Transportation Research Part A 46 (2012) 1291–1303

Contents lists available at SciVerse ScienceDirect

Transportation Research Part A

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

A dynamic cordon pricing scheme combining the Macroscopic Fundamental Diagram and an agent-based traffic model

Nan Zheng [a], Rashid A. Waraich [b], Kay W. Axhausen [b], Nikolas Geroliminis [a,*]

[a] Urban Transport Systems Laboratory, School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[b] Institute for Transport Planning and Systems, Swiss Federal Institute of Technology Zurich, Switzerland

ARTICLE INFO

Article history:
Received 23 September 2011
Received in revised form 3 May 2012
Accepted 16 May 2012

Keywords:
Cordon pricing
Macroscopic Fundamental Diagram
Agent-based model
Dynamic pricing

ABSTRACT

Pricing is considered an effective management policy to reduce traffic congestion in transportation networks. In this paper we combine a macroscopic model of traffic congestion in urban networks with an agent-based simulator to study congestion pricing schemes. The macroscopic model, which has been tested with real data in previous studies, represents an accurate and robust approach to model the dynamics of congestion. The agent-based simulator can reproduce the complexity of travel behavior in terms of travelers' choices and heterogeneity. This integrated approach is superior to traditional pricing schemes. On one hand, traffic simulators (including car-following, lane-changing and route choice models) consider travel behavior, i.e. departure time choice, inelastic to the level of congestion. On the other hand, most congestion pricing models utilize supply models insensitive to demand fluctuations and non-stationary conditions. This is not consistent with the physics of traffic and the dynamics of congestion. Furthermore, works that integrate the above features in pricing models are assuming deterministic and homogeneous population characteristics. In this paper, we first demonstrate by case studies in Zurich urban road network, that the output of a agent-based simulator is consistent with the physics of traffic flow dynamics, as defined by a Macroscopic Fundamental Diagram (MFD). We then develop and apply a dynamic cordon-based congestion pricing scheme, in which tolls are controlled by an MFD. And we investigate the effectiveness of the proposed pricing scheme. Results show that by applying such a congestion pricing, (i) the savings of travel time at both aggregated and disaggregated level outweigh the costs of tolling, (ii) the congestion inside the cordon area is eased while no extra congestion is generated in the neighbor area outside the cordon, (iii) tolling has stronger impact on leisure-related activities than on work-related activities, as fewer agents who perform work-related activities changed their time plans. Future work can apply the same methodology to other network-based pricing schemes, such as area-based or distance-traveled-based pricing. Equity issues can be investigated more carefully, if provided with data such as income of agents. Value-of-time-dependent pricing schemes then can also be determined.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

To alleviate traffic congestion in cities, congestion pricing has been proposed by researchers and policy makers as an effective traffic management policy, with direct applications to cities, e.g. London, Singapore and Stockholm. The intention of pricing is to change travelers' behavior choice, such as departure time, route or mode choice and to reduce congestion by

* Corresponding author. Tel.: +41 21 69 32481; fax: +41 21 69 35060.
E-mail address: nikolas.geroliminis@epfl.ch (N. Geroliminis).

0965-8564/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.tra.2012.05.006

charging users for the external costs they create. A comprehensive literature summary of congestion pricing models can be found in Yang and Huang (2005). The theoretical background of pricing has relied on the fundamental concept, first introduced by Pigou (1920) and followed by Vickrey (1963) and other researchers: if on each link of a network a toll is charged, which equals to the additional congestion cost imposed on other users by an extra traveler, the sum of consumer surplus and total revenue is maximized. In the traffic assignment literature tolls of this type belong to the first-best pricing and have been proposed to drive a user equilibrium pattern (Wardrop, 1952) towards a system optimum. Despite their idealized theoretical basis, first-best pricing models have been impractical and difficult to implement. Merchand (1968) investigated second-best tolls using a general equilibrium model. According to the second-best pricing models, e.g. Arnott et al. (1990), Small and Yan (2001), Verhoef (2002), tolls are charged in a subset of selected links where the bottlenecks are. Demand pattern changes, e.g. departure time choice or mode choice, as the "utility" of travelers changes for a given toll. For example, in Vickrey's model (1969) a traveler experiences a utility cost of waiting in the queue and a penalty, "schedule delay", which is the difference between the actual time passing the destination and the desired time; accordingly. The traveler may adjust his departure time to avoid high schedule delay. Equilibrium is obtained when no individual has an incentive to alter his departure time. Arnott et al. (1988) extended the schedule delay concept with heterogeneous travelers, however applied at a single bottleneck only and without pricing. The common inadequacies of the existing models include: (i) user heterogeneity is not well dealt with, e.g. "utility" is identical for all travelers, while in reality utilities varies in trip purpose, desired arrival time, willingness to pay, etc., (ii) demand elasticity is limited to departure time and route choice where in reality mode choice and travel or not travel are also important components and (iii) Pigouvian-type tolls assume a network supply curve (desired or input demand vs. average travel cost) which is not consistent with the physics of traffic (Geroliminis and Levinson, 2009).

The second issue of pricing is that charging individual links is difficult to implement. Instead, pricing schemes of aggregated links and networks have been developed and applied in different cities (Zhang et al., 2008). Recently, Maruyama and Sumalee (2007) compared the performance of cordon- and area-road pricing schemes regarding their efficiency and equity. Anderson and Mohring (1997) examined congestion on the Twin Cities road network having drivers face marginal rather than average costs to reflect optimal prices using a user equilibrium assignment for a single period. Yang and Huang (1998) examined the principle of marginal-cost pricing in a road network. The basic ambiguity in most of these models is that traffic conditions are considered stationary. Furthermore, the traditional network supply curve for congestion pricing modeling, relating input demand to average travel cost, is not consistent with the physics of traffic. This is because (Geroliminis and Levinson, 2009) for a given desired demand over a period of time, the total cost expressed in delay terms (i) is sensitive, during congested conditions, to small variations of flow within the given period and (ii) depends on the initial state of the system and the level of congestion. On the other hand, it has been broadly shown through simulation and field experiments (e.g., Munoz and Daganzo, 2003; Helbing et al., 2009; Geroliminis and Daganzo, 2008) that the linkage between pertinent variables flow, speed and density on a spatially disaggregated level (one link) is very scattered and does not follow a well-defined curve. One of the reasons is that traffic systems are not in steady-state conditions at a link level. Thus, the estimated congestion toll based on idealized versions of these curves may not be optimal and the system may be either still congested if under-priced or very uncongested if over-priced. According to the same research, a Macroscopic Fundamental Diagram (MFD) model can better capture traffic behavior on an aggregated level, say an urban network, without the detailed knowledge of conditions in individual links. It was recently observed from empirical data that by spatially aggregating the highly scattered plots of flow vs. density from individual loop detectors in networks (roughly) homogeneously loaded with traffic, the scatter almost disappeared and a well-defined Macroscopic Fundamental Diagram exists between space-mean flow and density. The homogeneity of traffic conditions assumed in each aggregated network is also consistent with the homogeneity that applies to flow and density when dealing with road congestion in urban economic models (Small and Chu, 2003; Arnott, 2007; Arnott and Inci, 2010).

Recently, Geroliminis and Levinson (2009) combined Vickrey's theory with a macroscopic traffic model to identify the equilibrium solution for a congested network in the no-toll case. A dynamic model of cordon-based congestion pricing (such as for the morning commute) for networks was also developed consistent with the physics of traffic. In comparison to the bottleneck model, in the network case the optimal length of the toll period was found to be smaller than the congestion period in the no-toll case and the total delay savings to be higher than the total toll paid. The above work assumes deterministic and homogeneous population characteristics. This might result in non-optimal estimated tolls. Agent-based models are possible solutions for representing demand elasticity and heterogeneity. This is because (i) activity-based demand generation is utilized. Change in travel cost does not only influence travel behaviors of an agent but also his daily plan, which is a more realistic and accurate representation, (ii) agents have different value of utilities when they perform different activities, which introduce heterogeneity among agents, (iii) one agent's behavior affects other agents' decisions (Zhang et al., 2008) and (iv) lots of efforts have been made on realizing individual heterogeneity, e.g. different value-of-times and schedule-delays to agents. If the output of an agent-based model shows the property of the MFD, we can develop a dynamic network-wide congestion pricing schemes controlled by a macroscopic tool. This approach is more robust since tolls are determined based on traffic flow dynamics, rather than the traditional models based on demand–supply curves and marginal cost, which are sensitive to demand fluctuations and non-stationary conditions.

Empirical and simulation studies have been carried out to test and evaluate the effectiveness of different congestion pricing schemes at network level. For example Seik (2000) investigated the advantages of Singapore's Electronic Road Pricing, in term of efficiency, flexibility, equity, practical convenience, reliability, etc. De Palma et al. (2005) investigated the welfare gains and the resulting behavioral changes (such as departure time shift) of first-/second-best pricing schemes in METROP-

OLIS by using cost-benefit analysis. Similar analysis can also be found in May and Mline (2000), de Palma and Lindsey (2006), etc. In this paper, we will carry out similar evaluations for testing the effectiveness of our approach.

Based on the discussion above, to test the effectiveness of different congestion pricing schemes such as cordon- or area-based pricing, one could (if the output of an agent-based model is able to represent collective traffic behavior as expressed by the MFD) utilize this macroscopic description to develop a congestion pricing scheme in the agent-based model. Advantages of using this macroscopic tool also lie in that it has lower collection and transaction costs than link-based or area-based tolling and is based on traffic models that are readily observable with existing monitoring technologies. Therefore, the objectives of this research are: (i) to investigate the existence of the MFD in an agent-based simulator, (ii) to develop and apply an MFD-controlled cordon-based pricing, (iii) to examine the effectiveness of the proposed pricing scheme and how trips with different purposes are affected. In our study, the utilized agent-based simulator is MATSim (www.matsim.org), a state-of-the-art multi-agents based simulator which was developed jointly by ETH Zurich and TU Berlin.

The rest of this paper is structured as follows: we describe the basic principles of the agent-based simulator in Section 2, while the main features of the MFD model are presented in Section 3. In Section 4 we explain the algorithm of an MFD controlled dynamic cordon-based pricing and we analyze how the necessary variables can be estimated from MATSim. Results of case studies are given in Section 5, where the existence of MFD and the effectiveness of the proposed pricing scheme are shown. In the final section we draw conclusions and point out future work.

2. Multi-agent based traffic simulator

We now introduce some basic concepts of the agent-based traffic simulation model MATSim, which has been widely applied for transport and land use studies, e.g. Axhausen (2008), Löchl and Axhausen (2010) and Davidsson et al. (2005), and travel behavior modeling, e.g. Waraich et al. (2009a), Vertic et al. (2010) and Nagel and Flötteröd (2009). The simulator integrates activity-based demand generation with dynamic traffic assignment. Activity-based demand generation (ABDG) models generate daily activities in sequence and trips connecting these activities for every "agent" in the network. Demand generation thus is embedded in a concept of daily activity demand from which the need for transport is derived. Random utility theory is used to generate plans of daily activities. Each agent is assigned with different utility functions when performing different activities therefore behavioral differences among the agents are realized. Besides, in the context of ABDG, the entire activity plan (mode choice, departure time choice and the activity sequence) is the unit of decision to iterate route assignments. These are superior to traditional demand generation models. The simulation structure of the system can be summarized as follows (Meister et al., 2010): for every agent, one initial activity plan is given. Input data such as population and land use data, as well as network data are processed to generate this initial demand. Each agent initially has only one plan. The selected activity plans are simulated along the timeline in the model representation of the physical world: (i) of loading the agent on the network link at which the previous activity is located at a given departure time, (ii) of moving the agent along a given route through the network, where it interacts with other agents under way, and (iii) unloading the agent from the network at the link of the destination activity. Different from a variety of models for the simulation of car traffic with discrete entities, including car-following models and lane-changing models, the approximation of traffic in MATSim is fulfilled as the following: each road segment is modeled as a First In First Out queue, with a minimum service time of the length of the road segment divided by the maximum travel speed. The maximum number of vehicles that a queue can discharge equals the road capacity, depending on the number of lanes, etc. The capacity is thus a predetermined value, as opposed to models with flow dynamics where the actual maximum outflow is influenced by the number of accumulated vehicles (density) and their interactions. The only concept related to flow dynamics integrated into the queuing simulation is a shockwave between vehicles traveling backwards at constant speed in the case of traffic jam discharge, bringing in some notion of kinematic waves represented in their full scale in macroscopic traffic models. In order to compare activity plans, they are evaluated with a measure of general utility, called score, including utilities for activities and penalties for undesired manners. The related scoring function describes the agent's preferences. The score of a daily activity plan U_{plan} is given by a utility equation, the detail of which can be found in Meister et al. (2010). Given the score, a replanning strategy is selected by probability such as change of mode choice and another plan is created and then executed; or agents just decides which plan to select from its existing memory for the next execution according to a Logit-type probability. This iteration cycle is stopped after the properties of the system fulfill a stopping criterion. Conceptually, the system has to run until the agents cannot significantly improve the score of the executed plans and when the agent-based stochastic user equilibrium is reached. The simulation process can be briefly summarized by Fig. 1. In order to decide, if the "equilibrium" has been reached, the best, average, executed and worst plan of all agents are compared after each iteration. As the scores of these plans converge and

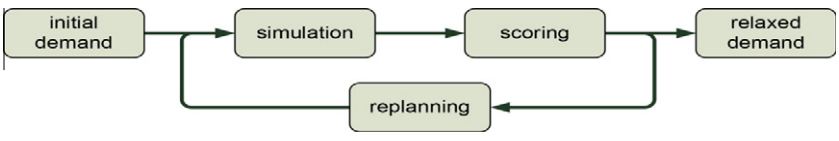

Fig. 1. Co-evolutionary simulation process of MATSim (Waraich et al., 2009b).

the improvements of the score between iterations decrease, the simulation is stopped. The execution time of such a simulation can take from hours to days for real-word scenarios depending on the population size and network traffic resolution (Waraich et al., 2009b).

3. The Macroscopic Fundamental Diagram

The idea of an MFD with an optimum accumulation belongs to Godfrey (1969) but the verification of its existence with dynamic features is recent (Geroliminis and Daganzo, 2007, 2008). These papers showed, using a micro-simulation and a field experiment, (i) that some urban regions approximately exhibit a "Macroscopic Fundamental Diagram" (MFD) relating the number of vehicles to space-mean flow of the region, (ii) there is a robust linear relation between the regions average flow and the rate with which vehicles reach their destinations (outflow) and (iii) the shape of the MFD is a property of the network itself (infrastructure and control) and not very sensitive to different demand patterns. Property (i) is important for modeling purposes as details in individual links are not needed to describe the congestion level of cities and its dynamics. Even though the flow-density plots for individual links exhibit considerable scatter, the scatter nearly disappears once individual loop detector data are spatially aggregated, and points lie along a well-defined curve. It can also be utilized to introduce simple control strategies to improve mobility in homogeneous city centers (see e.g. Daganzo, 2007; Geroliminis and Daganzo, 2007). The main logic of the strategies is that they try to decrease the inflow in regions with points in the decreasing part of an MFD. Property (ii) is important for monitoring purposes as flow can be easily observed with different types of sensors while outflow is more difficult. Property (iii) is important for control purposes as efficient active traffic management schemes can be developed without a detailed knowledge of O–D tables. For a review of other works related to MFD the reader could refer to Geroliminis and Sun (2011a).

Mathematically speaking, consider a region of a city, which traffic state is described by properties mentioned in the previous paragraph. Then the state of the system, $n(t)$, is governed by the mass conservation equation: $\frac{dn}{dt} = I'(t) - G(n(t))$, where n is the number of the vehicles in a road network, $I'(t)$ is the inflow to the network at time t, and G is the MFD as expressed by the total outflow from the system as a function of n. This equation simply explains that traffic systems are dynamic and to estimate the state of the system at time t, the knowledge of the input flow is not sufficient, but boundary conditions are needed, i.e. the state of the system at a prior time t'. Thus, a traffic model that estimates the average travel time based on a specific demand–cost curve ignores not only variations in the demand, but more importantly that this travel time will be different if the initial state of the system is in different traffic regimes.

Three traffic regimes can be observed for a unimodal-shape MFD, which are in conceptual agreement with the same regimes in the link fundamental diagram (lFD). For notation purposes, let's define I, II and III the MFD regimes and i, ii and iii the lFD regimes. Regime I represents under-saturated states where most of the links are in regime i of the lFD, i.e. queues are transient and the total number of vehicles served is smaller than the maximum possible. Regime II represents saturated states. Most of the links are in regime ii and filled part way with permanent queues and many traffic signals operate at capacity. There is a limit to vehicle accumulation corresponding to queues that fill the links. In regime ii, travel production (as expressed by vehicle–kilometers travelled per time unit) is approximately constant, but never larger than the quantity $\sum L \cdot g \cdot s$, where L is the link length, g is the duration of green phase and s the saturation flow of the signal of an intersection. Furthermore, the critical density is achieved during Regime II. The critical density is the density at which maximum outflow (or network capacity) is achieved and this is usually the objective for a successful traffic management strategy. In Regime III flow decreases with accumulation, corresponding to oversaturated states of the network and long queues or spillbacks are observed in many links, which results in states of regime iii for the lFD. These states cannot arise by increasing the input flow, but a restriction from downstream is necessary, for example if queues from downstream links block the departures during the green phase. Regime III consists of states where queues fill many links, vehicles are stopped or moving at less than saturation flows because of spillbacks. Congestion would be unevenly distributed over the network if states of individual link in regimes i and iii occurred simultaneously. This would create points beneath the curve, as we will show later. Nevertheless, to have a well-defined MFD, it does not mean that for the individual links should all belong in the same regime, as this is unrealistic. For a more detailed description of the distribution of link density of for low scatter MFDs, the reader can refer to Geroliminis and Sun (2011b). An example of the three regimes will be illustrated later.

As pointed out in Section 2, the agent-based simulator does not focus on modeling detailed disaggregated characteristics of traffic, like car-following, queue dynamics, etc. This is the main motivation for us to carry out the investigation to see if its outputs are consistent with the physics of traffic at an aggregated level, as expressed by a MFD. Apart from investigating the MFD, we also need to examine if spillback phenomena can be observed since this is the key reason of the existence of capacity decrease. It is known that a model with point queues (where vehicles take zero space), will produce a much slower (or non-existent) decreasing part of MFD (Regime III).

4. Methodology

4.1. Development of an MFD-controlled cordon pricing scheme

Assuming that the MFD and other traffic phenomena are well produced in the agent-based model, we would like to develop a macroscopic pricing scheme controlled by MFD. The idea, instead of using the traditional demand-supply curve, is to

determine a toll so that the targeted network operates at its saturated (maximum throughput) level or slightly less. These two levels correspond to the right hand side of Regime I and Regime II described in Section 3. In other words, once MFD traffic state drops to Regime III at time t and the traffic density of the network K_t exceeds the critical density K_{cr}, we apply a toll to shift agents' departure time in such a way so that the network state remains to Regimes I and II. Nevertheless, the center of a city experiences higher level of congestion than the periphery, so pricing should be applied in a smaller cordon for equity issues. The algorithmic steps of determining when, where and how a toll should be charged is the following:

(i) Define the initial pricing: cordon area, duration of tolling, amount of toll.
(ii) Obtain travel plans of agents including departure, mode, route choice, etc. and execute the plan and the traffic in MATSim until the agent-based equilibrium is achieved (After each iteration agents update their travel plans, the equilibrium will be achieved after 40–50 iterations.).
(iii) Obtain the resulting traffic performance and identify the duration of congested states by the MFD. Congested state is defined when network density K_t of the cordon area exceeds the critical network density K_{cr}, as illustrated by Fig. 2a.
(iv) Obtain the average network density of the congested states $\overline{K}_t$, which is the average of all the K_t in the congested states.
(v) Apply a proportional controller to update the toll by Eq. (1).
(vi) Start a new simulation and apply the updated toll in (v). Follow again from step (ii). This process will be repeated until there is no congested-states observed on the resulting MFD, as illustrated by Fig. 2b.

$$Toll_i = \max(0, Toll_{i-1} + c(\overline{K}_t - K_{cr})) \tag{1}$$

The proportional controller is a classic linear feedback control strategy. It is known in control theory that dynamical systems with well-defined properties (e.g. small errors in the state description) can be stabilized with a feedback strategy to a desired state by choosing an appropriate value of parameter c in Eq. (1). Detailed information on proportional controllers can be found in Ogata (2001). Basically the controller states that if K_t exceeds K_{cr}, an additional toll is charged and its value is proportional to the difference between K_t and K_{cr}. For practical consideration, time index t here refers to morning and evening peak and two different tolls are estimated accordingly. Parameter c is the constant proportion, which influences the rate of achieving the optimal toll. K_{cr} is a constant as well and estimated from the MFD directly, recalling that K_{cr} is a property of the network. These types of controllers have been widely-applied in freeway traffic flow management, e.g. the well-known ramp-metering algorithm ALINEA (Papageorgiou et al., 1991) utilizes it as the core control law to control flow entering from on-ramp to main freeway. Stability analysis has shown the global convergence of this law: For all choices of controller parameters the strategy preserves closed-loop stability and forces the actual K_t to reach the desired one (Kosmatopoulos and Papageorgiou, 2003).

The choice of the value of K_{cr} in Eq. (1) can be a policy decision. Some might argue that if K_{cr} belongs in Regime I the toll is very strict and over-charging might occur, as the system operates in a state less than its capacity. Nevertheless, this does not necessarily indicate over-charging, unless the total toll paid outweigh the total travel time savings. The reason is that the length of the toll period is shorter than the length of the congestion period without toll. This is because the network capacity, as expressed by the MFD, decreases for high values of network density and the system is not operating at the maximum network flow during congestion (this is not the case in the classical bottleneck problem of morning commute (Vickrey, 1969), where toll period is equal to the congestion period). Nevertheless, the system operator can choose the desired toll in a way to maximize the network flow and as a result to transfer the highest possible number of users. The value of toll which results in points in the MFD just without Regime III can be set as the lower bound, while the one which results the total toll paid just equal to total travel time savings can be set as the upper bound in order to avoid over-charging. This higher toll will push the system to operate at a smaller than the maximum outflow during the toll period, but with higher average speed and some potential savings in travel delay. This also implies that we should use a smaller value for K_{cr} in Eq. (1). For example, state for

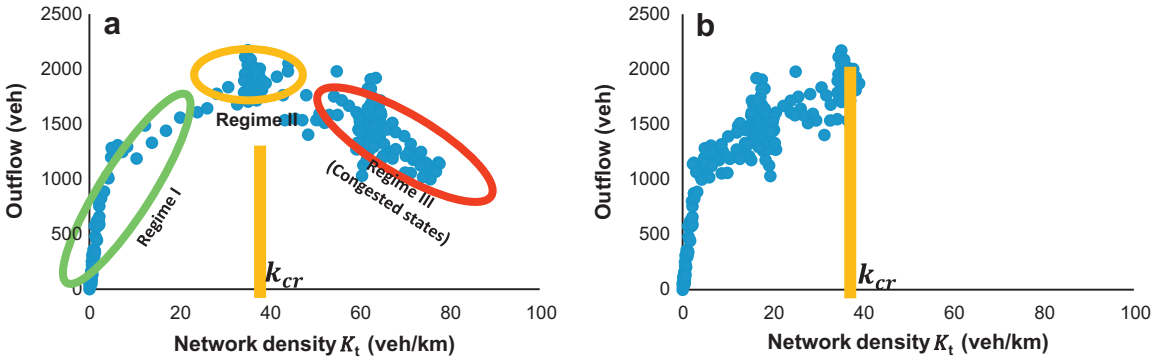

Fig. 2. An illustration of the MFD, density of network accumulated vehicles K_t vs. network outflow, critical density K_{cr} identified, congested states identified.

K_{cr} = 20 veh/km in Fig. 2 is a more reliable and less equitable state, because the average speed is higher (and more likely to be stable), but the system operates at a network outflow below the maximum. Thus, fewer people (presumably those with a higher value of time) pay the toll and travel in the rush hour. State for K_{cr} = 35 veh/km is more equitable (with a higher flow) but has a slower speed, so the total welfare may be smaller depending on the distribution of the value of time within the population. Also, state with smaller critical density is more reliable because given that users reaction in toll changes is not part of the model, a toll may not be efficient at all times, and allow the system to reach congested states with $K_t > K_{cr}$. In which state of the MFD a city should operate is a policy decision. Also, note that the city operates in the part of MFD with critical density. This means that most of the links during the peak hour operate in regime ii (capacity) and a few in regime i and iii. During the same period without pricing most of the links were in regime iii. While the objectives of existing operational pricing strategies are similar, the MFD gives the quantitative tools to meet these objectives and identify the efficiency and operability of a network.

4.2. Data derivation

To implement the macroscopic traffic analysis, one needs to obtain the traffic variables of space mean density (vehicle per kilometer) K_t and trip completion rate (outflow) A_t of a network for certain time intervals. K_t is the number of vehicles in a network, N_t, divided by the total length (lane–kilometers) of the network L at time t. Since the number of vehicles on one link equals to $k_i \cdot L_i$, K_t actually is a weighted space mean density and therefore reasonably reflects traffic states at an aggregated level. Trip completion rate (outflow) A_t is simply the sum of the number of finished trips a_i reaching destination at link i for the whole network. See Eq. (2).

$$K_t = \frac{N_t}{L} = \frac{(\sum_I (k_i \cdot l_i \cdot n_i))}{(\sum_I (l_i \cdot n_i))} \quad k_i = \frac{e_i - q_i}{l_i \cdot n_i}, \quad A_t = \sum_I a_i, \tag{2}$$

where $i \in I$ denotes an individual link i in the network; q_i is the number of vehicles leaving link i while e_i is the number of vehicles entering, k_i is the traffic density of link i, l_i is the length of link i and n_i is the number of lanes of link i. As demonstrated by Geroliminis and Daganzo (2008) there is a linear relation between A_t and space mean flow Q_t which can be calculated as K_t in Eq. (2) by using q_i instead of k_i. Then, one will also see that an MFD exists between Q_t and K_t. This can be observed later in Fig. 3.

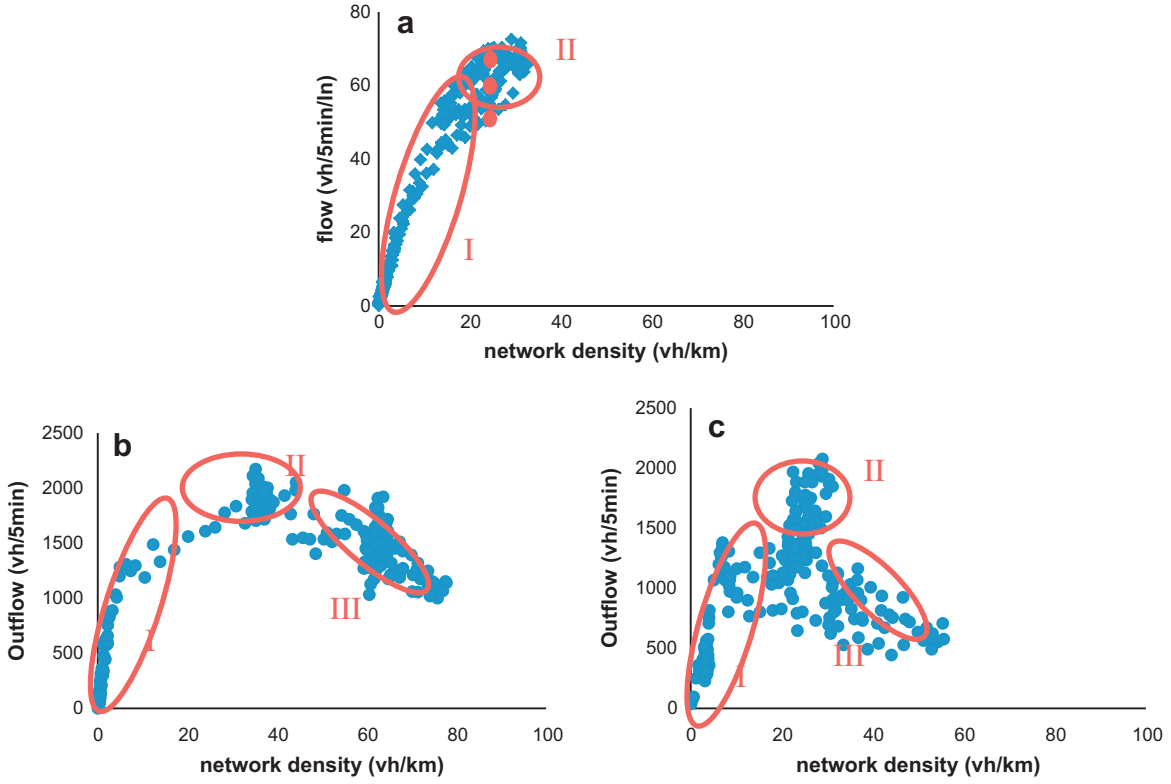

Fig. 3. (a) The MFD for a 4.5 km radius area (note that the value of *Y*-axis here is "flow" and different from the other figures) in Scenario 1; (b) the MFD of a 1 km radius area in Scenario 2; (c) the MFD of a 1 km radius area in Scenario 3.

4.3. Network partitioning

One of the main characteristics for a well-defined MFD is that the target network is homogeneously loaded and congestion is not unevenly distributed. For a big city with congestion heterogeneously distributed, Ji and Geroliminis (2011) developed an algorithm for partitioning the network into regions with well-defined MFDs. The motivation is to facilitate effective control for each region, as by aggregating congested and uncongested regions, the resulting MFD might exhibit high scatter and cannot identify the problematic region. We apply the same idea here and utilize a simpler filter to find a centered subnetwork where a cordon-based pricing can be easily implemented. To do this, we first calculate the density of every individual link of the network. We examine the distribution of these densities and zoom in on the region where congestion exists and the variance of link density (which is a proxy for heterogeneity) during peak periods is below a typical threshold.

5. Case studies and results

In this section, we investigate the applicability of the methodological framework with an agent-based simulation of Zurich urban road network and we present analyses for three different scenarios. Scenario 1 utilizes less population and a lower demand level (Mohit, 2010), compared to the other two scenarios. The resulting traffic condition is semi-congested. In Scenarios 2 and 3, higher demand is deployed (Meister et al., 2010), that caused significant level of congestion and states in the Regime III of the MFD. Since more agents are traveling the network, heavy congestion occurs in Scenarios 2 and 3 therefore we are able to see if pricing can be an efficient strategy to ameliorate the mobility in the network. To make congestion even heavier, lower values of link capacities are applied in Scenario 3. In all the scenarios only 25% of the population (agents) is used in the simulation. For a 100% population scenario, which has 7 million agents/day, one simulation run (40–50 iterations) requires a high performance computer 4 days to finish. While for a 25% population scenario it requires 1 day, including the optimization of tolls. By comparing two runs with 25% and 100% agents, the macroscopic variables of the simulation have almost identical values. For the sake of computational cost, we utilize the 25% population scenarios.

5.1. Investigation of traffic phenomena in the agent-based model

Before implementing the cordon-pricing strategy of Section 4.1, we investigate if a Macroscopic Fundamental Diagram exists for the study site and how the agent based simulator can reproduce traffic phenomena, like spillbacks and queue propagation.

5.1.1. The Macroscopic Fundamental Diagram

Density, flow and outflow are estimated according to Section 4.2 at 5 min resolution. For Scenario 1, The MFD for a 4.5 km-radius area of the center of city Zurich is shown in Fig. 3a. States in Regime I and in the beginning of Regime II are observed, indicating that the network on a macroscopic level is not heavily congested. Even if we focus on the center of the city that attracts most of the trips, the MFD obtained after filtering the network to a 1.5 km area around the center of the city does not contain more congested states. To observe a complete MFD with congested states, we look at Scenarios 2 and 3 where more agents are employed for simulations and consequently more traffic is generated. The MFDs are shown in Fig. 3b and c for the city center with a radius of 1 km (about 1000 links). We observe that (i) comparing to Scenario 1, Scenarios 2 and 3 have a complete Regime II in which network is operated at its capacity and (ii) congestion Regime III exists, where with higher density of agents in the network, the number of agents, which reach their destinations decreases. Note that the shape and scatter of the MFD is different in Fig. 3c. This is because in Scenario 3 smaller link capacities are utlized to create additional congestion. Therefore the dynamics of traffic are more complex and this results in higher scatter.

5.1.2. The fundamental diagram (FD) and spillback effects

We now provide a detailed investigation on how congestion propagates from one link to another. Flow rate q_i and density k_i are calculated for each link i. We study if spillback effects are present and long queues can decrease the output of upstream links. Fig. 4 shows the lFDs for four consecutive links along one of the arterial roads of Zurich in Scenario 1. The color[1] of the points in the figure indicates the corresponding link. Note that at a density of 40 veh/km, the flow rate reaches its maximum and remains the same value until a density of 150 veh/km. Then flow decreases as density increases, indicating that congestion happens at the latter point. Secondly, we see that congestion spills back from downstream to upstream, as the links in green and yellow operate in regime i and ii while the links upstream experience states in regime iii. These observations make clear that queues are growing from downstream to upstream and blocking effects are present in the simulator. We also observe that queues propagate from downstream to upstream, as congestion appears upstream at a later time. Note also that the individual links fundamental diagrams exhibit higher scatter, especially in the congested regime.

[1] For interpretation of color in Figs. 1–10, the reader is referred to the web version of this article.

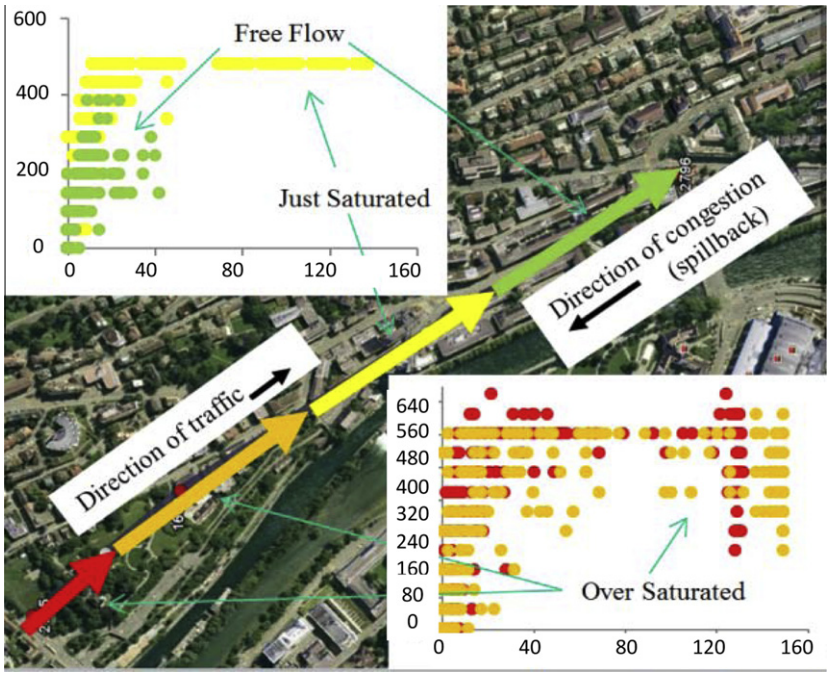

Fig. 4. An illustration of congestion propagation along consecutive links. Axis *x* is the density of the road and axis *y* is the flow rates. Note that the scenario utilized 25% of the population.

5.1.3. An explanation of scatter in the MFD

Now let us have a more careful look at Scenario 1 to provide an explanation for the scatter in the shape of its MFD. By zooming in Fig. 3a, we observe hysteresis loops in the (A_t, K_t) trajectory. Similar phenomena have been observed by Geroliminis and Sun (2011a,b) for freeway networks: higher network flows are observed for the same average network density in the onset and lower in the offset of congestion. This is because there are differences in the spatial distributions of congestion, and the variance among link densities, for the same level of average network density for different times. Fig. 5 shows a contour plot of link densities of all links in the network at three different times when the network holds the same amount of vehicles. At each link location the density value is shown as the diameter of a circle. The chosen times are 6h35, 9h15 and 11h40, correspondingly the highest point, the lowest point and the medium point in Fig. 3a. The time points refer to the onset of the morning peak from 6h30 to 9h30, and the offset from 9h30 to 11h30. In the figure, the size of the "bubble" is the value of density while *X* and *Y* axes are the coordinates of links. Note that traffic is more uniformly distributed at 6h35 than at 11h40, while at 9h15 densities are extremely high at some locations. The standard deviations of link density are 24, 29 and 32 veh/km respectively, while flow rates are 840 veh/h, 720 veh/h and 600 veh/h. In the next section, we show that even if the MFDs of Fig. 3 exhibit some scatter, the cordon-pricing schemes can significantly improve the mobility patterns.

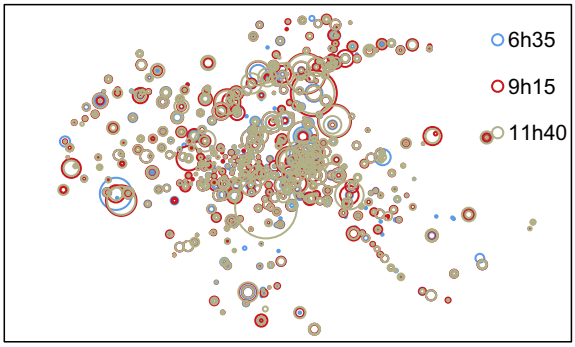

Fig. 5. Link density contour plots of the network at time 6h35, 9h15 and 11h40. Axis represent coordinates of links (*x*, *y*-axis represent coordinates).

5.2. Analysis of pricing

5.2.1. A cordon-based pricing controlled by MFD

The analysis in the previous section shows that the results of MATSim are consistent with previous studies with traffic simulation data or field experiments for MFDs. We now test the effectiveness of the MFD-based pricing scheme proposed in Section 4.1. The data of Scenario 2 are used, which experiences heavy congestion and scatter. In this way, we expect to provide a stricter test for the effectiveness/robustness of a cordon-based pricing scheme, given that our macroscopic approach is less accurate. The targeted cordon area is a circle of 1 km-radius, the area inside the red solid ring in Fig. 6. Agents who cross the border of the area, the red line, will pay a toll. Before we apply any toll on the cordon, we run one simulation until it achieves equilibrium. Network density over time is plotted in Fig. 7a, while the resulting MFD is plotted in Fig. 7b. Given the "no pricing" curve in these two figures, the following information is calculated, which is an input to Equation 2: the critical density K_{cr} (28 veh/km) and the periods for charging a toll (from 7:30 am to 9 am and from 4 pm to 8 pm).

To identify the optimal toll values, we apply the algorithm of Eq. (1). An initial toll of 1€ for the morning while 4€ for the evening is applied in the first trial. A value of 1 euro/density for the regulator parameter c is used. The optimal toll is achieved after four updates: a 2€ toll is charged for the morning peak, while 10€ toll is charged between 18:30 and 19:30 and an 8€ is charged for the rest part of the evening peak). We see in Fig. 7a that traffic congestion drops as the amount of toll increases; and in Fig. 7b, congestion states in Regime III disappear. Thus this aggregated approach for pricing, which does not consider individual link behavior, produces the desired results to reduce congestion and identifies the appropriate value of pricing tolls to meet the desired mobility goals. Furthermore, we look at the same graphs for the area outside the charging zone as defined in Fig. 6 (1.5 km distance from the cordon line, concentric). The motivation is to check if the traffic conditions in the periphery become worse. Fig. 7c shows that the density of the neighbor area slightly increases. The explanation is some agents who travelled through the cordon area now choose to detour within the neighbor area in order to avoid the toll. But if we look at the MFD for this area, which is shown in Fig. 7d, the entire area is operating in Regime I and II. Traffic volumes on the links in the cordon area are illustrated in Fig. 8, in which red colors represent congested links and green color uncongested or at capacity. By applying a toll, traffic decreases during peak hours and travel conditions are improved. We now provide a quantitative analysis to estimate delay savings after the implementation of tolls.

From an economic point of view, we look at travel time savings (as expressed by vehicle hours travelled) and total toll costs. Total travel time savings are estimated as the reduced travel time of the cordon area minus the increased travel time of the periphery. Results are summarized in Table 1. Considering the average value of travel time savings (VTTS) of the agents is 15€/h (Axhausen et al. (2007)), the total savings in the study network are larger than the total toll paid, by an amount of about 20%. Note that most congestion charging methodologies charge a toll equal to the delay cost, while in our case savings are significantly higher. This is a promising result at an aggregated level. Furthermore, we also investigate the savings at an individual user level, which is calculated as travel time savings per km travelled per trip. We see that the effectiveness ratio between travel savings per km travelled per trip inside cordon (positive) and outside cordon (negative) is 7, much higher than the same ratio at the aggregated level (in veh-hours), which is around 1.5. This indicates that a low amount of additional delay (0.13 min per km per trip) which generates almost no impact for the outside cordon, creates significant savings inside the cordon (almost 1 min per km per trip). Zhang and Levinson (2005) among the others argue that the value-of-time should be a non-linear function of travel delay, i.e. the effect of a longer delay (in €/min) is much more significant than the effect of a short delay. Furthermore, comparing the Zurich results to the cordon pricing case of London, the improvement in travel time savings in London is 0.7 min per km per agent (information from Transport for London, http://www.tfl.gov.uk), which shows that our pricing scheme is more effective.

Fig. 6. The targeted cordon area (area insider the solid line) and the neighbor area (between the solid line and the dotted line), the city of Zurich.

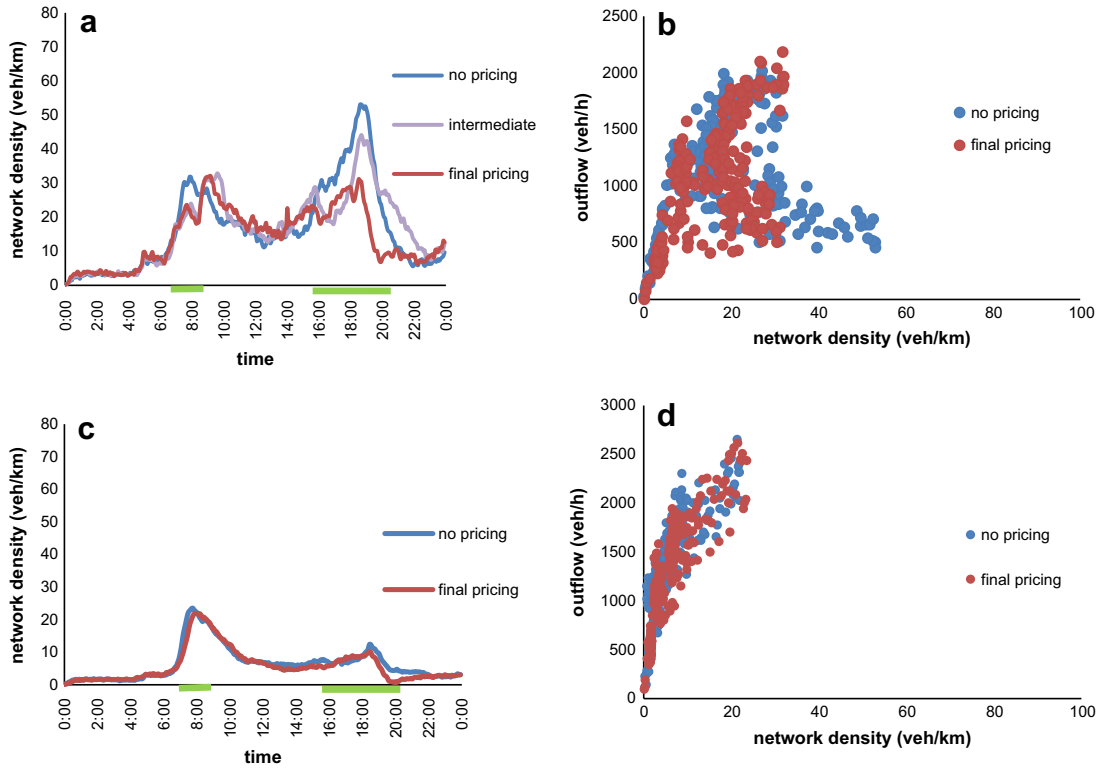

Fig. 7. For Cordon area (a) network density time series before and after final pricing; (b) the MFDs before and after the final update of pricing; for outside Cordon (c) density time series before and after the final update of pricing; (d) the MFDs before and after the final update of pricing. The green bars indicate toll period.

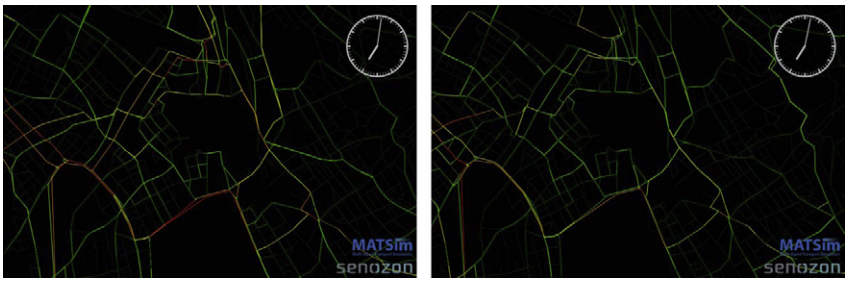

Fig. 8. Link traffic volumes in the no-toll scenario (left) and the 4th-toll scenario (right) at 19 pm.

Table 1
Summary of the social of the proposed pricing.

	TTT savings (in cordon)	TTT savings (Out cordon)	Effectiveness ratio
Aggregated social gain	6356 veh-h	−4541 veh h	1.5
	Travel savings per km travelled per trip (in cordon)	Travel savings per km travelled per trip (out cordon)	Effectiveness ratio
Disaggregated social gain	0.94 min	−0.13 min	7

5.2.2. Further investigation of behavioral shifts

In this part, we analyze the impact of pricing on shifting the time of departure of agent trips. One direct impact is that agents change the time they pass the cordon line. Fig. 9 shows a time series of the flow of agents crossing the cordon line with and without pricing. It is clear that some amount of agents avoid passing the cordon line during the tolling-period,

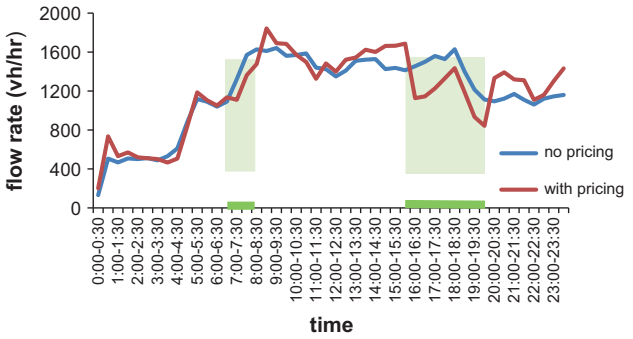

Fig. 9. Time series of the flow of agents passing the cordon.

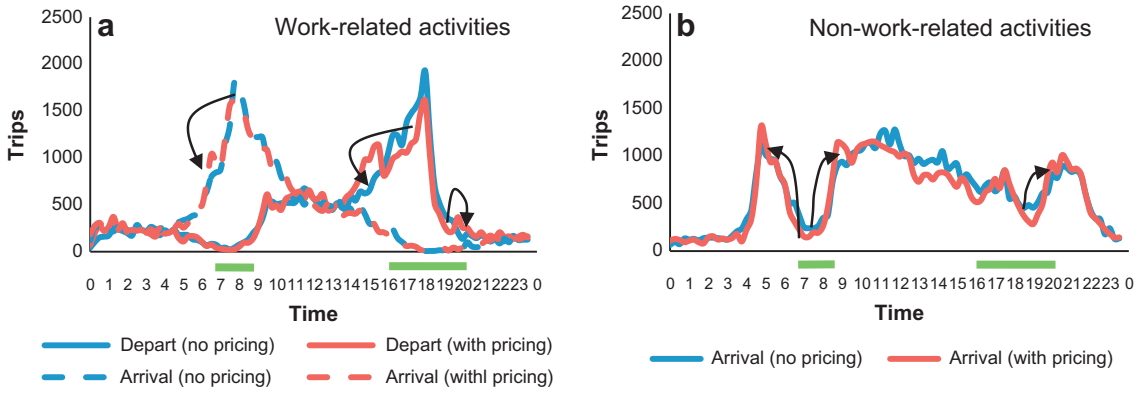

Fig. 10. Time shift of WA (a) and NWA (b) in the no pricing and the final pricing scenarios.

which is shown in green color. In the morning peak, agents tend to switch to later times while in the evening peak agents change their crossing times in both sides of the toll period. We now see which type of agents and to what extent agents change their behavior because of tolls.

In the current version of MATSim, two types of heterogeneity exist among agents: (i) a random error in the utility function as this is expressed by an extreme value distribution of the logit model and (ii) a consistent heterogeneity when agents perform different activities. This heterogeneity is determined by the duration of the performed activity. Given this information, we can identify the impact of pricing on different activities. We classify activities into two groups: purpose of going-to-work and leaving-from-work, as work-related activities (WA); and purpose of going-to-leisure as non-work-related activities (NWA). Fig. 10 shows the comparison of time shift of performing WA and NWA trips in the "no pricing" and the "with pricing" scenarios. The green bars mark the toll periods. Both groups of agents experience a time shift. Agents performing WA tend to switch the starting time of their trip to an earlier time, while agents performing NWA tend to switch to both earlier and later times. The explanation is that for work-related activities the penalty of earliness is lower than this of lateness. About 5% of WA trips shift during the morning toll period while the amount for NWA is 15.7%. During the evening toll period, 16.1% of WA and 19.5% of NWA shift their departure time. We can observe that WA are much less sensitive to small value of toll, as adjusting WA is less flexible than NWA. For higher tolls, the impacts are more substantial. An additional explanation for the observation that the morning shift is less than evening shift, is that starting time of the activities is fixed in the morning, therefore agents have to be on time to avoid a large penalty of not performing their trip (do nothing); while in the evening there is more flexibility. Given the optimal pricing, the number of agents who perform NWA by car outside toll period reduces, which is due to mode change. A deeper investigation on behavioral shift and equity can be done, if individual data such as income or value of time of agents are available, similar to Axhausen et al. (2007).

6. Conclusions

In this work, we introduce a new approach to develop a cordon-based congestion pricing scheme. The idea is to utilize a macroscopic traffic model, the Macroscopic Fundamental Diagram, to determine optimal tolls for an urban network. The methodology was evaluated with an agent-based model for the city of Zurich. We first investigate the feasibility of combining an agent-based model with an MFD network representation by examining the outputs of three simulation scenarios on

the urban road network of Zurich. We find that the outputs of the agent-based model are consistent with the physics of traffic. On the link level, the agent-based model is able to reflect spill-back effects once congestion is formed. On the macroscopic level, it shows the existence of an MFD linking network trip completion rate and network density. In addition, the cause of some scatter in the MFD can be explained by the uneven distribution of congestion, results which are consistent with empirical observations.

Given these findings, we test the proposed cordon-based pricing in the city of Zurich, utilizing a controller to update tolls of a 1 km-radius cordon area. By an iterative process (in total of four updates), congestion represented by Regime III on the MFD in the area, is eliminated. Meanwhile, there is almost negligible negative impact on the neighbor area outside the cordon area (between 1 km- and 2.5 km-radiuses). Total travel time savings by imposing the toll outweighs the total toll paid plus the increased travel time for the outside cordon area. The savings of time per distance travelled outweighs significantly the extra cost per distance travelled for the users that avoid traveling in the charging zone. Furthermore we examine the behavioral shift on performing work and non-work activities. Pricing has stronger impact on non-work-related activities, as more agents tend to change time plans to avoid tolls. Work trips also show time shift but are more stable even for higher toll, as expected.

We would conclude that the proposed pricing scheme is effective both from a traffic and an economic point of view. Determining tolls by MFD is a promising approach for congestion pricing. Following this work, future studies are recommended on: (i) how other macroscopic pricing schemes can improve mobility patterns, such as area-based pricing or distance based pricing, (ii) equity impacts, provided with the necessary data, e.g. income, car ownership, (iii) investigate the effectiveness of a time-dependent toll and a value-of-time-dependent toll. The fast convergence of the agent-based model in identifying the optimal toll (4 updates) shows that this methodology could be directly applied by cities without the need of time-consuming and expensive surveys on a large part of the population or tedious simulation experiments, if data from loop detectors or GPS are available. Nevertheless, survey information can be useful to create more equitable schemes.

Acknowledgement

This research was partially funded by Swiss National Science Foundation Grant (SNSF) of EPFL #200021_132501.

References

Anderson, D., Mohring, H., 1997. Congestion Costs and Congestion Pricing. The Full Costs and Benefits of Transportation: Contributions to Theory, Method and, Measurement.
Arnott, R., 2007. Congestion tolling with agglomeration externalities. Journal of Urban Economics 62 (2), 187–203.
Arnott, R., Inci, E., 2010. The stability of downtown parking and traffic congestion. Journal of Urban Economics 68 (3), 260–276.
Arnott, R., de Palma, A., Lindsey, R., 1988. Schedule delay and departure time decisions with heterogeneous commuters. Transportation Research Record: The Journal of Transport Research Board (1197), 56–67.
Arnott, R., de Palma, A., Lindsey, R., 1990. Departure time and route choice for the morning commute. Transportation Research Part B: Methodological 24 (3), 209–228.
Axhausen, K., 2008. Accessibility long TERM perspectives. Journal of Transport and Land Use 1 (2), 5–22.
Axhausen, K., Hess, S., Koenig, A., Bates, J., Bierlaire, M., Abay, G., 2007. State-of-the-art estimates of swiss value of travel time savings. In: The 86th Annual Meeting of Transportation Research Board, Washington, DC.
Daganzo, C., 2007. Urban gridlock: macroscopic modeling and mitigation approaches. Transportation Research Part B: Methodological 41 (1), 42–69.
Davidsson, P., Henesey, L., Ramstedt, L., Törnquist, J., Wernstedt, F., 2005. An analysis of agentbased approaches to transport logistics. Transportation Research Part C: Emerging Technologies 13 (4), 255–271.
De Palma, A., Lindsey, R., 2006. Modelling and evaluation of road pricing in Paris. Transport Police: The Official Journal of the World Conference on Transport Research Society (WCTRS) 13 (2), 115–126.
De Palma, A., Kilani, M., Lindsey, R., 2005. Congestion pricing on a road network: a study using the dynamic equilibrium simulator METROPOLIS. Transportation Research Part A: Policy and Practice 39 (7), 588–611.
Geroliminis, N., Daganzo, C., 2007. Macroscopic modeling of traffic in cities. In: The 86th Annual Meeting Transportation Research Board, Washington, DC.
Geroliminis, N., Daganzo, C., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transportation Research Part B: Methodological 42 (9), 759–770.
Geroliminis, N., Levinson, D., 2009. Cordon pricing consistent with the physics of overcrowding. Transportation and Traffic Theory, 219–240.
Geroliminis, N., Sun, J., 2011a. Properties of a well-defined macroscopic fundamental diagram for urban traffic. Transportation Research Part B: Methodological 45 (3), 605–617.
Geroliminis, N., Sun, J., 2011b. Hysteresis phenomena of a macroscopic fundamental diagram in freeway networks. Transportation and Traffic Theory, 213–228, and (2011). Transportation Research Part A: Policy and Practice 45(9), 966–979.
Godfrey, J., 1969. The mechanism of a road network. Journal of Traffic Engineering and Control 11 (7), 323–327.
Helbing, D., Treiber, M., Kesting, A., Schonhof, M., 2009. Theoretical vs. empirical classification and prediction of congested traffic states. European Physical Journal B 69 (4), 583–598.
Ji, Y., Geroliminis, N., 2011. Spatial and temporal analysis of congestion in urban transportation networks. In: The 90th Annual Meeting Transportation Research Board, Washington, DC.
Kosmatopoulos, E., Papageorgiou, M., 2003. Stability analysis of the freeway ramp metering control strategy ALINEA. In: The Proceeding of the 11th IEEE Mediterranean Conference on Control and Automation, Rhodes.
Löchl, M., Axhausen, K., 2010. Modeling hedonic rents for land use and transport simulation while considering spatial effects. Journal of Transport and Land Use 3 (2), 39–63.
Maruyama, T., Sumalee, A., 2007. Efficiency and equity comparison of cordon- and cordon-based road pricing schemes using a trip-chain equilibrium model. Transport Research Part A: Policy and Practice 41 (7), 655–671.
May, A., Mline, D., 2000. Effects of alternative road pricing systems on network performance. Transport Research Part A: Policy and Practice 34 (6), 407–436.
Meister, K., Balmer, M., Ciari, F., Horni, A., Rieser, M., Waraich, R., Axhausen, K., 2010. Large-scale agent-based travel demand optimization applied to Switzerland, including mode choice. In: The 12th World Conference on Transportation Research, Lisbon.
Merchand, M., 1968. A note on optimal tolls in an imperfect environment. Econometrica 36, 575–581.

Mohit, S., 2010. Activity-based Travel Demand Modelling including Freight and Cross-border Traffic with Transit Simulation, Project Report, IVT of ETH Zürich, Zürich.
Munoz, P., Daganzo, C., 2003. Structure of the transition zone behind freeway queues. Transportation Science 37 (3), 312–329.
Nagel, K., Flötteröd, G., 2009. Agent-based traffic assignment: going from trips to behavioral travelers. In: Paper Presented at the 12th International Conference on Travel Behaviour Research (IATBR), Jaipur.
Ogata, M., 2001. Modern Control Engineering, fourth ed., pp. 24–27.
Papageorgiou, M., Hadj-Salem, H., Blosseville, J., 1991. ALINEA: a local feedback control law for on-ramp metering. Transportation Research Record: Journal of Transportation Research Board 1320, 58–64.
Pigou, A., 1920. Wealth and Welfare, first ed.
Seik, F., 2000. An advanced demand management instrument in urban transport: electronic road pricing in Singapore. Cities: The International Journal of Urban Policy and Planning 17 (1), 33–45.
Small, K., Chu, X., 2003. Hypercongestion. Journal of Transport Economics and Policy 37 (3), 319–352.
Small, K., Yan, J., 2001. The value of "value pricing" of roads: second-best pricing and product differentiation. Journal of Urban Economics 49 (2), 310–336.
Verhoef, E., 2002. Second-best congestion pricing in general networks: heuristic algorithms for finding second-best optimal toll levels and toll points. Transport Research Part B: Methodological 36 (8), 707–729.
Vertic, M., Schuessler, N., Erath, A., Axhausen, K., 2010. The impacts of road pricing on route and mode choice behaviour. Journal of Choice Modelling 3 (1), 109–126.
Vickrey, W., 1963. Pricing and resource allocation in transportation and public utilities. The American Economic Review 72 (3), 452–465.
Vickrey, W., 1969. Congestion theory and transport investment. The American Economic Review 59 (2), 251–260.
Waraich, R., Galus, M., Dobler, C., Balmer, M., Andersson, G., Axhausen K., 2009a. Plug-in hybrid electric vehicles and smart grid: investigations based on a micro-simulation. In: Paper Presented at the 12th International Conference on Travel Behaviour Research (IATBR), Jaipur.
Waraich, R., Charypar, D., Balmer, M., Axhausen, K., 2009b. Performance improvements for large scale traffic simulation in MATSim. In: Paper Presented at the 9th Swiss Transport Research Conference, Ascona.
Wardrop, J., 1952. Some theoretical aspects of road traffic research. Proceedings of the Institute of Civil Engineers 1 (3), 325–362.
Yang, H., Huang, H., 1998. Principle of marginal-cost pricing: how does it work in a general network. Transportation Research Part A: Policy and Practice 32 (1), 45–54.
Yang, H., Huang, H., 2005. Mathematical and Economic Theory of Road Pricing. Elsevier Science Inc., New York.
Zhang, L., Levinson, D., 2005. Balancing efficiency and equity of ramp meters. ACSE Journal of Transportation Engineering 131 (6), 447–481.
Zhang, L., Levinson, D., Zhu, J., 2008. Agent-based model of price competition, capacity choice and product differentiation on congested networks. Journal of Transport Economics and Policy 42 (3), 435–461.