Route Advice Based on Subnetwork Accumulations —
Control based on the macroscopic fundamental diagram

Victor L. Knoop, Delft University of Technology, v.l.knoop@tudelft.nl
Hans (J.W.C.) van Lint, Delft University of Technology, j.w.c.vanlint@tudelft.nl
Serge P. Hoogendoorn, Delft University of Technology, s.p.hoogendoorn@tudelft.nl
Keywords: Macroscopic Fundamental Diagram, Traffic Control, Routing

1 Introduction

Whereas research into (and application of) freeway traffic control in the previous century predominantly focussed on local applications (e.g. ramp metering), in the 21st century the focus of both researchers and practitioners is firmly on the coordination of traffic control over corridors (see e.g. Kotsialos et al. (1997); Papamichail et al. (2010)) or even larger mixed motorway-urban networks, as for instance presented by Van den Berg et al. (2004). Network control in urban networks have been around a little longer, e.g. (Lowrie, 1982; Robertson and Bretherton, 1991; Dinopoulou et al., 2006), but hybrid and hierarchical control of mixed networks is one of the biggest challenges in the coming years.

As an alternative approach to centralised or fully communicating traffic control systems, one can introduce multi-level control. This limits the amount of information needed. The control on the lower level can be detailed with detailed information. However, the control the higher level, only aggregate information of the lower level is used and and decentralized control architectures on the basis of escalation and coordination are used. As long as traffic conditions are mild and local controllers do their work sufficiently, no coordination is required. As soon as this is the case, local controllers escalate “control” to a control agent on a higher level (e.g. a corridor or subnetwork), which in turn answers with instructions based on the traffic conditions on that level.

The information needs in this case may be much lighter than in a centralized model. Theoretically, the information needs may be even very light. The macroscopic fundamental diagram (MFD) or network fundamental diagram as purported by Daganzo (2007) and Geroliminis and Daganzo (2008) summarizes the state of an entire traffic network into just two (in principle measurable) quantities: the accumulation and production of a traffic network. In case an area – which might be a network under control or a subnetwork thereof – reaches a critical accumulation, a (sub)network level control agent may communicate this to its neighboring agents, which in that case could instruct their lower level agents to adapt their controllers such as to minimise the inflow into this (sub)network which operates at its limits. Unfortunately, the quantities in the macroscopic fundamental diagram are not easily acquired, since these would necessitate a dense network of sensors on all links in a network. Moreover, the two-state representation may not tell the entire story. Recent work by Cassidy et al. (2011) suggests that only in case networks are evenly loaded, the macroscopic fundamental diagram provides an informative measure for the state in such a network. Section 2 gives an overview of the work on the MFD up to now.

In this paper we explore how alternative but still “light” information may be utilised for traffic control, and in particular the effectiveness of routing. This information includes the standard deviation of vehicle accumulation and incorporates intrinsically the spatial distribution of congestion in a (sub) network. We do so by simulating a network and applying different control strategies. The experimental setup is given in section 3, and the routing strategies in section 4. Section 5 presents the results and the paper concludes by the conclusions and the discussion in section 6.

2 Literature review

In the past five years the theory of a macroscopic fundamental diagram (MFD) has been developed. Concepts were already proposed by Godfrey (1969), but only when Daganzo (2007) reintroduced the concept, more studies
Traffic and Granular Flow 2011, Moscow

Table 1: Overview of the papers discussing the macroscopic fundamental diagram

<table>
<thead>
<tr>
<th>Paper</th>
<th>data</th>
<th>network</th>
<th>insight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daganzo (2007)</td>
<td>theory</td>
<td>none</td>
<td>Overcrowded networks lead to a performance degradation – the start of the MFD</td>
</tr>
<tr>
<td>Geroliminis and Daganzo (2008)</td>
<td>real</td>
<td>Yokohama</td>
<td>MFDs work in practice, and there is a relation between the average flow (production) and the arrival rate (performance)</td>
</tr>
<tr>
<td>Daganzo and Geroliminis (2008)</td>
<td>data &amp; simulation</td>
<td>Yokohama &amp; San Francisco</td>
<td>The shape of MFDs can be theoretically explained</td>
</tr>
<tr>
<td>Buisson and Lavier (2009)</td>
<td>real</td>
<td>urban + urban motorway</td>
<td>There is scatter on the FD if the detectors are not ideally located or if there is inhomogeneous congestion</td>
</tr>
<tr>
<td>Ji et al. (2010)</td>
<td>simulation</td>
<td>urban + motorway</td>
<td>Hybrid networks give a scattered MFD; inhomogeneous congestion reduces flow, and should therefore be considered in network control</td>
</tr>
<tr>
<td>Cassidy et al. (2011)</td>
<td>real</td>
<td>3 km</td>
<td>MFDs on motorways only hold if stretch is completely congested or not; otherwise, there are points within the diagram</td>
</tr>
<tr>
<td>Mazloumian et al. (2010)</td>
<td>simulation</td>
<td>urban grid – periodic boundary</td>
<td>Spatial variability of density is important in deriving the performance</td>
</tr>
<tr>
<td>Geroliminis and Ji (2011)</td>
<td>real</td>
<td>Yokohama</td>
<td>Spatial variability of density is important in deriving the performance</td>
</tr>
<tr>
<td>Knoop et al. (2011)</td>
<td>simulation</td>
<td>urban grid – periodic boundary</td>
<td>Standard deviation of subnetwork accumulation can be used as measure for spread of congestion</td>
</tr>
<tr>
<td>Wu et al. (2011)</td>
<td>real</td>
<td>900m arterial</td>
<td>There is an arterial fundamental diagram, influenced by traffic light settings</td>
</tr>
<tr>
<td>Daganzo et al. (2011)</td>
<td>simulated</td>
<td>grid</td>
<td>Equilibrium states in a network are either free flow, or heavily congested. Rerouting increases the critical density for the congested states considerably.</td>
</tr>
<tr>
<td>Gayah and Daganzo (2011)</td>
<td>simulated</td>
<td>grid/bin</td>
<td>Hysteresis loops exist in MFDs due to a quicker recovery of the uncongested parts; this is reduced with rerouting.</td>
</tr>
</tbody>
</table>

started. This section, similar to the literature overview in Knoop and Hoogendoorn (in print), gives an overview of the most important publications on the MFD. Table 1 gives an overview.

The best-known studies are the ones by Daganzo (2007) and Geroliminis and Daganzo (2008). Geroliminis and Daganzo (2008) show the relationship between the number of completed trips and the performance function which is defined as a weighted average of the flow on all links. This means that the network performance can be used as a good approximation of the utility of the users for the network, i.e., it is related to their estimated travel time. Furthermore, after some theoretical work, Geroliminis and Daganzo (2008) were the first to show that MFDs work in practice. With pioneering work using data from the Yokohama metropolitan area, an MFD was constructed with showed a crisp relationship between the network performance and the accumulation.

Also, theoretical insights have been gained over the past years. Daganzo and Geroliminis (2008) have shown that rather than to find the shape of the MFD in practice or by simulation, one can theoretically predict its shape. This gives a tool to calculate the best performance of the network, which then can be compared with the actual network performance.

One of the requirements for the crisp relationship is that the congestion should be homogeneous over the network. Buisson and Lavier (2009) were the first to test the how the MFDs change if the congestion is not homogeneously distributed over the network. They showed a reasonably good MFD for the French town Toulouse in normal conditions. However, one day there were strikes of truck drivers, driving slowly on the motorways, leading to traffic jams. The researchers concluded that that leads to a serious deviation from the MFD for normal conditions. The inhomogeneous conditions were recreated by Ji et al. (2010) in a traffic simulation of a urban motorway with several on-ramps (several kilometers). They found that inhomogeneous congestion leads to a reduction of flow. Moreover, they advised on the control strategy to be followed, using ramp metering to create homogeneous traffic states. Cassidy et al. (2011) studied the MFD for a motorway road stretch. They conclude, based on real data, that the MFD only holds in case the whole stretch is either congested or in free flow. In case there is a mix of these conditions on the studied stretch the performance is lower than the performance which would be predicted by the MFD.

The effect of variability is further discussed by Mazloumian et al. (2010) and Geroliminis and Ji (2011). Contrary to Ji et al. (2010), both papers focus on urban networks. First, Mazloumian et al. (2010) show with simulation that the variance of density over different locations (spatial variance) of density (or accumulation) is an important aspect to determine the total network performance. So not only too many vehicles in the network in total, but also if they are located at some shorter jams at parts of the networks. The reasoning they provide is that “an inhomogeneity in the spatial distribution of car density increases the probability of spillover, which substantially
This finding from simulation and reasoning is confirmed by an empirical analysis by Geroliminis and Ji (2011), using the data from the Yokohama metropolitan area. Knoop et al. (2011) shows that the standard deviation of subnetwork accumulation can also be used as approximation for the spread of congestion. In a separate paper (Knoop and Hoogendoorn, in print) the shape of the MFD as function of accumulation and spread of congestion is discussed.

A final theoretical explanation for the phenomenon of the influence of the spatial variance of the accumulation is given by Daganzo et al. (2011). He shows that turning at intersections is the key reason for the drop in performance with unevenly spread congestion. Gayah and Daganzo (2011) then use this information by adding dynamics to the MFD. If congestion solves, it will not solve instantaneous over all locations. Rather, it will solve completely from one side of the queue. Therefore, reducing congestion will increase the spatial variance of the accumulation and thus (relatively) decrease the performance. This means that the performance for a system of dissolving traffic jams is under the equilibrium state, thus under the MFD. This way, there are hysteresis loops in the MFD, as also noted by Ji et al. (2010). Note that these loops are an effect by themselves and are different from for instance the capacity drop (Hall and Agyemang-Duah, 1991; Cassidy and Bertini, 1999).

### 3 Model

This research uses a similar simulation setup as is used in Knoop and Hoogendoorn (in print). For completeness, this section restates this setup. It first describes what will be simulated in terms of network and demands. Then, section 3.2 describes the model used for this simulation. Section 3.3 describes the output of the simulator that is used later in the paper. The routing, which is different from Knoop and Hoogendoorn (in print) is discussed separately in section 4.

#### 3.1 Experimental settings

In the paper an urban network is simulated, since this is the main area where MFDs have been tested. We follow Geroliminis and Ji (2011) and choose a Manhattan network with periodic boundary conditions. This means that the nodes are located at a regular grid, for which we choose a 16x16 size. Then, one-way links connect these nodes. The direction of the links changes from block to block, i.e. if at $x=2$ the traffic is allowed to drive in the positive $y$ direction, at $x=1$ and at $x=3$ there are one-way roads for traffic to drive in the negative $y$ direction.

Furthermore, periodic boundary conditions are used, meaning that a link will not end at the edge of the network. Instead, it will continue over the edge at the other side of the network. An example of a smaller grid network with periodic boundary conditions is given in figure 1. Traffic can continue in a direct line from node 13 to node 1.
Table 2: The variables used

<table>
<thead>
<tr>
<th>Symbol</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>Node</td>
</tr>
<tr>
<td>c</td>
<td>Cell in the discretised traffic flow simulation</td>
</tr>
<tr>
<td>L_c</td>
<td>Length of the road in cell c</td>
</tr>
<tr>
<td>q_c</td>
<td>Flow in cell c</td>
</tr>
<tr>
<td>k_c</td>
<td>Density in cell c</td>
</tr>
<tr>
<td>φ_{ij}</td>
<td>Flux from link i to link j</td>
</tr>
<tr>
<td>S</td>
<td>The supply of cell c</td>
</tr>
<tr>
<td>D</td>
<td>The demand from cell c</td>
</tr>
<tr>
<td>i</td>
<td>The links towards node r</td>
</tr>
<tr>
<td>j</td>
<td>The links from node r</td>
</tr>
<tr>
<td>C</td>
<td>The capacity of node r in veh/unit time</td>
</tr>
<tr>
<td>α</td>
<td>The fraction of traffic that can flow according to the supply and demand</td>
</tr>
<tr>
<td>β</td>
<td>The fraction of traffic that can flow according to the demand and the node capacity</td>
</tr>
<tr>
<td>γ</td>
<td>The fraction of the demand that can flow over node r</td>
</tr>
<tr>
<td>Ψ</td>
<td>The split for a destination s at a node r</td>
</tr>
<tr>
<td>t</td>
<td>Time period</td>
</tr>
<tr>
<td>Π</td>
<td>Number of iterations in the probit assignment</td>
</tr>
<tr>
<td>ω</td>
<td>Path</td>
</tr>
<tr>
<td>π</td>
<td>Iterations in the probit process</td>
</tr>
<tr>
<td>αω_j</td>
<td>Link incident matrix: 1 if link j in pad ω, 0 otherwise</td>
</tr>
<tr>
<td>κ</td>
<td>Extent to which drivers adapt their routing due to new information</td>
</tr>
</tbody>
</table>

1 or from node 5 to node 8. This way, all nodes have two incoming and to outgoing links and network boundaries have no effect.

The destinations are randomly chosen from all points in the network. In the network, there are 19 nodes chosen as destination nodes. There are no origin nodes. Instead, at the beginning of the simulation, traffic is put on the links. Vehicles are assigned to a destination, and for this distribution is equal over all destinations.

When the cars have reached their destination, they will not leave the network, but instead they are assigned a new destination. We use a macroscopic model (see section 3.2), hence we can split the flow of arriving traffic equally over the 18 other destinations. The number of cars in the network is hence constant. This number will be a parameter setting for the simulations, but throughout one simulation, it is constant. The demand level is expressed as the density on all links at the start of the simulation, as fraction of the critical density. Figure 3a shows the network used under initial conditions.

3.2 Traffic flow simulation

This section describes the traffic flow model. The variables used in this section and further in the paper are listed in table 2. For the traffic flow modelling we use a first order traffic model. Links are split into cells with a length of 250 meters (i.e., 4 cells per link). We use the continuum LWR-model proposed by Lighthill and Whitham (1955) and Richards (1956) that we solve with a Godunov scheme (Godunov, 1959). Lebacque (1996) showed how this is used for traffic flows, yielding a deterministic continuum traffic flow simulation model. The flux from one node to the next is basically restricted by either the demand from the upstream node (free flow) or by the supply from the downstream node (congestion):

\[ \phi_{c,c+1} = \min \{D_c, S_{c+1}\}; \]  
(1)

At a node r we have inlinks, denoted by i which lead the traffic towards node r and outlinks, denoted by j which lead the traffic away from r. At each node r, the demand D to each of the outlinks of the nodes is calculated, and all demand to one link from all inlinks is added. This is compared with the supply S of the cell in the outlink. In case this is insufficient, a factor, α, is calculated which show which part of the demand can continue.

\[ \alpha_r = \arg\min_{[j \text{ leading away from } r]} \left\{ \frac{S_j}{D_j} \right\} \]  
(2)

This is the model developed by Jin and Zhang (2003). They propose that all demands towards the node are multiplied with the factor α, which gives the flow over the node.

This node model is adapted for the case at hand here. Also the node itself can restrict the capacity. In our case, there are two links with a capacity of 3000 veh/h as inlinks and two links with a capacity of 3000 veh/h as outlinks. Since there are crossing flows, it is not possible to have a flow of 3000 veh/h in one direction and a...
flow of 3000 veh/h in the other direction. To overcome this problem, we introduce a node capacity (see also for instance Tampère et al. (2011)). The node capacity is the maximum of the capacities of the outgoing links. This means that in our network, at maximum 3000 veh/h can travel over a node. Again, the fraction of the traffic which can continue over node $r$ is calculated, indicated by $\beta$:

$$\beta_r = \frac{C_r}{\sum_{i \in o_r} D_i}$$ (3)

The demand factor $\gamma$ is now the minimum of the demand factor calculated by the nodes and the demand factor due to the supply:

$$\gamma = \min \{\alpha_r, \beta_r, 1\}$$ (4)

Similar to Jin and Zhang (2003), we take this as multiplicative factor for all demands to get to the flux $\phi_{ij}$, i.e. the number of cars from one cell to the next over the node:

$$\phi_{ij} = \gamma D_{ij}$$ (5)

### 3.3 Variables

In this paper, several traffic flow variables will be used. In this section we will explain them and show the way to calculate them.

Standard traffic flow variables are flow, $q$, being the vehicle distance covered in a unit of time, and density, $k$, the number of vehicles per unit road length. The network is divided into cells, which we denote by $c$, which have a length $L_c$. Flow and density in cells are denoted by $q_c$ and $k_c$.

Furthermore, the accumulation $N$ in an area $X$ is the weighted average density:

$$N_X = \sum_{c \in X} \frac{k_c \cdot L_c}{L_c}$$ (6)

Similarly, the production $P$ in an area $X$ is the weighted average flow:

$$P_X = \sum_{c \in X} \frac{q_c \cdot L_c}{L_c}$$ (7)

Since the cell length are the same for all links in the network, the accumulation and production are average densities and flows. Recall that there is a strong relationship between the production and the number of completed trips, as shown by Geroliminis and Daganzo (2008).
Table 3: Characteristics of the routing strategies used

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>1 – Fixed</th>
<th>2 – Speed-based</th>
<th>3 – Subnetwork based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routing type</td>
<td>Destination-specific, node specific split fractions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Update frequency</td>
<td>fixed</td>
<td>15 minutes</td>
<td>15 minutes</td>
</tr>
<tr>
<td>Basis</td>
<td>Distance</td>
<td>Time</td>
<td>distance/(subnetwork speed)</td>
</tr>
<tr>
<td>Model</td>
<td>Analytical</td>
<td>Probit, 3 draws</td>
<td>Probit, 3 draws</td>
</tr>
<tr>
<td>Compliance</td>
<td>100%</td>
<td>50% per round</td>
<td>50% per round</td>
</tr>
</tbody>
</table>

The 16x16 (street) block network is split up into subnetworks. Here, we choose subnetworks of 4x4 nodes, as shown in figure 2.

This paper also studies the variations in densities and accumulations. The standard deviation of the cell density is found by considering all cell densities for one moment in time, and calculate the standard deviation of these numbers. Similarly, the standard deviation of the subnetwork accumulation can be calculated. To this end, the accumulation in all subnetworks is calculated for one moment in time. The standard deviation of this list of numbers is the standard deviation of subnetwork accumulation. Note that this does not involve the internal distribution of the densities within one subnetwork and only the subnetwork accumulations are required. Therefore, the calculation of the standard deviation of the subnetwork accumulation requires much less data than the calculation of the standard deviation of the densities, which requires all cell densities.

4 Control strategies

There are three routing scenarios considered in this paper:

1. Fixed routing
2. Speed-based routing
3. Subnetwork speed based routing

Details of these strategies follow below; a summary of the characteristics of the strategies is found in table 3.

For the initial situation, the route choice is determined based on distance to the destination. Traffic will take the shortest route towards the destination. For intersections where both directions will give the same path length towards a destination, the split of traffic to that direction is 50-50. Note that for the initial conditions the distances are proportional to the times, since traffic is loaded in at under-critical conditions and the traffic is at free flow speeds at the whole network.

For the case with fixed routing, the initial routes are used throughout the whole simulation period. Routing strategies 2 and 3 are adaptive strategies which vary with the travel times in the network. Each 15 minutes there is a route update. Both are based on a probit assignment (Daganzo, 1979; Sheffi, 1985). The strategies differ in the utility function which is used as basis for the probit assignment.

For strategies 2 and 3 there is dynamic information which is used for the adaptive routes. In strategy 2, the routes are determined based on the speed on the links. Strategy 3 uses the average speed in a subnetwork as representative for the speed of all links in the subnetwork.

The travel times which result from these interpretations are disturbed with an error of 10% to mimic user interpretation. For each node, the shortest path (in time) to each of the destinations is determined, indicated with $\tilde{\omega}^*$. On the node, this leads to a single decision: turn or straight, indicated with $\tilde{\Psi}^*$. This process is repeated 3 times, which gives 3 decisions for the route from node $r$ to destination $s$. These all give a split ($\Psi$) at node $r$, which are averaged, which is denoted by $\Psi^+$:

$$\Psi^+_j = \frac{\sum_{i=1}^{n} \tilde{\Psi}^*_i}{n}$$ (8)

Note that all routing variables are destination specific, but the destination index is omitted for reasons of notational simplicity.

Then, this average is averaged with the split vector in the previous time period.

$$\Psi^+_j = (1 - \kappa)\Psi_{t-1} + \kappa\Psi^+_j$$ (9)
5 Results

Figure 3 shows the network state for different routing strategies at different times. The initial state is the same for all routing strategies. This situation, with the vehicles distributed evenly, is depicted in figure 3a. In case of no
routing, the congestion clusters more and more. The reason is that the flow in these areas is low, and the flow in
the lower density area is high. Vehicles from the uncongested, or less strong congested areas, can move quicker
and reach the area of heavy congestion, thus increasing the area of heavy congestion.

With routing based on speeds, routing 2, the congestion is much more spread, as is depicted in figure 3c and
d. However, in the situation with route advice, the traffic is adaptive and the vehicles will avoid the areas with
the strong congestion. Therefore, there is not such a clustered area of congestion, but there are several local
bottlenecks. These are “fed” by vehicles which are routed around the strong congestion. This situation does not
change considerably after this spread congestion has set in, as the evolution from 1.5 (figure 3c) to 3 hours (figure
3d) shows.

This level of control is only possible with detailed information of traffic speeds at each cell of each link. How-
ever, proper traffic control is possible with much less information, namely subnetwork aggregated information, as
is shown by the network states in figure 3e and f. Also here, the congestion is spread, meaning there are different
bottlenecks active, and thus the performance is the sum of the capacities of these bottlenecks.

Figure 4 show the performance for the three routing strategies. It shows that the situation without routing
degrades to a situation with very low performance quickly, and continously. With routing on a link level, this
process is interrupted each time when a new advice is computed and communicated to the vehicles, every 30
minutes. This is visible in figure 4 by the decreasing performance for one route, followed by a sharp increase of
performance at each half hour after a route advice.

With routing based on subnetwork speeds, the control is almost as effective, but less sensitive to fluctuations.
The performance remains on average slightly under the average level with control based on the full information,
but there are less fluctuations. One reason is that the subnetwork average speeds, being the basis for control in this
routing strategy, will not fluctuate quickly.

Another measure of performance might not be the average flow (production, according to Geroliminis and
Daganzo (2008), but the arrival rate. This is what interests the end user in the end the most. There is a good
relationship between the network performance (the average flow) and the arrival rate at the destination, as is
shown in figure 5a. The simulation starts with all users distributed over the network, at undercritical densities. In
this situation, the performance is high, but the arrivals still low since drivers have to drive towards their destination.
This is the line at the right hand side of the figure. This line is equal for all routing strategies since in the first time
period of 30 minutes, the actual routing is the same. The routing strategies can be ordered as follows, in which the
first gives the best result: speed routing – subnetwork routing – no routing. This result holds both for the arrivals
as for the network performance, i.e. the average flow.

As shown by Mazloumian et al. (2010), and further analysed by Knoop and Hoogendoorn (in print), a two-
dimensional macroscopic fundamental diagram exists. The performance depends on the accumulation and the
spatial fluctuation of the density. Note that in the simulation we have a fixed number of vehicles, and thus have
the same accumulation.

We analyse this relationship again for the situation with control, and this is shown in figure 5b. The spatial
fluctuation of the density is expressed as the standard deviation of the cell density. The performance reduces with
an increase of spatial fluctuation of density, similar to what is found by Mazloumian et al. (2010) and Knoop
et al. (2011). However, with routing a higher performance is realised under similar spatial spread of denisites. Generally, that means that the form of the 2-dimensional network fundamental diagram changes as result of the network control.

6 Conclusions and discussion

In this paper the possibilities for traffic control are explored. In particular we focus at the possibilities for routing based on the speed in subnetworks. The first conclusions is that routing advice is effective and can prevent the breakdown of traffic which happens without control. It might not be as effective as routing based on actual speeds of all links, but is much more robust and the network performance varies less.

Secondly, we found that this performance increase can be found with the current speeds. Using the spatial spread of congestion, as suggested by Knoop et al. (2011) is not strictly needed, but might increase of the network performance even further. This is subject of further research.

Thirdly, the macroscopic fundamental diagram, even as function of two variables, accumulation and spatial distribution of accumulation, changes in shape under traffic control.

Acknowledgement This research was sponsored by a IP-CC subsidy from ICTregie/NWO in the project SI4MS, Sensor Intelligence for Mobility Systems, and by the foundation Next Generation Infrastructures in the project JAMS, Joint Approach for Multi-level Simulation.

References


