Traffic management for smart cities

Andreas Allström, Jaume Barceló, Joakim Ekström, Ellen Grumert, David Gundlegård and Clas Rydergren

Book Chapter

N.B.: When citing this work, cite the original article.


ISBN: 9783331944927 (Print), 97833319449241 (eBook)

DOI: http://dx.doi.org/10.1007/978-3-319-44924-1_11

Copyright: Springer

Available at: Linköping University Electronic Press
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-131631
Traffic Management for Smart Cities

Andreas Allström, Jaume Barceló, Joakim Ekström, Ellen Grumert, David Gundlegård, and Clas Rydergren

Abstract  Smart cities, participatory sensing as well as location data available in communication systems and social networks generates a vast amount of heterogeneous mobility data that can be used for traffic management. This chapter gives an overview of the different data sources and their characteristics and describes a framework for utilizing the various sources efficiently in the context of traffic management. Furthermore, different types of traffic models and algorithms are related to both the different data sources as well as some key functionalities of active traffic management, for example short-term prediction and control.

1 Introduction

One of the key components in smart cities of the future is the use of Advanced Traffic Management Systems (ATMS) and Advanced Traveler Information Systems (ATIS) for efficient management and control of traffic flows. The purpose of the ATMS/ATIS is to improve the overall traffic system performance, e.g. reducing emissions, noise and travel times.

In order to manage and control traffic flows, the conditions of the road traffic have to be captured. The road traffic state can be described using speed, flow and density on a specific segment of the road. The length of the segment might vary depending on the geometry of the road. When estimating the traffic state, different types of traffic models are commonly used. However, the models can not include all aspects of the real system, and in order to have a good representation of reality...
the models have to be combined with measured data of the traffic state, e.g. traffic
counts and speed/travel time measurements.

Today, most existing ATMS/ATIS rely on fixed point (eulerian) measurements
from loop and radar detectors. Eulerian sensors can collect observations in terms of
flow, speed and occupancy\(^1\), but are unable to provide any trajectory based mea-

\(\text{\textsuperscript{1}} \text{Road density is commonly approximated from measured occupancy.}\)

\[\text{\textsuperscript{1}} \text{Road density is commonly approximated from measured occupancy.}\]
2 Sensors

Traditionally, eulerian sensors have been used for road traffic observations. Sensors like loop detectors and radars are widespread in many major roads in cities throughout the world. These sensors observe or estimate one or several of flow, speed and occupancy for a specific location with a very high penetration rate. The last decades significant efforts have been made in the area of langrangian sensors and in many cities travel times or point speed observations from GPS devices has been used for real-time traffic state estimation, see e.g. [55, 68]. However, the sensors that enable a bridging between traffic demand models and real-time traffic prediction models are the sensors that maintain a user or vehicle identity over large periods of time and distance. These sensors are for example cellular network data, Bluetooth/Wi-Fi sensors, automatic license plate recognition systems and to some extent GPS equipped vehicles. These are all lagrangean sensors and independent of the traffic data type (e.g. travel times or travel demand) that is to be estimated, the performance depends on four main properties of the probe system; penetration level, sampling strategy, measurement type and measurement accuracy.

The penetration level is the number of equipped vehicles compared to the total number of vehicles in the area of interest. The sampling strategy refers to the frequency with which the probe measurements are recorded and how often these are sent to the traffic information server. The measurement type is mainly related to the type of sensors that the probe is equipped with, e.g. GPS or accelerometer, but also what kind of data that is supported by the transmission protocol. The measurement accuracy refers to which accuracy that can be achieved in each measurement, e.g. positioning accuracy. The importance of the different properties is depending on what type of traffic data that is estimated, but also what kind of traffic application that will make use of the data.

Lagrangean sensors can be used to estimate most traffic data types, like speed, flow, density, travel time and incidents. Some types of traffic data can be measured directly, and some needs to be estimated based on models that relate the different kinds of traffic data. Common for most probe systems available today is that the probes are measuring only the state of the probe itself. Few attempts have yet tried to measure the state of the road section that it traverses, but a good exception can be found in [94]. This means that the traffic state of the road section needs to be estimated based on the measurements made by the probes. Few probe systems are developed explicitly for traffic estimation applications, and thus the sampling strategy, measurement type and measurement accuracy are not decided with traffic estimation applications in mind but for e.g. a fleet management application. Common for most probe systems is that speed and position are available. Travel times can then easily be inferred from consecutive position measurements given that an identity of the probe is included, and traffic flow and density can be estimated from point speed or travel times using the fundamental diagram [104].

An important aspect of all vehicle probe systems is that we typically cannot control when and where we get access to state measurements. For critical traffic applications this can be problematic. Fixed sensors are placed where traffic planners
believe it will be most important and hence we get constant high quality measurements in these places. The main advantage of probe data is the fact that we utilize already existing data, which makes it cost efficient compared to fixed installations.

2.1 Cellular network data

Since the first trial to use cellular network data for road traffic estimation, the CAPITAL project [98], a lot of progress has been made. The CAPITAL project failed due to poor cellular location accuracy. However, since then the available data in the network, as well as the methods to process data, has changed dramatically. The data available in cellular networks related to road traffic estimation is described in detail in e.g. [96]. As for today, numerous projects have shown positive results and indicate a large potential for the data source [7, 100].

Early projects using cellular network data aimed at travel time estimation, mainly based on handover events [7, 47]. Lately, there has been an increasing interest in estimating OD matrices based on this data [20, 103]. A limiting factor of this data type has historically been the difficulty to get access to the data from the cellular operators. Recently Orange released a cellular network data set for research purposes and the interest was very high from researchers and practitioners all over the world [15]. This might increase the knowledge about the potential of the data source, but possibly also make it easier for other cellular operators to share data.

The penetration level for cellular network data depends on how the cellular network data is collected. The main aspect is related to whether cell phones that are (1) making phone calls, (2) that have a data connection or (3) are idle, provide data. For (3) the penetration level is equal to the operators’ market share, for (1) it is equal to a few percent of the vehicles and for (2) the penetration rate is somewhere in between, depending on the users’ data usage. The sampling strategy strongly depends on which type of cellular network and which interface that is monitored, see e.g. [46] for details.

Most of the research so far in the area has been using call detail record data, that is used for billing purposes in the operators’ network. However, this data is only a fraction of the data available in the cellular network. If the personal integrity aspect of using cellular network data can be handled efficiently, it is likely that cellular network data will become an important data source in the close future, both within travel time, traffic flow and OD estimation.

2.2 GPS-equipped devices

Alternative traffic data is available from different probe client types, e.g. navigation systems, fleet management clients and insurance black-box systems. The client type will affect the characteristics of the data that is collected and might introduce bias
in the traffic estimations. For example, professional drivers with good knowledge of the traffic system tend to avoid congested parts of the road network and can hence cause underestimation of the traffic congestion. Some probe client types, e.g., some municipal vehicles, are only available at certain time, other types, e.g., taxis, can use dedicated bus/taxi lanes and hence indicate lower travel times than for regular vehicles. Another possible problem is that some client types, e.g., heavy vehicle fleet management clients and buses are more restricted in terms of speed limits. The probe client type needs to be considered when estimating the traffic state.

The device type can also affect the characteristics of the traffic data. We can for example be relatively sure that a navigation device is located in a vehicle, which is not necessarily true for a cell phone device. Some devices have poor GPS-receivers and the location accuracy reported from these clients can potentially affect the results of travel time estimations, especially if the travel time is calculated over a relatively short distance. The main type of measurements from GPS-equipped vehicles are point speed and travel times. Some devices and/or client types does not support point speed measurements, which reduces the amount of traffic state related information that can be extracted from each probe significantly. If the id of the probe is collected and maintained it is still possible to collect travel times, otherwise we are left to try to estimate density based on probe locations, which with most realistic penetrations rates is challenging to do with reasonable accuracy.

The sampling strategy of the probes is also very important for the road traffic estimation result. A long sampling time causes two main problems. The first is that there is an inherent delay of the measurement equal to the sampling time when travel times are calculated. The second problem is that path inference becomes more difficult the longer the sampling time is. This is especially problematic in dense urban areas with a large number of route choices. Map matching and path inference for low sampling devices are further described in e.g., [61, 91]. The location sampling strategy for GPS-equipped devices is time-based sampling with a typical reporting interval between 30 seconds and two minutes. This type of sampling together with a high measurement accuracy enables a large number of traffic related applications.

### 2.3 Automatic Vehicle Identification

Automatic Vehicle Identification (AVI) is an aggregate name for data collection techniques where vehicle identities are captured in selected locations. The most common systems are based on license plate recognition (LPR) and re-identification of Bluetooth or Wi-Fi physical addresses. Bluetooth is for example used in hands-free devices and for communication between different devices in vehicles. The measurement type for AVI systems is similar as for cellular network data, i.e., space-time tuples with an identifier of the vehicle or the Bluetooth/Wi-Fi device. In Europe, conducted field trials, using Bluetooth technique, indicate a penetration rate around 30% [13, 69].
The measurement error relates to the location error of where the device is captured. For Bluetooth, it is possible to adjust the coverage of the sensors collecting Bluetooth data and research has shown that a lower coverage, which also means a lower number of captured vehicles, decrease the error in the travel time estimation [90]. Also, identifying the MAC address takes about 5 seconds on average but may take up to 10 seconds in some extreme cases [60], which will affect the travel time estimation. However, a number of field trials have reached the conclusion that travel times estimated from Bluetooth detectors, deployed along a motorway or arterial, are comparable to those estimated from GPS and toll tag readers [50, 66, 69, 90].

3 Traffic state estimation and prediction for ATMS/ATIS

The ability of predicting short term evolution of the current traffic state is an important basis for further applications on traffic management and control. Here we focus on methods applicable to ATMS/ATIS, in a traffic network equipped with traffic sensors supplying real-time traffic data. After an appropriate management of the data (e.g. filtering outliers), suitable statistical and traffic models are applied to either generate the local information necessary to estimate and predict the short term evolution of the current traffic state and associated variables, or to generate the input data to more sophisticated traffic models to support wide area management and control policies.

As a consequence of the advent of new sensor technology, which makes massive amount of new traffic data available, there is a need for new techniques and methodologies to combine the vast amount of data from different sources. A plausible technological scenario will combine:

- Point detection with discrete time resolution, as for instance conventional inductive loop detectors, and radars, measuring flows (veh/hour), occupancies (time %), spot speeds (km/hour), traffic mix (% light, heavy vehicles).
- Point detection with continuous time resolution. For examples magnetometers measuring time in/time out on the detector from which flow counts, spot speeds, occupancies or traffic mix can be estimated.
- Advanced AVI detectors recording time tag, vehicle/device identification and downstream re-identification, providing sample counts of travel time measurements.
- Continuous time-space detection. For example, by tracking GPS devices, or connected cars, supplying time tag, position (X, Y, Z coordinates) local speed and heading direction.
- Additionally every cell phone can, besides sharing the position data from its GPS, supply spatio-temporal observations associated with signaling in the cellular network.

From a methodological point of view, the main change from using new sensor technologies, consist on having to deal with massive amount of heterogeneous data.
Thus, in order to efficiently manage and utilize such massive amount of data in an ATMS/ATIS, complex statistical techniques (e.g. Kernel smoothing, Kalman filtering and Bayesian networks) have to be combined with mathematical models of the traffic system. This process is synthesized in the methodological diagram in Figure 1.

![Methodological Diagram](image)

**Fig. 1** Main methodological components for short term prediction of the traffic state.

Approaches for traffic state estimation and prediction can be divided into non-parametric and parametric prediction models. Non-parametric models require both parameters and model structure to be determined from data. Thus, they rely on a vast amount of historical data. Traffic dynamics can be captured although no knowledge of the traffic processes as such is included. Non-parametric models have the property that only already occurred traffic states can be predicted. Parametric models, on the other hand, include parameters with a predetermined structure. Still, the parameters must be calibrated according to empirical data. Parametric models have the property of only describing traffic phenomena which follows from the predetermined relationship between model parameters. Also, they rely on boundary conditions, e.g. traffic demand, which need to be predicted for the entire prediction horizon.

The remainder of this section gives an overview of models, filtering of traffic data, assimilation of traffic data with traffic models and fusion of heterogeneous types of data.

### 3.1 Non-parametric models

Non-parametric models are created from large amount of historical data, and make use of big data analysis approaches, such as linear time series, K-nearest neighbors, locally weighted regression, Fuzzy logic, Bayesian networks and Neural networks. In [57], an overview of commonly used methods is given. It is important to recognize that although non-parametric models can capture the traffic dynamics even though no knowledge of the traffic processes as such is needed, they all inherit the property that only traffic states already occurred can be predicted. Thus, they are appropriate for predicting recurring traffic conditions, but less appropriate for non-recurring
traffic conditions, such as congestion due to road work, events or incidents. Also, since they lack any knowledge of the traffic processes, they are of less interest when one wants to evaluate traffic management strategies.

### 3.2 Parametric models

Parametric models applied to traffic planning and estimation problems are commonly related to as “traffic models” or “transportation models” in the literature, and are based on mathematical modeling of the transportation systems. Different models describe the transportation system with different level of detail, depending of their application areas. Here we will focus on models appropriate for application within an ATMS/ATIS. Thus we will solely describe dynamic modeling approaches, which can describe both the spatial and temporal evolution of congestion. Static modeling approaches can still be relevant for an ATMS/ATIS in terms of providing an initial estimate of the OD matrix.

Dynamic traffic assignment approaches decompose the problem of route choice and network loading into two main components:

1. A method for determining the flow rates on the paths in the network
2. A dynamic network loading method, which determines how these path flows give raise to time-dependent link volumes, link travel times and route travel times.

The difference between pure dynamic traffic assignment (DTA) and dynamic user equilibrium (DUE) depends on how the first component is implemented, but both are based on a route choice model. In DTA, paths and path proportions are selected at each time interval according to a discrete choice model, that does not guarantee equilibrium, while DUE uses an iterative procedure to solve the variational inequality formulation of equilibrium [25].

In the network loading phase, the evolution of congestion can be modeled in detail by microscopic approaches, in which each single driver-vehicle unit is described in detail, with position, speed, acceleration, and driver behavior. When the density of a road segment increase, the interactions among the vehicles will result in lower speeds and either density or mean travel cost can be used as a measure of the level of congestion. There are also macroscopic approaches, in which the dynamic traffic conditions are described by aggregated measures (e.g. flow, density and speed) as a function of both space and time. Macroscopic models are based on mathematical relationships between flow, density and speed, and these relationships are used to describe how the traffic state will evolve in both space and time. This group of models include the Lighthill-Whitham and Richard (LWR) model [73], and its discretized version, the cell transmission model (CTM) [29]. A third group of dynamic modeling approaches can be placed somewhere between microscopic and macroscopic models, and are referred to as mesoscopic models. They usually model single vehi-
cles or group of vehicles but with simplified description of the interactions between them.

While dynamic modeling approaches provide mechanisms for high resolution modeling of both the spatial and temporal distribution of congestion, it is important to recognize that they heavily rely on the availability of high resolution data. Both in terms of boundary conditions and calibration. Boundary conditions are commonly the OD matrix, and calibration of models can be done using both eulerian or lagrangian measurements.

### 3.3 Filtering, fusion and assimilation

Filtering, fusion and assimilation are all referring to the process of improving the accuracy of a traffic state estimate by interpolation and removal of noise and bias in measurements. The main difference between the different concepts relates to the core methodology of improving the estimate. In plain filtering, there is typically only one modality of measurement included. In fusion, the focus is to combine measurements of different modalities and in assimilation the focus is to combine measurements with the output of a mathematical model. However, whereas assimilation can include multiple measurement types, fusion typically includes a model for the system evolution. In general, to turn the raw traffic data into a state estimate, a number of steps needs to be performed:

**Step 1.** Removing outliers and erroneous measurements, which could induce large errors and biases in the estimate.

**Step 2.** Filling gaps in data, since malfunctioning sensors and the removal of outliers will generate gaps, these gaps must be filled in order to have complete coherent series of observations.

**Step 3.** Fusing measurements from different sources, to extract the richest homogeneous information from the heterogeneous data available.

**Step 4.** Combining the measurements with a system or process model.

In some literature, especially in the area of weather predictions and hydrology, Step 4 is referred to as assimilation [34]. For traffic state estimation, typically Step 1 and 2 are jointly implemented as a preprocessing step and Step 3 and 4 are implemented as a joint fusion and assimilation step. In the first step, knowledge of the physical process that is observed and the sensor that produces the measurements can be used to remove outliers and obviously faulty measurements. In the second step the outliers are replaced with reasonable values. Although both fusion and more complex process models can be used in this step, often very basic system models are used separately for sensors with different modality. In Step 3, sensors with different modality are fused together to improve the estimate. The different modalities require models for translation between them, these are often referred to as measurement models, and an example can be the relationship between flow and density, i.e. the fundamental diagram. Different sensors are also separated in space and their
observations are separated in time, which requires models for the evolution of the state in both the spatial and temporal domain. The macroscopic models described in Section 3.2 are good examples of models that describes the traffic state evolution in space and time and can be used for assimilation.

The Kalman filter was first presented in [64, 65] and has successfully been used for all steps described above. The right part of Figure 1 depicts two examples, the output of a Kalman filter to estimate outliers and replace missing data for traffic flows (upper right part of Figure 1), and a space-time reconstruction using an assimilation technique (lower right part of Figure 1). For a more in-depth description of Kalman filtering, and its extensions, in traffic management and control, the reader is referred to [5, 74].

In [35] a survey is presented which shows that for traffic state estimation and prediction a large variety of methods and models have been evaluated. Autoregressive models, Bayesian frameworks, Kalman filters and neural networks are some of the methods that are mentioned. However, their conclusion is that none of the proposed methods produce accurate estimations and predictions except for some special conditions. The explanation for this is that the dynamics in traffic cannot be formalized by a single procedure. In order to achieve accurate results for a variety of network configurations and data sets available, a combination of different approaches is recommended.

A Single-Constraint-At-A-Time (SCAAT) Kalman filter that uses the single most recent speed measurement from any available sensor, in this case loop detectors and GPS probe data, is applied in [19]. The state is updated based on the characteristics of that particular sensor and the accumulated state estimation from the previous step. The results are promising but the limitation is that only speed measurements can be used.

In [6], different methods for fusing speed from loop detector data with travel times collected with Bluetooth are compared. This was done both on collected data and simulated data. Among the evaluated methods were neural network, measurement fusion Kalman filter, SCAAT Kalman filter, fuzzy integral, ordered weighted averaging and a simple convex combination. The results show that most data fusion techniques improve the accuracy over single type sensor approaches. Moreover, the measurement fusion Kalman filter, which is a multi-sensor multi-temporal Kalman filter [81], and the simple convex combination perform well in all scenarios and do often significantly improve the accuracy of the estimations.

The adaptive smoothing method is presented in [99], which is a nonlinear spatio-temporal low pass filter that uses the information from loop detectors to reconstruct the spatio-temporal traffic dynamics for a certain road network. The proposed method is based on a number of parameters and some of them are traffic flow related such as the propagation velocity in congestion and free-flow conditions. This method was later extended and generalized in [76] to also handle other types of data. The so-called Extended Generalized Treiber-Helbing filter (EGTF) can handle all kinds of data, no matter of their respective spatial and temporal resolution, as long as the data provides a mean to distinguish between free flow and congestion. In [75] the method is compared with piecewise, linear and quadratic speed-based
interpolation methods. As earlier concluded in [58], using kinematic wave theory in combination with filtering technique results in more accurate travel time estimations in comparison with interpolation methods.

In [104] the cell transmission model for velocities (CTM-v) is used together with an ensemble Kalman filter (EnKF) to fuse speed measurements from loop detectors and probe data. This method is modified for prediction and is described in more detail in a case study presented in Section 4.2. Later [80] evaluated almost 1000 scenarios where the CTM-v and EnKF was used to fuse probe and loop detector data and estimate travel times. The results clearly showed that when complementing loop detector data with probe vehicle data, better estimates for travel times were obtained. However, this implementation of the CTM-v and the EnKF could only handle point speed measurements, not travel times. In [45], the use of EnKF was extended to allow for fusion between point speed and travel time measurements and this is further described in the case study in Section 4.2.

### 3.4 Time-dependent OD estimation

The use of traffic models within an ATMS/ATIS raises an additional problem concerning the traffic data processing: the need to estimate time-dependent traffic patterns formulated in terms of OD matrices, which in recent years has become a field of intense research, see for instance [12, 21, 59, 62]. Taking into account the dynamic nature of traffic phenomena, it has been quite natural to develop estimators based on variants of Kalman filter approaches, exploiting real-time traffic measurements. However, although in general all these approaches have proven their robustness in terms of convergence to sound solutions, most of them have so cumbersome computational requirements that they are not applicable to support real-time decisions. Among the various factors determining the computational performance, the quality of the initial OD estimate proved to be critical [11]. A solution to the initialization problem could be to adjust OD matrices exploiting available traffic measurements and using a static off-line approach. Among the static off-line procedures to adjust OD matrices to traffic measurements, the bi-level optimization methods provide the most consistent results [77]. These approaches formulate and solve an upper level optimization problem. At the upper-level the objective function is usually a distance measure including both link counts, measured by sensors and estimated from the updated OD matrix, and the OD information itself, measured by the difference between the updated OD matrix and a target matrix (e.g. from a travel survey). At the lower level a traffic assignment is carried out, that estimates the values of the traffic variables as a function of the adjusted OD at the upper level of the current iteration. Figure 2 illustrates the logic of the process and the information exchange between both levels. The objective function at the lower level depends on what type of assignment that is carried out (e.g. dynamic or static), and is for both the dynamic and static cases a non-differentiable function.
In [27], a stochastic perturbation stochastic approximation (SPSA) method is proposed for solving the upper level problem with a DUE assignment carried out to account for the dynamics of the congestion propagation at the lower level. This approach has later on been improved in [22]. In the formulation proposed by [78] the bi-level problem to solve is:

\[
F(\tilde{g}^k, \tilde{v}^k, \tilde{tt}^k) = \gamma_1 F_1(\hat{g}^k, g^k) + \gamma_2 F_2(v^k, \tilde{v}^k) + \gamma_3 F_3(tt^k, \tilde{tt}^k)
\]

subject to \((\tilde{v}^k, \tilde{tt}^k) = \text{assignment}(\hat{g}^k)\)

\[\hat{g}^k \geq 0.\]

Where \(\hat{g}^k, \tilde{v}^k, \tilde{tt}^k\) are, respectively, the estimates of the OD matrix \(g\), the traffic volumes \(v\), and the travel times between pairs of Bluetooth detectors \(tt\), at iteration \(k\). The functions \(F_1, F_2\) and \(F_3\) are distance functions. The output of this off-line procedure generates an off-line data base of target OD matrices that can be used to initialize an on-line procedure which estimate the real-time OD matrix to be used in dynamic traffic model.

\[\tilde{v}, \tilde{tt} \quad \text{Upper level (Objective function evaluation)} \quad \text{Lower level (Traffic assignment)} \]

\[\tilde{g} \quad \text{Fig. 2 Computational scheme of the bi-level OD estimation approach.}\]

### 4 Case studies of traffic state estimation and prediction

We will present two case studies where new sources of data are used for providing information of both the current and the future traffic state. The first one is related to on-line time dependent OD matrix estimation, here with an example from the city of Vitoria in Spain, combining traditional loop detector data with AVI Bluetooth detectors. The second one is related to fusion of different data sources and assimilation
with a traffic model, to provide both real-time traffic state estimation and prediction for a part of the Stockholm motorway system in Sweden.

4.1 On-line OD estimation in Vitoria

Application of dynamic traffic assignment models in an ATMS/ATIS rely on known information of demands, in terms of OD matrices. Section 3.4 outline a methodology for off-line OD estimation. Here we provide an example of how such an historical (known) OD matrix can be used for updating a current OD matrix, using traffic counts and AVI Bluetooth detectors.

4.1.1 A Kalman filtering approach

A Kalman filter approach, with the recursive linear Kalman filter state space formulation [11, 12], is adapted to exploit traffic counts collected by AVI sensors and conventional detection technologies, and travel times observed by the AVI sensors. The formulation uses deviations of OD path flows as state variables, calculated with respect to DUE based historic OD path flows. A subset of the most likely OD path flows identified from a DUE assignment is used. The number of paths to take into account is a design parameter. A list of paths going through each sensor is automatically built for each AVI sensor from the OD path description, AVI sensor location and the network topology.

The proposed approach initially assumes flow counting detectors and AVI sensors located in a cordon at each possible point for flow entry and AVI sensors located at intersections in urban networks covering links to/from the intersection. Flows and travel times are available from AVI sensors for any time interval greater than one second. Trip travel times from origin entry points to sensor locations are measures provided by the detection layout. Therefore, they are no longer state variables but measurements, which simplify the model and make it more reliable.

The time-varying dependencies between measurements (sensor counts of Bluetooth equipped vehicles) and state variables (deviates of Bluetooth equipped OD path flows), are used for estimating discrete approximations of travel time distributions. Since the approach uses the AVI travel time measurements from Bluetooth equipped vehicles, the non-linear approximations can be replaced by estimates from a sample of vehicles. Then no extra state variables for modeling travel times and traffic dynamics are needed, since sampled travel times are used to estimate discrete travel time distribution, see [11, 12] for details.
4.1.2 Results

Computational experiments were conducted with urban networks of various sizes, see [4] for details. For the sake of completeness we report here those for the medium-size network from the city of Vitoria (Spain), depicted in Figure 3, including 57 centroids, 3249 OD pairs, 2800 intersections and a modeled network of about 600 km. This network resembles a reasonable sized real-life network, with representative congestion levels and route choice dimension as found in many large urban areas. Two different sets of sensors have been identified in the Vitoria network:

- 389 standard loop detectors, located as depicted in the left-hand side of Figure 3, providing flows, speeds and occupancies, related to all detected vehicles at the loop.
- 50 AVI sensors, located as shown in the right-hand side of Figure 3. Notably, the AVI sensors are deployed following a layout strategy, whose details can be found in [9], to optimize the capturing of Bluetooth equipped vehicles and provide effective travel time measurements between AVI sensors.

Various methods were tested in the referenced paper [4] depending on different values for the design factors, such as initial a priori OD estimates, and penetration rates of Bluetooth equipped vehicles. Results for the modified Kalman filtering approach (denoted Method 5 in [4]), assuming 100% Bluetooth penetration rate and prior OD scenarios D7 (low-demand), D8 (medium demand) and D9 (high-demand), are presented in Table 1 for selected goodness of fit measures: normalized root mean squared error (NRMSE), Theil’s U coefficient and $R^2$. Good performance of the modified Kalman filtering approach is reflected in high $R^2$ fit (above 85%) that is obtained for the overall OD pairs and demand levels, but mostly for the most important OD flows (i.e. those in the 4th and 3rd quantiles). The fit of true versus estimated OD flows for all considered OD pairs (for the aggregated one-hour period) and for a scenario initialized with prior demand levels D7, D8 and D9, show coefficients of determination of the simple regression line of almost 90%. Figure 4
presents the evolution of estimated OD flows (for OD pairs 221 and 343 in Figure 4) per departure time interval for prior high OD demand scenario (D9), demonstrating also how Method 5 is able to recover from an initial point characterized by overestimated prior OD flow with respect to the true OD flow.

![Fig. 4](image)

**Fig. 4** Filtered values throughout 1h for prior OD scenario D9 for the modified Kalman filtering approach.

<table>
<thead>
<tr>
<th>Initial matrix</th>
<th>Theil’s coefficient</th>
<th>NRMSE (%)</th>
<th>R² (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q1</td>
<td>Q2</td>
<td>Q3</td>
</tr>
<tr>
<td>D7</td>
<td>0.16</td>
<td>0.15</td>
<td>0.14</td>
</tr>
<tr>
<td>D8</td>
<td>0.15</td>
<td>0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>D9</td>
<td>0.15</td>
<td>0.13</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Overall, results presented show that the potential of information from advanced traffic measurements to improve OD demand estimation can be fully exploited only using OD estimation methods capable to correct for biases in spatial and temporal OD pattern given by prior OD matrix, especially in congested networks. In this respect, the computational experiments presented in this work prove the robustness and quality of the OD estimates exploiting AVI measurements. Computational times of less than two minutes on standard laptop and MATLAB implemented algorithms make them applicable for real-time operations.
4.2 Traffic state estimation and prediction in Stockholm

To provide an accurate estimation of the current traffic condition in Stockholm, a Kalman filtering approach has been adopted for estimation and short term prediction of the traffic state. The filtering approach can be used both for fusion of different data sources, and for assimilation of traffic sensor observations, or predictions of such observations, with macroscopic traffic flow model outputs. Here we present the outline of the methodology, and an evaluation of traffic state estimation and prediction for a 7km long section of the Stockholm motorway, just north of central Stockholm, for traffic going towards the central parts. The motorway section is illustrated in Figure 5, in which each subsection is a link with different characteristics and thus different fundamental diagram parameters. For this section three different type of sensors are available:

- 20 fixed point radar detector, measuring speed and flow. In Figure 5 these detectors are marked by “+”.
- 1500 equipped taxis (for the whole of Stockholm county), sending their position every one to two minute.
- AVI/Bluetooth measurements from seven subsections, providing mean travel times. In Figure 5 the approximate locations of the Bluetooth detectors are marked by “BT”.

In the evaluation process the AVI measurements is mainly used for calibration and evaluation purposes. In this section we give the outline of the core components, and present the main results from the evaluations. For further details on both model components, the calibrations procedure and the results we refer the interested reader to [3, 45].

4.2.1 Core components for traffic state estimation and prediction

An EnKF approach is used as core for data fusion and assimilation with a macroscopic traffic model. This approach was developed for data fusion and traffic state estimation within the Mobile Millennium project [14, 104], and adopts the CTM-v
as system model in the EnKF. Data fusion and assimilation based on the CTM-v and the EnKF is appealing for many reasons; it is developed to run in real-time and for large networks, it can fuse different types of eulerian point speed measurements, and it enables the possibility to include lagrangian travel time measurements.

The state-space model of the system is

\[ v_n = M(v_{n-1}) + \eta^n \]
\[ y_k^n = h_k(v^n) + \chi_k^n \]
\[ \eta^n \sim (\mu_{\text{mod}}, Q^n) \]
\[ \chi_k^n \sim (\mu_{\text{obs}}, R_k^n) \]

where \( v^n \) is the state vector in time step \( n \), including the speed for each part of the road network (cell) according to the spatial resolution of the system. \( M(\cdot) \) is the system model, here the CTM-v. \( y^n \) is the observation vector in time step \( n \) and \( h_k(\cdot) \) is the observation model for observation type \( k \). \( \eta \) and \( \chi \) are the possibly time-varying model uncertainty (with mean \( \mu_{\text{mod}} \) and covariance \( Q \)) and observation noise (with \( \mu_{\text{obs}} \) and covariance \( R \)), respectively.

The observation model for eulerian speed observations is simply a mapping of speed from the point speed sensor location to a cell in the road network. For lagrangian sensors measuring travel time, a corresponding naïve solution would be to allocate the travel time to the cell traversed. This would, however, assume that the travel time is equally distributed over the traversed stretch. A sounder method would be to estimate the most likely speed profile for the travel time observation, given some external input and use that speed profile as input to the traffic estimation. The external input can be information about the road network, traffic models and other observations (see e.g. [45]).

While the standard Kalman filter is designed for linear problems, the EnKF can handle both non-linear and non-differentiable state space equations, which is the case implied by the use of CTM-v. For further reading about the EnKF we refer to [33].

The CTM-v is based on the kinematic wave theory and the discretization [29] of the LWR model [73]. Traditionally, these models use density as a measure of state. However, due to the characteristics of the data available in the Mobile Millennium project, new models were developed that use velocity as the state. The new models are referred to as the LWR-v and the CTM-v. The velocity state estimation problem is solved using the EnKF. For a more in depth description of the models the reader is referred to [104, 105]. The CTM-v require boundary conditions as input, in terms of inflow. This could either be determined by using the OD estimation procedure presented in Section 3.4 and 4.1, or by using a non-parametric model. Here a simple approach has been applied by taking the mean of historical inflows, clustered by day of week, for 15 minute periods.
4.2.2 Results

Two studies on traffic state estimation and prediction have been carried out using the proposed framework.

In [45] the framework is used for data fusion of travel time measurements from taxis with fixed point speed measurements from radar sensors. The results presented in Table 2 show mean average percentage error (MAPE) values, with AVI travel times used as ground truth, for different combinations of radar and taxi data. MAPE values are computed for four days in March 2013. The inclusion of taxi travel times improves the travel time for all evaluated days.

Table 2  Travel time MAPE for different combinations of taxi and radar measurements

<table>
<thead>
<tr>
<th>Scenario</th>
<th>March 19</th>
<th>March 20</th>
<th>March 21</th>
<th>March 22</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTM-v model only. Sources and sinks predicted using historic observations.</td>
<td>0.701</td>
<td>0.583</td>
<td>0.262</td>
<td>0.667</td>
</tr>
<tr>
<td>CTM-v model and taxi observations</td>
<td>0.233</td>
<td>0.228</td>
<td>0.213</td>
<td>0.263</td>
</tr>
<tr>
<td>CTM-v model and two radar observations in start and end of test site.</td>
<td>0.134</td>
<td>0.154</td>
<td>0.178</td>
<td>0.151</td>
</tr>
<tr>
<td>CTM-v model, two radar observations and taxi observations.</td>
<td>0.119</td>
<td>0.126</td>
<td>0.154</td>
<td>0.126</td>
</tr>
<tr>
<td>CTM-v model and all radar observations</td>
<td>0.049</td>
<td>0.067</td>
<td>0.058</td>
<td>0.053</td>
</tr>
<tr>
<td>CTM-v model, all radar observations and taxi observations.</td>
<td>0.049</td>
<td>0.052</td>
<td>0.055</td>
<td>0.047</td>
</tr>
</tbody>
</table>

In [3] the framework is used for short term travel time prediction. Two approaches are evaluated. One is based on running the CTM-v forward in time, from an estimated traffic state, the second approach is a hybrid between a non-parametric prediction of radar sensor measurements and the CTM-v model. Here the EnKF framework is not only used for estimating the current traffic state, but also for combining the two type of predictions. Results, in terms of MAPE values comparing predicted travel times with AVI measurements from the 21st of March 2013, are provided in Table 3. Note that a prediction horizon of 15 minutes corresponds to a trip taking place 15 minutes into the future, making use of predicted travel times 15 to 30 additional minutes into the future, depending on the level of congestion in the network. For 5, 15 and 30-minute prediction horizons, the hybrid prediction has shown better results in comparison with only using the CTM-v. For the presentation of the complete results from the study we refer to [3].
Table 3  Travel time MAPE for EnKF estimation, and for CTM-v and hybrid prediction

<table>
<thead>
<tr>
<th>Prediction/estimation method</th>
<th>Horizon 0</th>
<th>Horizon 5</th>
<th>Horizon 15</th>
<th>Horizon 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation</td>
<td>4.8%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CTM-v prediction</td>
<td>5.1%</td>
<td>9.2%</td>
<td>10.1%</td>
<td>10.9%</td>
</tr>
<tr>
<td>Hybrid prediction</td>
<td>5.5%</td>
<td>6.8%</td>
<td>7.7%</td>
<td>7.9%</td>
</tr>
</tbody>
</table>

5 Active traffic management

If the main objective is to support traffic management and information decisions, the traffic state estimation and prediction techniques discussed in Section 4 are not sufficient. Early research on traffic management and control topics, see for example [16, 17, 28, 67], propose decision support systems whose architectures have two main components: a rule based knowledge system and a dynamic traffic simulation model, usually a mesoscopic or macroscopic model [8]. The integrated corridor management program, prompted by the U.S. Federal highway administration, has contributed to the consolidation of these architectures, whose guidelines can be found in [2].

5.1 Proposal of an logical scheme for an ATMS/ATIS

The ATMS/ATIS logical scheme is depicted in Figure 6, and is split into four components:

A. The off-line generation of candidate target matrices for the initialization of the on-line procedure. It combines an initial OD matrix or similar practical applications with the historical traffic data clustered in profiles for different times of the day (e.g. raining Tuesday from 6:00 am until 8:00 am, and so on). This is likely to come from a static model OD demand calibration procedure. For most metropolitan areas, such models should already be available as they have been part of the strategic planning tools for several decades. A heuristic procedure, as the one described in [10], split the static demand in time slices in order to create an initial OD demand in terms of the traffic profiles. These time slices, along with a network model, are the input to the described OD matrix adjustment process in Section 3.4. The repetitions of the process for the different time intervals generate a database of candidate target matrices off-line.

B. An on-line selection of the most likely OD matrix for the initialization, given the current traffic conditions. The real-time traffic data from the traffic monitoring system provides input used for identification of the profile that best fits the current situation. Resulting in the most likely OD matrix for that time interval is selected.
C. The real-time estimation and prediction of the OD matrix. The selected OD matrix is the initial matrix used by a Kalman filter model, as is illustrated in Section 4.1, to estimate and predict the expected OD matrix for the next time period. This OD matrix will be the input for the traffic state estimation process, to estimate the network state and to predict its short-term evolution.

D. On-line event detection and selection of management strategies. These strategies are evaluated in a traffic model, based on performance indicators such as travel time and emissions.

![Logical scheme of an ATMS/ATIS.](image)

**Fig. 6** Logical scheme of an ATMS/ATIS.

### 5.2 Traffic control strategies

Control of traffic volumes (demand based strategies) and homogenization of traffic flows (supply based strategies) are two key concepts for improving the traffic situa-
tion in a congested traffic network. Concepts based on these control strategies could be part of the management strategies database included in component D in Figure 6. The aim (measured in terms of performance indicators) could for example be to reduce emissions, to increase safety or to maximize throughput. The two concepts address different problems.

Control of traffic volumes can be done by:

1. Pricing and information strategies aiming to shift the demand to less congested time periods, or to other modes of transportation. Examples of such strategies are congestion pricing [31, 32] and real-time traffic information systems [72].
2. Re-routing the traffic to improve system performance, either by providing real-time information direct to the car drivers [84], or by using variable message signs.
3. Gating the inflow to areas sensitive to congestion by using perimeter control, for example using signal timings (see e.g. [30]).

Example of a combination of demand strategy 2 and 3 above is illustrated in Figure 7, which include both re-routing and gating.

![Flow control in an urban area through re-routing and perimeter control.](image)

When the traffic flow on the road is approaching capacity levels, and above capacity levels, the conditions on the road usually become unstable resulting in incidents, big variations in speed levels and resulting stop-and-go conditions. Further, it is a well-known phenomenon that the capacity in unstable conditions tend to drop, resulting in lower throughput levels of the traffic flow than what can actually be achieved. Several empirical studies has identified the drop, see e.g. [26, 95, 110]. Flow harmonization can be used to keep the traffic flow stable and thereby prevent
the capacity drop to occur. Also, when an incident has occurred flow harmonization can be used to prevent further breakdown and to keep the traffic flow stable under the incident situation. Flow harmonization can be achieved by:

1. Ramp metering, which distribute the inflow to a motorway in such a way that the total flow on the motorway does not exceed its capacity. It can also be viewed as a perimeter control for motorways, allowing the traffic at the motorway to flow, and storing the vehicles in the secondary network. However, ramp metering systems commonly discontinue the ramp metering whenever the queue from the ramp spills over to the surrounding traffic network, which is not the case for perimeter control.

2. Variable speed limits, which by altering the speed limits on roads close to congested areas or to an incident, tries to keep the flow levels as high as possible by preventing a breakdown before it has happened.

In the remainder of this section we provide an overview of traffic control strategies based on gating and variable speed limits.

5.2.1 Traffic control on motorways

Ramp metering
Ramp metering tries to limit the on-ramp flow entering a freeway by the use of green-red signal timing. The aim is to keep the traffic flow as close to capacity levels as possible by allowing as much vehicles as possible without causing a drop in capacity. This is similar to perimeter control strategies, and thereby ramp metering can be considered as a local perimeter control strategy.

Numerous ramp metering strategies have been presented in the literature. See [87] for an overview of the different strategies. The ramp metering strategies can be either fixed-time or traffic-responsive. The fixed-time ramp metering strategies derive signal patterns off-line based on historical data which is dependent on the time of the day, and the freeway is divided into segments consisting of at a maximum one on-ramp. A static model is used to derive the traffic flow entering and leaving each segment of the freeway. The traffic-responsive ramp metering strategies are based on real-time measurements and the aim is to keep the traffic conditions close to some user defined values, and the measurements are usually collected through detectors on the road. Traffic-responsive ramp metering strategies can further be divided into local ramp metering and coordinated ramp metering. Local ramp metering strategies make use of traffic measurements collected where the on-ramp is entering the freeway. The flow control could be based on a feed-forward control, see e.g. [79], or feed-backward control as for example ALINEA [86]. Other strategies make use of fuzzy logic such as in [97] and neural network, see e.g. [108]. Coordinated ramp metering tries to optimize the traffic flow over a larger area including multiple on-ramps. The coordinated ramp metering strategies can be divided into multivariable ramp metering strategies and optimal control strategies. The multivariable strategies, see for example [41, 85], make use of measurements from a larger area and try
to coordinate the traffic flow based on all on-ramps included in the area. These algorithms are based on linear equations of the traffic flow models and might therefore be inefficient during congestion. Optimal control strategies are better in modeling the traffic flow but they are usually complex and difficult to implement in field. See [42, 51, 88, 109] for example of optimal control strategies. Due to the complexity of optimal control strategies a third category of coordinated ramp metering has evolved, called heuristic rule-based strategies. An overview of such strategies can be found in [18]. More recent examples are presented in [49, 89].

**Variable speed limit systems**

Many of the existing variable speed limit (VSL) systems [101, 102] are mainly focusing on safety, i.e. incident detection systems. Detector stations are used to measure flow and/or speed. The VSL is communicated to the vehicles via a VSL sign. When an incident occur, measured as very low speed and/or flow levels, the system is triggered by applying lower speed limits. Thereby, the risk of further breakdown is limited and the system tries to resolve the existing congestion. In incident detection systems the speed limit is usually lowered substantially. The speeds in the VSL system can be recommended, as in the Swedish motorway control system [102], or mandatory as in the UK system [56].

For flow harmonization systems the main purpose is to keep the flow levels as high as possible by preventing a breakdown before it has happened. In this case the speed limit is not necessarily activated in low speed and/or flow situations. Activation can occur when there is a high diversity in speed levels among the vehicles, or at high speed levels with high traffic flow. The new speed limit is based on an evaluation of the risk of a breakdown. This can be done by prediction of the future traffic states using current and/or historic information. The speed limit is lowered in small steps compared to the more abrupt change in the incident detection systems.

Combined systems making use of both flow harmonization and incident detection exists in implemented systems, such as in the system used in the United Kingdom [56]. Flow harmonization systems only exist the literature, see for example [23].

VSL systems presented in the literature can be categorized as predictive models and reactive models. A common approach for the predictive models is to make use of model predictive control [36, 82]. In predictive models the traffic evolution is described with a dynamic traffic model, a history of past control actions and an optimization cost function, in order to predict the future traffic states. The cost function is constructed based on the aim and the purpose of the system. The aim could be to minimize the total travel time spent as in [52], or to minimize the emission levels, or a combination of the two as in [106]. In the reactive models only current information about the traffic state given from road detectors, road design and current operational conditions are needed in order to determine a new speed limit. Examples of reactive models can be found in [23, 53, 71].

A simulation based study [44] of four reactive VSL systems show that the accuracy of the information from detector stations, the prediction of the traffic states, the estimation of when an incident occur and over which stretch the incident reaches are essential for the performance of the VSL system. Further, reactions to changed conditions on the road which are not necessarily related to an incident, such as high
variations of speed levels, high flow levels, are also important for the efficiency of the system.

Data from connected vehicles can for VSL applications provide a new source of data. For example, speed, distance to an updated VSL and position on the road, could be used in order to improve the VSL system performance. Also, by having a traffic system with a high rate of connected vehicles the VSL systems does not have to be dependent on detector stations and VSL signs, which are expensive to maintain. Instead, the systems could use communication between vehicles and between vehicles and the infrastructure (and some control center) to give individualized speed limits to the drivers [43]. This could presumably maximize the throughput on the road even further.

**Combinations of ramp metering and variable speed limits**

Ramp metering is an efficient tool to use to increase efficiency during congestion since it limits the flow to the capacity levels of the freeway. The main drawback is the spill-back to other arterial roads due to queue propagation from the on-ramp. This is especially observable during heavy congestion. On the other hand, VSLs cannot reduce the inflow. The goal is to make the best of the traffic situation by trying to maximize efficiency and/or reduce the effect of an incident without changing the conditions of the inflow levels. During moderate congestion the VSLs can usually increase the efficiency, especially if it is set to harmonize the traffic flow rather than to increase safety. Although, for heavy congestion, and for long time periods with congestion, the traffic system tends to become unstable even though VSLs are applied.

Therefore, a combination of ramp metering and VSLs could potentially increase the efficiency on the freeway even further, see e.g. [24, 54, 93, 107]. In order to maximize the throughput even more than when only one measure is applied, the measures can be applied simultaneously, or after each other.

### 5.2.2 Perimeter control in urban networks

Perimeter control has been suggested for reducing congestion in urban areas and to improve travel times. The non-linear nature of congestion, where a few additional vehicles on a road can create a traffic breakdown, is difficult to deal with using local control algorithms, since they only consider local measurements but not the overall network (or subnetwork) performance. Instead, by using signal timings in order to coordinate and control the traffic flow on the border of a controlled region, or in the case of multiple controlled regions, in-between regions, overcrowding of a specific region can be avoided. Thus, the traffic flow can be maintained at capacity. The concept of perimeter control for urban regions was first presented for a single region in [30], and later extended to two regions in [39, 48] and multiple regions in [1, 92].

In perimeter control algorithms, macroscopic fundamental diagrams (MFD), sometime referred to as network fundamental diagrams, plays an important role. They can be viewed as an area-wide extension of fundamental diagrams, and in [38]
it was shown that such a MFD relationship can be found for neighborhood-sized sections of a city. Traditionally, fundamental diagrams describe the relationship between flow, density and speed at a specific road segment. The MFD instead describe the relationship between network (or subnetwork) wide densities and mean flows (or out flows from the subnetwork). Used in an automatic control loop, the MFD provide a bridge between the control, usually signal timing, and the traffic state. The MFD can provide this, without requiring microsimulation of the traffic, which makes it appealing for large scale on-line applications.

A prerequisite to applying perimeter control is a suitable way of defining subnetworks, for which a well behaved MFD can be found [40] (i.e. low scatter MFD). A determining factor for finding an MFD has shown to be areas of a city with homogenous densities on the roads. One methodology for finding such areas, based on network partitioning, is suggested in [63]. To determine the current traffic state in the subnetworks, both fixed location sensors (commonly loop detectors), as well as vehicle probes [37, 83], can be used. This can include both eulerian and lagrangian observations, and estimation of MFDs based on the two types of observations are compared in [70]. Their results highlight that a better estimation of the MFD can be achieved by combining the two types of data.

The key concept of perimeter control includes sensors and a control algorithm. The sensors provide real-time measurements of the density and the control algorithm makes use of these measurement, through the MFD, to determine the most suitable signal timings. Two main control algorithms have been proposed; multivariable feedback regulators, which is an optimal open-loop control, and model predictive control algorithms, which is an optimal closed-loop control. A linear-quadratic multivariable feedback regulator is presented in [1], and model predictive control algorithms are proposed in [39, 92]. While an open-loop control cannot take the error between model and reality, i.e. the MFD scatter, into account, a closed-loop control approach, like the model predictive control approach, can take this error into account. The model predictive control approach also has the benefit of handling noise in both travel demands and the MFDs. For a model predictive control approach a prediction of future traffic demand is required, which could for instance be provided by the Kalman filtering approach presented in Section 4.1.

6 Conclusions

A large number of heterogeneous sensors for road traffic observation are already in place in many cities around the world. The large amount of data from these sensors does not by itself improve information to the road users (e.g. travel time information) or provide means for traffic control to the road authorities. It is through the use of filtering techniques and models that this data enables new possibilities for on-line estimation and prediction of the traffic state, and for wide-area control in urban areas.
The emerging road traffic sensors described in this chapter enable a better understanding of long term travel demand and travel pattern dynamics, but also makes it possible to estimate and predict travel demand and wide-area traffic densities more accurately in real-time. This will enable the introduction of wide-area traffic control, which is a research area that has gained a lot of interest in recent years. Also, by making use of new data sources shared by communication between vehicles, and between vehicles and the infrastructure, so called cooperative systems, there is a great potential to improve the performance of existing traffic control strategies as well.

References

[9] Barceló, J., Gilliéron, F., Linares, M., Serch, O., Montero, L.: Exploring link covering and node covering formulations of detection layout problem. Trans-


[98] Transportation Studies Center: Final evaluation report for the CAPITAL-ITS operational test and demonstration program (1997)


