

Active traffic management on road networks: a macroscopic approach

BY ALEX A. KURZHANSKIY* AND PRAVIN VARAIYA

*Electrical Engineering and Computer Sciences, University of California,
Berkeley, CA 94720-1770, USA*

Active traffic management (ATM) is the ability to dynamically manage recurrent and non-recurrent congestion based on prevailing traffic conditions in order to maximize the effectiveness and efficiency of road networks. It is a continuous process of (i) obtaining and analysing traffic measurement data, (ii) operations planning, i.e. simulating various scenarios and control strategies, (iii) implementing the most promising control strategies in the field, and (iv) maintaining a real-time decision support system that filters current traffic measurements to predict the traffic state in the near future, and to suggest the best available control strategy for the predicted situation. ATM relies on a fast and trusted traffic simulator for the rapid quantitative assessment of a large number of control strategies for the road network under various scenarios, in a matter of minutes. The open-source macrosimulation tool AURORA ROAD NETWORK MODELER is a good candidate for this purpose. The paper describes the underlying dynamical traffic model and what it takes to prepare the model for simulation; covers the traffic performance measures and evaluation of scenarios as part of operations planning; introduces the framework within which the control strategies are modelled and evaluated; and presents the algorithm for real-time traffic state estimation and short-term prediction.

Keywords: active traffic management; macroscopic traffic model; hierarchical control; set-valued estimation

1. Introduction

Traffic congestion is a source of productivity and efficiency loss, wasteful energy consumption and excessive air pollution. Continued travel demand growth and budget constraints have made it difficult for transportation agencies to increase roadway capacity in major metropolitan areas. Active traffic management (ATM) is the ability to dynamically manage recurrent and non-recurrent congestion based on prevailing traffic conditions without capacity growth. It makes use of automated systems and human intervention to manage the traffic in order to maximize the effectiveness and efficiency of road networks. This paper proposes a certain structure of ATM and focuses on some of its components. It stems from the authors' engagement in the Tools for Operational Planning (TOPL) research

*Author for correspondence (akurzhan@cal.berkeley.edu).

One contribution of 10 to a Theme Issue 'Traffic jams: dynamics and control'.

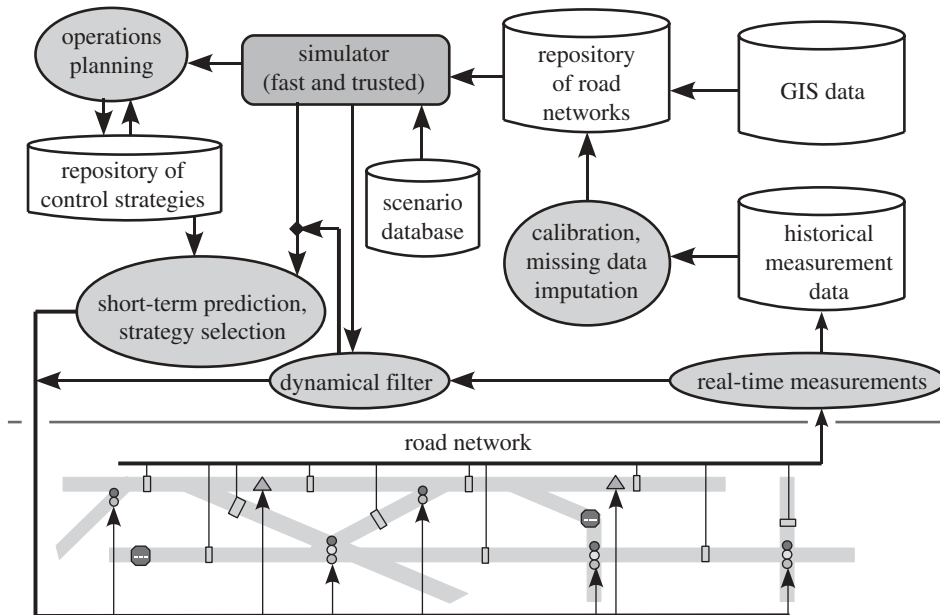


Figure 1. Active traffic management workflow diagram.

project conducted at Berkeley since 2006 (<http://path.berkeley.edu/topl>). TOPL seeks to assist the California Department of Transportation in improving the operation of California freeways.

ATM is the cycle consisting of: (i) continuous traffic measurement and measurement data analysis, without which any attempt to manage a road network would be blind; (ii) operations planning, which includes evaluating the road network performance under various scenarios, such as demand increase, lane closures, special events, etc., developing control strategies that improve performance and testing these strategies in terms of their cost and the benefits they bring; (iii) implementing the most effective of these control strategies by installing necessary hardware and software in the field; and (iv) running the decision support system in real time, which includes filtering the measurement data, providing short-term prediction of the traffic state and selecting the best available control strategy for the next 1–2 h.

The ATM workflow diagram is presented in figure 1. It relies on a communication network that feeds real-time measurements from the field elements in the road network to the traffic control centre and transfers commands generated by the control software (and human operators) to the actuators in the field. The central element of the ATM workflow is the ‘fast and trusted simulator’. The simulator is trusted because it is founded on sound theory of traffic flow; it is parsimonious, only including parameters that can be estimated; and it is tested for reliability. As a candidate for such a simulator, we propose AURORA ROAD NETWORK MODELER (RNM; <http://code.google.com/p/aurorarnm>), an open-source tool set for modelling road networks that can include freeways and arterials with signalized intersections. The underlying *macroscopic* dynamical traffic model and the way it is built using geographical information systems (GISs) and historical measurement data are explained in §2.

The simulator has three modes of operation. In the first or *operations planning* mode, a large number of simulations are run to evaluate scenarios and test potential operational improvements on the already prepared and calibrated road network. Scenarios incorporate known events such as a football game or a confirmed accident, or future events considered plausible on the basis of statistical inference and learning techniques that combine historical data with the current estimate of the state of the system (e.g. likelihood of an accident occurring in a certain location, conditional on the current congestion state, rain or fog conditions). The configurations for the implemented control strategies are stored in the repository, readily accessed by the real-time decision support system. Section 3 touches on the subject of operations planning, while §4 focuses on the control structure and modelling. The second mode of operation is the *dynamical filter*: the simulation, with some uncertainty in system parameters and inputs, runs in real time as the noisy measurements arrive from the traffic sensors. By filtering the measurement data through the simulation, the traffic state is estimated and fed back into the traffic responsive control algorithms. The third mode of operation is the *short-term prediction and strategy selection*: with the initial conditions coming from the filtered measurements and predicted with some uncertainty in short-term future inputs, the simulator runs a number of plausible near-term scenarios with available control strategies and calculates the resulting congestion and potentially serious stresses. The strategy promising the greatest benefits is deployed by sending the corresponding commands to the actuators in the field. Section 5 describes the traffic state estimation and prediction.

2. Dynamical model of traffic

(a) Model description

We start by introducing the traffic model employed by our trusted simulator, AURORA RNM (<http://code.google.com/p/aurorarm>), which is based on the cell transmission model (Daganzo 1994, 1995).

The road network consists of directed links and nodes, where links represent stretches of roads and nodes connect the links. Denote by \mathcal{L} the set of links, and by \mathcal{N} the set of nodes in the network. A node must always have at least one input and at least one output link. A link is called an *ordinary link* if it has both begin and end nodes. A link with no begin node is a *source link* or *source*, and a link with no end node is a *destination link* or *sink*. Each link $l \in \mathcal{L}$ is characterized by its length Δx_l , number of lanes k_l and the fundamental diagram (capacity F_l , free-flow speed v_l and congestion wave speed w_l , figure 2),¹ which takes into

¹The fundamental diagram is a density-flow function, usually concave. In our case, it has the triangular shape of figure 2, defined by three values: the capacity (maximum number of vehicles per hour that the link can let through) F_l ; the free-flow speed (average vehicle speed in the link measured under low traffic density conditions) v_l ; and congestion wave speed (speed with which the congestion wave propagates backwards) w_l . Alternatively, one could specify the triangular fundamental diagram as the triplet: capacity F_l ; critical density (number of vehicles per mile in the link at which the capacity is reached) ρ_l^c ; and jam density (number of vehicles per mile in the link at which traffic can no longer move) $\hat{\rho}_l$. Although in this paper we imply a triangular fundamental diagram, other shapes, e.g. parabola (Greenshields 1934), trapezoid (Munõz *et al.* 2003) or discontinuous, the so-called ‘inverse lambda’ (Kerner 1998; Orosz *et al.* 2009), could be used just as well.

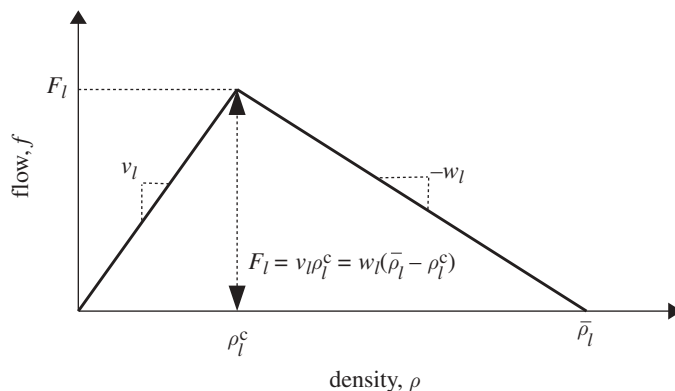
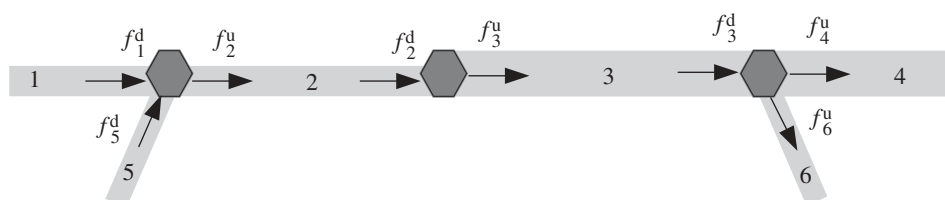
Figure 2. Fundamental diagram associated with some link l .

Figure 3. Simple road network: links are numbered from 1 to 6, nodes are shown as hexagons.

account the number of lanes. Sources are the links through which the vehicles enter the system, and therefore have demand profiles assigned to them. Each node $v \in \mathcal{N}$ is characterized by a split ratio matrix \mathcal{B}_v that determines how the incoming flows are distributed among the output links. Nodes may be used not only to represent intersections, merge or diverge points, but also to break up long links into smaller ones.

In the simple example of the single-directional freeway of figure 3, nodes are places where on-ramps merge into and off-ramps diverge from the freeway, or where freeway characteristics, such as the number of lanes, change; ordinary links are the stretches of freeway going from node to node; sources are on-ramps; and sinks are off-ramps.

The state of the road network at time t is described by the vehicle density in each link, $\rho_l(t)$. Given some initial condition, which usually comes from measurements, $\rho_l(t_0) = \rho_l^0$, the system evolves in time according to

$$\rho_l(t + \Delta t) = \rho_l(t) + \frac{\Delta t}{\Delta x_l} [f_l^u(t) - f_l^d(t)], \quad (2.1)$$

where Δt is the size of the time step, satisfying the condition $\Delta t \leq \min_l \{\Delta x_l / v_l\}$, and $f_l^u(t)$ and $f_l^d(t)$ are upstream (entering link l) and downstream (exiting link l) flows, respectively. For sources, $f_l^u(t) = r_l(t)$, where $r_l(t)$ is the value of demand—the flow that desires to enter the system through source link l at time t . For sinks, $f_l^d(t) = v_l \rho_l(t) \min\{1, F_l / v_l \rho_l(t)\}$. For ordinary links, $f_l^u(t)$ is determined by the begin node and $f_l^d(t)$ is determined by the end node of link l .

A node with m input and n output links has an $m \times n$ split ratio matrix $\mathcal{B}_v(t) = \{\beta_{ij}(t)\}_{i=1, \dots, m}^{j=1, \dots, n}$. This matrix is non-negative, its elements lie in the interval $[0, 1]$ and the sum of the elements in each row equals 1. The element $\beta_{ij}(t)$ defines the portion of the vehicle flow coming from input link i that has to be directed to the output link j at time t . Input flows $f_i^d(t)$ and output flows $f_j^n(t)$, $i = 1, \dots, m$, $j = 1, \dots, n$, for the node are computed in steps 1–7 that follow.

1. Compute supply for each output:

$$s_j(t) = \min\{F_j, w_j(\bar{\rho}_j - \rho_j(t))\}, \quad j = 1, \dots, n, \quad (2.2)$$

where F_j is the capacity, w_j is the congestion wave speed, $\bar{\rho}_j$ is the jam density and $\rho_j(t)$ is the current density of the output link j .

2. Set index $q = 0$.
3. Compute input demands:

$$\tilde{d}_i^{[q]}(t) = v_i \rho_i(t) \min\left\{1, \frac{F_i}{v_i \rho_i(t)}\right\}, \quad i = 1, \dots, m, \quad (2.3)$$

where F_i is the capacity, v_i is the free-flow speed and $\rho_i(t)$ is the current density of the input link i . Quantity $\tilde{d}_i^{[0]}(t)$ represents the desired flow from the input link i .

4. Compute output demands:

$$d_j^{[q]}(t) = \sum_{i=1}^m \beta_{ij}(t) \tilde{d}_i^{[q]}(t), \quad j = 1, \dots, n. \quad (2.4)$$

Quantity $d_j^{[0]}(t)$ represents the total flow that desires to enter the output link j .

5. For $q = 1, \dots, n$, repeat.

- (i) Scale down input demands to satisfy the output supply if necessary:

$$\tilde{d}_i^{[q]}(t) = \begin{cases} \tilde{d}_i^{[q-1]}(t), & \text{if } \beta_{iq}(t) = 0, \\ \tilde{d}_i^{[q-1]}(t) \min\left\{1, \frac{s_q(t)}{d_q^{[q-1]}(t)}\right\}, & \text{otherwise,} \end{cases} \quad i = 1, \dots, m. \quad (2.5)$$

- (ii) Recompute output demands $d_j^{[q]}(t)$, $j = 1, \dots, n$, according to equation (2.4).

This step implements the proportional priority rule for merging links and the first-in/first-out rule for diverging links as they are stated in Daganzo (1995).²

²Proportional priority rule means that each output link accommodates vehicles from the input links proportionally to the input demands. First-in/first-out rule means that the input-to-output flows in the node should always be in proportion to each other defined by the split ratio matrix.

6. Flow leaving the input link i is

$$f_i^d(t) = \tilde{d}_i^{[m]}(t), \quad i = 1, \dots, m. \quad (2.6)$$

7. Flow entering the output link j is

$$f_j^u(t) = \sum_{i=1}^m \beta_{ij}(t) \tilde{d}_i^{[m]}(t), \quad j = 1, \dots, n. \quad (2.7)$$

(b) *Building the model*

Building a model that is ready for simulation consists in: (i) putting together a road network by creating nodes and links with correct lengths and lane counts; (ii) calibrating the system, that is, assigning a fundamental diagram to each link; and (iii) defining time-varying demand functions for the source links and split ratio matrices for the nodes. This is generally a time-consuming process, as no single data source provides the information necessary for all three tasks.

The starting point of the process is obtaining the GIS data about the roads of interest from available commercial (Navteq Corporation, <http://www.navteq.com>; Tele Atlas, <http://www.teleatlas.com>) or free (OpenStreetMap Project, <http://www.openstreetmap.org>) sources. GIS data come in the form of shape files with information about the street segments that can be converted into a link–node description. AURORA RNM (<http://code.google.com/p/aurorarnm>) has a utility called GIS IMPORTER that performs this task. In special cases, for example, freeways of California or freeways of Portland region in Oregon, we are lucky to have single sources, PeMS (<http://pems.eecs.berkeley.edu>) and PORTAL (<http://portal.its.pdx.edu>) for the freeway geometry and for the traffic measurement data.

Measurement data from the freeway detectors provided by information systems such as PeMS or PORTAL allows fundamental diagrams to be estimated for the corresponding freeway links. The full description of the calibration algorithm is given in Dervisoglu *et al.* (2009). Here we provide a summary.

1. For each *reliable*³ detector on the selected stretch of freeway, extract available historical density and flow measurements. The distance between detectors is often larger than the link size in our model. Hence, the retrieved detector data may apply (and usually do) to more than one link.
2. Find the maximum measured flow value. Usually, this will be the capacity value F_l .
3. Use the least-squares method to estimate the free-flow speed v_l . Practice shows that free-flow density–flow pairs give a good fit. Set the critical density $\rho_l^c = F_l/v_l$.
4. Use the constrained least-squares method to determine the congestion wave speed w_l .

When detector data are good, steps 1–4 produce a decent result (figure 4*a*). If, on the other hand, detector data are poor owing to a malfunctioning detector or just because the capacity is never reached at this point of the freeway

³PeMS, for example, provides day-by-day health status for each detector.

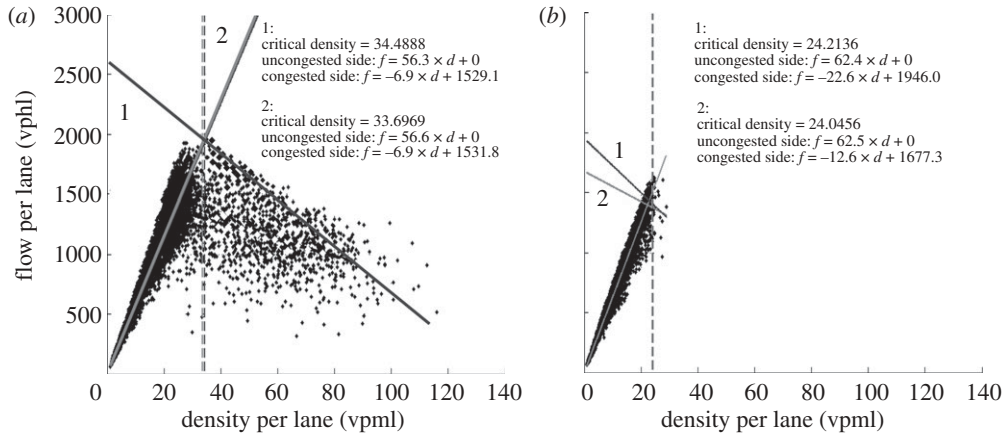


Figure 4. Estimating fundamental diagram: (a) good data; and (b) poor data. Black line (1), v_1 ; grey line (2), v_2 .

(figure 4b), then we can either use fundamental diagrams from the neighbouring links or impute the missing data as suggested in Chen *et al.* (2003) and repeat steps 2–4. Figure 4 illustrates the fact that measurements on the free-flow side of the fundamental diagram are usually represented by a straight line, whereas measurements in the congested region tend to be more scattered, which may justify the case for alternative forms of the fundamental diagram (e.g. trapezoid, inverse lambda).

The most difficult part of calibrating urban streets or arterials is to estimate their capacities. Once the capacity of an arterial link is determined from available measurements or intelligent guess, the free-flow speed can be set to the speed limit assigned on that street, and the jam density can be derived from estimating the number of vehicles this link can store and dividing this number by the length of the link.

Finally, it remains to define the demand functions for the sources and split ratio matrices for the nodes. For freeways, systems like PeMS and PORTAL should provide on-ramp flow data that could be used to construct the demand functions. The split ratios should be computed from the measurements of the mainline and off-ramp flows: in the example shown in figure 3, these would be the measurements of flows f_4^u and f_6^u . In practice, however, on- and off-ramp flows may not be available, which is currently the case for some California freeways. The demands and split ratios must then be imputed so that, when used in the model, the model produces mainline flows that match the measurements. As for most inverse problems, the problem of the demand and split ratio imputation is ill-posed in the sense of Hadamard: its solution is not unique. One has to impose certain restrictions on the demand and split ratio values and demand variations. The imputation problem is thoroughly covered in Muralidharan & Horowitz (2009).

Obtaining demand and split ratio data for arterials is a challenge. Sources of this information vary from city to city. In California, we have to rely on regional planning agencies such as MTC in San Francisco Bay Area, or SANDAG in San Diego.

3. Operations planning

The objective of operations planning is to develop strategies that improve traffic performance on congested road networks. The foundation of operations planning is the monitoring of traffic in the areas of interest and analysis of the measurement data. Such analysis pinpoints the weaknesses in travel corridor operations, indicates bottlenecks and accident hot spots, and provides hints about possible strategies of congestion relief.

The devised strategies may involve expanding capacity at the bottleneck by adding extra lane(s), or such operational techniques as: *demand management*, which focuses on reducing the excess demand during peak hours; *incident management*, which targets resources to alleviate incident hot spots; providing *traveller information*, which seeks to reduce traveller buffer time, the extra time that travellers must add to their average travel time when planning trips to ensure on-time arrival; *traffic flow control*, which implements ramp metering (RM) at freeway on-ramps near locations where significant reductions of congestion are likely to occur; imposing a *variable speed limit* (VSL) on freeways to homogenize the flow during peak hours; and optimizing *signal timing plans* at signalized intersections. Operations planning needs quick quantitative assessment of the performance benefits that can be gained from the congestion relief strategies, in order to rank them and, combined with a separate estimate of the deployment cost of these strategies, to select the most promising of them based on benefit/cost ratios or the magnitude of benefits.

General link performance measures are listed below.

- *Traffic speed* (V), measured in miles per hour (mph):

$$V_i(t) = \frac{f_i^d(t)}{\rho_i(t)}. \quad (3.1)$$

- *Instantaneous travel time* (ITT), measured in hours—the travel time through the link that would occur if traffic speed in the link stayed constant at its value at the current time t , $V(t)$:

$$ITT_i(t) = \frac{\Delta x_i}{V_i(t)}. \quad (3.2)$$

- *Actual travel time* (ATT), measured in hours—the travel time computed using actual speed values past the current time t :

$$ATT_i(t) = T_i(t)\Delta t, \quad (3.3)$$

where

$$T_i(t) = \arg \max_{\tau} \left\{ \sum_{\tau'=0}^{\tau-1} V_i(t + \tau')\Delta t \leq \Delta x_i \right\}. \quad (3.4)$$

- *Vehicle miles travelled* (VMT) is the measure of the throughput of the link during the current time step:

$$VMT_i(t) = \rho_i(t) V_i(t) \Delta x_i \Delta t. \quad (3.5)$$

- *Vehicle hours travelled (VHT)* reflects the number of vehicles in the link during the current time step:

$$\text{VHT}_l(t) = \rho_l(t) \Delta x_l \Delta t. \quad (3.6)$$

- *Delay (D)* measured in vehicle-hours:

$$D_l(t) = \text{VHT}_l(t) - \frac{\text{VMT}_l(t)}{v_l}. \quad (3.7)$$

- *Productivity loss (PL)*, measured in lane-mile-hours—the degree of under-utilization of the link lanes due to congestion:

$$PL_l(t) = \begin{cases} 0, & \text{if } V_l(t) = v_l, \\ \left(1 - \frac{f_l^d(t)}{F_l}\right) k_l \Delta x_l \Delta t, & \text{otherwise.} \end{cases} \quad (3.8)$$

Note that productivity loss depends on the number of lanes. For example, given the same link output flow $f_l^d(t)$ and the same link capacity F_l , the productivity loss of a two-lane link is twice the productivity loss of a single-lane link.

All these performance values, except for the actual travel time, can be computed at run time of the dynamical system. Computation of the actual travel time requires knowledge of the whole dynamical system trajectory. Knowing the actual travel time, VMT, VHT, delay and productivity loss for each individual link, one can compute the same quantities for any specified route in the network.

Arterial links whose end nodes are signalized intersections have additional performance measures: *delay per cycle*, number of vehicle-hours spent waiting at the signal; *queue size*, number of vehicles in the link; *phase utilization*, percentage of the green phase time used during cycle; *cycle failure*, percentage of vehicles waiting for more than one red light; *flow-to-capacity ratio*, which characterizes the utilization of the available capacity; and *progression quality*, percentage of vehicles arriving during the green phase. Knowing these performance measures for each incoming link of an intersection, one can assess the overall performance of the intersection.

To evaluate the road network performance under different scenarios, such as incidents, lane closures, special events, demand fluctuations, etc., using the macroscopic traffic model (2.1)–(2.7), we use ‘switches’ in the model parameters (fundamental diagrams at links and split ratio matrices at nodes) or inputs (demands at source links) at specified times.

Switches in fundamental diagrams model incidents and lane closures. For example, suppose the original fundamental diagram of some link l is $F_l = 6000$ vehicles per hour (vph), $v_l = 60$ mph, $w_l = 15$ mph, and an accident blocking half of the lanes occurs at 10.00 and lasts until 10.30 when the road is cleared. To model this accident, one has to specify two switches in the fundamental diagram: $F_l(10) = 3000$, representing the capacity drop, and $F_l(10.5) = 6000$, representing the return to normal operation. Evaluating the impact of faster reaction times (how much better would the system perform if the accident were cleared in

20 min instead of 30 min?) is part of incident management. One can also use switches in the fundamental diagrams to study the benefit of opening shoulders as general-purpose lanes near bottlenecks during peak hours.

Special events and the impact of providing traveller information are modelled by switches in the split ratio matrices at given nodes. A change in the proportion of traffic flow directed into a certain output link of a node may create a bottleneck and cause a congestion upstream of this node. Demand fluctuations are modelled by applying some non-negative coefficients to the original demand functions assigned to given sources.

In AURORA RNM, these switches in parameters and inputs are implemented as events that can be generated by the user and triggered at user-specified times during the simulation. The impact of each scenario can be quickly assessed, as AURORA RNM SIMULATOR computes the general performance measures for links and routes.

Another important aspect of the operations planning, as well as real-time decision support, is the evaluation of the traffic flow control technologies, considered next.

4. Control of traffic flow

Among the components of the congestion relief strategies mentioned in the previous section, demand management, incident management and traveller information influence the traffic indirectly: they affect the inputs and the parameters of the system in a way that can be estimated but not measured exactly. Direct and measurable influence over traffic behaviour is achieved through traffic flow control.

In the macroscopic link–node model (2.1)–(2.7), the link state evolves in time according to the law of conservation, for any link l with an end node, whereas the flow f_l^d exiting this link can be potentially restricted by the node. (By definition, the flow exiting a destination link is restricted only by the capacity of that destination link.) So nodes are the network components where traffic flow control can be imposed.

We propose the controller hierarchy that is summarized in figure 5.

- *Local* controller is assigned to a particular input link of a node and controls only the flow coming from that link. An example of a local controller is a ramp meter shown in figure 5*a*. Here, the node has two input links, uncontrolled freeway link 1 and controlled on-ramp 2.
- *Node* controller is assigned to a node and controls the flows coming from all the input links of that node. An example of a node controller is a signalized intersection shown in figure 5*b*. Here, only two non-conflicting input flows (1 and 5, 2 and 6, 3 and 7, or 4 and 8) are allowed at any one time, while the others are blocked. The individual input flows are still controlled each by its own local controller, but now the local controllers are synchronized by the centralized node control.
- *Complex* controller operates on multiple local, node, or even other complex controllers, coordinating their action towards some common objective. An example of a complex controller is a coordinated RM system shown in figure 5*c*. This is a freeway with two identified bottlenecks, which divide

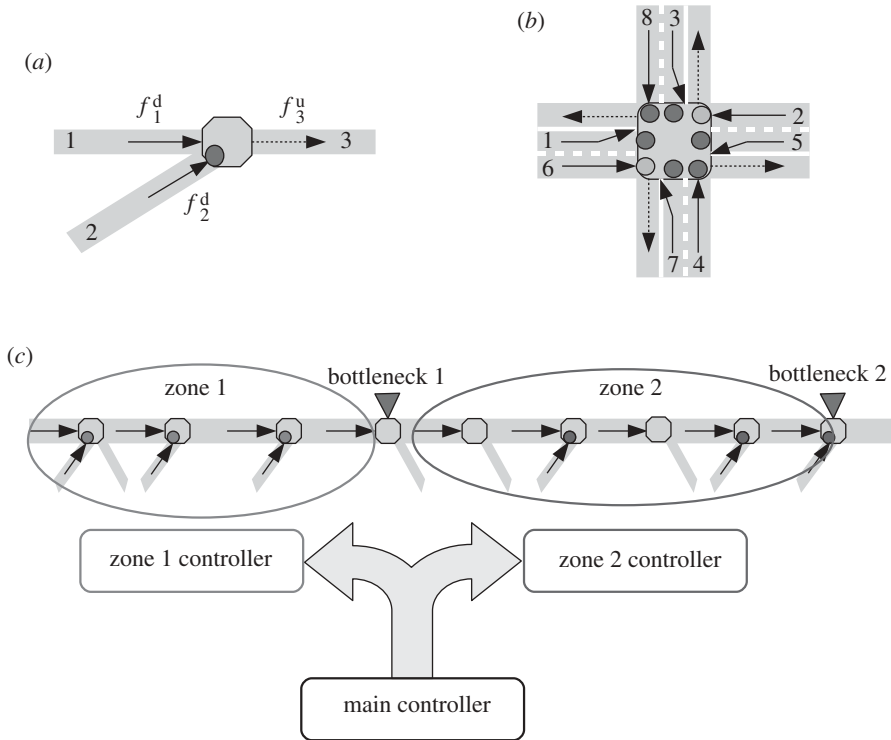


Figure 5. Controller hierarchy: examples of (a) local controller (ramp meter); (b) node controller (signalized intersection); and (c) complex controller (coordinated RM).

the freeway into zone 1 and zone 2, each ending at the corresponding bottleneck. Bottlenecks are expected to move throughout the day. Zone controllers are responsible for coordinating local ramp meters at the on-ramps within their respective zones, whereas the main controller keeps track of the bottleneck locations and zone configuration. Main and zone controllers are both complex controllers.

The model (2.1)–(2.7) incorporates a controller action by replacing the formula (2.3) for input demands with

$$\tilde{d}_i^{[0]}(t) = v_i \rho_i(t) \min \left\{ 1, \frac{F_i}{v_i \rho_i(t)}, \frac{C_i(t, \boldsymbol{\rho}(t))}{v_i \rho_i(t)} \right\}, \quad i = 1, \dots, m, \quad (4.1)$$

where m is the number of input links in a node, $\boldsymbol{\rho}(t)$ in bold represents the entire system state of densities in all the links in the network and $C_i(t, \boldsymbol{\rho}(t))$ is the control function for the flow entering the node from input link i .

The control may be open loop, $C_i(t, \boldsymbol{\rho}(t)) = C_i(t)$, or closed loop. The simplest example of an open-loop controller is the time-of-day local ramp meter, which restricts the flow coming from an on-ramp link by some constant value that changes several times during the day. A pre-timed signal with fixed offset, time cycle and phases, operating on an arterial intersection, is an example of an open-loop node controller.

Closed-loop control responds to the traffic state and can potentially adapt to the special situations such as incidents. An example of a closed-loop controller is ALINEA (Papageorgiou *et al.* 1991), a local ramp meter that for the configuration in figure 5a is defined by the formula

$$A_2(t, \rho_3(t)) = A_2(t - \Delta t, \rho_3(t - \Delta t)) + v_3[\rho_3^c - \rho_3(t)], \quad (4.2)$$

where subscripts ‘2’ and ‘3’ refer to links 2 and 3, respectively, and A_2 is the ALINEA flow rate. It is generally a good idea for a controller such as ALINEA to work in conjunction with a queue controller that prevents the vehicle queue on the ramp from growing too much, causing traffic spillback further upstream. The most common queue controller uses the queue override algorithm

$$Q_2(t, \rho_2(t)) = f_2^u(t) + v_2[\rho_2(t) - \rho_2^c], \quad (4.3)$$

where Q_2 is the flow rate prescribed by the queue override controller. Having both ALINEA and queue override active at the same time, the local ramp meter computes its rate as

$$C_2(t, \boldsymbol{\rho}(t)) = \max\{A_2(t, \rho_3(t)), Q_2(t, \rho_2(t))\}. \quad (4.4)$$

Coordinated ALINEA, known as HERO (Papamichail *et al.* 2010), whose idea is first to apply the ALINEA algorithm (4.2) to the on-ramp closest to the bottleneck until the queue limit is reached, then to turn on ALINEA at the next on-ramp upstream until the queue limit is reached there and so on, is a closed-loop complex controller. An example of a closed-loop complex controller that coordinates multiple node controllers operating on signalized arterial intersections is TUC (Diakaki *et al.* 2000).

VSL control also influences the traffic flow from the nodes. If some arbitrary VSL controller calculates the desired speed limit v_l^* for link l , the flow rate prescribed by the corresponding local controller is

$$C_l(t, \rho_l(t)) = v_l^* \rho_l(t). \quad (4.5)$$

It is often the case that, while being tested in simulations, closed-loop controllers greatly improve the system performance, reducing delay and productivity loss, but after deployment of these controllers in the field, the results are disappointing. The reason is the poor quality of the feedback signal from noisy or malfunctioning sensors, or the lack of feedback altogether in the real world. For example, PeMS (<http://pems.eecs.berkeley.edu>) detector health reports indicate that up to 30 per cent of California freeway loop detectors are not fully functional throughout the year.

To simulate feedback control algorithms and test their performance in a setting that resembles the real-world situation, we propose to incorporate sensors into the traffic model. We call this concept *virtual sensors*. Virtual sensors model the work of measurement devices, reporting vehicle counts and/or speeds from particular road network locations based on simulated link data. Within each such virtual sensor, one can adjust the noise level and choose the sensing quality in the range from excellent to unsatisfactory. The data reported by these virtual sensors are used by the closed-loop traffic flow control. Figure 6 summarizes the idea on the high level. In the ideal setting (figure 6a), the controller has access

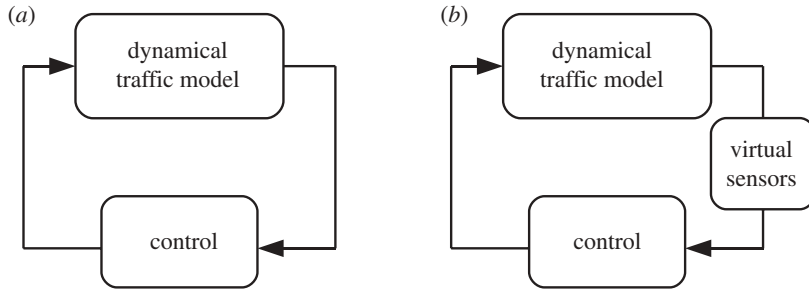


Figure 6. Feedback control system: (a) state feedback; and (b) measurement feedback.

to the current system state directly from the state equation (2.1). Introducing virtual sensors, we replace the *state feedback* with the *measurement feedback* in the model.

Among the available sensing techniques, we distinguish between *point sensors* such as loop detectors and wireless sensors, *mobile sensors* such as GPS (Global Positioning System) equipped vehicles and automatic vehicle location techniques, and *space sensors* such as aerial photography and satellite data.

- Point sensors are fixed in location along a roadway and measure vehicles passing through this location throughout the time for which they are active.
- Mobile sensors in vehicles move with the traffic flow in space–time and collect the time-stamped position (and speed) of the vehicles.
- Space sensors can take snapshots of traffic at a given instant of time and repeat such snapshots at multiple time instants.

We model point sensors and mobile sensors. A point sensor is assigned to a link and its position within this link is defined. A link may have multiple point sensors assigned to it. The specific sensor model may implement either loop detector or a Sensys wireless sensor (<http://www.sensysnetworks.com>). The measurements provided by a loop detector model are vehicle counts and, possibly, speeds. Data from Sensys sensors can be processed to obtain traffic density in a link, as described in Papageorgiou & Varaiya (2009). A mobile sensor is assigned to a route between the given origin and destination. It represents a probe vehicle. These probe vehicles are ‘phantoms’ in the sense that they do not affect the density and flow quantities produced by the original traffic model. A mobile sensor reports its current link together with its position and speed within this link. The intended use of mobile sensors is to compute the actual travel time for certain routes as the system evolves in time.

Another purpose of the virtual sensors is to serve as interfaces to the real measurement devices communicating directly to the equipment in the field or to the traffic control centre and collecting raw measurement data in real time, while the model (2.1)–(2.7) is used as a dynamical filter for these data before feeding back to the closed-loop traffic controllers. This dynamical filter is described in §5.

5. Traffic state estimation and prediction

The objectives of dynamic filtering of the measurement data are (i) to provide a better quality feedback for closed-loop controllers, (ii) to detect faulty sensors as early as possible, and (iii) to help determine initial conditions for the model (2.1)–(2.7) used for the short-term prediction in the real-time decision support system.

The set-valued estimation of freeway traffic density using the cell transmission model was discussed in Kurzhanskiy (2009). This result can be extended to the case of a road network. We assume that demands $r_l(t)$ at source links are known with some uncertainty; more precisely, they are constrained by a box $r_l^-(t) \leq r_l(t) \leq r_l^+(t)$ ($r_l^-(t)$ and $r_l^+(t)$ are known). The other assumption is that the link capacities F_l lie within given intervals $F_l^- \leq F_l \leq F_l^+$ (F_l^- and F_l^+ are known). Accordingly, denote

$$\bar{\rho}_l^+ = \frac{F_l^+}{w_l} + \frac{F_l^+}{v_l} \quad \text{and} \quad \bar{\rho}_l^- = \frac{F_l^-}{w_l} + \frac{F_l^-}{v_l}. \quad (5.1)$$

Noisy measurements of the output flow

$$y_l^{(f)}(t) = f_l^d(t) + \omega_l^{(f)}(t) \quad (5.2)$$

and speed

$$y_l^{(V)}(t) = V_l(t) + \omega_l^{(V)}(t) \quad (5.3)$$

are available at each link. Here, $\omega_l^{(f)}(t) \in [-\omega_l^{0,(f)}(t), \omega_l^{0,(f)}(t)]$ is the flow measurement noise, $\omega_l^{(V)}(t) \in [-\omega_l^{0,(V)}(t), \omega_l^{0,(V)}(t)]$ is the speed measurement noise and bounds $\omega_l^{0,(f)}(t)$ and $\omega_l^{0,(V)}(t)$ are known. Thus, for each link, we get an estimate of the density coming from the measurements:

$$\hat{\rho}_l^-(t) \leq \hat{\rho}_l(t) \leq \hat{\rho}_l^+(t), \quad (5.4)$$

where

$$\hat{\rho}_l^-(t) = \frac{y_l^{(f)}(t) - \omega_l^{0,(f)}(t)}{y_l^{(V)}(t) + \omega_l^{0,(V)}(t)} \quad \text{and} \quad \hat{\rho}_l^+(t) = \frac{y_l^{(f)}(t) + \omega_l^{0,(f)}(t)}{y_l^{(V)}(t) - \omega_l^{0,(V)}(t)}. \quad (5.5)$$

Recall state update equation (2.1), and define state bounds update equations for each link l :

$$\rho_l^-(t + \Delta t) = \rho_l^-(t) + \frac{\Delta t}{\Delta x_l} [f_l^{u^-}(t) - f_l^{d^+}(t)] \quad (5.6)$$

and

$$\rho_l^+(t + \Delta t) = \rho_l^+(t) + \frac{\Delta t}{\Delta x_l} [f_l^{u^+}(t) - f_l^{d^-}(t)]. \quad (5.7)$$

For sources,

$$f_l^{u^-}(t) = r_l^-(t) \quad \text{and} \quad f_l^{u^+}(t) = r_l^+(t).$$

For sinks,

$$f_l^{d^-}(t) = v_l \rho_l^-(t) \min\{1, F_l^- / v_l \rho_l^-(t)\} \quad \text{and} \quad f_l^{d^+}(t) = v_l \rho_l^+(t) \min\{1, F_l^+ / v_l \rho_l^+(t)\}.$$

Otherwise, $f_l^{u^-}(t)$ and $f_l^{u^+}(t)$ are determined by the begin node, and $f_l^{d^-}(t)$ and $f_l^{d^+}(t)$ are determined by the end node.

Lower input/output flow bounds $f_l^{d-}(t)$ and $f_l^{u-}(t)$ are computed using steps 1–7 from §2a with slight modifications. For a node with input links $i = 1, \dots, m$ and output links $j = 1, \dots, n$, these modified steps are as follows.

1. Compute lower supply bound for each output (recall formula (2.2)):

$$s_j^-(t) = \min\{F_j^-, \max\{0, w_j(\bar{\rho}_j^- - \rho_j^+(t))\}\}, \quad j = 1, \dots, n. \quad (5.8)$$

2. Set index $q = 0$.
3. Compute lower input demand bounds:

$$\tilde{d}_i^{-[q]}(t) = v_i \rho_i^-(t) \min \left\{ 1, \frac{F_i^-}{v_i \rho_i^-(t)}, \frac{C_i(t, \boldsymbol{\rho}^-(t))}{v_i \rho_i^-(t)} \right\}, \quad i = 1, \dots, m. \quad (5.9)$$

4. Compute lower output demand bounds:

$$d_j^{-[q]}(t) = \sum_{i=1}^m \beta_{ij}(t) \tilde{d}_i^{-[q]}(t), \quad j = 1, \dots, n. \quad (5.10)$$

5. For $q = 1, \dots, n$, repeat.

- (i) Scale down lower input demand bounds according to the lower output supply bounds if necessary:

$$\tilde{d}_i^{-[q]}(t) = \begin{cases} \tilde{d}_i^{-[q-1]}(t), & \text{if } \beta_{iq}(t) = 0, \\ \tilde{d}_i^{-[q-1]}(t) \min \left\{ 1, \frac{s_q^-(t)}{d_q^{-[q-1]}(t)} \right\}, & \text{otherwise,} \end{cases} \quad i = 1, \dots, m. \quad (5.11)$$

- (ii) Recompute lower output demand bounds $d_j^{-[q]}(t)$, $j = 1, \dots, n$, according to equation (5.10).

6. Lower bound for flow leaving the input link i is

$$f_i^{d-}(t) = \tilde{d}_i^{-[n]}(t), \quad i = 1, \dots, m. \quad (5.12)$$

7. Lower bound for flow entering the output link j is

$$f_j^{u-}(t) = \sum_{i=1}^m \beta_{ij}(t) \tilde{d}_i^{-[n]}(t), \quad j = 1, \dots, n. \quad (5.13)$$

Upper input/output flow bounds $f_l^{d+}(t)$ and $f_l^{u+}(t)$ are obtained through the same procedure by replacing ‘-’ superscripts with ‘+’ and vice versa.

By definition of $\rho_l(t)$ in equation (2.1), $\rho_l^-(t)$ in equation (5.6) and $\rho_l^+(t)$ in equation (5.7), if $\rho_l^-(0) \leq \rho_l(0) \leq \rho_l^+(0)$, then $\rho_l^-(t) \leq \rho_l(t) \leq \rho_l^+(t)$ for $t \geq 0$. Note that $\rho_l^-(t)$ and $\rho_l^+(t)$ must be restricted so that $0 \leq \rho_l^-(t)$ and $\rho_l^+(t) \leq \bar{\rho}_l^+$. These restrictions are not satisfied automatically in equations (5.6) and (5.7) and must be imposed explicitly every time step. Boundary trajectories $\rho_l^-(\cdot)$ and $\rho_l^+(\cdot)$ are used to filter the incoming measurement data.

Systems (5.6) and (5.7) start evolving at time $t = 0$ with initial conditions $\rho_l^-(0) \leq \rho_l^+(0)$ for every link l , possibly the result of previous estimates. If no initial conditions are available, take the flow and speed measurements $y_l^{(f)}(0)$

and $y_l^{(V)}(0)$ for each link, and determine $\hat{\rho}_l^-(0)$ and $\hat{\rho}_l^+(0)$ from equations (5.5). Systems (5.6) and (5.7) evolve in time until time step $\tau > 0$ when the results of new measurements, $\hat{\rho}_l^-(\tau)$ and $\hat{\rho}_l^+(\tau)$, arrive. Density bounds $\rho_l^-(\tau)$ and $\rho_l^+(\tau)$ are adjusted according to these measurements:

$$\rho_l^-(\tau) \leftarrow \rho_l^-(\tau) \max \left\{ 1, \frac{\hat{\rho}_l^-(\tau)}{\rho_l^-(\tau)} \right\} \quad \text{and} \quad \rho_l^+(\tau) \leftarrow \rho_l^+(\tau) \min \left\{ 1, \frac{\hat{\rho}_l^+(\tau)}{\rho_l^+(\tau)} \right\}. \quad (5.14)$$

These corrections make sense only if

$$[\rho_l^-(\tau), \rho_l^+(\tau)] \cap [\hat{\rho}_l^-(\tau), \hat{\rho}_l^+(\tau)] \neq \emptyset. \quad (5.15)$$

Condition (5.15) must be true in theory. Otherwise, it would mean that we made wrong assumptions about the range of measurement noise. In reality, however, empty intersections do occur. What to do in case condition (5.15) is not satisfied depends on the specific situation. If we trust our model more than the measurements, we should skip the correction (5.14). Moreover, we can use our dynamical filter to detect faulty measurement sensors. If, on the other hand, we believe that the measurement data are reliable, correction (5.14) should be modified:

$$\rho_l^-(\tau) \leftarrow \hat{\rho}_l^-(\tau) \quad \text{and} \quad \rho_l^+(\tau) \leftarrow \hat{\rho}_l^+(\tau). \quad (5.16)$$

Once the upper bound $\rho_l^+(\tau)$ is corrected according to equation (5.14) or (5.16), we must check that for each link $\rho_l^+(\tau) \leq \bar{\rho}_l^+$. If for some link this inequality does not hold, there are two alternative ways to proceed. The first is to modify $\rho_l^+(\tau)$:

$$\rho_l^+(\tau) \leftarrow \hat{\rho}_l^+(\tau). \quad (5.17)$$

The alternative is to modify the upper capacity bound of the link:

$$F_l^+ \leftarrow F_l^+ \frac{\rho_l^+(\tau)}{\bar{\rho}_l^+}, \quad (5.18)$$

and then adjust jam density $\bar{\rho}^+$ according to equation (5.1). Once the corrections (5.14)–(5.17) are applied, the density bounds $\rho_l^-(\tau)$ and $\rho_l^+(\tau)$ in every link can be treated as new initial conditions, and starting at these initial conditions, systems (5.6) and (5.7) should run until the next set of measurements arrives, and so on.

Traffic state prediction involves computing only the bounding trajectories $\rho_l^-(\cdot)$ from equation (5.6) and $\rho_l^+(\cdot)$ from equation (5.7) without corrections (5.14). Since the difference $\rho_l^+(t) - \rho_l^-(t)$ gets larger as t increases, it makes sense to limit the prediction to 1–2 h into the future, continuously recomputing it as time goes by and new measurements arrive.

Prediction results for 25 July 2009, for 25 miles of the I-210 West freeway in Southern California from Baseline Road (postmile 52) to the junction with I-710 (postmile 27), are shown in figure 7. Here, 06.00 is the current moment.

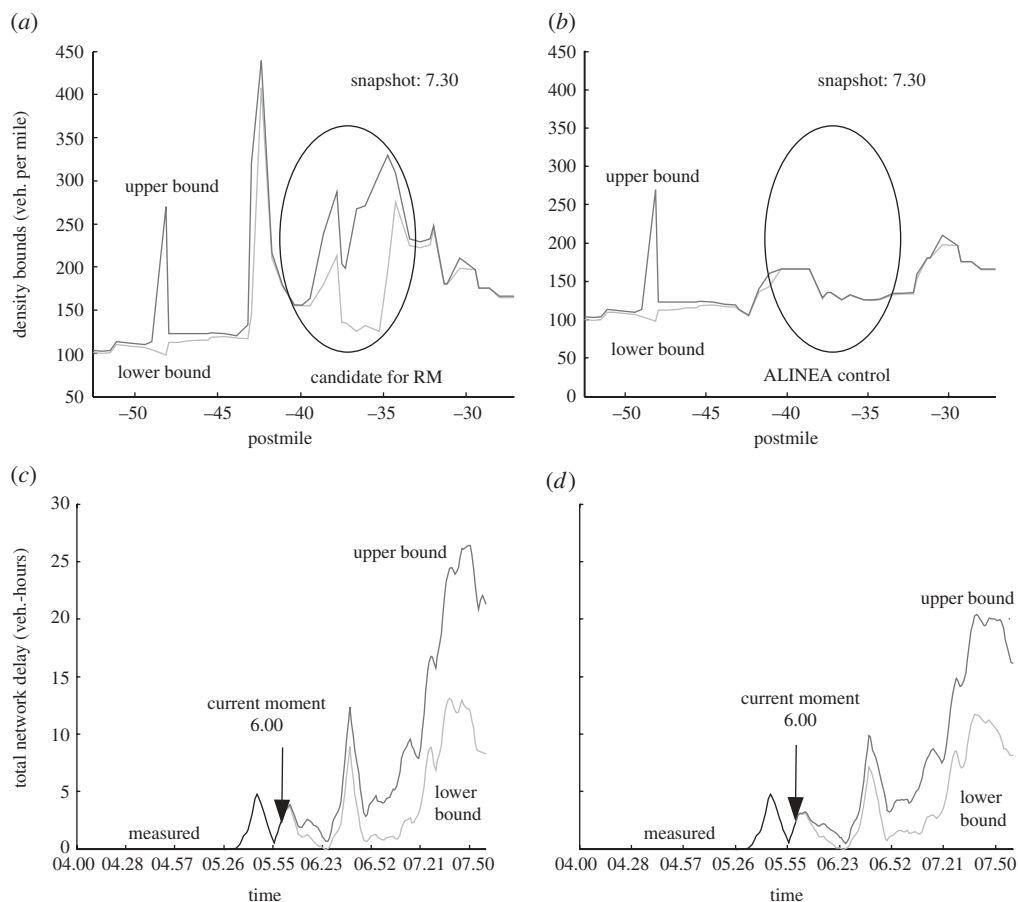


Figure 7. Short-term prediction from 06.00 to 08.00 for I-210 West: (a) projected density bounds for 07.30 without RM; (b) projected density bounds for 07.30 with ALINEA ramp control; (c) projected bounds for total network delay without RM; and (d) projected bounds for total network delay with ALINEA ramp control.

Data before 06.00 come from measurements and state estimation. At 06.00, traffic behaviour is predicted for the next two hours, until 08.00: lower and upper density bounds are computed for all the freeway links.

Figure 7a is a snapshot of the projected density bounds along the freeway at 07.30 (1.5 h into the future) computed with no traffic flow control at nodes. At the location marked by the ellipse (between postmiles 40 and 35), the projected density uncertainty interval is large, with its upper bound exceeding the critical density and the lower bound staying in the free-flow state, indicating that the system may develop congestion if left to its own devices, or stay in free flow if managed properly. Such locations are primary candidates for aggressive RM.

Figure 7b shows a snapshot of projected density bounds at the same time if ALINEA RM (4.2) were applied within the marked freeway segment. The size of the projected density interval is significantly reduced: *the controlled system is more predictable than the uncontrolled one*. Figure 7c,d shows the evolution

of the total network delay (including on-ramp queue delay) for the uncontrolled and ALINEA-controlled cases, respectively. From 04.00 to 06.00, the delay is computed from the measurements (black), and from 06.00 to 08.00, from the predicted delay bounds. One can see that ALINEA RM potentially yields a noticeable delay reduction.

This is just one example. In reality, the traffic operator should test several potential control strategies and apply the most effective one. A macroscopic traffic simulator such as AURORA RNM (<http://code.google.com/p/aurorarnm>) is able to run hundreds of simulations with preprogrammed scenarios in a matter of several minutes.

6. Conclusion

Adequate historical and real-time traffic data that support a fast and trusted simulation engine form the basis for constructing the analytical capability the ATM needs. High simulation speed allows the operator to analyse tens of potential control strategies in a matter of minutes. Therefore, microsimulation is not up to the task. We advocate macroscopic simulation, which while being fast can also be trusted because (i) it adequately captures the dynamics of traffic flow and (ii) all simulation model parameters can be reliably estimated from traffic data. Calibration of the proposed macroscopic model parameters is straightforward when the corresponding data are available. The frequent lack of ramp flow measurements can be partially compensated by imputation of on-ramp demands and off-ramp split ratios.

The system performance under various scenarios and control strategies is evaluated using such measures as traffic speed and actual travel time, which can be computed per link or per route, and VMT, VHT, delay and productivity loss, which can also be computed for the whole network. For signalized arterial intersections, additional performance measures, such as delay per cycle, queue size, phase utilization, cycle failure, flow-to-capacity ratio and progression quality, apply.

Traffic control in our model is actuated at nodes. The controller hierarchy is formed by local controllers operating on individual input links, node controllers operating on all input links in individual nodes and complex controllers operating on multiple local, node or other complex controllers. To simulate a realistic model of a closed-loop traffic control system, we replace the state feedback with the measurement feedback by introducing the concept of virtual sensors. A virtual sensor models a measurement device whose quality may be anywhere between excellent and unsatisfactory. Significantly, it can serve as an interface with field measurement devices during real-time operation.

The simulator has three modes of operation. In the operations planning, mode scenarios are tested, and control strategies are assessed in terms of their cost and benefits, so that reliable decisions may be made about which strategies to implement. The second mode is the dynamical filter, used in real time to improve the quality of the feedback for the traffic controllers. The third mode is short-term prediction, which estimates the uncertainty in the future traffic state based on current measured conditions and the projected demands for the near future (1–2 h) under a variety of implemented control strategies for the purpose

of selecting the most suitable one for the predicted situation. An example of the I-210 West freeway in Los Angeles illustrates some ATM features based on the presented framework.

The dynamical traffic model, some elements of the model-building process (e.g. GIS data importing), the control framework together with selected local node and complex controllers and estimation/prediction algorithms are implemented in AURORA RNM (<http://code.google.com/p/aurorarnm>) and available for download.

Finally, we would like to mention some directions for future research.

For the calibration process, it is important to develop a mechanism that would determine the quality of the measurement data, since only when based on reliable data can the model parameters be properly identified. Even an elaborate information system such as PeMS (<http://pems.eecs.berkeley.edu>), which provides detector health reports, often displays faulty detectors as healthy. Other, less sophisticated measurement data sources do not provide detector health information at all.

For the operations planning, the challenge is to incorporate heterogeneous traffic, which includes different vehicle types, such as high-occupancy, heavy-weight (trucks) and transit (buses) that may or may not use special lanes, into the macroscopic model in a realistic way. Other tasks include calculation of emissions and fuel consumption, and studying the impact of dynamic tolls.

For the traffic flow control, the goal is the evaluation of existing controversial technologies such as VSL in terms of throughput improvement, and design of new coordinated strategies that would control a freeway and surrounding arterials as one system.

The crucial part of the fully functional dynamical filter is the real-time evaluation of the sensing devices and fault detection. In case the condition (5.15) fails, the decision should be made as to which to trust, the model or the measurements. Each sensor must be evaluated not as a standalone, but together with its neighbours and their past behaviour, and in conjunction with data from other sources such as emergency calls reporting an accident, etc.

This research was supported by National Science Foundation Award CMMI-0941326 and the California Department of Transportation. We are grateful to members of the TOPL research project (<http://path.berkeley.edu/topl>), especially G. Dervisoglu, G. Gomes, R. Horowitz, A. Muralidharan and R. Sanchez.

References

- Chen, C., Kwon, J., Skabardonis, A. & Varaiya, P. 2003 Detecting errors and imputing missing data for single loop surveillance systems. *Transport. Res. Rec.* **1855**, 160–167. (doi:10.3141/1855-20)
- Daganzo, C. F. 1994 The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transport. Res. B* **28**, 269–287. (doi:10.1016/0191-2615(94)90002-7)
- Daganzo, C. F. 1995 The cell transmission model II: network traffic. *Transport. Res. B* **29**, 79–93. (doi:10.1016/0191-2615(94)00022-R)
- Dervisoglu, G., Gomes, G., Kwon, J., Horowitz, R. & Varaiya, P. P. 2009 Automatic calibration of the fundamental diagram and empirical observations on capacity. In *Transportation Research Board, 88th Annual Meeting, Washington, DC, 11–15 January*. Compendium of Papers DVD. See <http://pubsindex.trb.org/view.aspx?id=882131>.

- Diakaki, C., Papageorgiou, M. & McLean, T. 2000 Integrated traffic-responsive urban corridor control strategy in Glasgow, Scotland. *Transport. Res. Rec.* **1727**, 101–111. (doi:10.3141/1727-13)
- Greenshields, B. D. 1934 A study of highway capacity. *Proc. Highway Res. Board* **14**, 448–474.
- Kerner, B. S. 1998 Experimental features of self-organization in traffic flow. *Phys. Rev. Lett.* **81**, 3797–3800. (doi:10.1103/PhysRevLett.81.3797)
- Kurzhanskiy, A. A. 2009 Set-valued estimation of freeway traffic density. *Proc. 12th IFAC Symp. on Control in Transportation Systems, Redondo Beach, CA, September*.
- Munöz, L., Sun, X., Horowitz, R. & Alvarez, L. 2003 Traffic density estimation with the cell transmission model. *Proc. 2003 American Control Conf., 4–6 June, Denver, CO*, vol. 5, pp. 3750–3755. New York, NY: American Automatic Control Council, IEEE. (doi:10.1109/ACC.2003.1240418)
- Muralidharan, A. & Horowitz, R. 2009 Imputation of ramp data flow for freeway traffic simulation. *Transport. Res. Rec.* **2099**, 58–64. (doi:10.3141/2099-07)
- Orosz, G., Wilson, R. E., Szalai, R. & Stepan, G. 2009 Exciting traffic jams: nonlinear phenomena behind traffic jam formation on highways. *Phys. Rev. E* **80**, 046205. (doi:10.1103/PhysRevE.80.046205)
- Papageorgiou, M. & Varaiya, P. 2009 Link vehicle-count—the missing measurement for traffic control. *Proc. 12th IFAC Symp. on Control in Transportation Systems, Redondo Beach, CA, September*.
- Papageorgiou, M., Hadj-Salem, H. & Blosseville, J. M. 1991 ALINEA: a local feedback control law for on-ramp metering. *Transport. Res. Rec.* **1320**, 58–64.
- Papamichail, I., Papageorgiou, M., Vong, V. & Gaffney, J. 2010 HERO coordinated ramp metering implemented at the Monash Freeway. In *Transportation Research Board, 89th Annual Meeting, Washington, DC, 10–14 January*. Compendium of Papers DVD. See <http://pubsindex.trb.org/view.aspx?id=909624>.