Working Paper

Congestion pricing based on dynamic features of the Macroscopic Fundamental Diagram

Author(s):
Simoni, M.D.; Pel, A.J.; Waraich, Rashid A.; Hoogendoorn, S.P.

Publication Date:
2014

Permanent Link:
https://doi.org/10.3929/ethz-b-000083502

Rights / License:
In Copyright - Non-Commercial Use Permitted
Congestion pricing based on dynamic features of the Macroscopic Fundamental Diagram
M.D.Simoni, A.J. Pel, R.A.Waraich, S.P.Hoogendoorn

Abstract: Studies into optimal congestion pricing are typically based on analytical representations of travel choices and network traffic flows. In this paper we show how congestion pricing schemes (or more precisely, cordon tolls) can be derived from the concept of the Macroscopic Fundamental Diagram (MFD). This approach is consistent with the transport economic notion of marginal cost pricing, while it also enables to realistically account for the dynamics and heterogeneity in the (aggregated) network traffic flow conditions. To this end, we first analyze the hysterical dynamic relations between the macroscopic fundamental traffic variables of network speed, network production, network accumulation, and spatial variance in the network accumulation. Subsequently, we show how through these relations captured by the generalized MFD one can compute the marginal costs of a surplus of traffic inside the cordon area, which is then used as a basis to design a flat toll and step toll pricing scheme. Furthermore, a third tolling scheme is designed to explicitly account for the temporal hysterisis loops in the MFD and the influence of the spatial variance in network accumulation. The pricing framework is implemented in the agent-based simulation model MATSim and applied to a case study of the city of Zurich. The alternative tolling schemes are evaluated against a number of traffic performance indicators to show how these deal with the hysteresis and spatial variance in the network accumulation.

Keywords: Macroscopic Fundamental Diagram, Spatial spread of congestion, Hysteresis, Dynamic Congestion Pricing, Cordon-based Pricing Schemes

1. Introduction

The Macroscopic Fundamental Diagram (MFD) is an innovative concept that allows to model traffic flows in large (urban) regions at an aggregate level. It describes the relationship between the traffic density of vehicles in the network (or accumulation), the aggregated traffic flow (or production), and the aggregated traffic speed. This way, the MFD 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 MFD has been growing rapidly.

At a theoretical level, the fundamental studies by Daganzo (2007) and Geroliminis and Daganzo (2008) were soon followed by several investigations that led to important findings on the properties and conditions to derive the MFD. Perhaps the most important finding consists of the influence of inhomogeneous conditions of traffic on the performance of the network. As a result of the uneven distribution of congestion the MFD shows scatter and hysteresis loops occur in the accumulation-production diagram (Mazloumian and Geroliminis, 2010; Geroliminis and Sun, 2011). 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 network performance by means of a generalized MFD (gMFD).

At an applied level, other studies have utilized the concept of MFD 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 measures (Keyvan-Ekbatani et al., 2012; Geroliminis et al., 2013) and congestion pricing schemes (Geroliminis and Levinson, 2009; Zheng et al., 2012; Gonzales et al., 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 second-best 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 capacity is exceeded.

On the same page, this paper addresses the design of cordon-based congestion pricing schemes based on the notion of the generalised Macroscopic Fundamental Diagram. The purpose of this paper is threefold. First of all, we analyse the time-dynamics of the generalized MFD variables to show the relation between the hysterisis loops in the accumulation-production diagram and the evolution of the spatial spread of accumulation in the morning and evening peak periods. Second of all, we derive the gMFD and fit a continuous polynomial plane in order to compute the marginal costs of a surplus of traffic inside the cordon area, that are then used as a basis to design a time-varying pricing scheme. Third of all, a second tolling scheme is designed to explicitly avoid tolling drivers for travel time delays that are due to (uncontrolled) changes in the spatial spread of congestion inside the cordon area. The study is conducted on the urban road...
network of the city of Zurich in Switzerland. The gMFD data are collected through the agent-based simulation model MATSIM. The two tolling schemes are 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. 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 MFD 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 in an advanced traffic simulation characterized by variegated network topology and realistic travel behaviour choices (in terms of route, time-of-departure and mode choice). Based on these analyses we will finally derive a gMFD that includes the spatial distribution of traffic 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.

In the second part of this study, we will apply our findings about macroscopic properties of the (heterogeneous) network to derive three alternative congestion-pricing schemes controlled by the gMFD within MATSim. The following study aims at extending the approach by Zheng et al. (2012), characterized by 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 solid and transparent model consistent with both the economic (Pigouvian tax) and engineering (network-wide macroscopic modelling) theories. Hence, we derive three different cordon-based tolls: a basic uniform toll (Flat Toll); 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 (Spread Toll).

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 (Simoni, 2013). Additional applications of MATSim are described at www.matsim.org.

This paper is organized as follows. Section 2 reviews the concept of the MFD and describes the theoretical derivation of aggregated traffic flow properties in a road network. In Section 3 the macroscopic traffic relations are analyzed and the effects of spatial distribution of congestion are investigated in order to derive the gMFD. Sections 4 and 5 describe the design of three alternative congestion-pricing schemes controlled by the gMFD and analyse their impacts on traffic conditions. In Section 6 we summarize our findings and we discuss the main implications for practice and further research.

2. The Macroscopic Fundamental Diagram

This section provides the background of the MFD, starting with an overview of the main studies on the MFD in Section 2.1 and then continuing in Section 2.2 with an explanation of the variables of the MFD and how these are derived.

2.1 State of the art

The MFD extends the concept of the Fundamental Diagram (FD) that relates traffic flux and traffic density on (continuous) road sections. The MFD expresses the aggregated flow of 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, different traffic regimes can be identified on the diagram. The left branch where production and accumulation have quasi-linear dependence corresponds to the free-flow regime. While the region characterized by slower increase of production until capacity is reached is typically referred to as capacity regime. And the congested regime can be identified on the right branch and is characterized by a decrease of production with increase of accumulation.
Godfrey (1969) introduced the concept of the MFD, although 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). It is claimed in these studies, though, that a well-defined MFD applies under specific “regularity conditions” concerning the homogeneity of links and the possibility of reaching the Wardrop Equilibrium. Buisson and Ladier (2009) were the first to relax these conditions by analyzing traffic data of the city of Toulouse. The fact that the MFD presented high scatter forming hysteresis loops was attributed to the spatially heterogeneous evolution of congestion and it was soon formalized in other studies (Geroliminis and Sun, 2011). Additional research mainly (co)authored by Daganzo and Geroliminis have shed light on a series of aspects concerning the nature of the MFD (Daganzo et al., 2011; Gayah et al., 2011; Geroliminis and Sun, 2011). In these studies, the scatter and the presence of hysteretic patterns in the aggregated accumulation-production diagram have been associated with the presence of inhomogeneous traffic conditions, e.g. at the onset and offset of congestion. In particular, Mazloumian and Geroliminis (2010) observed that congestion in urban networks is by nature inhomogeneous in time and space and that a unique monotonously falling relationship exists between the average flow and the spatial distribution of link densities in the network. On the same page, Geroliminis and Sun (2011) confirmed that this property plays a key role in identifying the MFD. Another recent contribution came from Knoop and Hoogendoorn (2013) who described the deviation in density of the different links through the network with the spatial spread of density indicator and included it as additional property in what they called generalized Macroscopic Fundamental Diagram (gMFD).

2.2 Derivation of macroscopic traffic variables

The MFD 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) and average speed as average of all vehicle trajectories according to Edie’s generalized definitions (Edie, 1965). First, the average density $k_j$ and average flow $q_j$ for each link are derived for 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:

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

$$\Delta k_j = \frac{e_j - o_j}{l \cdot n}$$ (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 $j$, 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:

$$u_j = \frac{q_j}{k_j}$$ (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:

$$K = \frac{\sum^{n}_{i=1}(k_i \cdot l_i \cdot n_i)}{\sum^{n}_{i=1}(l_i \cdot n_i)}$$ (4)
The average network speed $U$ can be derived as:

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

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 MFD 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:

$$\gamma = \sqrt{\frac{\sum_{i=1}^{n} l_i c (k_i - K)^2}{\sum_{i=1}^{n} l_i}} \quad (9)$$

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.

### 3. Dynamics features of the Macroscopic Fundamental Diagram

In Section 3 we use the variables derived in the previous section to describe the relation between the network production, accumulation, and spatial distribution of congestion. We particularly focus on the phenomenon of hysteresis loops in Section 3.1, and derive the gMFD for the city of Zurich in Section 3.2.

#### 3.1 Macroscopic traffic relations

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). 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.

![Figure 1: City centre of Zurich (adapted from GoogleEarth)](image)

The road network used in the simulation consists of a high-resolution navigation network including about one million road segments (links) and five hundred thousand junctions (nodes).

The available transportation modes in the simulation are car, public transport, bike and walk (freight transport is not included). 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 MFD. 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.

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

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>total length</td>
<td>175.5 km</td>
</tr>
<tr>
<td>number of links</td>
<td>1224</td>
</tr>
<tr>
<td>number of intersections</td>
<td>550</td>
</tr>
<tr>
<td>simulation time</td>
<td>00:00-24:00</td>
</tr>
<tr>
<td>aggregation time</td>
<td>5 minutes</td>
</tr>
</tbody>
</table>

The accumulation-production relation (Figure 2) illustrates the traffic conditions of the network and it highlights the optimal throughput and the critical accumulation (capacity). 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 and Saberi (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.

![Figure 2: accumulation-production relationship](image)

The accumulation-average speed relation reported in Figure 3 describes the deterioration of average speed as a function of vehicles inside the network (accumulation). This relationship is typically characterized by a region with constant or slightly decreasing speed corresponding to the free-flow state and a clear decreasing slope corresponding to the congested state. Despite the presence of some outliers, the diagram shows an average free-flow speed of about 70 km/hour and a decreasing curve until 15 km/hour. The presence of some arterial roads, characterized by high speed is probably the main reason for such a high value.
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 4. 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.

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) and Mahmassani and Saberi (2013) who discussed the phenomenon of bifurcation of the MFD, 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 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 5. This picture corresponding to a “snapshot” of link densities (vertical axis) inside the cordon (horizontal plane) taken at 07:30 (Figure 5a) and 18:30 (Figure 5b) shows that whereas the aggregated value of density is similar (about 14-15 veh/km/lane), the nature of congestion is rather different. Indeed, during the morning peak the links are more homogeneously 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.

Figure 5a. Snapshot of density inside the cordon at 07:30

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 6. 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). Physically this can be associated to decreases of production generated by clusters of congestion rather than increases of users in the network.

Figure 5b. Snapshot of density inside the cordon at 18:30
The issue of inhomogeneity is potentially problematic because travel conditions cannot be identified straightforwardly through the MFD, 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) a generalized Macroscopic Fundamental Diagram (gMFD) that relates accumulation, production and spatial spread of density is represented Figure 7. Figure 7 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 MFD would show up under any circumstance, whereas in our study the diagram exhibits high scatter. The gMFD 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, a gMFD 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 MFD 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.

4. Macroscopic traffic flow dynamics in congestion pricing models

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 three alternative cordon-based tolling schemes, which are consistent with main principles from the field of Transport Economics and Traffic Flow Theory. Finally, the eventual fares are briefly discussed 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). A possible way to look at congestion pricing models is to focus on their capability to explain the dynamic properties of traffic flows. 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, 1377). 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 modelling 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, 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 very uncongested 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 MFD 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 MFD 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 MFD 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 came recently from Zheng et al. (2012) who advanced this approach by combining macroscopic modelling 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. The main advantage of the model lies in its consistency with the engineering and economic perspective as it reproduces the dynamics of congestion well and it achieves high levels of realism in terms of travel behavior and users’ heterogeneity.

4.2 Development of three alternative cordon-based schemes based on the (g)MFD

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 MFD. 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.

Three alternative cordon-based schemes (Flat Toll, Step Toll and Spread Toll) controlled by the aggregated traffic relationships of the network and characterized by different conceptual approaches are presented. The Flat Toll aims at operating the system below a threshold value corresponding to the critical accumulation value in the MFD and is derived by means of an iterative control process similarly to Zheng et al. (2012). Conceptually different, the Step Toll consists of a time-varying toll that aims at eliminating delays in the cordon estimated by means of the MFD. The rationale behind this scheme is to implement a more dynamic and flexible approach in order to limit traffic demand more smoothly. The Spread 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 the (g)MFD. The latter scheme represents a complementary approach to derive control strategies that explicitly accounts for (uneven) distribution of congestion.

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 regardless of their direction. The cordon-based tolls involve both the inbound and outbound trips. Although penalizing trips exiting the area might seem counterintuitive, this setup has as main advantage that it penalizes all drivers in the cordon area, regardless of whether their trip is inbound or outbound – that seems appropriate as both inbound and outbound trips contribute to the network accumulation and performance. Furthermore, this creates an additional penalty for drivers who cross the cordon area. 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 three alternative schemes are described in detail in Sections 4.2.1 through 4.2.3 respectively.

4.2.1 The Flat Toll

The first scheme corresponds to 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 until the average density is below a threshold value (equal to the critical accumulation). 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. 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).

4.2.2 The Step Toll

The second scheme corresponds to a time-varying toll in discrete intervals of half an hour, referred to as Step Toll. In this case, the price is derived such that the new users are charged for the additional delay they create by travelling inside the cordon. 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 MFD it is rather straightforward to estimate the delays by measuring the changes of average speed and the corresponding increases of accumulation. The approach is illustrated by Figure 8 and expressed by means of the following equations. The time loss per user determined by a decrease of speed corresponds to:
\[ \Delta d = t_1 - t_0 = \frac{s}{u_1} - \frac{s}{u_0} \quad (9) \]

where \( s \) corresponds to the average trip distance travelled inside the cordon that is equal to:

\[ s = \frac{u_1 + u_2}{2} \cdot \Delta t \quad (10) \]

where \( \Delta t \) corresponds to the time interval between two measurements (5 minutes). The total delay for users on the network inside the cordon is given by:

\[ \Delta D = \Delta d \cdot K \cdot L \quad (11) \]

where \( L \) corresponds to the total length of the network. The number of additional users \( N \) is derived from the change of average density:

\[ N = \Delta K \cdot L \quad (12) \]

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.

\[ \tau = \frac{VOT \cdot \Delta d}{N} \quad (13) \]

The final toll has been determined by means of an iterative process where at the end of each simulation, the delays and the time-varying toll were derived and updated until significant drops of production were eliminated. Tolls have 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.

\[
\begin{align*}
\Delta d = t_1 - t_0 = \frac{s}{u_1} - \frac{s}{u_0} \\
\Delta d = \frac{u_1 + u_2}{2} \cdot \Delta t \\
\Delta D = \Delta d \cdot K \cdot L \\
N = \Delta K \cdot L \\
\tau = \frac{VOT \cdot \Delta d}{N}
\end{align*}
\]

The spread toll has been determined by means of an iterative process where at the end of each simulation, the delays and the time-varying toll were derived and updated until significant drops of production were eliminated. Tolls have 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.

4.2.3 The Spread Toll

The third scheme 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 it is possible to see from Figure 9, 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” MFD 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.

![Figure 8: Macroscopic derivation of delays inside the cordon. Adapted from Geroliminis and Daganzo (2008)](image-url)
Figure 9: Changes of production and corresponding changes of accumulation and spatial spread of density

The Spread Toll explicitly considers this issue and it applies a charge that internalizes only the delay due to the increase of density. The resulting fares and improvements will be lower than the other two tolling schemes 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 “fairer” approach as it “punishes” users only for the actual drop of performance they determine by entering the cordon. For additional insights about the economic impacts of the alternative tolling schemes please refer to Simoni (2013).

In order to derive the Spread Toll, the magnitude of the drop in throughput specifically determined by the increases in density and spread needs to be identified. Following the example of Knoop and Hoogendoorn (2013), the following polynomial form is adopted to express the relationship between accumulation, spread of density and production:

$$Q(k, \gamma) = a \cdot k + b \cdot k^2 + c \cdot k^3 + d \cdot \gamma$$ (13)

Where $Q$ corresponds to the total production, $k$ corresponds to the accumulation and $\gamma$ corresponds to the 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 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,

$$a = 127.6; b = -5.61; c = 0.082; d = -11.75$$

with RMSE=13.25 and coefficient of determination $\hat{\rho}^2 = 0.69$. Although the $\hat{\rho}^2$ is not particularly high, the function gives a reasonable approximation (around 8% of average error) for congested and nearly congested traffic conditions.

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 (g)MFD. Further research will be needed to identify a sound 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 $\nabla Q(k, \gamma) = \frac{dq}{dk} + \frac{dq}{d\gamma}$, to identify the variations of outflow “strictly” due to the variations in density and spread. Hence, a toll aimed to internalize solely the decrease in performance determined by additional users can be identified. The initial toll is updated with the same mechanism of the Step Toll until all the delays determined by additional users are eliminated.
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 gMFD, but require further validation before the results can be used to support policy decisions.

The final Flat Toll corresponds to an amount of 1.5 CHF from 06:30 to 08:30 and 2 CHF from 16:30-19:00. The final Step Toll reaches a maximum amount of 1.3 CHF between 07:30-08:00 and 2.5 CHF between 18:00-18:30. The final step toll accounting for spread (Spread Toll) entails significantly lower fares (less than 1 CHF) and it applies during shorter intervals (07:00-08:00 and 18:00-19:30). The resulting fares are in line with the expectations, especially the Spread Toll that is by definition a “milder” approach when drops of performance are determined by clusters of congestion. The time period between 19:00-19:30 where the Spread Toll is higher than the Step Toll represents the only exception and it can be explained by different behavioral responses of drivers in the previous intervals. It is worth mentioning that the levels of fares in the
evening peak are higher than in the morning peak, even though the overall levels of traffic are lower. This outcome is the result of the different nature of congestion of the two peaks. The morning peak is characterized by more severe congestion, but also by more homogeneous traffic conditions than the evening peak. The fact that the network performance deteriorates, regardless of the overall increases in accumulation during the evening peak, then burdens fewer users more heavily.

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.

5.1 Accumulation-production relationships

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 seems to yield higher performances of the network, as it does not show any congested branch (Figure 12b). However, upon closer inspection it presents differences from the Flat Toll scheme. The free-flow branch presents higher scatter that can be explained by hysteresis phenomena characterized by frequent loading and unloading cycles (Figure 13). At every cycle the performance of the system seems to progressively deteriorate. 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.

The accumulation-production diagram resulting from the Spread Toll still exhibits considerable decreases in performance (Figure 12c). All in all the traffic improvements seem to be less than that for the other two tolling schemes. A more careful analysis reveals that the combination of hysteresis and increases in spread of

Figure 12: Accumulation-production relationship determined by the Flat Toll (a), Step Toll (b) and Spread Toll (c) and resulting accumulation variations during the day (d).
density play an important role in the drop in performance. Such a result appears reasonable as the overall amount of the charge is lower than in the other two schemes and it raises the question whether it would be preferable to apply a “stricter” scheme with higher traffic performances or “milder” ones with lower improvements. Further investigations concerning the economic and social welfare impacts of the three charges are discussed in Simoni (2013).

As to the accumulation, Figure 12d shows that both the morning and evening peaks are significantly smoothened by the 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 Spread Toll produces a significant decrease in accumulation only during the morning peak, while no appreciable reduction is achieved during the evening peak.

Figure 12d: Accumulation comparison between the three tolling schemes

5.2 Traffic performance enhancements

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).

Table 2: Traffic enhancements determined by the alternative tolling schemes

<table>
<thead>
<tr>
<th></th>
<th>flat toll</th>
<th>step toll</th>
<th>spread toll</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic efficiency (veh*km/hours) m.p. change (%)</td>
<td>29.4</td>
<td>26.7</td>
<td>-5.8</td>
</tr>
<tr>
<td>Traffic efficiency (veh*km/hours) e.p. change (%)</td>
<td>-17.9</td>
<td>-21.6</td>
<td>-21.2</td>
</tr>
<tr>
<td>Traffic demand (veh-km) m.p. change (%)</td>
<td>-13.6</td>
<td>-9.7</td>
<td>-7.7</td>
</tr>
<tr>
<td>Traffic demand (veh-km) e.p. change (%)</td>
<td>-21.7</td>
<td>-20.9</td>
<td>-13.0</td>
</tr>
<tr>
<td>Travel delays (veh<em>loss</em>hours) m.p. change (%)</td>
<td>-89.9</td>
<td>-75.9</td>
<td>-26.8</td>
</tr>
<tr>
<td>Travel delays (veh<em>loss</em>hours) e.p. change (%)</td>
<td>-71.2</td>
<td>-44.0</td>
<td>54.1</td>
</tr>
<tr>
<td>Heaviness of congestion (km*hours) m.p. change (%)</td>
<td>-85.1</td>
<td>-73.8</td>
<td>-47.8</td>
</tr>
<tr>
<td>Heaviness of congestion (km*hours) e.p. change (%)</td>
<td>-83.8</td>
<td>-73.8</td>
<td>-12.3</td>
</tr>
</tbody>
</table>

1 m.p stands for morning peak
2 e.p. stands for evening peak
In order to investigate the optimal utilization of the network intended as a trade-off between network utilization and network performance, the indicator Traffic efficiency is calculated adopted from Brilon et al. (2005) as:

\[ E = Q \cdot U \cdot T \]  

(14)

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 \( V \) 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 an 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).

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 (even comparable to those from the Flat Toll in the case of Step Toll) by favoring the rescheduling of trips rather than the modal shift. This result is consistent with studies based on the bottleneck model (Arnott et al., 1993; Van den Berg, 2012) and on queue-based models (de Palma et al., 2005), where the time varying toll resulted more beneficial than the uniform 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 reach achieve substantial benefits. Here only the Step Toll, which determines a similar decrease of traffic demand to the Flat Toll, produces considerable traffic improvements.

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 Spread Toll, which on the contrary addresses only those delays ascribable to increases in accumulation, becomes totally inefficient.

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 Spread 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 similar to the case of the Step Toll.

### 6. Concluding remarks

In the first part of this study, we shed light on the dynamics of spatio-temporal congestion patterns in large urban networks characterized by inhomogeneous traffic conditions by means of an agent-based model. In
particular, it emerged that the spatial spread of congestion, corresponding to clustering of congestion can affect the performance of the network and determine considerable drops of production, even for moderate levels of demand. This is one of the main causes of scatter in the accumulation-production diagram. In order to account for this property, a generalized macroscopic fundamental diagram (gMFD) that includes the spatial spread as an additional dimension has been derived. Furthermore, the occurrence of the phenomenon of hysteresis and its dependence to the inhomogeneous conditions of congestion has been confirmed by additional investigations. These findings are in line with previous studies that found urban systems to be hysteretic, dynamic and path dependent systems (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 case), since frequent cycles of network loading-unloading seem to deteriorate its production. In the second part of this study, we derived congestion pricing schemes that explicitly consider the identified macroscopic properties of congestion within an activity-based transport model. Two alternative schemes controlled by macroscopic traffic properties of the network have been implemented in order to integrate fundamental aspects from the classic economic theory, such as the marginal cost pricing, with the macroscopic dynamic features of traffic networks. This approach allowed higher levels of realism and higher consistency with the Traffic Flow Theory than traditional analytical methods. 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). The Spread Toll instead, did not achieve substantial improvements because of the lower fare. However, this limitation might be obviated by linearly increasing the levels of fares or assuming higher value-of-time. The most important finding lies in the lack of efficiency of all the typologies of cordon-based schemes in addressing congestion issues when the distribution of traffic (inside the cordon) is not even.

This study has led to a series of speculations of practical relevance as well. Above all, understanding and considering the phenomenon of spatial spread of density clearly appeared as an overriding concern when implementing traffic and mobility management policies. The inability of traditional cordon-based schemes to cope with clustering of congestion represents a main limitation of this demand management measure. As a consequence, it is strongly advisable to combine it with traffic control measures aimed at spreading users more evenly over the network like gating, traffic signal control, variable-message signs and in-vehicle route guidance systems to achieve larger improvements. This way future studies may consider an integrated approach where tolling schemes are optimized jointly with other traffic flow control and pricing measures, such as parking and public transit fares. If the pricing solution is pursued, then approaches based on the partitioning of network in homogeneous areas or similar categories of road might be more effective. Factors related to the urban morphology like the road network and urban layout might affect the distribution of traffic and ultimately the effectiveness of this demand management measure. Furthermore, the influence of fluctuations of demand that are responsible for hysteresis loops and decreases in performance suggests that a smoothening control of traffic demand might be more beneficial from a traffic flow perspective.

Designing congestion-pricing charges through macroscopic traffic variables seems to be a solid and rather practical approach as it allows the control of large complex networks by means of few indicators. The main barrier to the implementation of the proposed schemes consists in the collection of data (loop detectors or floating car data) necessary to build a (g)MFD, since monitoring resources are often scarce and many cities do not have access to the large amount of data required to build it. However, efforts are currently being made in developing techniques to derive an accurate MFD from a limited fraction of links of the network and the results seem to be encouraging (Ortigosa et al., 2013).

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

References


Simoni, M.D. 2013. Congestion pricing schemes controlled by the gMFD: a comprehensive design and appraisal to bridge the engineering and economic perspective. Master Thesis, Delft University of Technology. Accessible from repository of Delft University of Technology (repository.tudelft.nl)


