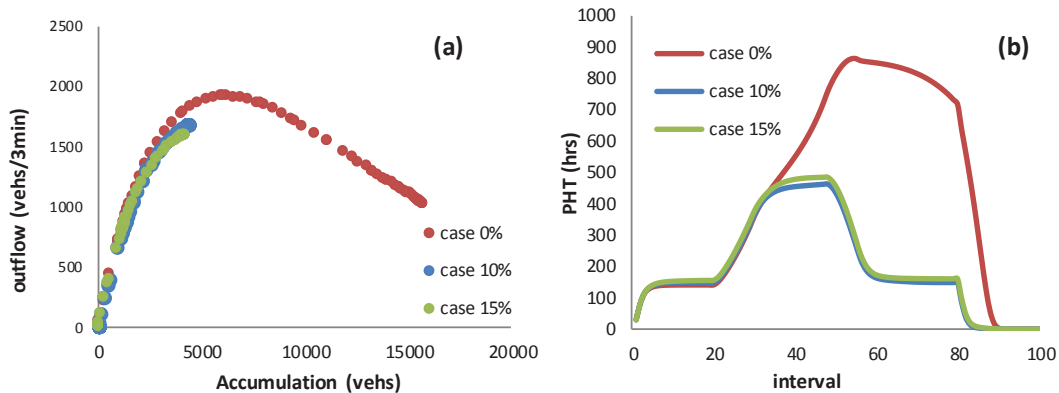

Fig. 5. An illustration of the performance of the system with two cases of space allocation to buses:
(a) The car MFD in center region (cases 10%, 15% cars only, case 0% cars and buses mixed), (b) PHT for both modes over time
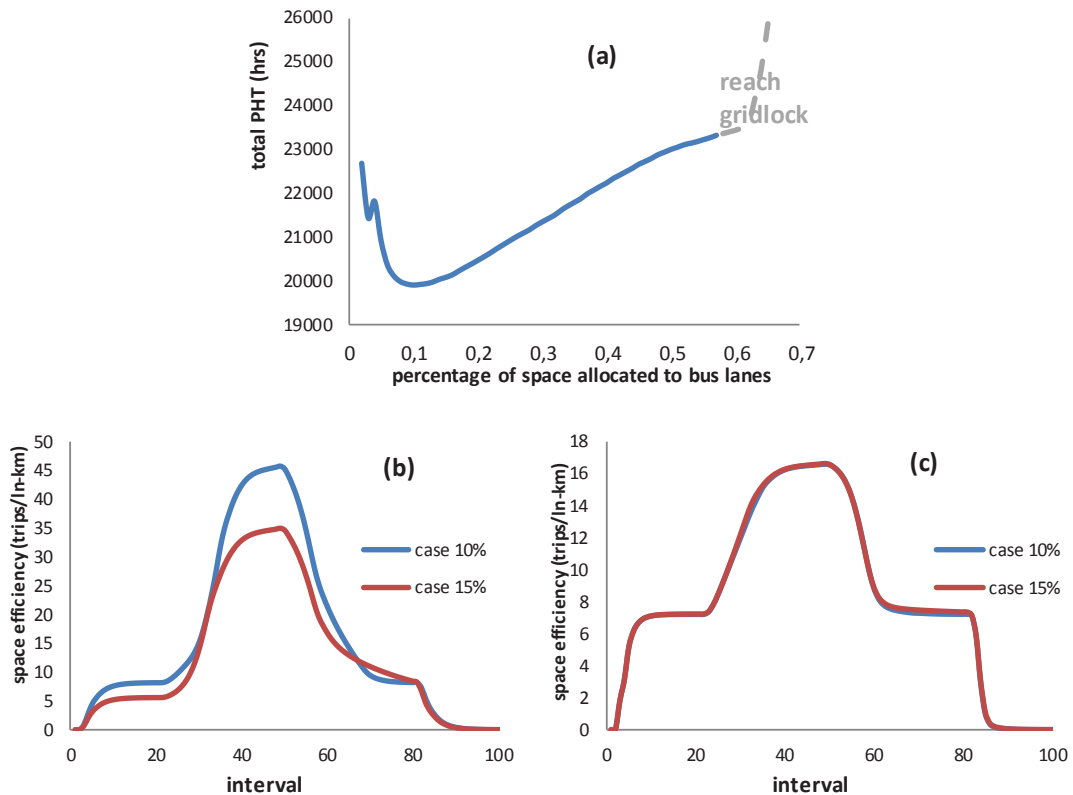

Fig. 6. (a) (b) efficiency of bus lanes given difference space allocation (c) efficiency of the whole space with different space allocation of bus lanes

Let us check more carefully the efficiency of the allocated space, which is defined as the ratio between the passenger trips completed during the total simulation time and the total lane kilometers of the space for buses. Space allocations of 10% and 15% cases are compared, where efficiency of bus lanes is shown in Fig. 6b and efficiency of space of the whole center region in Fig. **6**c. If we look at the space efficiency of the bus lanes, it is clear that with the increase of allocated space for buses, the trip completion rate per ln-km actually decreases. Note that for the whole region (Fig. 6c), the two cases are almost identical. The reason is that: More space for buses allows to transport more persons on bus lanes, but at the same time it constraints the capacity of car network and car users travel at a slower rate. Also, higher number of bus passengers increases the dwell times and the benefit of more space is smoothed off. Given the fact that the cost of operating dedicated bus lanes is high, the optimal solution should still hold at the lower value of 10%.

4.2   Dynamic space allocation

It is clear in Fig. 6b that the utilization of bus lanes is higher during peak hour, while it is much less during the onset and the offset of congestion. A dynamic space allocation strategy, where space is allocated in a time-dependent way is expected to increase spatial efficiency and further minimize the total travel cost. Nevertheless, given the associated infrastructure, the change of space should not happen many times during the day as implementation will be difficult. In this section, we investigate the optimal solution for a four-variable, three-period dynamic space allocation strategy. The three periods are the peak hour and the off-peak before and after the peak hour. The four optimized variables are: (i) the starting time t_1 and (ii) the ending time t_2 of the high space allocation, (iii) space allocated to bus lane during the off-peak periods pi_1 and (iv) during the peak period pi_2.

We apply the algorithm described in Section 3.4. The iteration and the solution of 10 executions with different random initial solutions out of the 1000 are shown in Fig. 7. The dotted line in both figures indicates the best solution among these 10 examples. We see that by applying a multiple random initial search, we are indeed able to avoid local minima, for example note a solution that gives 30% of the space to buses during the peak period.
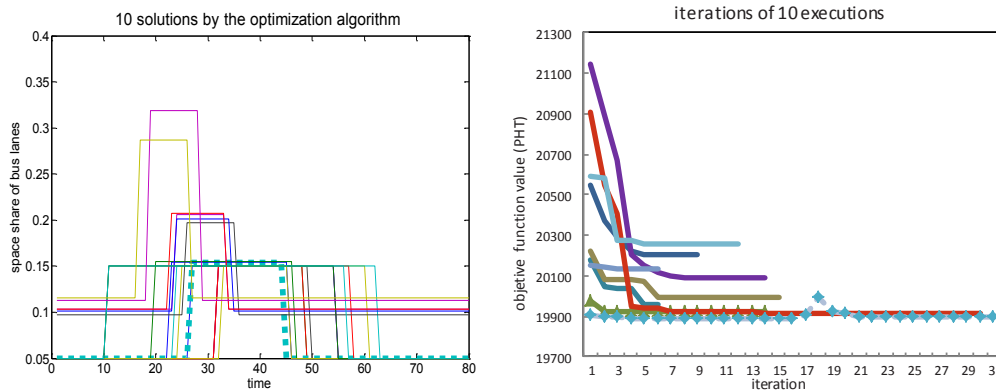

Fig. 7. (a) 10 executions of the SQP optimization algorithm for different starting points and (b) the corresponding optimal solutions after a number of iterations for each execution

Fig. 8 shows the optimal solution for this dynamic space allocation strategy. The line "pi" is the resulting space share of bus lanes over time, while "demand" is the demand profile. Note that the optimal strategy allocates less space when demand is low, but three times more during the "identified" peak period. The starting time and the ending time of the identified peak period makes physical sense, as it starts earlier than demand rate reaches its maximum and finishes later than demand rate starts decreasing. In such a way, the strategy "forces" travelers to choose to travel by bus proactively, before the beginning of peak hour, therefore all travelers avoid experiencing high travel cost. It also "tolerates" travelers to choose cars as the main mode of travel after the end of the peak hours. The optimal PHT is 19715 hours, better than the optimal PHT by a static space allocation, which is 20216

hours (10% further improvement from the $\pi = 0$ solution). Note that this improvement is succeeded with 15% less total space allocated for buses, as expressed by the total lane-km-hrs of individual bus lanes.
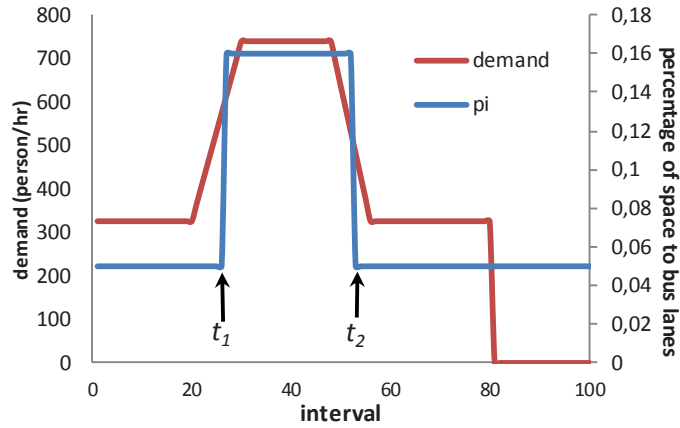

Fig. 8. The optimal solution of a time-dependent bus lane space allocation strategy

Now let us compare the efficiency of allocated bus lanes and the bus occupancy during the peak hour between the static strategy and the dynamic strategy. The results are shown in Fig. 9. During the off-peak period, the dynamic strategy allocates less space to buses, dropping from 10% to 5%, which increases the efficiency of bus lanes (as expressed by passengers served per lane-kilometer of bus space). While during the peak period, the dynamic strategy has lower utilization rate when the allocated space increases from the constant 10% to 16%, nevertheless the dynamic strategy actually serves more demand for buses during the period (shown by the right figure). Note that two sharp changes for the dynamic case are observed during the space changes at times t_1 and t_2 (off-peak to peak, and peak to off-peak). Nevertheless, the number of served passengers is smooth vs. time (see Fig. 9b). These changes represent the transition of operation between the two space allocation scenarios.
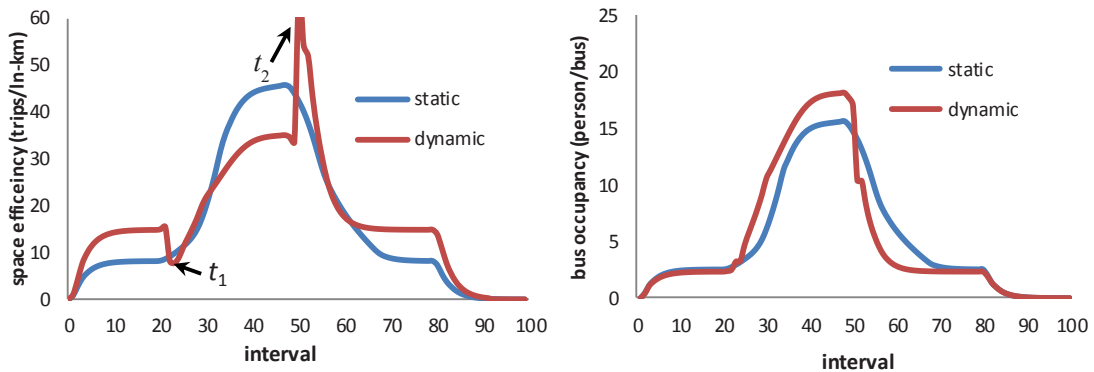

Fig. 9. Comparison of static and dynamic allocation: (a) space efficiency vs. time and
(b) Average bus occupancy vs. time

As a high utilization rate indicates that a large amount of demand chooses to travel by bus, one might question if this strategy will create cases of buses full of passengers (at capacity C_{bus}), which is neither comfortable nor realistic. This will not happen in the developed model, because Equation (11) ensures that mode choice of buses will reduce significantly when buses are crowded. If we check the occupancy of buses, as shown in Fig. 9b, we see for both the static strategy and the dynamic strategy that buses are occupied less than half of their capacity (40 persons/bus). In fact, if more demand shifts from cars to buses, while buses are given enough space to travel with high speed, PHT could be further reduced. This is investigated in the next section.

4.3   Dynamic space allocation with congestion pricing

Following the discussion, we will investigate measures to improve the occupancy of buses. To trigger mode shift from cars to buses, we apply an area-based pricing in the peak-hour, where car users have to pay when they drive in the center of the city. The cost of pricing is added into Equation 10 and influences the mode choice. Since toll is in money units, we transform it to time by multiplying with a value-of-time (VOT). In our application we utilize a value of 16 Swiss Franc (CHF) per hour, which is estimated based on an analysis by Axhausen et al. [46] for Swiss travelers. The start and end of pricing is assumed to be the same as of the peak-hour space allocation, between t_1 and t_2. The amount of toll charged c_1 is an additional variable to be optimized (5 in total). In our analysis we assume that toll will only affect the mode choice and not the departure time of the trip. While an analysis of the morning commute for multiple modes is analyzed for a simpler system [24], departure time is not included in this current work as it will make the modeling part very tedious and intractable.

We execute this five-variable optimization with the SQP algorithm with multiple initial searches. The total PHT is 18215hrs, which is about 12% smaller than the PHT without pricing and dynamic space allocation. The space allocation during the peak period decreases from $\pi = 16\%$ without pricing to 14%, while values for t_1 and t_2 are identical for the two cases. The savings in PHT are about the same with the total toll paid divided by VOT. While this is the case in the single bottleneck model [47], in systems governed by a variable capacity (like the MFD model of this paper), savings can be much larger, because congested states are avoided [48]. But, the no toll case operates close to capacity, because of an efficient space allocation, while the toll paid is similar to delay savings. Nevertheless, as we will show later, an increase in the demand of travel in a city can create additional congestion and given that a new design of space allocation is not an easy solution (due to high infrastructure cost), variations in pricing can significantly improve the state of the system and avoid congestion.

The occupancy of buses over time is compared in Fig. 10a where "dynamic with pricing" is the resulting bus occupancy of the pricing strategy. As expected, the buses are more occupied after pricing. Besides, it can be observed the occupancy goes to zero earlier when approaching the end of the simulation (interval 81), indicating that passengers are able to finish their trips earlier. This shows that a faster travel speed is achieved for all users and it corresponds to a reduction in PHT. Fig. 9b shows the MFD states for cars in the center of the city. The two different maximum values for the dynamic case represents the two different space allocations. With pricing the system operates at a more reliable state as this is less than the critical accumulation that maximizes outflow (n=4000veh). As we will show later, in this case the system can absorb increase in demand or small stochastic fluctuations without entering the congested regime, while in the case without pricing, a demand increase will also be associated with higher congestion.

The pricing strategy also succeeds higher demand shift and utilization rates of the buses during peak hours (Fig. 9c). Fig. 9d shows how the utility of bus and car change over time for two different O-D pairs (center to center and periphery to center). Note that when bus share is minimum car utility (during the off-peak) is significantly higher, while during the other times the two utilities are about the same. Thus, Equation (12) succeeds in identifying a dynamic equilibrium of mode choice.
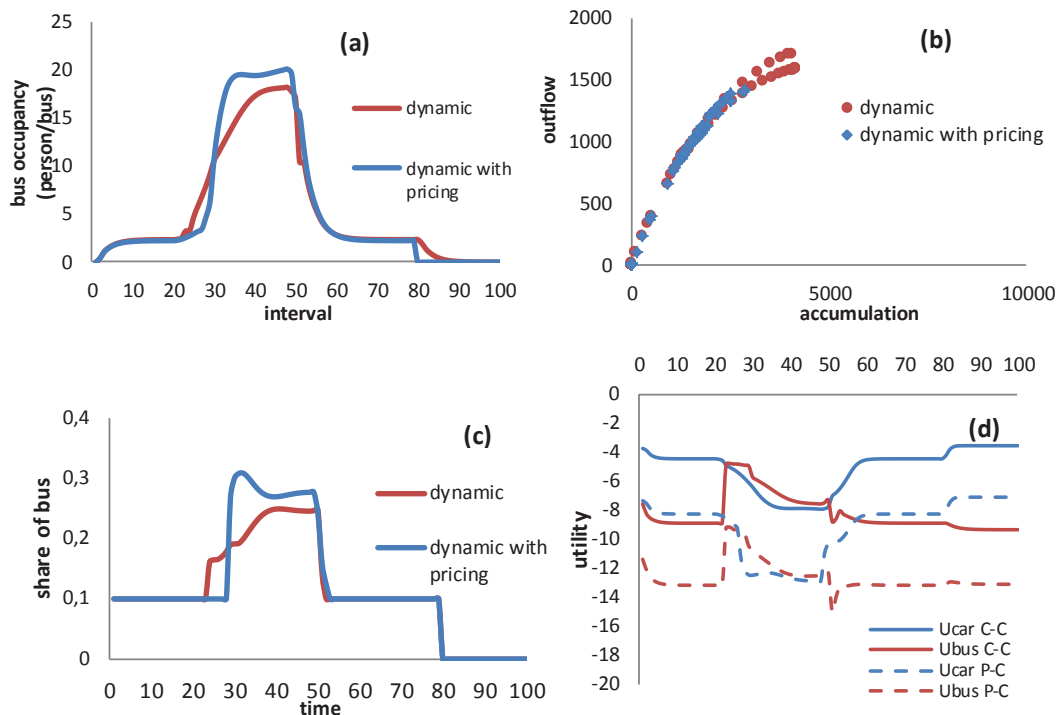
Fig. 10. Comparison of the dynamic allocation with and without pricing: (a) occupancy of buses vs. time; (b) MFDs of the center region for cars; (c) mode share of bus vs. time and (d) utilities of car and bus vs. time for trips from center to center (C-C) and periphery to center (P-C) for the "dynamic with pricing" case

4.4   Sensitivity analysis

In this section we discuss how sensitive is the resulting PHT, if parameters of mode choice β_1 and β_2 in (12) and discomfort parameter γ in Equation 11 are given different values. We also investigate how an increase in the demand will create additional congestion and if this can be resolved without a new re-distribution of the urban space. Table

Table 2 shows the results of a sensitivity test. The basic scenario is the one where the solution of section 4.2 is applied. For each of the three parameters, values vary by percentages as shown in the first row of the table. When one parameter is changing, the other two remain constant and the corresponding change of PHT is calculated.

Table 2 sensitivity analysis on $\beta_1\beta_2$ and γ

	change of parameters	-75%	-50%	-25%	Base	+25%	+50%	+75%	+100%	+200%
change	β_1	0.7	-0.03	0.1	0	0.02	0.01	0.2	0.3	0.6
of PHT	β_2	0.2	0.08	0.04	0	-0.09	0.1	-0.2	-0.2	-0.4
(%)	γ	-1.4	-1.5	-0.8	0	0.7	1.4	2	3	5

We see that the changes of β_1 and β_2 have small impact on the resulting PHT, while the change of γ has relatively greater impact. This result shows a good property of β_1 and β_2. For γ the result is consistent with reality. A higher value of γ prevents users from using buses therefore worse off the whole system significantly, while a low γ encourages users and consequently reduces PHT. But when γ is too low, buses attract too much demand, the dwell times increase and reduce the speed of buses therefore PHT starts to increase again. Note that for β_1 and β_2, an extreme high value will result to an "all-or-nothing" mode choice: all users choose to travel with the better mode but then the travel cost of this mode becomes huge suddenly and result a complete choice

of the other mode in the next time period, creating unrealistic instability and oscillations. For γ, a high value will prevent travelers from choosing buses, which not only underutilizes the bus space but also results in heavier congestion for cars. In our model, we tune the parameters via trial-and-error, with objective of having a coherent and stable mode choice process. As mentioned already, these parameters need to be calibrated when the model is applied in a real city.

During the optimization process, perfect information about the demand profile was assumed. In reality demand can experience stochastic fluctuations or erroneous measurements, as it might not be straightforward to accurately measure, even at the aggregate level. To investigate how the model reacts under demand uncertainty, the effect of unbiased and biased errors are introduced for the optimal solution, obtained by the trapezoidal demand. With respect to the unbiased demand, we consider fluctuations from the deterministic trapezoidal demand as a standard normal distribution $\mathcal{N}(0,1)$, multiplied by a degree of error λ, i.e. $\hat{Q}_i^k(t) = Q_i^k(t) \cdot \left(1 + \lambda \cdot \mathcal{N}(0,1)\right)$. We perform multiple runs for values of $\lambda \in [0,0.25]$ and the predetermined space allocation and tolls. Even for large degrees of error, the total passenger hours travelled (PHT) do not change more than 5%, which highlights that the approach is robust and not very sensitive to random demand fluctuations.

In the case of biased error, for example because of a demand increase the analysis is more complicated as PHT can significantly increase. The reason is that the base case represents a high demand case, which without efficient space allocation creates significant congestion, while the optimized solution operates at the network capacity in the center region for the roads devoted to cars. Thus, this demand increase will create states in the congested part of the MFD and increase the PHT. Nevertheless, a new redistribution of the urban space might not be a feasible solution due to high infrastructure costs and it should not be performed frequently. Instead, a change in the price of toll is easy to be implemented and for some range of demand increase can still lead the system to an efficient state and avoid states in the congested part of the MFD, which results in network capacity loss. To quantitatively analyze the above, we estimate the total PHT of the two-region model as demand increases for the optimal space allocation and tolls of the base scenario. Afterwards, we keep the space allocation identical and optimize the system with respect to the value of the toll in the peak hour for different demand levels.
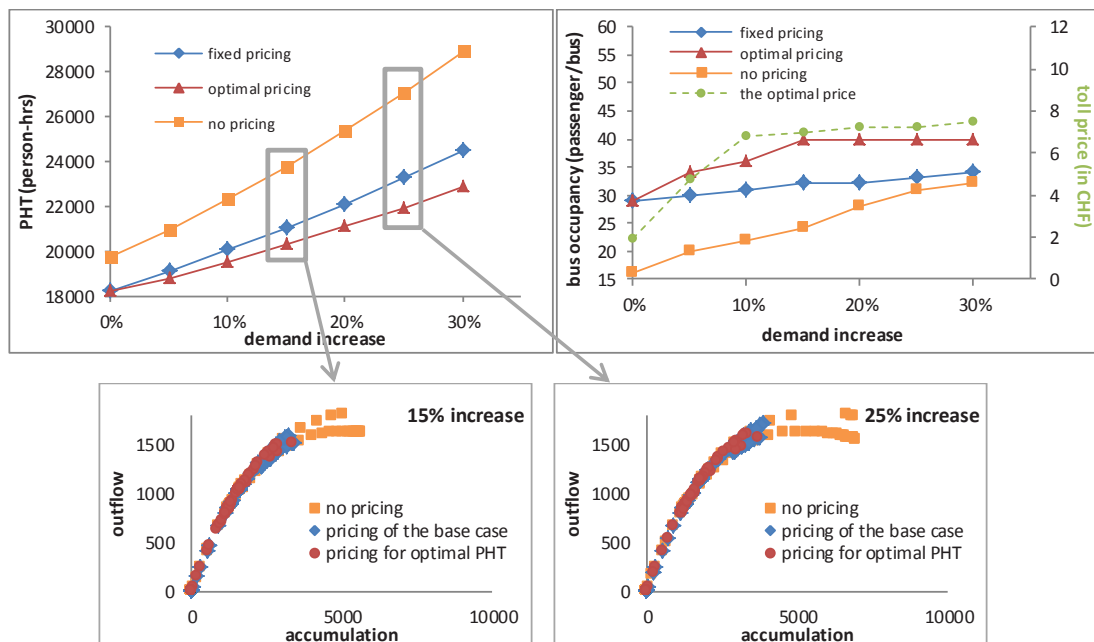

Fig. 11. Comparison of system performances under different pricing scenarios for different levels of demand increase

The results are summarized in Fig. 11, where PHT for different demand increase is estimated for (i) the dynamic allocation without toll, (ii) the dynamic allocation with fixed toll of the base scenario and (iii) the dynamic allocation with toll re-optimization for each demand level. We also show the average bus occupancy (in passengers/bus) during the peak hour for the three scenarios and the price of the optimal toll. Note that with price re-optimization, bus utilization is much higher while congested states are avoided for demand increase up to 25% (see the associated MFDs for cars in the city center). For higher increase (>25%) buses become full and congestion is unavoidable. The price of toll stabilizes for higher demand as buses are full and cannot attract more passengers. A solution in this case is to introduce more frequent buses in the system, which will allow for lower bus occupancy, but it might decrease the efficiency of the bus network.

5   DISCUSSION

In this paper, we present a macroscopic approach for allocating road space among modes of transport with the objective of minimizing the total travel time of travelers, PHT. We extend the single-mode MFD to multimodal cases, and utilize it to reproduce the traffic dynamics of a multimodal multi-region urban transport system. The extended model also considers the effect of scheduled stops on the operation of public transport, as it quantifies their effect on the average speed of the mode. This traffic model is consistent with the physics of traffic and the operational characteristics of a multimodal system. It can be analytically estimated and it can be practically observed with fewer requirements or data. Furthermore, we propose a dynamic mode choice model in order to investigate how the change of road space influences mode choice behavior and traffic performance. Given relevant information of a city (its road space allocation strategy and an aggregated O-D profile), the performance of the city can be evaluated and optimal space strategy can be provided with our approach.

We apply the approach on a two-region bi-modal city case study, and investigate the performance of two space allocation strategies for the center region of the city where demand is high and heavy congestion exists for static and dynamic case. Comparing the two optimal solutions, we find that the dynamic allocation strategy manages to minimize the PHT in a more efficient way as it utilizes the bus lane space better during the off-peak period and serves a higher amount of passengers during peak period. In practice policy makers may need to make trade-off between the operation cost of bus lanes and the savings in PHT, as it is not always efficient to give space to bus lanes. By implementing pricing during the peak period, more demand shift from cars to buses is succeeded increasing the occupancy of buses and further reducing the PHT. The result of a sensitivity analysis shows the robustness of our approach towards the fluctuations in the inputs and model parameters.

The results of the proposed approach are promising. Ongoing work focuses on testing the space allocation strategies in microscopic traffic models and investigates if the findings by macroscopic models are consistent with the outputs by traditional microscopic models. Further research direction is to investigate (i) deeper analysis of multiple regions, where regional route choice might change (ii) incorporate parking space in the developed framework and consider the effect of limited parking space into account for mode choice, (iii) combine space allocation with perimeter control of traffic signals and bus priority (iv) integrate additional modes like taxis, tram and metro lines. Another future direction is how the output of this macroscopic optimization framework can be implemented in reality (i.e. how to choose which roads should be devoted to public transport). The multimodal urban modeling should be also investigated by considering additional heterogeneity (i) among users, with respect to their mode choices and the trip length and (ii) among different regions of a city.

References

[1] Chen, C., Varaiya, P. and Kwon, J. (2005). An empirical assessment of traffic operations, *the 16th International Symposium on Transportation and Traffic Theory,* Maryland, USA.

[2] Helbing, D. (2001). Traffic and related self-driven many-particle systems, *Reviews of Modern Physics* 73, 1067-1141.

[3] Gipps, P. (1981). A behavioural car following model for computer simulation, *Transportation Research Part B* 15, 105-111.

[4] Nagel K. and Schreckenberg, M. (1992). A cellular automaton model for freeway traffic, J.Phys. I France 2, 2221-2229.

[5] Daganzo C.F., (1994). The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory, *Transportation Research Part B* 28 (4), 269-287.

[6] Helbing D. and Treiber, M. (1998). Gas-kinetic-based traffic model explaining observed hysteretic phase transition, *Physical Review Letters* 81, 3042-3045.

[7] Lebacque, J. P. (1996). The godunov scheme and what it means for first order traffic flow models, *the Proceedings of the International Symposium on Transportation and Traffic Theory*, Lyon, France, 647-678.

[8] Leclercq, L., Laval, J., Chevallier, E. (2007). The Lagrangian coordinate system and what it means for first order traffic flow model, *the proceedings of the 17th International Symposium on Transportation and Traffic Theory* , New York, USA, 735-753.

[9] Jiang, Y.Q., Wong, S.C., Ho, H.W., Zhang, P., Liu, R.X., Sumalee, A. (2011). A dynamic traffic assignment model for a continuum transportation system. *Transportation Research Part B* 45 (2), 343-363.

[10] Wirasinghe, S., Hurdle, V. and Newell, G. (1977). Optimal parameters for a coordinated rail and bus transit system, *Transportation Science*, 11(4), 359-374.

[11] Sparks, G. and May, A. (1971). A mathematical model for evaluating priority lane operations on freeways, *Transportation Research Record: Journal of Transportation Research Board* 363, 27-42.

[12] Dahlgren, J., (1998). High occupancy vehicle lanes: Not always more effective than general purpose lanes, *Transportation Researh Part A* 32(2), 99–114.

[13] Daganzo, C. and Cassidy, M.J. (2008). Effects of high occupancy vehicle lanes on freeway congestion, *Transportation Research Part B* 42(10), 861–872.

[14] Vuchic, V. (2007). Urban transit systems and technology. Published by John Wiley & Sons. Inc., New Jersey.

[15] Gonzales E., Geroliminis N., Cassidy M., Daganzo C.F., (2010). On the allocation of city space to multiple transport modes, *Transportation Planning and Technology* 33 (8), 643-656.

[16] Ceder, A. and Wilson, N. (1986). Bus network design, *Transportation Research Part B: Methodology,* 20 (4), 331-344.

[17] Radwan, A. and Benevelli, D. (1983). Bus priority strategy: Justification and environmental aspects, *Journal of Transportation Engineering* 109(1), 88-106.

[18] Black, J.; Lim, P. and Kim, G. (1992). A traffc model for the optimal allocation of arterial road space: A case study of Seoul's first experimental bus lane, *Transportation Planning and Technology* 16(3), 195-207.

[19] Currie, G.; Sarvi, M. and Young, W. (2004). A new methodology for allocating road space for public transport priority, in: C.W. Brebbia and L.C. Wadhwa (Eds.) *Urban Transport X: Urban Transport for the 21st Century.*

[20] Pushkarev, B. and Zupan, J. (1977). Public Transportation and Land Use Policy. Indiana University Press.

[21] Pickrell, D. (1985). Estimates of rail transit construction costs, *Transportation Research Record: Journal of Transportation Research Board* 1006, 54-60.

[22] Kenworthy, J., and Laube, F. (2001). Millenium Cities Database for Sustainable Transport: Analysis and Recommendations, *International Union of Public Transport (UITP)*, Brussels, Belgium.

[23] Cameron, I.; Kenworthy, J. and Lyons, T. (2005). Understanding and predicting private motorized urban mobility, *Transportation Research Part D* 8(4), 267-283.

[24] Gonzales, E. and Daganzo, C. (2012). Morning commute with competing modes and distributed demand: User equilibrium, system optimum, and pricing, *Transportation Research Part B* 46 (10), 1519–1534.

[25] Mesbah, M., Sarvi, M., Ouveyse, I. and Currie, Gl (2011). Optimization of transit priority in the transportation network using a decomposition methodology, *Transportation Research Part C,* 19 (2), 363-373.

[26] Godfrey, J. (1969). The mechanism of a road network, *Traffic Engineering and Control*, 11 (7), 323-327.

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

[28] Mahmassani, H., Williams, J. and Herman, R. (1987). Performance of urban traffic networks, *the 10th International Symposium On Transportation and Traffic Theory*, Amsterdam, The Netherlands.

[29] Daganzo, C. (2007). Urban gridlock: macroscopic modeling and mitigation approaches*, Transportation Research Part B* 41(1), 49-62.

[30] Daganzo, C., Geroliminis, N. (2008). An analytical approximation for the macroscopic fundamental diagram of urban traffic, *Transportation Research Part B* 42(9), 771-781.

[31] Geroliminis, N. and Boyaci, B. (2012), The effect of variability of urban systems characteristics in the network capacity, *Transportation Research Part B* 46 (10), 1607–1623.

[32] Buisson, C. and Ladier, C. (2009). Exploring the impact of homogeneity of traffic measurements on the existence of Macroscopic Fundamental Diagrams. *Transportation Research Record: Journal of the Transportation Research Board*. 2124, 127-136.

[33] Ji, Y., Daamen, W., Hoogendoorn, S., Hoogendoorn-Lanser, S. and Qian, X. (2010). Macroscopic fundamental diagram: investigating its shape using simulation data, *Transportation Research Record: Journal of the Transportation Research Board* 2161, 42-48.

[34] Mazloumian, A., Geroliminis, N. and Helbing, D. (2010). The spatial variability of vehicle densities as determinant of urban network capacity, *Philosophical Transactions of Royal Society A* 368 (1928), 4627-4648.

[35] Daganzo, C.; Gayah, V. and Gonzales, E. (2011). Macroscopic relations of urban traffic variables: bifurcations, multivaluedness and instability. *Transportation Research B* 45(1), 278-288.

[36] Geroliminis, N. and Sun, J. (2011). Properties of a well-defined macroscopic fundamental diagram for urban traffic. *Transportation Research Part B* 45(3), 605-617.

[37] Knoop, V., Hoogendoorn, S. and Van Lint, H. (2012). Routing Strategies based on the Macroscopic Fundamental Diagram, *the 91st Annual Meeting Transportation Research Board*, Washington D.C., USA.

[38]   Saberi, M., Mahmassani, H., 2012. Exploring properties of network-wide flow-density relations in a freeway network, *the 91st Annual Meeting Transportation Research Board*, Washington D.C., USA.

[39]   Gonzales, E.; Chavis, C.; Li, Y. and Daganzo, C. (2011). Multimodal Transport in Nairobi, Kenya: Insights and Recommendations with a Macroscopic Evidence-Based Model, *the 90th Annual Meeting Transportation Research Board*, Washington, D.C., USA.

[40]   Boyaci, B. and Geroliminis, N. (2011). Estimation of the network capacity for multimodal urban systems, *the 6th International Symposium of Highway Capacity*, Stockholm, Sweden.

[41]   Ji, Y. and Geroliminis, N. (2012). On the spatial partitioning of urban transportation networks, *Transportation Research Part B* 46 (10), 1639–1656.

[42]   Geroliminis, N. and Daganzo, C. (2007). Macroscopic modeling of traffic in cities, *the 86th Annual Meeting Transportation Research Board*, Washington D.C.

[43]   Little, J. (1961). A proof of the queueing formula L=λW, *Operations Research* 9 (3), 383-387.

[44]   Geroliminis, N. and Daganzo, C. (2008). Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. *Transportation Research B* 42(9), 759-770.

[45]   Nocedal, J. and Wright, J. (2006). Numerical Optimization, Second Edition. Springer Series in Operations Research, Springer Verlag.

[46]   Axhausen, K.; Hess, S.; Koenig, A.; Bates, J.; Bierlaire, M. and Abay, G (2007). State-of-the-Art estimates of Swiss value of travel time savings, *the 86th Annual Meeting Transportation Research Board*, Washington D.C., USA.

[47]   Vickrey, W. (1969). Congestion theory and transport investment, *American Economic Review* 59, 251-260.

[48]   Geroliminis, N. and Levinson, D. (2009). Cordon pricing consistent with the physics of overcrowding, in *Transportation and Traffic Theory*, Springer, 219-240.

.