# Marginal cost congestion pricing based on the network fundamental diagram

**Marginal cost congestion pricing based on the Network Fundamental Diagram**

M.D.Simoni[a,c], A.J. Pel[a], R.A.Waraich[b], S.P.Hoogendoorn[a]

[a] Department of Transport and Planning, Delft University of Technology, the Netherlands
[b] Institute for Transport Planning and Systems, Swiss Federal Institute of Technology Zurich, Switzerland
[c] Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, USA

**Abstract:** Congestion pricing schemes have been traditionally derived based on analytical representations of travel demand and traffic flows, such as in bottleneck models. A major limitation of these models, especially when applied to urban networks, is the inconsistency with traffic dynamics and related phenomena such as hysteresis and the capacity drop. In this study we propose a new method to derive time-varying tolling schemes using the concept of the Network Fundamental Diagram (NFD). The adopted method is based on marginal cost pricing, while it also enables to account realistically for the dynamics of large and heterogeneous traffic networks. We derive two alternative cordon tolls using network-aggregated traffic flow conditions: a step toll that neglects the spatial distribution of traffic by simply associating the marginal costs of any decrease in production within the NFD to the surplus of traffic; and a step toll that explicitly accounts for how network performance is also influenced by the spatial variance in a 3D-NFD. This pricing framework is implemented in the agent-based simulation model MATSim and applied to a case study of the city of Zurich. The tolling schemes are compared with a uniform toll, and they highlight how the inhomogeneous distribution of traffic may compromise the effectiveness of cordon tolls.

**Keywords:** *Network fundamental diagram, Spatial spread of congestion, Marginal cost congestion pricing, Time-varying cordon toll, Agent-based transport simulation*

## 1. Introduction

The NFD extends the concept of the Fundamental Diagram (FD) and it expresses the aggregated flow along all the links in the network (production or flow) and the total number of vehicles in the network (accumulation or density) by means of a concave function. Like in the FD, the free-flow regime and congested regime can be identified respectively on the left branch and right branch of the diagram, while the region characterized by slower increase of production until capacity is reached is typically referred to as capacity regime.

Although the concept of network traffic relations can be traced back to the 1960's (Smeed, 1966; Wardrop, 1968; Godfrey, 1969) and further developed during the 1970s and 1980s (Zahavi, 1972; Mahmassani et al., 1984), only recently the existence of an invariant macroscopic relation between network average flow, average density and average speed has been confirmed and formalized by Daganzo (2007) and Geroliminis and Daganzo (2008). The NFD (also referred to as Macroscopic Fundamental Diagram) considerably eases the understanding of complex traffic phenomena and the implementation of effective traffic management measures. For this reason the body of literature on theoretical insights and applications of the NFD has been growing rapidly.

Perhaps the most important theoretical finding consists of the influence of inhomogeneous conditions of traffic on the performance of the network. It is claimed in these studies that a well-defined NFD applies under specific "regularity conditions" concerning the homogeneity of links and the possibility of reaching the Wardrop Equilibrium. On the contrary, as a result of the uneven distribution of congestion the NFD shows scatter and hysteresis loops occur in the NFD diagram (Buisson and Ladier, 2009; Mazloumian et al., 2010; Geroliminis and Sun, 2011; Gayah et al., 2011; Saberi and Mahmassani, 2012). With regard to this issue, a recent contribution came from Knoop and Hoogendoorn (2013) who quantified the effect of spatial distribution (also termed "spread") of congestion on the performance of a freeway network by means of a generalized NFD (gNFD).

At an applied level, other studies have utilized the concept of NFD to derive efficient strategies for travel demand and traffic control. For the topic of this paper, the most relevant applications are on the development of perimeter control-gating measures (Keyvan-Ekbatani et al., 2012; Geroliminis et al., 2013; Keyvan-Ekbatani et al., 2013), motorway management (Chow, 2015) and congestion pricing schemes (Geroliminis and Levinson, 2009; Zheng et al., 2012; Gonzales and Daganzo, 2012). In particular, the application of a macroscopic representation of network traffic conditions to the design of congestion pricing models

represents a valuable approach to overcome some limitations of the traditional analytical congestion pricing schemes. Above all, the description of the supply curve (representing the cost related to the traffic volume) as a function of the network demand is consistent with the dynamic properties of traffic that are characterized by a drop of traffic throughput when the flow exceeds the capacity.

In this paper we address the design of cordon-based congestion pricing schemes based on macroscopic traffic variables. Above all, we discuss how the NFD and its generalizations can in various ways be used to compute fixed and variable congestion tolls that are consistent with the economical theory of marginal cost pricing. To this end we first discuss the way in which the spatial distribution of congestion influences the performance of a large urban network, here tested by means of agent-based simulations. Based on these findings, we propose two new methods to derive time-varying tolling schemes based on the NFD and its extension, a three dimensional NFD (3D-NFD) accounting for the effect of spatial spread. The first step toll depends on the marginal cost of a surplus of traffic inside the cordon area associated to the aggregated delay identified on the NFD. The second step toll avoids tolling drivers for travel time delays that are due to (uncontrolled) increases in the spatial spread of accumulation inside the cordon area. For the latter, we derive a 3D-NFD and fit a polynomial plane to quantify the extent of decrease of production determined respectively by the increase of accumulation and the deviation of spread from its natural increase. The two tolling schemes are finally evaluated on traffic flow performance indicators and compared with a uniform toll that operates the system at capacity by means of an offline iterative control.

In the first part of this study we will analyze the macroscopic traffic characteristics of the city of Zurich with the agent-based simulator MATSim (www.matsim.org), a state-of-the-art multi-agent model developed jointly by ETH Zurich and TU Berlin. In order to verify the appropriateness of using an agent-based approach to study, the main physical properties of traffic are tested. Since the boundaries of cordon-tolls are naturally defined by the existing constraints of the road network (ring roads or bridges) rather than by network partitioning techniques (Ji and Geroliminis, 2012), the area investigated will be likely characterized by heterogeneous conditions. For this reason, before designing the tolling schemes we will investigate dynamic features of the network-wide traffic, including the instability due to the presence of clusters of congestion, its relationship with scatter in the NFD and the hysteresis phenomena. In particular, we seek for additional evidence that even for low values of density, the traffic performance is compromised by unevenly distributed traffic. Based on these analyses we will investigate the relation between accumulation and spatial spread of traffic and finally derive a 3D-NFD that includes the deviation from the natural increase of spread as an additional dimension in order to consider the unstable conditions of the traditional aggregation-production relation. Along the same lines of Knoop and Hoogendoorn (2013), this approach is intended to represent an approach to deal with large heterogeneous networks, as alternative to the practice of network partition.

Subsequently we apply our findings about macroscopic properties of the (heterogeneous) network to derive two time-varying congestion-pricing schemes within MATSim. The following study aims at extending the approach by Zheng et al. (2012), who developed a uniform toll (Flat Toll) controlled by the NFD through an "offline" feedback control process, by introducing an analytical derivation of the levels of charge based on the marginal cost of surplus of traffic in the cordon. The main rationale is to design a methodologically sound and tractable model consistent with both the economic (Pigouvian tax) and engineering (network-wide macroscopic modeling) theories. Hence, we propose two different cordon-based tolls: a time-varying toll that changes in discrete time-intervals (Step Toll); and a time-varying toll that explicitly accounts for the property of spatial distribution of congestion (Hybrid Toll). The effects of these two schemes are finally compared to those derived from the Flat Toll by means of a series of performance indicators.

The agent-based model MATSim has been adopted in this study because it allows high levels of realism of the pricing model in terms of users' heterogeneous route, mode and departure time decisions in large-scale complex road networks with several thousand agents. Furthermore, thanks to its high level of disaggregation it is possible to investigate more in depth issues such as distributional impacts of congestion pricing schemes (Kickhofer et al., 2011). Additional applications of MATSim are described at www.matsim.org.

This paper is organized as follows. Section 2 describes the derivation of aggregated traffic flow properties in the urban road network of Zurich by means of the agent-based model MATSim. In Section 3 the effects of spatial distribution of congestion and the relation between accumulation and spatial spread of density are investigated. Sections 4 and 5 describe the design of three alternative congestion-pricing schemes controlled by the NFD and 3D-NFD and analyze their impacts on traffic conditions. In Section 6 we summarize our findings and discuss the main implications for practice and further research.

## 2. Network-wide traffic properties of Zurich city centre

This section provides an explanation of the main variables of the NFD in Section 2.1 and it continues with a description of the simulation scenario in Section 2.2.

### 2.1 Macroscopic traffic relations

The NFD can be estimated by means of different methods, both analytical and experimental. In this study we derive traffic accumulation (space mean density) and traffic production (outflow or exit rates) derived from the outputs of the mesoscopic simulator JDEQSim (see Section 3.1 for further description of the traffic model). First, the average density $k_j$ and average flow $q_j$ for each link are computed over intervals of typically 15 seconds and aggregated into larger intervals of 5 minutes. The average density $k_j$ for a single link is derived as:

$$k_j = k_{j-1} + \Delta k_j \quad (1)$$

Where $k_{j-1}$ corresponds to the density derived at time*j-1* and $\Delta k_j$ is the change of density that occurs during the time interval between *j-1* and *j*. Such a variation is calculated with the following formula that explicitly recalls the Cell Transmission Model (Daganzo, 1994):

$$\Delta k_j = \frac{e_j - o_j}{l \cdot n} \quad (2)$$

Where $e_j$ indicates the number of vehicles that has entered the link,$o_j$ indicates the number of vehicles that left the link, $l$ corresponds to the length of the link, and $n$ to the number of lanes of the link.

The outflow $q_j$ is simply estimated as the rate of vehicles leaving the link during the time interval between *j-1* and *j* (15 seconds), and estimated from $o_j$.

The average speed for single links is calculated by means of the well-known relationship of traffic flow theory with density and outflow:

$$u_j = \frac{q_j}{k_j} \quad (3)$$

This method holds only in case of space mean speed (average speed over a length of roadway), since using time mean speed (e.g. derived from loop detectors) would lead to systematic bias (Knoop et al., 2009). The fact that speed is derived from the outflow might determine little imprecision when during a certain interval a vehicle drives into the link, but it exits during the following one. However, it should be noted that, this problem entails particularly long links (over the hundred meters) that are a minority. Furthermore, the problem is resolved by aggregating the measurements in larger intervals.

Finally, the average density and the average outflow of the network composed by *n* links can be determined as the following weighted averages of the *i* individual link values (Zheng et al.; 2012):

$$K = \frac{\sum_i^n (k_i \cdot l_i \cdot n_i)}{\sum_i^n (l_i \cdot n_i)} \quad (4)$$

$$Q = \frac{\sum_i^n (q_i \cdot l_i \cdot n_i)}{\sum_i^n (l_i \cdot n_i)} \quad (5)$$

The average network speed U can be computed as Edie (1965) assuming that the traffic flow fundamental identity Q=K·U holds a with reasonable approximation:

$$U = \frac{Q}{K} \quad (6)$$

It has emerged from Section 2 that the deviation in density of the different links through the network plays also an important role in the shape of the NFD and it is considered as a cause of scatter by several scholars (Buisson and Ladier, 2009; Knoop and Hoogendoorn, 2013). Hence, the additional traffic variable called spatial spread of density, representing the "distribution" of congestion inside the cordon is introduced. Similarly to Knoop and Hoogendoorn (2013), the spread of density is estimated as the square root of the weighted variance of densities in all sections:

$$\gamma = \sqrt{\frac{\sum_i l_i \cdot (k_i - K)^2}{\sum_i l_i}} \quad (7)$$

Where $l_i$ corresponds to the length of the link $i$, $k_i$ to the density of the link $l$, and $K$ to the average density of the network.

## 2.2 Simulation scenario

The simulation scenario (from Meister et al., 2010) consists of an area of 30 km around the city of Zurich (Greater Zurich Area). Agents residing outside the study area, but entering at some time during the day are also included in the simulation (Waraich and Axhausen, 2012).

The road network used in the simulation consists of a high-resolution navigation network including about 1,000,000 road segments (links) and 500,000 junctions (nodes).

The available transportation modes in the simulation are car, public transport, bike and walk (freight transport is not included). The traffic simulator used in MATSim is the "Java Deterministic Event-Driven Queue-Based Traffic Simulation" (JDEQSim), which is an extension of classic queue-based micro-simulations similar to METROPOLIS (De Palma and Marchal, 2002). The main property of these models is that intersections alone determine the main features of traffic according to a series of constraints such as capacity, free speed travel time, intersection precedence and space available on the next link (Charypar et al., 2007). Only cars are "physically" simulated along the roads, while the other modes are "teleported" from the origin to the destination. The duration of the public transport trips is constant and corresponds to the origin-destination (OD) travel time derived from a matrix estimated for the whole metropolitan area of Zurich (Waraich and Axhausen, 2012). The duration of walk and bike trips is calculated by means of normative operational speeds.

The studied network consists of an area of 1.5 km around the city center (Figure 1). Table 1 provides a synthetic overview of the major quantitative characteristics of the analyzed data. From a qualitative point of view, the studied network is characterized by different typologies of road (arterial, local, and collector roads) and by the presence of several bottlenecks like bridges and tunnels. Hence, such a typology of network will hardly satisfy the homogeneity conditions formulated by Geroliminis and Daganzo (2008) necessary to derive a well-defined NFD. However, instead of looking for the optimal partitioning of the network in regions characterized by homogeneous traffic conditions (reservoirs) we explicitly consider this issue by including the property of distribution of traffic in our analysis.

Normally, each agent needs to travel at least once per day to execute his plans. Instead of simulating the full population, a sample of 10%, equivalent to 180,000 agents, is used for the experiments of this study. In order to deal with smaller samples, it is common practice to downscale link capacities to match these with the sample size.

Place Figure 1

**Table 1: Key characteristics of the data analyzed within the cordon**

| | |
|---|---|
| **total length** | 175.5 km |
| **number of links** | 1224 |
| **number of intersections** | 550 |
| **simulation time** | 00:00-24:00 |
| **aggregation time** | 5 minutes |

Since scaling might represent an issue for reliable representation of traffic dynamics under congestion, a series of analyses have been performed to verify the consistency of the traffic simulation with important traffic flow phenomena such as spillback effects, blocking effects and shockwaves. In order to provide additional evidence of the reliability of such approach, we provide in Figure 2 a comparison of the accumulation-production relationship corresponding to three different sample sizes: 10%, 20% and 50%. Despite some differences ascribable to the calibration process, it is possible to see that the upper bound is similar. The free flow and capacity point correspond to each other, whereas the congested branch looks steeper with increasing sample size.

While the obtained capacity (around 700-800 veh/h/lane) corresponds with the outputs of previous similar studies (Daganzo et al. (2011), Mazloumian et al. (2010), Mahmassani et al. (2013)), the identified critical accumulation (around 15 veh/km/lane) is 30-40% lower than in other studies. However, this outcome does not constitute a problem since every network is characterized by a different "shape" with its own specific threshold values, depending on several factors such as the typology of roads, the mix of traffic and the traffic control systems (although in our simulation no signal control at junctions is implemented). More important, the diagram exhibits the typical concave shape, but the plot is not as clear as for the experimental relationship determined in homogeneous conditions. In this case, when density approaches values closer to capacity, a "bifurcation" takes place and multiple levels of performance can correspond to the same value of accumulation.

Place Figure 2

## 3.The effects of inhomogeneous distribution of traffic

In order to investigate the effects of spatial distribution of traffic on network throughput, we illustrate the relation between accumulation and spatial spread of density in Figure 3. The diagram highlights a parabolic trend and it shows that different values of spread might correspond to the same value of accumulation. The "physical interpretation" of this fact can be that the same amount of users in the system (density) could be evenly distributed on the network or more concentrated in specific zones (higher spread).

Interestingly, only an upward trend is observed, while after a certain level of density a decrease of spread would be expected as the network becomes congested everywhere. However, since in this simulation during the most congested period only 25% of the links are congested (the network is far away from being "uniformly" congested) the resulting pattern of spatial spread seems plausible.

Place Figure 3

The previous analyses show that when the regularity conditions are relaxed, the accumulation-production relation is characterized by a high amount of scatter. Like Daganzo et al. (2011), Mahmassani et al. (2013) and Saberi and Mahmassani (2013) who discussed the phenomenon of bifurcation of the NFD, and Geroliminis and Sun (2011) who identified a correspondence between clockwise patterns in the density-outflow diagram; we also investigate this property in the network of Zurich. The main peculiarities of our study consist of: a vast and highly heterogeneous urban network (including local roads, main arterials), the spatial heterogeneity of origins and destinations, the realistic travel behavior of the simulation, and the time-span of the simulation (full day).

The relationship between scatter in the aggregation-production diagram with the inhomogeneous traffic conditions inside the cordon is confirmed by the fact that the spatial spread is not unique for a specific level of traffic accumulation, and by a qualitative analysis of traffic distribution inside the cordon network reported in Figure 4. This picture corresponding to a "snapshot" of link densities (vertical axis) inside the cordon (horizontal plane representing the geographic coordinates) taken at 07:00 (Figure 4a) and 18.30 (Figure 4b) shows that whereas the aggregated value of density is similar (about 14-15veh/km/lane), the nature of congestion is rather different. Indeed, during the morning peak the links are more homogenously congested, while during the evening peak, there are fewer locations with more severe congestion. A possible explanation for the difference between the two peak periods could be the generation of traffic flows from concentration of facilities with concurrent closing time.

Place Figure 4

Traffic conditions during the evening peak are clearly affected by local spillbacks that reduce the overall capacity of the network even for relatively low aggregated densities. The presence of different states in the network (congested and uncongested) and the instability of its traffic conditions can also be related to the phenomenon of hysteresis loops illustrated in Figure 5. As is possible to see below, a clockwise loop in the accumulation-production relation (left) corresponds to an anticlockwise and spatial spread-production (right) relation during the evening peak. Here the inhomogeneous distribution of congestion plays a role in the performance conditions even at below critical densities. Indeed, during certain time intervals drops of production occur together with rapid increases of spread whereas the accumulation remains almost constant

(see "vertical" slopes in the black circles).

Place Figure 5

Physically this can be associated to decreases of production generated by clusters of congestion rather than increases of users in the network. The issue of inhomogeneity is potentially problematic because travel conditions cannot be identified straightforwardly through the NFD, but it is necessary to know the pattern followed before reaching the current point in the accumulation-production plane, i.e. at onset (upper part of the loop) or offset (lower part of the loop) of congestion. This is the reason why it is necessary to consider the spatial distribution of traffic in order to accurately describe traffic in large-scale complex urban networks characterized by inhomogeneous traffic conditions. In a similar way to the study by Knoop and Hoogendoorn (2013) accumulation, production and spatial spread of density are represented together in Figure 6, which relates to the output from 5 simulations with (slight) variations in the origin-destination patterns (determined by a random seed).

Under homogenous conditions, a well-defined invariant NFD would show up under any circumstance, whereas in our study the diagram exhibits high scatter. The 3-dimensional relation confirms that the scatter resulting from aggregated patterns is not the result of randomness, but it is connected with the spatial distribution of congestion. In fact, as is possible to see from the colored plot, higher accumulation values correspond to higher spread values and more importantly, for the same value of accumulation a lower production is caused by a higher spread. Furthermore, this property is consistent with the hysteresis phenomena characterized by different levels of production (and clustering of traffic) at the onset and offset of congestion.

So far we demonstrated that the simulated urban network system is hysteretic and that traffic conditions cannot be simply described by the 2-dimensional accumulation-production relation given the inhomogeneous distribution of traffic. Indeed, without information about the followed path it is not possible to know the current state of the system (Geroliminis and Sun, 2011). For this reason, the 3D-NFD might provide a more complete explanation of the network-wide traffic conditions. In the next sections the discussion will be extended to the role that dynamic features of NFD play in the formulation of alternative congestion pricing schemes. Furthermore, a series of analyses will show through a set of performance indicators how the issue of inhomogeneity is potentially problematic for the efficiency of the proposed schemes.

Place Figure 6

## 4. Marginal cost pricing tolling schemes based on traffic flow dynamics

In Section 4 the previous insights on the aggregated properties of traffic flows are applied to the design of congestion pricing schemes. To this end, Section 4.1 reviews traffic flow dynamics in (existing) congestion pricing models. After which Section 4.2 presents two alternative time-varying cordon-based tolls that are both consistent with marginal cost pricing theory, but address the spatial spread of density differently. Finally, the eventual fares are briefly discussed and compared to a uniform toll in Section 4.3.

### 4.1 Enhancing dynamic features in congestion pricing models: earlier studies

Road congestion pricing as a measure to alleviate congestion is well known among transport economists, traffic engineers and policy-makers. Since the first study by Pigou (1920), an extensive body of knowledge about this topic has constantly grown (De Palma and Fosgerau, 2010;De Palma and Lindsey, 2011). For the sake of this study, we consider congestion pricing models with respect to their ability to account for dynamic traffic flow properties. This perspective highlights the difference between "static models" that treat congestion as constant over time, and "dynamic models" that account for the evolution of traffic systems over time.

An important category of models also known as static Marginal-Cost Pricing (MCP) models originate from Pigou's principle that the "costs of traffic congestion are borne by travelers collectively but, because individual travelers impose delays on others, they do not pay the full marginal social cost of their trips and therefore create a negative externality" (De Palma and Lindsey, 2011). Then, an "optimal toll" equal to the marginal external cost of congestion cost could be set such that the external costs generated by each traveler is internalized. The main drawback of static MCP models is the invalidity of the assumption of a constant

demand-supply relationship, as in reality travel times (and costs) during peak and off-peak periods differ significantly from each other (De Palma and Fosgerau, 2010).

In order to overcome this limitation (and other ones concerning the lack of modeling of departure time choice) a new category of models were introduced called Bottleneck Models where congestion is defined as a queue at a bottleneck (Vickrey, 1969; Arnott et al., 1993). In the last decades several studies have been conducted in order to explore and include additional features to improve the realism of the model in terms of elasticity of demand, heterogeneity of users, and real world applications (second-best models). The interested reader may refer to Small and Verhoef (2007) for a comprehensive overview of these studies.

The main drawback of Bottleneck Models lies in the lack of consistency with Traffic Flow Theory, particularly in case of urban networks, since the average flow (corresponding to the demand) and the related delays do not reproduce realistic congestion behavior, e.g. that bottleneck throughput decreases when traffic breaks down and queues form. As a result, the estimated congestion toll based on idealized versions of these curves may not be optimal and the system may be either still congested if underpriced or under-utilized if over-priced (Zheng et al., 2012).

An innovative solution came in recent years from the engineering field and consisted of a combination of the "classic" Bottleneck Model with macroscopic traffic models. As mentioned in Section 2, the NFD is capable of describing aggregated traffic conditions in urban regions. Geroliminis and Levinson (2009) applied these findings to include in the traditional bottleneck model a supply curve determined by the NFD and evaluated the trade-off between efficiency and equity of a cordon-based toll. Gonzales and Daganzo (2012) used a similar approach to study the morning commute problem for a network served by both car and public transit sharing the same space in order to find optimal toll and transit fares. The concept of NFD is finally gaining ground also in the field of economics where some dynamic features of congestion (decrease of traffic throughput when capacity is exceeded) are included in analytical models (Arnott, 2011; Fosgerau and Small, 2013).

An important contribution for its level of realism and consistency with traffic dynamics came recently from Zheng et al. (2012) who combined macroscopic modeling of congestion with an agent-based model to develop a cordon-based toll. In this study, the authors derive a fixed charge (Flat Toll) for the city center of Zurich by means of a feedback linear control process where the fare is proportionally updated until the average density is below a threshold value (equal to the critical accumulation). Liu et al. (2013) proposed a similar approach to determine the toll by means of a revised genetic algorithm to reach optimal levels of average speed inside the cordon.

### 4.2 Development of two alternative (bi-directional) time-varying cordon-based schemes based on macroscopic traffic variables

The accuracy of congestion pricing models in representing correctly the mechanism of congestion and ultimately the relation between demand and supply is paramount to derive realistic and efficient tolls. Recent findings from the Traffic Flow Theory have allowed improvements of traditional analytical models and simulation-based models. In the first part of the study, we investigated the macroscopic properties of traffic in a large-scale complex urban road network and we identified, in particular, the issue of inhomogeneity of traffic conditions as a critical aspect in the dynamics of the NFD. Based on these observations, we derive network-wide tolls that apply the macroscopic dynamic properties of traffic to the theory of MCP in order to provide a more transparent and theoretically sound approach than simulation-based models.

Two alternative cordon-based schemes (Step Toll and Hybrid Toll) controlled by the aggregated traffic relationships of the network and characterized by different conceptual approaches are presented. The Step Toll consists of a time-varying toll that aims at eliminating delays in the cordon estimated by means of the NFD. The rationale behind this scheme is to implement a more dynamic and flexible approach in order to limit traffic demand more smoothly than uniform tolls. The Hybrid Toll is a variation of the Step Toll where the inhomogeneous distribution of traffic and its influence on performance are considered as well when setting the fares by means of a 3D-NFD. The latter scheme represents a complementary approach to derive control strategies that explicitly accounts for (uneven) distribution of congestion. Evidently, as the toll level is here set based on the partial derivative towards the density (instead of the gradient towards density and spread) the tolls are lower and hence the reduction in congestion level will be lower as well. Nevertheless the fact that this toll will effectuate a smaller improvement in the network performance, we believe that this alternative way of computing the step toll may still be a valid option from an equity perspective as it only charges users based on the additional network density they create, but does not internalize (i.e. charge) the

costs associated with changes in network performance due to changes in the spatial spread of density. That the changes in network performance due to changes in the spatial spread of density (while the absolute density level remains similar) can be substantial is shown in Fig 6-8.

A ring of 1.5 km around the city center corresponding to the area analyzed in the first part of the study is identified as cordon where agents will be charged. At this stage of the study, a bi-directional toll, which involves both inbound and outbound trips is compared with the traditional one-directional toll that charges only in-coming agents. On the one hand penalizing trips entering the area seem to be a more user-friendly and practical approach. On the other hand a bi-directional charge could affect a larger portion of drivers in the cordon area, regardless of whether their trip is inbound or outbound. In order to provide a basic comparison between the effectiveness of the two schemes, two flat uniform tolls are derived by means of linear-feedback control process until the accumulation is below the "critical" value of 15 veh/lane/km. Further explanations about the methodology and the algorithmic steps used are provided in the following subsections.

In order to bring the accumulation below the critical level, the levels of the charge need to be higher in case of one-directional toll: 2.2 CHF instead of 1.5 CHF during the morning peak and 3.2 CHF instead of 2.0 CHF during the evening peak. The lower elasticity of traffic during the evening peak could be explained by the nature of traffic in Zurich itself that is characterized by prevalent inbound flows during the morning and outbound ones during the evening. As Figure 7 shows, for the same levels of charge, congestion is considerably higher in case of one-directional toll, particularly during the evening peak, probably because this scheme can affect only entering users and part of the crossing traffic.

Hence a bi-directional approach seems to be more appropriate as outbound trips considerably contribute to the network accumulation and performance, especially during the evening peak. Furthermore, this creates an additional penalty for drivers who cross the cordon area. The Stockholm congestion charge, where the toll is levied in both directions, is the real-world evidence of the feasibility of this configuration (Eliasson et al., 2009). Note that in the model setup chosen here, agents facing a toll may choose to cancel their trip, change their destination, change their mode of transport, change their time of day to undertake their trip, and change their route.

The two alternative schemes are described in detail in Sections 4.2.1 and 4.2.2 respectively. In Section 4.2.3 we briefly describe a uniform toll (Flat Toll) that aims at operating the system in non-congested conditions and is derived by means of a control process similar to that by Zheng et al. (2012).

Place Figure 7

*4.2.1 The Step Toll*

The Step Toll corresponds to a time-varying toll in discrete intervals of half an hour. The idea is to determine the level of fares such that the new users are charged for the additional delay they create by travelling inside the cordon derived from the aggregated accumulation-production relation. This principle is in line with the MCP approach where the external costs generated by each traveler are internalized by means of a charge. Thanks to the NFD it is rather straightforward to estimate the delays by measuring the changes of average speed and the corresponding increases of accumulation. The approach is expressed by means of the following equations. The time loss per user determined by a decrease of speed during the time interval *j* corresponds to:

$$\Delta d_j = \frac{s_j}{U(t+\Delta t)} - \frac{s_j}{U(t)} \quad (8)$$

Where $s_j$ corresponds to the average trip distance travelled inside the cordon during the interval time interval j and it is equal to:

$$s_j = \frac{U(t+\Delta t) + U(t)}{2} \cdot \Delta t \quad (9)$$

Where $\Delta t$ corresponds to the time interval between two measurements (5 minutes). The change of total delay for all network users inside the cordon $\Delta D_j$ during the time interval *j* is given by:

$$\Delta D_j = \Delta d_j \cdot K \cdot L \quad (10)$$

Where $L$ corresponds to the total length of the network. The number of additional users $\Delta N_j$ is derived from the change of average density $\Delta K_j$:

$$\Delta N_j = \Delta K_j \cdot L \quad (11)$$

Finally, the toll is derived by dividing the product of total delay and average value of time (25 CHF/h) by the number of additional users.

$$\Delta \tau_j = \frac{VOT \cdot \Delta D_j}{\Delta N_j} \quad (12)$$

$\Delta \tau_j$ has been estimated by aggregating five minutes intervals over longer spans of thirty minutes that are set as a time constraint to derive the steps. Such a constraint has been chosen for reason of "transparency towards the users" with reference to the currently operating systems in Singapore and Stockholm where the charges vary according to intervals of half an hour, respectively one hour.
The final toll has been determined by means of an iterative process as follows.

i. Identify the duration of the toll, time intervals, initial levels of fare (equal to zero) and cordon area
ii. Perform the MATSim simulation consisting of 50 iterations during which agents' can adjust their travel behavior choices (travel departure time, mode, route) until an agent-based stochastic user equilibrium is reached
iii. Calculate the aggregated Marginal Cost of congestion for each interval and update the toll for each interval as toll(new) = toll(old) + $\Delta\tau$(new)
iv. Return to step ii.

The iterative process is stopped when delays are eliminated for each interval.

*4.2.2 The Hybrid Toll*

The Hybrid Toll consists of a variation of the Step Toll that accounts for the issue of uneven distribution of traffic by means of the spread of density. In practice, as it has been already shown, the accumulation-production relation is characterized by scatter and is determined by the distribution of congested links. For example, as is possible to see from Figure 8, representing the variation of accumulation-production-spatial spread in intervals of 5 minutes, all the quadrants are characterized by high scatter, whereas in a situation of "crisp" NFD a univocal correspondence could be identified. Under these conditions, we can see that considerable drops in performance occur even for low increases in accumulation and might be associated to clustering of congestion identified by high increase in spread of density (red colored dots in the lower-right quadrant). The Step Toll applies a charge regardless of the loss of performance due to the heterogeneity of traffic conditions inside the cordon. As a result, it cannot properly internalize the cost of delay related to new entrants. Indeed, few entrants might pay high tolls only because clusters of congestions (of users already entered) have occurred during the same time interval.

Place Figure 8

The Hybrid Toll explicitly considers this issue and it applies a charge that internalizes only the delay due to the increase of users inside the cordon. We know in advance that the resulting fares and consequently the reduction of congestion will be lower than the Step Toll as the uneven distribution of congestion proved to be a major cause of the decrease in performance of the system. On the other hand, from a social perspective this scheme represents a more accountable approach as it charges users only for the actual drop of performance they personally cause by entering the cordon and it might be more beneficial in terms of increased social welfare. In order to derive the Hybrid Toll, the magnitude of the drop in throughput specifically determined by the increases in density and spread needs to be identified.

From the previous analyses, we could infer that the spatial spread of congestion increases together with growing accumulation. Indeed, the increase of traffic within the restricted area will necessarily lead to further spread of congestion later in time once those vehicles continue their trip within the restricted area. From this perspective, uneven distribution of traffic (clusters) might be interpreted as positive deviations from such a natural increment. Hence, we assume that a part of the increment of spread, the one that increases naturally with rise of traffic, can be represented by the lower envelope in Figure 9 and be expressed by the following polynomial function:

$$\gamma(k) = a \cdot k + b \cdot k^2 \quad (13)$$

Where $\gamma$ corresponds to the spread of density and $k$ corresponds to the accumulation. The coefficients a, b are estimated by means of a weighted polynomial regression where the root mean square error is minimized. This way 80% of the 1440 measurements (from the 5 simulation runs) is used for calibration, while the

remaining 20% is used to compute the goodness-of-fit. This leads to the following estimates,

$$a = 1.143\,; b = 0.015$$

Place Figure 9

In order to account for additional increases of spread due to uneven distribution of traffic, we introduce the indicator called "deviation from spread" ($\sigma$) that is derived as:

$$\sigma = \gamma' - \gamma(k) \quad (14)$$

Where $\gamma'$ represents the measured spread. Basically, this indicator shows how much the increase of spread deviates from its "natural" increase due to the accumulation.

Then, we derive the following polynomial form to express the relationship between accumulation, deviation of spread of density and production:

$$Q(k,\gamma) = a \cdot k + b \cdot k^2 + c \cdot k^3 + d \cdot \sigma \quad (15)$$

Where Q corresponds to the total production, $k$ corresponds to the accumulation and $\sigma$ corresponds to the deviation from the natural increase of spread of density.

The coefficients *a*, *b*, *c*, *d* are estimated by means of a weighted polynomial regression where the root mean square error with the data points from the same dataset is minimized. This leads to the following estimates,

$$a = 114.8\,; b = -5.88\,; c = 0.0856\,; d = -11.74$$

The function gives a reasonable approximation: RMSE=13.25 and around 8% of average error for congested and nearly congested traffic conditions, particularly when compared to 3rd degree polynomial function (NFD), which is characterized by RMSE= 57.14 and around 14%, of average error during congestion.

Although the free-flow regime looks slightly overestimated, the polynomial function fairly reproduces the congested regime (Figure 10). We admit that the assumed relationship is only an approximation that seems to provide a good estimation of decreases in performance due to variations of accumulation and spread in this specific study, but it may not necessarily be the best functional form to capture the 3D-NFD. Further research will be needed to identify the most appropriate generic form to express the relationship between the three variables.

Finally, once the production is expressed as a polynomial function of density and spread, it is possible to compute the gradient as a composition of partial derivatives $\nabla Q(k,\gamma) = \frac{dQ}{dk} + \frac{dQ}{d\sigma}$, to identify the variations of outflow "strictly" due to the variations in density and deviation of spread. Hence, a toll aimed to internalize solely the decrease in performance determined by additional users can be identified. The same algorithmic steps used for the Step Toll are applied to determine the levels of fare, with the only difference that the Marginal Cost is computed from the partial derivative of the 3D-NFD with respect to the density k.

Place Figure 10

*4.2.3 The Flat Toll*

The Flat Toll is a uniform toll during the morning and evening peak hours so that the network (inside the cordon) operates at its maximum throughput level. The charge is derived by means of a feedback control process similar to Zheng et al. (2012) where the toll is updated with a constant proportion until the average density is below a threshold value (equal to the critical accumulation). In this case the control variable is the density. The initial fare is progressively increased at the end of each simulation during which travel plans of agents are updated by departure time, mode and route choice. The toll is updated proportionally with a fix step of 0.1 CHF until all the values of accumulation are below the critical value. Two different threshold values for the accumulation during morning and evening peaks are chosen as the "drop of performance" occurs at lower density during the evening (in part due to the inhomogeneous conditions).

### 4.3 Resulting fares

The resulting fares have been estimated by means of an iterative estimation process where at the end of every simulation the levels of the charge were updated based on the delay estimated through the aggregated traffic relationships. This feedback process was carried out until the aggregated delays were eliminated and resulted in the fares illustrated in Figure 11. We would like to point out that the case study setting has been calibrated towards observed aggregated mode shares (Meister et al., 2010), while agents' willingness-to-pay coefficients are only face-validated (i.e., within the range of willingness-to-pay estimates reported in literature). Hence, the following analyses serve the purpose of illustrating the effects of the tolling schemes on the dynamic features of the 3D-NFD, but require further validation before the results can be used to support policy decisions.

Place Figure 11

## 5. Results

Section 5 illustrates the impacts derived from the different tolling schemes. In Section 5.1 the accumulation-production relations are presented. In Section 5.2 an evaluation of the schemes will be carried out by means of a set of commonly used traffic flow performance indicators.

Place Figure 12

### 5.1 Accumulation-production relationships

The accumulation-production relations resulting from the implementation of the three alternative tolling schemes are consistent with the findings emerged from the previous discussion. In particular, the impacts of the alternative tolling schemes on the accumulation-production patterns seem to be strongly affected by hysteresis phenomena. The Flat Toll scheme produces a significant improvement of performance. The congested branch disappears and it exhibits no drop of production (Figure 12a). Also the Step Toll and Hybrid Toll seem to yield higher performances of the network, as it does not show any congested branch (Figure 12b and 12c).

As to the overall reduction of accumulation, Figure 12d shows that all the schemes can considerably smooth the morning peak, whereas the evening peak is reduced only in case of Flat and Step Toll. In line with the previous analytical studies (Van den Berg, 2012), the Flat Toll generates a higher reduction in overall demand (-10%) than the Step Toll (-7%) that is instead characterized by more rescheduled trips. Interestingly, the Step Toll seems to generate a shift in trips to the lunch period that however does not create any decrease in performance. This additional peak might be related to a slight increase in car trips mainly directed to shopping-leisure activities late in the morning. The Hybrid Toll produces a significant decrease in accumulation only during the morning peak, while no appreciable reduction is achieved during the evening peak.

Upon closer inspection, it emerges that the spatial distribution of congestion and the related phenomenon of hysteresis, affect considerably the effectiveness of pricing schemes. In case of the Step Toll and Hybrid

Toll, the free-flow branch presents higher scatter than the one derived from the Flat Toll, because during loading and unloading cycles the performance of the system seems to progressively deteriorate (Figure 13). This result suggests that hysteresis phenomena might be a reason of decrease in the performance of the system as well and that fluctuations of demand should be minimized by means of smoothing traffic demand and control strategies.

Place Figure 13

### 5.2 Traffic performance enhancement

In order to examine the performance of a traffic network from different perspectives, the following aspects are considered: traffic efficiency, travel time savings, decrease in travel demand, heaviness of congestion, and queue length (Table 2). Additional analyses concerning the overall decrease of demand, modal shift and total collected revenues are included in order to give a fuller picture of the effects of the tolls (although

further calibration of willingness-to-pay coefficients is required in order to draw conclusive policy implications).

As the optimal utilization of the network is intended as a trade-off between network utilization and network production, the indicator Traffic Efficiency is adopted from Brilon et al. (2000). Traffic Efficiency expresses the performance of the network as production per time unit, rather than considering demand, capacity and quality of the flows as separate indicators. According to this indicator, the more (veh x km) are produced by a traffic facility per hour, "the greater efficiency with which the potential of the existing infrastructure is exploited" (Brilon, 2000). The Traffic Efficiency is defined as:

$$E = Q \cdot U \cdot T \quad (16)$$

Where $Q$ represents the average production (veh/h) and $U$ the average speed (km/h) over the network during a time interval $T$ (h) of 5 minutes. The factors $Q$ and $U$ are obtained from equations 5 and 6. Time savings are expressed as reduction of travel delay estimated as vehicle-loss hours where the reference free-flow speed is assumed to be 70 km/h. The traffic demand is expressed as total travelled kilometers by all vehicles inside the cordon (veh-km). In order to represent the extent of congestion both in space dimension (by means of queue length) and in time dimension (by means of duration of congestion) the indicator heaviness of congestion is introduced. The index is calculated as a product between the total length of congested links (with density above a threshold value of 35 veh/km) and the time interval of 5 minutes (0.083 hour). The queue length is estimated as a sum of links with density higher than 35 veh/km during time intervals of 5 minutes.

The traffic improvements derived from the schemes during the morning and evening peak hours (06:00-10:00 and 16:00-22:00) are reported in Table 2. The several adopted indicators suggest a coherent interpretation of the effects of the schemes.

All the cordon-based schemes seem to be more beneficial in the morning rather than in the evening. While in the morning the congestion is severe and the decrease in performance large, in the evening traffic is presumably not so heavy to compensate the lower utilization of the network and justify the application of such a demand management measure (traffic efficiency might be intended as a proxy indicator of the optimal usage of transport infrastructure). From this perspective, the Hybrid Toll, despite the lower levels of fare, leads to a better optimization of the network usage than the other two schemes.

However, this is just a single perspective to evaluate the effects of the tolls, as the improvements in terms of travel delays, heaviness of congestion and queue length show. For example, during the morning peak the two time-varying tolls achieve important improvements, comparable to those from the Flat Toll, by favoring the rescheduling of trips rather than the modal shift. Like in those studies based on the bottleneck model (Arnott et al., 1993; Van den Berg, 2012) and on queue-based models (de Palma et al., 2005), users seem to be more willing to reschedule their trips and accept to pay a toll rather than switching to public transit in case of time-varying congestion charges. Indeed, the modal shift over the entire day is almost null in case of Step Toll, whereas it accounts for about 40% of the reduced demand in case of Flat Toll.

This situation seems to hold only when traffic conditions are homogeneous. Indeed, during the evening peak, a higher decrease in demand is necessary to the time-varying tolls in order to achieve substantial benefits. Here the Step Toll, which determines a similar decrease of traffic demand to the Flat Toll, produces larger traffic improvements than the Hybrid Toll.

Hence, it seems that not only the level of fare, but also the spread of density and hysteresis can significantly affect the performance of the schemes. Above all, the uneven distribution appears to be the main reason of lack of the efficiency of cordon-schemes that only by reducing the inflow with higher flows are able to produce benefits. Under these circumstances, the Hybrid Toll, which on the contrary addresses only those delays ascribable to increases in accumulation, determines smaller improvements.

This is an interesting outcome, since the two schemes are characterized by a different conceptual approach and might determine different impacts in terms of economic gains, distributional effects and public acceptability. The fact that the Hybrid Toll determines a substantially lower reduction of users compared to the other two schemes might justify the lower traffic enhancements.

On the other hand, the lower reduction of demand and smaller improvements are not necessarily a negative, as they might be compensated with higher economic gains (for example, people might benefit from a "softer" approach). Furthermore, the occurrence of hysteresis loops related to loading-unloading cycles seems to progressively deteriorate the production in both the two time-varying schemes.

The Flat Toll generates the highest amount of revenues (18,111 CHF, 6,887 payments) thanks to its higher levels of fare, whereas the Step Toll and the Hybrid Toll are characterized by higher number of payments

(respectively 8,970 and 7,970) and lower collected revenues (14,196 CHF and 5,504 CHF). These results are in line with the rest of the traffic enhancement indicators: because of their higher flexibility, the two time-varying tolls allow users to reschedule their trips within the peak hours rather than stop entering the cordon. Although the absolute values are not conclusive given that agents' price elasticities have not been calibrated towards tolling measures, these analyses do suggest that time-varying tolls in general can be seen as fairer from an economic perspective. Where the Flat Toll is a coarser measure, the Step Toll allows for a better distribution of (lower) fares across a larger number of agents, while the Hybrid Toll specifically charges (lower) fares targeted at a smaller number of agents excluding agents and tolls associated with network performance reduction due to changes in the spatial spread of congestion.

**Table 2: Traffic enhancements determined by the alternative tolling schemes**

| | Flat Toll | Step Toll | Hybrid Toll |
|---|---|---|---|
| Traffic efficiency (veh*km/hours) m.p.[1] change (%) | 29.4 | 26.7 | 21.3 |
| Traffic efficiency (veh*km/hours) e.p.[2] change (%) | -17.9 | -21.6 | -4.2 |
| Traffic demand (veh-km) m.p. change (%) | -13.6 | -9.7 | -12.1 |
| Traffic demand (veh-km) e.p. change (%) | -21.7 | -20.9 | 11.5 |
| Travel delays (veh*loss*hours) m.p. change (%) | -89.9 | -75.9 | -80.5 |
| Travel delays (veh*loss*hours) e.p. change (%) | -71.2 | -44.0 | -52.4 |
| Heaviness of congestion (km*hours) m.p. change (%) | -85.1 | -73.8 | -66.6 |
| Heaviness of congestion (km*hours) e.p. change (%) | -83.8 | -73.8 | -33.8 |
| Queue length (km) m.p. change (%) | -82.6 | -72.4 | -67.4 |
| Queue length (km) e.p. change (%) | -81.9 | -70.3 | -53.9 |
| Total collected revenue (CHF) | 18,111 | 14,196 | 5,504 |
| Total demand reduction (%) | -10.0 | -7.0 | -3.4 |
| Modal shift from car to public transport (% of trips) | +2.3 | -0.9 | +1.5 |

## 6. Concluding remarks

In this study, we proposed two alternative cordon-based tolls controlled by macroscopic traffic properties of the network in order to apply the Marginal Cost Pricing theory to large urban networks. This approach allowed higher levels of realism and higher consistency with the Traffic Flow Theory than traditional analytical methods. The idea is to provide a methodology to derive tolling schemes that thoroughly reflects the nature of urban traffic congestion and that, at the same time provides a better understanding than data-driven approaches.

Traffic control strategies and demand management measures have thus far been mainly applied to networks or motorway stretches under specific "regularity conditions". In our study instead, we deal with a complex urban network characterized by heterogeneous features in space (different typologies of roads) and time (traffic demand patterns within the day). Hence, by means of the Hybrid Toll we propose an approach to derive tolling schemes that fully accounts for the consequences of uneven spatial distribution of traffic inside the cordon. We believe that describing the macroscopic traffic properties of the network by means of a 3D-NFD that includes the deviation of spatial spread as an additional dimension, allows acknowledges this aspect.

The analysis of traffic improvements has shown that a more flexible toll (Step Toll) might determine benefits comparable to ones of a uniform toll (Flat Toll), by higher rescheduling of trips rather than by modal shift and it confirms with a novel approach what was previously demonstrated through the economic bottleneck models. The fact that the Hybrid Toll yields a lower reduction of congestion than the other two

[1] m.p stands for morning peak
[2] e.p. stands for evening peak

schemes could be ascribed to the lower fare in the first place. However, more detailed analyses show that, when traffic is more uniformly distributed over the network (during the morning peak) this scheme can also determine substantial improvements of traffic conditions (in terms of decrease of delays, heaviness of congestion and queues) by means of a relatively low charge. Moreover, as the traffic efficiency indicator shows, the Flat Toll and the Step Toll can achieve satisfactory results, only by considerably reducing the traffic demand, and ultimately yielding a sub-optimal utilization of the network. Thus, if the priority of policy-makers is to achieve the most efficient utilization of the network, the Hybrid Toll represents the most valuable option in case of unevenly distributed congestion. In future research, additional efforts could be made to investigate the influence of the spatial spread of traffic on the network performance and improve the quality of the fitting 3D-NFD function, from which the tolling method depends on.

Part of this study was devoted to the analysis of the dynamics of spatio-temporal congestion patterns in large urban networks characterized by heterogeneous traffic conditions by means of an agent-based model. From the investigations it emerged that the distribution of traffic expressed by the spatial spread of density has a negative effect on the performance of cordon-based tolls. This finding is in line with previous studies that found urban systems to be hysteretic, dynamic and path dependent (Geroliminis and Sun, 2011a). Moreover, qualitative analyses suggest that the hysteresis itself might be a reason behind the decrease of the system performance (Step Toll and Hybrid Toll case), since frequent cycles of network loading-unloading seem to deteriorate its production, particularly when the system approaches its critical value of accumulation. Further investigations are suggested in order to explore the influence of macroscopic characteristics and to provide effective solutions to control the dynamics of congestion.

Some practical implications as well can be drawn based on this study. The inability of traditional cordon tolls to cope with clustering of congestion represents a main limitation of this demand management measure. When traffic is not homogeneously distributed, cordon tolls are not an efficient solution. As a consequence, it is strongly advisable to pursue a network-wide traffic coordination of control measures aimed at spreading congestion more evenly over the network (gating, traffic signal control, variable-message signs) and travel demand measures aimed at reducing the overall amount of users (pricing measures, public transit improvements). If the simple pricing solution is pursued, then approaches based on the partitioning of network in homogeneous areas or similar categories of road might be more effective, even though less practical.

The implementation of the proposed methods to design congestion-pricing charges through macroscopic traffic variables seems to be a solid and practical approach. It allows the control of large complex networks by means of few indicators and without any information about the origin-destination and travel patterns. Furthermore, the proposed agent-based approach seems to represent appropriately congestion dynamics also when smaller samples are used. For this reason, upon additional tests, the proposed strategies might become soon ready for the implementation of practitioners. The main barrier to the implementation of the proposed schemes, consists in the collection of data necessary to build the NFD, but encouraging results are coming from recent studies focused on deriving accurate NFD from limited fractions of links of the network (Ortigosa, 2014).

**Acknowledgements**

The authors thank Kay W. Axhausen, Nan Zheng, Erik T. Verhoef and Mehdi Keyvan-Ekbatani for their valuable comments and suggestions with respect to the research presented here.

## References

Arnott, R., 2013. A bathtub model of downtown traffic congestion. Journal of Urban Economics 76: 110-121

Arnott, R, A. Palma, Lindsey R., 1993. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. The American Economic Review: 161-179.

Brilon, W., 2000, June. Traffic flow analysis beyond traditional methods. In Proceedings of the 4th International Symposium on Highway Capacity: 26-41.

Buisson, C., 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: 127-36.

Charypar, D., Nagel, K., Axhausen, K.W., 2007. An event-driven queue-based microsimulation of traffic flow Transportation Research Record, 2003: 35–40

Chow, A. H. F., 2015. Optimization of dynamic motorway traffic via a parsimonious and decentralised approach. Transportation Research Part C: Emerging Technologies. In press.

Daganzo, C.F., 1994. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. Transportation Research Part B: Methodological 28 (4): 269-287.
Daganzo, C.F., 2007. Urban gridlock: macroscopic modeling and mitigation approaches. Transportation Research Part B: Methodological 41 (1): 49-62.
Daganzo, C.F., 2008. An analytical approximation for the macroscopic fundamental diagram of urban traffic. Transportation Research Part B: Methodological 42 (9): 771-81.
Daganzo, C.F., Gayah, V.V., and Gonzales., E.J., 2011. Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability. Transportation Research Part B: Methodological 45 (1): 278-88.
Daganzo, C.F., and Geroliminis., N., 2008. An analytical approximation for the macroscopic fundamental diagram of urban traffic. Transportation Research Part B: Methodological 42 (9): 771-81.
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.
De Palma, A., Fosgerau, M., 2010. Dynamic and static congestion models: A review. Cahier de recherche, Ecole Polytechnique, CNRS
De Palma, A., Lindsey, R., 2011. Traffic congestion pricing methodologies and technologies. Transportation Research Part C: Emerging Technologies 19 (6): 1377-99.
De Palma, A., Marchal F., 2002. Real cases applications of the fully dynamic METROPOLIS tool-box: an advocacy for large-scale mesoscopic transportation systems. Networks and spatial economics, 2(4): 347-369.
Eliasson, J., Hultkrantz, L., Nerhagen, L., Rosqvist, L. S. 2009. The Stockholm congestion–charging trial 2006: Overview of effects. Transportation Research Part A: Policy and Practice, 43(3), 240-250.
Fosgerau, M., and K.A. Small. 2013. Hypercongestion in downtown metropolis. Journal of Urban Economics 76: 122-134.
Gayah, V.V., Daganzo, C.F., 2011. Clockwise hysteresis loops in the macroscopic fundamental diagram: an effect of network instability. Transportation Research Part B: Methodological 45 (4): 643-655.
Geroliminis, N., Daganzo, C.F., 2007. Macroscopic modeling of traffic in cities. Paper presented at Transportation Research Board 86th Annual Meeting.
Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. Transportation Research Part B: Methodological 42 (9): 759-70.
Geroliminis, N., Levinson, D.M., 2009. Cordon pricing consistent with the physics of overcrowding. In Transportation and traffic theory: Golden jubilee. 219-240 Springer.
Geroliminis, N., Sun, J., 2011. Hysteresis phenomena of a macroscopic fundamental diagram in freeway networks. Transportation Research Part A: Policy and Practice, 45(9), 966-979.
Geroliminis, N., Sun, J., 2011.Properties of a well-defined macroscopic fundamental diagram for urban traffic.Transportation Research Part B: Methodological 45 (3): 605-17.
Geroliminis, N., Haddad, J., Ramezani, M., 2013. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach. IEEE Transactions on Intelligent Transportation Systems 14 (1): 348-359.
Godfrey, J., 1969. The mechanism of a road network. Traffic Engineering and Control , 11(7): 323–327
Gonzales, E.J., Daganzo, C.F., 2012. Morning commute with competing modes and distributed demand: User equilibrium, system optimum, and pricing. Transportation Research Part B: Methodological 46 (10): 1519-1534
Ji, Y., and Geroliminis, N., 2012. On the spatial partitioning of urban transportation networks. Transportation Research Part B: Methodological 46 (10): 1639-1656.
Keyvan-Ekbatani, M., Kouvelas, A., Papamichail, I., Papageorgiou, M., 2012. Exploiting the fundamental diagram of urban networks for feedback-based gating. Transportation Research Part B: Methodological, 46(10), 1393-1403.
Keyvan-Ekbatani, M., Papageorgiou M., Papamichail, I., 2013. Urban congestion gating control based on reduced operational network fundamental diagrams. Transportation Research Part C: Emerging Technologies 33, 74-87.
Kickhöfer, B., Dominik G., Nagel, K., 2011. Income-contingent user preferences in policy evaluation: application and discussion based on multi-agent transport simulations. Transportation 38.6: 849-870.
Knoop, V. L., Hoogendoorn, S. P., Van Lint, J.W., 2012. Routing strategies based on macroscopic fundamental diagram. Transportation Research Record: Journal of the Transportation Research Board, 2315(1): 1-10.
Knoop, V. L., Hoogendoorn, S.P., 2013. Empirics of a generalized macroscopic fundamental diagram for urban freeways. Transportation Research Record: Journal of the Transportation Research Board 2391(1): 133-141.
Knoop, V.L., Hoogendoorn, S.P., van Zuylen, H., 2009. Empirical differences between time mean speed and space mean speed. Traffic and Granular Flow'07. Springer Berlin Heidelberg: 351-356.
Liu, Z., Meng, Q., and Wang S., 2013. Speed-based Toll Design for Cordon-Based Congestion Pricing Scheme, Transportation Research Part C, 31, 83-98.
Mahmassani, H.S., Williams, J. C., Herman, R., 1984. Investigation of network-level traffic flow relationships: some simulation results. Transportation Research Record: Journal of the Transportation Research Board, 742: 121-130
Mahmassani, H.S., Saberi, M., Zockaie, A., 2013. Urban network gridlock: Theory, characteristics, and dynamics. Transportation Research Part C: Emerging Technologies 36: 480-497..

Mazloumian, A., Geroliminis N., Helbing, D., 2010. The spatial variability of vehicle densities as determinant of urban network capacity. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368 (1928): 4627-47.

Meister, K., Balmer, M., Ciari, F., Horni, A., Rieser, M., Waraich, R.A., Axhausen, K.W., 2010. Large-scale agent-based travel demand optimization applied to Switzerland, including mode choice, paper presented at the 12th World Conference on Transportation Research.

Ortigosa, J., Menendez, M., Tapia, H., 2014. Study on the number and location of measurement points for an MFD perimeter control scheme: a case study of Zurich. EURO Journal on Transportation and Logistics, 3(3-4), 245-266.

Pigou, A.C. 1920. The economics of welfare, London: Macmillan.

Saberi, M., Mahmassani, H.S., 2012. Exploring the properties of network-wide flow-density relations in freeway networks. Transportation Research Record: Journal of the Transportation Research Board, 2315(1): 153-163.

Saberi, M., Mahmassani, H.S., 2013. Empirical Characterization and Interpretation of Hysteresis Phenomena in Freeway Networks. Transportation Research Record: Journal of the Transportation Research Board, 2391:44-55

Small, K.A., Verhoef, E.T., 2007.The economics of urban transportation. Routledge.

Smeed, R.J., 1966. Road capacity of city centers. Traffic Engineering and Control 8 (7), 455–458.

van den Berg, V.A.C., 2012. Step-tolling with price-sensitive demand: Why more steps in the toll make the consumer better off. Transportation Research Part A: Policy and Practice 46.10: 1608-1622.

Vickrey, W. S., 1969.Congestion theory and transport investment. The American Economic Review 59 (2): 251-60.

Waraich, R.A., Axhausen, K.W., 2012. Agent-Based Parking Choice Model. Transportation Research Record: Journal of the Transportation Research Board, *2319*: 39-46.

Wardrop, J.G. 1968. Journey speed and flow in central urban areas. Traffic Engineering and Control, 11(7): 528-532

Zahavi, Y., 1972. Traffic performance evaluation of road networks by the a-relationship. Parts I and II. Traffic Engineering and Control14 (5 and 6), 228–231, 292–293.

Zheng, N., Waraich, R.A., Axhausen, K.W., Geroliminis, N., 2012. A dynamic cordon pricing scheme combining the macroscopic fundamental diagram and an agent-based traffic model. Transportation Research Part A: Policy and Practice 46 (8): 1291-303.