Traffic State Estimation and Prediction using the Cell Transmission Model with Implicit Mode Switching and Dynamic Parameters

Chris M.J. Tampère, Katholieke Universiteit Leuven, Traffic & Infrastructure, Kasteelpark Arenberg 40, 3001 Leuven. T +32 16 321673; F +32 16 321976; E chris.tampere@bwk.kuleuven.be; I www.kuleuven.be/traffic

Lambertus H. Immers, Katholieke Universiteit Leuven, Traffic & Infrastructure, Kasteelpark Arenberg 40, 3001 Leuven. T +32 16 321669; F +32 16 321976; E ben.immers@bwk.kuleuven.be; I www.kuleuven.be/traffic

Abstract. This paper presents a traffic state estimation and prediction model based on the first order cell transmission model (CTM). The nonlinear CTM is transcribed in a closed analytical state-space form, so that it can be used within a general extended Kalman filtering framework that was recently proposed. The main advantage of the approach is that the state-space CTM switches implicitly between numerous possible linear modes, so that standard linear Kalman filter theory can be applied. The paper also provides measurement models for the traffic state, model parameters and boundary conditions that are needed to automatically estimate the traffic conditions and the model parameters in an online context. These can be used to obtain short-term predictions of traffic conditions and travel times. The viability of the approach is illustrated in a simulated case study of an incident on a motorway. The online model is capable of tracking the sudden capacity reduction, the congestion that builds up and the corresponding increase in travel times.

1 INTRODUCTION

Real-time traffic information provision and dynamic traffic management are considered indispensable tools for managing congested traffic flows both on motorways and urban networks. In order to perform their tasks adequately, traffic information and traffic control centers depend on techniques for traffic state estimation and prediction. Such techniques utilize real-time data from installed sensors like inductive loop detectors, traffic monitoring and surveillance cameras, vehicle re-identification systems, and floating car data systems. However, traffic data are subject to substantial measurement error and noise. Moreover, even if these can be removed from the data, they offer a spatially fragmented snapshot image of traffic conditions that tend to change rapidly in space and time.

Online traffic flow models can be used to remedy these limitations of crude data. On the one hand, they provide a means of filtering noise and error from error-prone traffic sensors. Because the traffic flow model captures some fundamental properties of traffic flow, it improves consistency between measurements and filtering that obeys physical principles like conservation of vehicles on links and over nodes. Secondly, online traffic models overcome the spatially fragmented character of traffic data, since they provide a state estimation for every part of the network that is considered. This interpolation is done consistently with well-known traffic phenomena like congestion propagation in wave fronts. Thirdly, online traffic flow models provide a single state estimation for every part of the network, regardless of how many different data sources are used to estimate the traffic state. As such, online models provide a natural framework for data fusion of different sensor systems. Finally, online traffic models are very well suited to extrapolate state estimations into short term traffic state predictions.

So far, the majority of work on online traffic models has been done for motorway networks (e.g. 1, 2, 3, 4, 5, 6). Such models only simulate traffic flows, like shockwave propagation, and approximate route choice in a simplified way, for instance through splitting fractions at nodes. Online traffic models based on dynamic traffic assignment models explicitly model route choice behavior and route flows, and are therefore in principle better suited for online modeling of urban or mixed urban-motorway networks. The principal achievements in this area are (7) and (8). However, it is recognized that such models are by far more complex than those that purely model traffic flows (see for instance the problems reported in 2).

Although the ultimate goal of our efforts is online modeling of urban (or mixed urban-motorway) networks, we limit the complexity of the model for now by only considering the traffic flow component and by not treating route choice explicitly. We borrow the general approach of Wang & Papageorgiou (6), who summarize and generalize earlier efforts in this field by respected authors like Payne and Cremer. Their approach is based on a discrete second order traffic flow model, which they calibrate online to real-time traffic data using extended Kalman filtering techniques (see section 2). The authors suggest an elegant approach for online estimation of the principal parameters and boundary conditions of the traffic flow model. It is based on a random-walk model,
adapted to realtime traffic conditions by a simple Kalman filter. Because of its transparency and straightforward and rapid calculation, we prefer this approach for now over valuable alternatives, provided by for instance mixture Kalman filter (\cite{5}), switching mode model (\cite{10}), Fuzzy logic (\cite{11}), or particle filters (\cite{4}).

However, whereas a second order traffic flow model might be needed to capture for instance stop and go waves on motorways, it is unnecessarily complex for urban networks, in which a first order approach (e.g. the cell transmission model (CTM), \cite{12}) is more than sufficient. Because CTM also has less parameters to calibrate, we decided to apply the approach of Wang & Papageorgiou to this first order model.

The aim of this paper is twofold:

- to show how the cell transmission model (CTM) can be included in the general extended Kalman filtering (EKF) framework of Wang & Papageorgiou; the key issue here is to linearize the non-linear CTM model around its current state, which we will do implicitly by introducing an appropriate formulation of CTM;

- to show the capability of the combined CTM-EKF model to capture (rapid) changes of important modeling parameters like the capacity.

The latter property will be very important when – in the next stages of our research – we will apply the CTM-EKF model to urban traffic conditions. The essence there will be to adequately capture changes of capacity attributed by traffic lights at intersections and parameters like split fractions at nodes. However, in this paper we limit ourselves to a case study on a motorway, which keeps more closely to the applications considered by the authors from whom we adopted our modeling approach. On the other hand, the case – an incident causing congestion – bears some resemblance to traffic flow interrupted by a traffic light.

2 TRAFFIC STATE ESTIMATION THROUGH (EXTENDED) KALMAN FILTERING

In this section, we briefly summarize the general traffic state estimation approach by (\cite{6}). For a detailed description, we refer to the original reference. For more background on (extended) Kalman filtering, we refer to (\cite{13} and \cite{14}).

Consider a traffic network, discretized spatially into \( N \) segments \( i=1..N \). Time is discretized in time steps of duration \( T \) with indices \( k=0,1,2,\ldots \). The state estimation problem consists in finding for any segment \( i \) and time \( k \) the state variable(s) that define the traffic state unambiguously. Traffic flow is represented as a single commodity macroscopic flow, so that any two state variables from: \( \rho \) (density), \( q \) (flow), and \( v \) (average speed) together with the fundamental relation \( q = \rho v \) suffice. In (\cite{6}), a second order traffic flow model is used, which leaves two independent traffic state variables to be determined. In this paper, we model traffic flow through first order traffic flow theory, so that one state variable – in our case \( \rho \) – is sufficient. The speed and flow are then obtained through the fundamental diagrams of speed and flow versus density. We denote the state of a segment \( i \) as \( s_i = [s^1_i \ldots s^5_i] \), and the vector combining the states of all segments as \( s = [s_1 \ s_2 \ldots s_i \ldots s_N]^T \). We assume the uncertainty or model error of any state variable \( j \) of segment \( i \) to be white noise with the value \( \xi^{s^j_i} \), which combines to \( \xi^s_i = [\xi^{s^1_i} \cdots \xi^{s^5_i}] \) for a segment \( i \) and to \( \xi^s = [\xi^{s^1} \xi^{s^2} \cdots \xi^{s^5}]^T \) for the entire network.

Before proceeding to the Kalman filter formulation, let us introduce some more notations. For one, any traffic flow model defined on the segments of the traffic network, has some parameters. Let us denote \( p \) the vector of all parameters in the model, consisting of \( p = [p_1 \ p_2 \ldots p_i \ldots p_N]^T \). Each segment \( i \) is represented in this parameter vector, by its own parameter set \( p_i = [p^1_i \cdots p^{10}_i] \), which contains for instance the capacity, free speed and other parameters of the fundamental diagram. Secondly, we need to define the boundaries of the traffic network. These include for instance the evolution of demand upstream of boundaries crossing the main carriageway of the motorway, demand at on-ramps and split fractions determining the ratio of exit flows at off-ramps over the main flow. Without going into detail, let us denote the vector defining all boundaries as \( d \).

The state vector \( s \), the state noise vector \( \xi\), the parameter vector \( p \), and the boundary vector \( d \) all depend on time, which we denote by adding the time index \( (k) \) to their notation. A traffic model \( a \) (in state-space formulation, referred to as ‘system model’ in standard Kalman filter literature) is then defined as any (nonlinear)
differential vector function that determines the state vector $s$ at time step $k+1$, based on the state, parameters, boundaries, and state noise at time $k$:

$$s(k+1) = a\left(s(k), p(k), d(k), \xi_1(k)\right)$$  \hspace{1cm} (1)

In (1), it is proposed to treat the parameters and boundaries of the model as additional model state variables, the value of which will be estimated in the Kalman filtering process. To that end, their system model is defined as a random walk:

$$p(k+1) = p(k) + \xi_2(k)$$  \hspace{1cm} (2)

$$d(k+1) = d(k) + \xi_3(k)$$  \hspace{1cm} (3)

Equations (2) and (3) express that the model assumes parameters and boundaries to be constants subject to white noise. The purpose of the Kalman filter will be to adapt the value of these constants, dependent on the real-time measurements. Combining the original and additional state vectors into an augmented state vector

$$x = \begin{bmatrix} s^T & p^T & d^T \end{bmatrix}^T$$

and the corresponding noise components into an augmented noise vector $\xi = \begin{bmatrix} \xi_1^T & \xi_2^T & \xi_3^T \end{bmatrix}^T$, the augmented state-space traffic model $A$ is the combination of system models (1), (2), and (3), which we write as:

$$x(k+1) = A(x(k), \xi(k))$$  \hspace{1cm} (4)

The idea of Kalman filtering for state estimation is that a set of $M$ real-time measurement $z = [z_1 ... z_M]^T$ gives information about the state variables of the system, even though these measurements might contain noise and might not be direct observations of system state variables. We therefore need to define the (possibly nonlinear) relationship $H$ between the state vector $x$ and the measurements $z$, as well as the measurement noise $\eta = [\eta_1 ... \eta_M]^T$, which results from a combination of measurement errors and errors in the measurement model $H$:

$$z(k) = H(x(k), \eta(k))$$  \hspace{1cm} (5)

As we will show in the next sections, we formulate our traffic model (CTM) as a set of linear models, from which the appropriate one is chosen implicitly at every time step. As a result, application of extended Kalman filter theory, reduces in our case to applying linear Kalman filter theory, once the appropriate linear models $A(k)/B(k)$ and $H(k)$ have been selected. We thus rewrite the system and measurement models (4) and (5) as follows:

$$x(k+1) = A(k)x(k) + B(k) + \xi(k)$$  \hspace{1cm} (6)

$$z(k) = H(k)x(k) + \eta(k)$$  \hspace{1cm} (7)

With these definitions, and the notations $P$, $Q$ and $R$ for the covariance matrices of the augmented state vector $x$ and white noise vectors $\xi$ and $\eta$ respectively, we set up the recursive linear Kalman prediction and correction cycle. To distinguish the actual value of the state vector $x(k)$ from estimated values, we introduce the notations $\hat{x}(k)$ and $\hat{x'}(k)$ for the estimations prior to and after Kalman correction respectively.

Correspondingly, we use $P'(k)$ and $P(k)$ for the covariance matrix of $x(k)$ prior and after Kalman correction. The first step is the prediction. Given the state estimate $\hat{x}(k-1)$ and covariance $P(k-1)$ for the uncertainty over this estimate in the previous time step $k-1$, a prediction of the state and covariance at $k$ are obtained by applying the system model $A$ (under the assumption that there was no noise involved):

$$\hat{x}^- (k) = A(k-1)\hat{x}(k-1) + B(k-1)$$  \hspace{1cm} (8)

$$P^-(k) = A(k-1)P(k-1)A^T(k-1) + Q$$  \hspace{1cm} (9)

These a priori estimates are now corrected in the second step, through:
\[
K(k) = P^\ast(k)H^\ast(k)\left[H(k)P^\ast(k)H^\ast(k) + R\right]^{-1}
\]
(10)
\[
\dot{x}(k) = \hat{x}^-(k) + K(k)\left[z(k) - H(k)\hat{x}^-(k)\right]
\]
(11)
\[
P(k) = \left[I - K(k)H(k)\right]P^\ast(k)
\]
(12)

We see that the a priori estimate \(\hat{x}^-(k)\) is corrected proportional to the Kalman gain \(K(k)\) on the one hand, and to the difference between the measurement vector \(z(k)\), and \(H(k)\hat{x}^-(k)\), i.e. what we would have measured, had our a priori state estimate, the measurement and the measurement model been free of error.

If we can provide the actual state-space traffic model \(A(k)/B(k)\), the actual measurement model \(H(k)\), initial estimates for \(\hat{x}(0)\) and \(P(0)\), and quantifications of the covariance matrices \(Q\) and \(R\), equations (8) – (12) unambiguously define a recursive traffic state estimation model. The remainder of this paper shows how CTM can be rewritten so as to provide the actual state-space traffic model \(A(k)/B(k)\), and provides definitions for the actual measurement model \(H(k)\), both for correcting the traffic state variables \(s(k)\), and for correcting the augmented state variables \(p(k)\) and \(d(k)\).

3 STATE-SPACE FORMULATION OF THE CELL-TRANSMISSION MODEL WITH IMPLICIT MODE SWITCHING

In this paper, we use first order traffic flow theory as the model for traffic flow on network links. In continuous form, the first order traffic flow model reads:

\[
\frac{\partial \rho}{\partial t} + \frac{\partial Q(\rho)}{\partial x} = q_{in}
\]
(13)

The density \(\rho\) and the net inflow function \(q_{in}\) both depend on the continuous time and space coordinates \(t\) and \(x\), which is omitted here for notational convenience. Although the fundamental diagram \(Q(\rho)\) can in theory have any (concave) shape, we assume in this paper a triangular relationship between flow and density:

\[
Q(\rho) = \min\left[F \rho, w(J - \rho)\right]
\]
(14)

This flow-density relationship has three parameters: the free speed \(F\), the backward wave speed \(w\), and the jam density \(J\). The capacity \(C\) and critical density \(\rho_c\) can be shown to be:

\[
C = \frac{wJF}{w + F} \quad \rho_c = \frac{C}{F}
\]
(15)

However, because of observability, we prefer to consider capacity as the independent parameter and calculate the jam density \(J\) using equation (15).

Considering again the spatial discretization in segments \(i\) of length \(L_i\) and the time discretization in time steps of duration \(T\) indexed by \(k\), a discretized version of equation (13) reads:

\[
\rho_i(k+1) = \rho_i(k) + \frac{T}{L_i}Q_{in}(k) - \frac{T}{L_i}Q_{i-\sigma}(\rho_{i-\sigma}(k)) - Q_{i+\sigma}(\rho_{i+\sigma}(k))
\]
(16)

In equation (16), \(Q_{in}\) stands for the integration over \(L_i\) of the net inflow function \(q_{in}\). Moreover, we have applied the – slightly unconventional – notation \(i \pm \frac{1}{2}\) to refer to values that need to be evaluated at the segment boundary between \(i-1\) and \(i\), or \(i\) and \(i+1\) respectively. As can be seen, both the density \(\rho\) and the (parameters of the) fundamental diagram need to be evaluated at the segment boundaries. The cell-transmission model \((12)\) defines the flow over a segment boundary as follows:
\[ Q_{i+1} (\rho_{i+1}) = \min(S_i, R_{i+1}) \]  

(17)

The symbols \( S \) and \( R \) denote the sending and receiving flow respectively. They are defined by (using (14)):

\[ S_i = \min\left(C_i, F_i \rho_i\right) \]  

(18)

\[ R_i = \min\left[C_i, w_i \left(J_i - \rho_i\right)\right] \]  

(19)

Due to the presence in every part of its definition of min operators, the CTM is difficult to write in a linear state-space form, as required by equation (6). We solve this issue by using the relation between the min operator and the Heaviside function \( \Xi(y) \). With its definition:

\[ \Xi(x) = \begin{cases} 
0 & x \leq 0 \\
1 & x > 0 
\end{cases} \]  

(20)

we have the following equivalence:

\[ \min (a, b) = a\left[1 - \Xi(a - b)\right] + b\Xi(a - b) \]  

(21)

Substituting this for example in eq. (18), we may write the sending flow of segment \( i \):

\[ S_i = F_i \rho_i \Xi(C_i - F_i \rho_i) + C_i \left[1 - \Xi(C_i - F_i \rho_i)\right] \]

\[ = F_i \rho_i \delta_i + C_i (1 - \delta_i) \]  

(22)

In this equation, we have introduced the notation \( \delta_i \) as a Boolean indicating whether or not segment \( i \) is in free flowing regime. Correspondingly, we can write the receiving flow for segment \( i+1 \) according to eq. (19):

\[ R_{i+1} = C_{i+1} \delta_{i+1} + w_{i+1} \left(J_{i+1} - \rho_{i+1}\right)(1 - \delta_{i+1}) \]  

(23)

In order to rewrite definition (17) for the flux over the interface between segments \( i \) and \( i+1 \), we introduce four more Booleans in addition to \( \delta_i \):

\[ \alpha_i = \Xi(C_{i+1} - F_i \rho_i) \quad \text{free inflow into } i+1 \]

\[ \beta_i = \Xi(w_{i+1} \left(J_{i+1} - \rho_{i+1}\right) - F_i \rho_i) \quad \text{free outflow from } i \]

\[ \gamma_i = \Xi(w_{i+1} \left(J_{i+1} - \rho_{i+1}\right) - C_i) \quad \text{capacity } i \text{ limits reduced inflow } i+1 \]

\[ \delta_i = \Xi(C_i - F_i \rho_i) \quad \text{free flow within } i \]

\[ \epsilon_i = \Xi(C_{i+1} - C_i) \quad \text{capacity } i \text{ limits free inflow into } i+1 \]  

(24)

Note that these Booleans all depend on time because \( \rho_i \) and \( \rho_{i+1} \) are all time-dependent (and possibly also the parameters \( F, C, w - \) and through eq. (15) also \( J \)). Therefore, we should in principle add \((k)\) to the notation of each one of them, which we have omitted for readability. These definitions enable us to rewrite eq. (17) without any min operators:
\[ Q_{r_{i+1}}(\rho_{r_{i+1}}) = \min (S_i, R_{r_{i+1}}) \]
\[ = \delta_{i+1} \left[ \alpha_i F_i \rho_i + (1-\alpha_i) C_{i+1} \right] \]
\[ + \delta_i \left[ (1-\delta_{i+1}) \left[ \beta_i F_i \rho_i + (1-\beta_i) w_{i+1} (J_{i+1} - \rho_{i+1}) \right] \right] \]
\[ + (1-\delta_{i+1}) \delta_i \left[ \varepsilon_i C_i + (1-\varepsilon_i) C_{i+1} \right] \]
\[ + (1-\delta_i) (1-\delta_{i+1}) \left[ \gamma_i C_i + (1-\gamma_i) w_{i+1} (J_{i+1} - \rho_{i+1}) \right] \] (25)

Note that – once the Booleans of eq. (24) have been calculated for time step \( k \) – equation (25) is linear in the state variables \( \rho \) and \( \rho_{i+1} \). After lengthy but straightforward transcription of all terms in eq. (16) (analogously to what we have shown with eq. (25)), we end up with a state-space form for the discretized CTM of eq. (16):

\[ \mathbf{p}(k+1) = \mathbf{A}(k) \mathbf{p}(k) + \mathbf{B}(k) \] (26)

The matrix \( \mathbf{A}(k) = [a_{i,j}(k)] \) in this equation has a band structure, with only elements on the diagonal, upper and lower diagonal, i.e. \( a_{i,j}(k) = 0 \) \( \forall j \not\in \{i-1, i, i+1\} \). The non-zero elements of the matrix are given by:

\[ a_{i,j} = \frac{T}{L_j} \left[ (\Delta_1 + \Delta_2) \alpha_{i-1} + (\Delta_3 + \Delta_4) \beta_{i-1} \right] F_{i-1} \] (27)

\[ a_{i,j} = 1 - \frac{T}{L_j} \left[ (\Delta_1 + \Delta_2) \alpha_i + (\Delta_3 + \Delta_4) \beta_i \right] F_j \]
\[ - \frac{T}{L_j} \left[ (\Delta_3 + \Delta_4) (1-\beta_{i-1}) + (\Delta_5 + \Delta_6) (1-\gamma_{i-1}) \right] w_i \]
\[ a_{i,j} = \frac{T}{L_j} \left[ (\Delta_2 + \Delta_6) (1-\beta_i) + (\Delta_5 + \Delta_8) (1-\gamma_i) \right] w_{i+1} \] (28)

Again, we have omitted the dependency on time \( k \) for notational convenience. The Boolean symbols \( \Delta_1 \) up to \( \Delta_8 \) are combinations of the Booleans \( \delta_{i-1}, \delta_i, \delta_{i+1} \) and are defined as follows:

\[ \Delta_1 = \delta_{i-1} \delta_i \delta_{i+1} \]
\[ \Delta_2 = \delta_{i-1} \delta_i (1-\delta_{i+1}) \]
\[ \Delta_3 = \delta_{i-1} (1-\delta_i) \delta_{i+1} \]
\[ \Delta_4 = \delta_{i-1} (1-\delta_i) (1-\delta_{i+1}) \]
\[ \Delta_5 = (1-\delta_{i-1}) \delta_i \delta_{i+1} \]
\[ \Delta_6 = (1-\delta_{i-1}) \delta_i (1-\delta_{i+1}) \]
\[ \Delta_7 = (1-\delta_{i-1}) (1-\delta_i) \delta_{i+1} \]
\[ \Delta_8 = (1-\delta_{i-1}) (1-\delta_i) (1-\delta_{i+1}) \] (30)

The vector \( \mathbf{B}(k) = [b_i(k)]^T \) in the state-space CTM formulation (26) is given by:
\[ b_i = \frac{T}{L_i} Q_{in} \]
\[ + \frac{T}{L_i} \left[ (\Delta_j + \Delta_k) \gamma_{r-1} + (\Delta_5 + \Delta_6) \varepsilon_{r-1} \right] C_{r-1} \]
\[ + \frac{T}{L_i} \left[ (\Delta_1 + \Delta_2)(1 - \alpha_{r-1}) - (\Delta_4 + \Delta_5) \gamma_i \right] C_i \]
\[ + \frac{T}{L_i} \left[ (\Delta_3 + \Delta_6)(1 - \varepsilon_{r-1}) - (\Delta_3 + \Delta_7) \varepsilon_i \right] C_i \]
\[ - \frac{T}{L_i} \left[ (\Delta_4 + \Delta_5)(1 - \alpha_i) + (\Delta_3 + \Delta_7)(1 - \varepsilon_i) \right] C_{r+1} \]
\[ + \frac{T}{L_i} \left[ (\Delta_3 + \Delta_4)(1 - \beta_{r-1}) + (\Delta_5 + \Delta_6)(1 - \gamma_{r-1}) \right] w_i J_i \]
\[ - \frac{T}{L_i} \left[ (\Delta_2 + \Delta_6)(1 - \beta_i) + (\Delta_4 + \Delta_8)(1 - \gamma_i) \right] w_{i+1} J_{i+1} \]

(31)

It is interesting to note that at each time \( k \), only one of the 8 \( \Delta \)'s in equations (27) – (31) is different from zero. As a result, many terms in these equations reduce to zero and the matrix \( A \) and vector \( B \) are actually very simple combinations of the parameters \( F, C, w, \) and \( J \) – in spite of their quite extensive general notations. The general notation is extensive because it combines in one closed analytical expression a very large number of potential linear modes, between which the nonlinear CTM model may switch at any time step.

At this point, we have rewritten the nonlinear CTM traffic model in a state-space formulation. Once the Booleans \( \alpha \) to \( \varepsilon \) and \( \Delta_1 \) to \( \Delta_8 \) have been calculated for each segment \( i \) in a time step \( k \), equations (27) – (31) implicitly perform mode-switching so as to provide a locally linearized traffic flow model around the actual traffic state. This is the linear model that is used in equations (8) and (9) of the Kalman filtering procedure.

## 4 TREATMENT OF BOUNDARY CONDITIONS IN THE STATE-SPACE CTM TRAFFIC MODEL

We consider three types of boundaries so far: upstream boundaries, downstream boundaries, and boundaries at on/off-ramps. The former two types are boundaries on the main road of the motorway. We choose their position so that they coincide with the location of a traffic detector. The latter type is modeled through the source term \( Q_{in} \) in equation (31).

For upstream boundaries, a time series of upstream demand flows \( \{ Q_0(k) \} \) needs to be available. We then define an origin segment with index \( i = 0 \) that is actually outside the physical network. For that segment, the following definitions apply \( \forall k \):

\[ C_0 = \infty \]
\[ F_0 = e^{\omega t} \]
\[ \rho_0(k) = \frac{Q_0(k)}{F_0} \]
\[ \delta_0(k) = 1 \]

(32)

The capacity and free speed have no physical meaning. They are just intended to enable calculation of equations (27) – (31). Traffic congestion can spill back into the origin segment without hampering the calculation, however, the definitions of (32) do not keep track of a vertical stack of vehicles. A generalization of the origin segment concept that would track vertical queues is however straightforward.
For downstream boundaries, we define a destination segment with index $N+1$, which also lies outside the physical domain. This segment represents the traffic conditions on the downstream traffic detector, from which two time series $\{Q_{N+1}(k)\}$ and $\{V_{N+1}(k)\}$ need to be available. We define a speed threshold $V_{\text{cong}}$, below which the conditions on the downstream detector are considered to be congested. Dependent on this, the following definitions apply $\forall k$:

$$
C_{N+1}(k) = \begin{cases} 
C_N(k) & V_{N+1} \geq V_{\text{cong}} \\
Q_{N+1}(k) & V_{N+1} < V_{\text{cong}} 
\end{cases}
$$

$$
\delta_{N+1}(k) = \begin{cases} 
1 & w_{N+1} = 0 \text{ (any value will do)}
\end{cases}
$$

Note that any value for the backward wave speed will do, since it occurs in eqs. (27) – (31) only in combinations with $\Delta$’s that are zero due to the fact that $\delta_{N+1}$ of the downstream boundary cell is by definition set equal to 1. Definition (33) imposes an outflow constraint (modeled as a capacity constraint equal to the observed flow) in the case of a congested measurement, and no constraint otherwise (unconstraint outflow because $\delta_{N+1}$ forces free flow in the destination segment and its capacity is set equal to that of the preceding segment).

Boundaries through on/off-ramps are not elaborated. For now, the simplest solution is used: we consider them as a source term, to be added as $Q_{in}$ in eq. (31) (both positive and negative flows possible). Note that this does not necessarily prevent negative density in the corresponding segment, which is simply circumvented by testing for non-negativity of $\rho$ and replacing $\rho$ by zero if necessary. A more elegant solution for modeling outflows as a fraction of the flow in the adjacent segment is provided by (6) and can be implemented in a straightforward manner in our framework.

5 MEASUREMENT MODELS FOR THE STATE-SPACE CTM TRAFFIC MODEL

We consider only measurements by local detectors (spot measurements like inductive loop, infrared, radar detectors), thus excluding vehicle identification systems and floating car data. It is assumed that the location of the measurement always coincides with the boundary between two segments (actually, the definition of the segments can always be chosen, so that this requirement is fulfilled).

In general, spot measurements may provide flow, occupancy, and/or speed data. It is not advisable to use flow measurements directly for the Kalman filter because the CTM model is based on density state estimations and flows can correspond in general to two regimes with different density: free flowing or congestion. Neither can occupancy measurements be used directly, since occupancy is not a state variable of the model, nor is it analytically linked to the density through the fundamental diagram or through the fundamental relation $q = \rho v$.

However, an unambiguous relationship between occupancy and density can be established, if one makes assumptions on the average vehicle length. For that reason, we do not treat occupancy data separately, but assume that beforehand, they have been transformed into (estimated) density data.

We do similarly for flow measurements: through the fundamental relation $q = \rho v$ we convert them – together with the speed data – into estimates of the density. The error that is introduced by this approximation should be accounted for by an appropriate choice of the measurement error covariance matrix $\mathbf{R}$.

5.1 Measurement model for (approximated) density data

Since the measurement is taken on the segment boundary between segment $i$ and $i+1$, the flow over this boundary is given by equation (25). It is easy to verify from that equation that the corresponding density that would be observed over the segment interface is:
\[ \rho_{i+\frac{1}{2}} = \delta_i \delta_{i+1} \left[ \alpha_i \rho_i + (1-\alpha_i) \frac{C_{i+1}}{F_{i+1}} \right] \]

\[ + \delta_i (1-\delta_{i+1}) \left[ \beta_i \rho_i + (1-\beta_i) \rho_{i+1} \right] \]

\[ + (1-\delta_i) \delta_{i+1} \left[ \varepsilon_i \frac{C_i}{F_i} + (1-\varepsilon_i) \frac{C_{i+1}}{F_{i+1}} \right] \]

\[ + (1-\delta_i)(1-\delta_{i+1}) \left[ \gamma_i \frac{C_i}{F_i} + (1-\gamma_i) \rho_{i+1} \right] \]

(34)

We see from that equation that the density measurement \( m \) at \( i+\frac{1}{2} \) only relates to elements of the traffic state in the adjacent segments (this was already obvious without considering the equation). As a consequence, all elements \( h_{m,j} \) with \( j \notin \{i,i+1\} \) of the measurement model matrix \( H \) must be equal to zero. It turns out that if measurement \( m \) is a density measurement taken at interface \( i+\frac{1}{2} \), the \( m \)-th row in the measurement model reads:

\[ z_m^p = \rho \left[ \delta \delta_{i+1} \left[ \alpha_i \rho_i + (1-\alpha_i) \frac{C_{i+1}}{F_{i+1}} \right] \right] \]

\[ + \rho_{i+1} \left[ \delta_i (1-\delta_{i+1}) \left[ \beta_i \rho_i + (1-\beta_i) \rho_{i+1} \right] \right] \]

(35)

For the sake of completeness, we mention that in principle some constants should be added to eq. (35). That is: the equation should contain some critical densities \( C/F \) multiplied by some Boolean values, leading to a measurement model \( z(k) = H(k) x(k) + D(k) + \eta(k) \) instead of equation (7). However, for the Kalman procedure, only the sensitivity \( H(k) \) of the measurements with respect to variations in the state variables is relevant, which is why we omit the constant term here.

In order for the state estimation to be robust against missing data, we specify that in that case, all elements \( h_{m,j} \) on row \( m \) of the measurement model \( H \) should be zero. As a result, the corresponding row in the Kalman gain \( K(k) \) will be zero and no correction will take place; i.e. the CTM prediction will interpolate the missing value.

### 5.2 Measurement model for speed data

As in the case of density data, we observe that the flow over the boundary between segments \( i \) and \( i+1 \) where the measurement was taken is given by equation (25). Using the fundamental relation \( q = \rho v \), (25) can be converted into an equation for the speed that would be observed over the segment interface:

\[ v_{i+\frac{1}{2}} = \delta_i \delta_{i+1} \left[ \alpha_i F_i + (1-\alpha_i) \frac{J_{i+1} - \rho_{i+1}}{\rho_{i+1}} \right] \]

\[ + \delta_i (1-\delta_{i+1}) \left[ \beta_i F_i + (1-\beta_i) \rho_{i+1} \right] \]

\[ + (1-\delta_i) \delta_{i+1} \left[ \varepsilon_i F_i + (1-\varepsilon_i) \frac{J_{i+1} - \rho_{i+1}}{\rho_{i+1}} \right] \]

\[ + (1-\delta_i)(1-\delta_{i+1}) \left[ \gamma_i F_i + (1-\gamma_i) \rho_{i+1} \right] \]

(36)

The speed over the segment interface is not linearly dependent on the state variable \( \rho \). As in the case for density data, we argue that only the sensitivity to changes of the state variables is required. Because only the density in segment \( i+1 \) occurs in (36), it is clear that all elements \( h_{m,j} \) with \( j \notin \{i+1\} \) of the measurement model (sensitivity or Jacobian) matrix \( H \) must be equal to zero. After applying a first order Taylor approximation of (36), we find that if measurement \( m \) is a speed measurement taken at interface \( i+\frac{1}{2} \), the \( m \)-th row in the measurement model reads:
At this point, the measurement models required by the state estimation procedure of section 2 have been established, and the Kalman filter procedure can be applied. Before turning to a simulated case study, we will specify how the parameters of the CTM can be calibrated online.

5.3 Some remarks on the update frequencies of measurement and system models

For the CTM traffic model to provide accurate simulations, it requires a relatively fine discretization in space and time. Typically the order of magnitude of the spatial discretization is 100 to 500 meters, requiring a maximum time increment $T$ of 3 to 15 seconds for free speeds around 120 km/h. Measurements are typically aggregated in time intervals $T_m$ of 1 minute or more. This means that the system model has to come up with updates (prediction steps) at a higher frequency, than the availability of new measurement data.

One option would be to apply the Kalman correction at the same frequency of the system model. The latest measurements could then be repeatedly used until updated measurements come available. In our opinion, this has two disadvantages:

- the Kalman correction is a computationally expensive procedure, since it requires matrix inversion in (10). One should therefore refrain from too frequent calculations, unless this is justified by the presence of new data.
- As time elapses, the latest measurement data becomes outdated. Actually, at the time $t$ when it came available, the information contained in the data was already ‘old’ (it is an average value of the interval $[t-T_m, t]$, and its ‘age’ is therefore $T_m/2$). The Kalman correction therefore more or less ‘resets’ the system state by the same time shift $T_m/2$. We would only increase this numerical error if we would repetitively use the data from $[t-T_m, t]$ until $t+T_m$.

For these two reasons, we perform the correction step only once every $T_m$ seconds. We make an average $\hat{\rho}^- (k)$ of all the system updates …, $\hat{\rho}^- (k-1), \hat{\rho}^- (k)$ since the latest correction (i.e. between $t-T_m$ and $t$). Whereas in the Kalman correction (11), the measurement model (35) and (37) is evaluated using $\hat{\rho}^* (k)$, we apply this correction term $K(k)[z(k) - H(k)\hat{\rho}^* (k)]$ to the latest available system model prediction $\hat{\rho}^- (k)$, so that the modified Kalman correction equation reads:

$$\hat{\rho} (k) = \hat{\rho}^- (k) + K(k)[z(k) - H(k)\hat{\rho}^* (k)]$$

(38)

6 ONLINE PARAMETER ESTIMATION

For the online estimation of parameters and boundary flows, we follow the approach outlined by (6) and summarized in section 2. Parameters and boundary flows are added there as additional variables. Formally this was done by augmenting the state vector $s$ (in the CTM this corresponds to $\rho$) with the parameter and boundary vectors $p$ and $d$ respectively to yield the augmented state $x = [s^T \ p^T \ d^T]^T$.

In contrast to the formal notation in one single augmented state $\chi$, we detach the estimation of traffic state and that of the parameter and boundary vectors $p$ and $d$ in practice for two reasons:

- the augmented state $\chi$ has much larger dimensions than the original state $\rho$. This also increases the dimensions of the matrix to be inverted in equation (10), leading to much higher computation times. However, this is not necessary, since the random walk system models for $p$ and $d$ (eqs. (2) and (3)) have no interaction with the traffic system model. Therefore, the recursive Kalman filters for $p, p$ and $d$ can all be decomposed into distinct models of lower dimension.
- Corrections for $p$ and $d$ need not necessarily be performed at the same frequency as corrections to $\rho$. Capacity estimates (‘measurements’) for instance, might be obtained through an independent online capacity estimator, running at its own frequency. It is therefore advisable to detach the estimation processed.
In the next subsections, the measurement models for the parameters and boundary conditions are specified.

6.1 Parameter measurement models

In each segment \(i\), there are three independent parameters: \(C_i\), \(w_i\), and \(F_i\). The backward wave speed \(w_i\) will not be calibrated online, because it is well-known that it is a constant that varies very little \((15, 16)\).

6.1.1 Measurement model for the free speed

The free speed \(F_i\) may vary with traffic composition, weather or light conditions, or as a response to variable speed limit signs. Since we assume a piecewise linear fundamental diagram \((14)\), free speed is observed whenever we are in free flow mode. There are roughly two ways in which to determine free flow or congested traffic operations: based on the actual measurement or based on the online model state estimate. We use a combination of both methods here.

On the one hand, we use the speed measurement \(v_{i+1/2}\) of detector \(m\) on segment interface \(i+1/2\) to distinguish between free flow or congested mode around the detector. This speed is compared against a threshold \(v^*\) (for instance 70% of the prevailing speed limit). A higher speed is considered a free flow speed measurement.

On the other hand, we use the state estimation by the online model. In order to do that, we first need to identify the segments \(i' = i - n_m^-, i - (n_m^- - 1), \ldots, i, \ldots, i + (n_m^+ - 1), i + n_m^+\) around segment interface \(i+1/2\), for which the speed measurement at \(m\) may be a representative sample of the free flow speed \(F_{i'}\). The numbers \(n_m^-\) of upstream segments and \(n_m^+\) of downstream segments associated with \(m\) determine the ‘influence range’ of measurement point \(m\) and need to be configured depending on the network under consideration. It means that the free speeds in \(n_m = n_m^- + n_m^+ + 1\) segments around interface \(i+1/2\) will be adapted together throughout the simulation. Although this approach does not require each segment \(i\) to be only associated with strictly one measurement \(m\) (influence ranges of successive detectors might in principle overlap partly), we have no experience other than with disjoint detector influence ranges and do not see arguments to consider overlapping influence ranges. Let us now define for each measurement \(m\) a free speed association vector

\[
h_m^F = \begin{bmatrix} h_{m1}^F & \cdots & h_{mi}^F & \cdots & h_{mN}^F \end{bmatrix},
\]

a row in the measurement model \(H^F\) for the free speed:

\[
h_{m,i'}^F = \begin{cases} \frac{1}{n_m} & i' \in \{i - n_m^-, i - (n_m^- - 1), \ldots, i, \ldots, i + (n_m^+ - 1), i + n_m^+\} \\ 0 & \text{otherwise} \end{cases}
\]

With these definitions and the notations \(\hat{F}^F = \left[ \hat{F}_i^F \cdots \hat{F}_j^F \cdots \hat{F}_N^F \right]^T\) for the a priori estimates of the free speeds of the segments and \(v = [v_1, \ldots, v_i, \ldots, v_N]^T\) for the average actual speeds \(v_i = \min\left(F_i, \frac{w_i(J_i - \rho_i)}{\rho_i}\right)\) within the segments, the influence range around measurement \(m\) is in free flow mode according to the online model if:

\[
h_m^F v = h_m^F \hat{F}^F
\]

Equation (40) is equivalent to saying that in each segment in the influence range of \(m\), there is free flowing traffic according to the online model. A Kalman correction to the free speeds in these segments will then be performed if both the measurements and the online model indicate free flow mode, i.e. if measurement \(v_{i+1/2}\) is below threshold \(v^*\) and condition (40) is fulfilled. If this is not the case, the row \(h_m^F\) in \(H^F\) is replaced by zeros. With this \(H^F\) as the measurement matrix, and the random walk (2) as the system model, equations (8) – (12) can be used to set up an iterative Kalman prediction – correction cycle for the free speed vector \(F\).
6.1.2 Measurement model for capacity

The capacity $C_i$ of a segment may change due to weather conditions, incidents, traffic management measures etc. Given a flow measurement $q_{i+\frac{1}{2}}$ and speed measurement $v_{i+\frac{1}{2}}$ observed by detector $m$ at the segment interface $i+\frac{1}{2}$, the challenge is: (a) to recognize whether the observed flow value can be considered a capacity flow measurement (or at least, a better approximation of capacity than the current a priori estimates $\hat{C}_i^{-}$), and (b) if we think we measure a capacity flow, to determine with which segment it should be associated.

Let us start with the latter question: if we measure capacity at $i+\frac{1}{2}$, which segment’s capacity should be corrected? Just like equation (39) for the free speed measurement model, assume that the user has configured a matrix $H^C$, indicating the association between segments $i$ and capacity measurement locations $m$. Each row $h^C_m$ of this matrix contains non-zero entries only for segments within an ‘influence range’ (this time for capacity) around segment $i$ that sum up to one. If measurement $m$ can be considered a capacity observation, all segments within this influence range will be corrected simultaneously.

This brings us to the criteria that determine whether or not we measure capacity flow at detector $m$. Any external online capacity estimation model can be used for this purpose. In absence of that, we propose three criteria, any of which is sufficient to identify flow measurement $q_{i+\frac{1}{2}}$ as a capacity flow observation. The first one recognizes underestimation of capacity, simply because higher flows are measured. The second one recognizes queue discharge, which happens at capacity flow according to first order traffic flow theory. The third one recognizes the presence of a bottleneck because less traffic is observed downstream than can be expected by propagation of traffic that was observed earlier upstream. The three criteria are specified as follows:

1. The observed flow $q_{i+\frac{1}{2}}$ is higher than the actual a priori estimates $\hat{C}_i^{-}$ within the influence range for capacity. This is the case if:

   $$q_{i+\frac{1}{2}} > h^C_m \hat{C}_i^{-} \quad (41)$$

2. The speed measured at $m$ is a free flow value, whereas that of the measurement $m-1$ upstream of $m$ is congested. We check this by comparing the measured values against the average free flow speeds in the respective influence ranges around $m-1$ and $m$:

   $$v_{i+\frac{1}{2}} > p^\text{free}_m h^C_m \hat{C}_i^{-} \quad \text{AND} \quad v_{i+\frac{1}{2}} < p^\text{cong}_m h^C_{i-1} \hat{C}_{i-1} \quad (42)$$

   In this equation the percentages $p^\text{free}$ and $p^\text{cong}$ are parameters determining the free flow and congested speed thresholds respectively (e.g. 90% and 60%)

3. During $N^C$ (e.g. 4) consecutive measurement intervals, the flow $q_{i+\frac{1}{2}}$ observed at $m$ is consistently lower than values observed upstream in earlier time steps and propagated downstream by the online model\(^1\). This means that consistently the following condition is true:

   $$q_{i+\frac{1}{2}} < Q_{i+\frac{1}{2}} \left( \rho_{i+\frac{1}{2}} \right) - p^\text{cap}_m h^C_m \hat{C}_i^{-} \quad (43)$$

   This condition requires that the observed flow $q_{i+\frac{1}{2}}$ is lower than the expected observation $Q_{i+\frac{1}{2}} \left( \rho_{i+\frac{1}{2}} \right)$ according to (25) minus a certain fraction $p^\text{cap}$ of the average capacity in the influence range around $m$.

If any of the criteria (41) – (43) holds for measurement $m$, the associated row $h^C_m$ of the capacity measurement matrix $H^C$ is as defined before, otherwise it is replaced by a row or zeros. With this $H^C$ as the measurement

\(^1\) Note that actually this is an online incident detection algorithm that uses the forward kinematic wave for rapid incident detection. Because of the high probability of ‘false alarms’ (that would lead to unnecessary capacity corrections), it is required here that the incident is consistently detected during $N^C$ consecutive time steps.
matrix, and the random walk (2) as the system model, equations (8) – (12) can be used to set up an iterative Kalman prediction – correction cycle for the capacity vector $C$.

### 6.2 Boundary measurement and prediction models

We assume for now that the downstream boundary guarantees free outflow and imposes no outflow restriction. It is imaginable however to apply EKF-CTM to cases where recurrent congestion from downstream enters the network under consideration each peak period. In that case, the downstream boundary can be estimated and predicted similarly to what is described in this section.

We thus focus on upstream boundaries or unknown inflows that are modeled as source terms (see also section 4). There are two possibilities: either the flow is observed by some measurement point (observed boundary flow $q_{dq}^{obs}$), or it is unobserved ($q_{dq}^{unobs}$), but can be derived as a sum or difference of other observed flows. We update the estimation of these boundaries with a period $T_m$ greater than or equal to the update frequency $T_m$ of measurement data. At each measurement time tick $k^m$, a new observation (direct or sum/difference of direct observations) of $q_d$ comes available, with which we update a moving average. Denoting the number of measurement updates since the latest boundary correction by count index $j$, we can write:

$$
\bar{q}_d(k^m) = \left(1 - \frac{1}{j}\right) \bar{q}_d(k^m - 1) + \frac{1}{j} q_d(k^m)
$$

The latest value of $\bar{q}_d(k^m)$ at the time of a new correction of the boundaries is denoted with time index $k^d$. Suppose that the user has access to a database of standard time series $q_d^{k^d}$ of the boundary flow $q_d$, that represents for instance the average pattern in a weekday. Such time series is not a stationary process, and hence not very well represented by the random walk system model of equation (3). However, the deviation from the standard time series is much more stationary:

$$
f_{q_d}(k^d) = \frac{\bar{q}_d(k^d)}{q_d^{k^d}}
$$

This deviation factor can be estimated using the random walk model (3) and the standard Kalman filter equations (8) – (12). Finally, in order to obtain predictions of the boundary flows, this factor is applied to future values of the standard time series: for each future time tick $k^d + \Delta k$, the predicted boundary is:

$$
q_d(k^d + \Delta k) = f_{q_d}(k^d) q_d^{k^d}(k^d + \Delta k)
$$

This equation assumes that the actual deviation $f_{q_d}(k^d)$ from the standard time series will remain constant in the near future, whereas the boundary flow does follow the general trend contained in the standard time series. Note that this approach is also applicable to situations in which one does not dispose of a time series of standard boundary values. They can be simply replaced by a constant, that will be scaled by $f_{q_d}(k^d)$ to obtain an actual estimate of the boundary condition. Eq. (46) then assumes that this condition will remain constant in the near future.

### 7 A SIMULATION CASE STUDY OF THE EKF-CTM ONLINE TRAFFIC STATE ESTIMATION AND PREDICTION MODEL

This case study is intended to illustrate the behaviour or the EKF-CTM online traffic flow model. We consider a homogeneous motorway of 15 km, equipped with detectors at longitudinal positions: 1.8, 5.0, 8.2, 11.4, and 14.6 km (each 3.2 km). The simulation covers a period of 4 hours. After 45 minutes, an incident just upstream of detector 14.6 km blocks one of 3 lanes during 105 minutes, after which the road is cleared. Demand is stochastic, with a typical rush hour evolution: in the first 30 minutes, demand rises sharply from 0 to approximately 5500 veh/h. It remains on that level in the next 2 hours, after which it gradually reduces over 1 hour to a level of 3000 veh/h.
The scenario was simulated with a microsimulation model (Aimsun2). Synthetic speed and flow loop detector data were collected and aggregated in 1 minute periods at the detector positions. These data were fed into the EKF-CTM model that was discretized in 75 segments of 200 meters and a time step of 6 seconds.

In Figure 1a and b space-time plots of the density are given for the reference (microsimulation) and the online model estimation (this is called the ex-post estimation to make a distinction with predicted traffic conditions). Gridlines indicate the start and end of the incident (vertical) and the position of the detectors (horizontal). The extent and severity of the congested region are reproduced well by the EKF-CTM model. Only the position of the incident differs noticeably. Thanks to condition (43), the online model tracks the capacity reduction after a slight delay (Figure 1c) and between the correct detector pair. However, the model cannot estimate the exact location of the incident in between the detectors, and assumes it halfway. Notice that the exact location (just upstream of a detector) is the worst case, maximizing this position error. Figure 1d shows the travel times in the reference micro model and in the online model. They were calculated in two steps. First, cell travel times for each minute were derived from the number of vehicles present, divided by the outflow rate. Secondly, we need to distinguish between instantaneous and experienced travel time. The former can be estimated by summing cell travel times from one time instant, as if a vehicle drives through the jam while position, length, and severity of the jam remain unchanged in time with respect to the moment of entering the road upstream. We see that the instantaneous travel time estimates by the online model coincide well with those from the reference model: there is a slight delay and underestimation of travel time due to the delayed tracking of the capacity reduction and due to the position error of the queue head respectively. Notice however from Figure 1d that the difference between the instantaneous and experienced travel times is substantial in this case. The latter is defined as the time difference between the start and end point of a trajectory drawn through the space-time plot. Because travel times are high (up to 40 minutes) the conditions experienced throughout the journey change substantially (i.e. shock wave fronts with speeds between 0 and 20 km/h can travel large distances).

Figure 2 shows results of traffic state predictions obtained by the online model. They are obtained through simple extrapolation of the state estimation, using the CTM model equations and boundary conditions fixed at the latest estimation. Figure 2a and b show the references of the micro simulated ‘virtual’ world and the online estimation respectively. Figure 2c and d show predictions over 15 minutes (shifted over the prediction horizon to allow for easy comparison with the references – a perfect prediction would look identical in this way). Of course, the occurrence of the incident cannot be predicted, yielding a too optimistic prediction for the first 15 minutes of the incident. However, we see that – once the capacity reduction has been tracked, the predictions compare nicely to the reference. A similar phenomenon is observed at the end of the incident. Figure 2c shows the prediction if the model has to track the capacity increase by itself. Naturally, it cannot predict the clearance of the road and the prediction is by far too pessimistic for the first 15 minutes after clearance. However, once it catches the restored capacity, dissolution of congestion is predicted correctly. Figure 2d shows the prediction, given that the incident response team is able to forecast the time needed to clear the road at least 15 minutes beforehand. The correspondence between the 15 minute prediction and ‘reality’ is now striking. Figure 2e and f show similar results for a 30 minute prediction. Apart from the previous observations that hold a fortiori in this case, it is also clear that the length of the queue is overestimated by 2.5 km. This is because there is a decreasing trend in traffic demand, which the predictions do not account for (the predictions are made assuming a constant upstream inflow of traffic, see eq. (46)).

Finally, Figure 3 shows how the traffic state predictions of Figure 2 translate into travel time predictions. Remember that predicted travel times are obtained as instantaneous travel times calculated in the predicted traffic states. We also compare them as such to the corresponding instantaneous travel times of the reference situation. The reader should therefore add to the prediction error, the error between predicted and experienced travel times, as was discussed in the context of Figure 1d. Figure 3 confirms that – after the initial delay due to the unpredictability of the incident – travel times are calculated well. However, the figure also shows the importance of appropriate and timely forecast of incident clearance: the 30 minute travel time prediction for time 165 min otherwise rises to 60 minutes, whereas this reduces to 40 minutes with knowledge of incident clearance (by that time, experienced travel times have dropped to 25 minutes).

8 SUMMARY, CONCLUSIONS AND FUTURE WORK

This paper showed how a discretized version of the first order traffic flow model (cell-transmission model) can be applied in an online setting using (extended) Kalman filtering. The main contribution consists in finding a closed analytical state-space formulation for the strongly non-linear CTM model that implicitly switches between numerous linear modes. The resulting model matches seamlessly with standard linear Kalman filter
theory. Another contribution is the formulation of measurement models for important model parameters and boundary conditions, so that these can also be automatically estimated online. Moreover, these estimation processes are numerically detached from the main traffic state estimation problem, herewith improving numerical efficiency and allowing different update frequencies for the different estimation processes.

The case study analyses the behaviour of the EKF-CTM online model with reference to an independently (micro)simulated virtual world. It shows the potential of the online model to track and predict traffic conditions, including unforeseen (large) variations of important model parameters like capacity. However, it also shows inherent limitations imposed by unpredictability of incidents and the importance of distinguishing between instantaneous and experienced travel times. The prediction for a longer horizon (30 min) showed the importance of the prediction of boundary conditions (like upstream inflow). One can expect that this importance increases as the prediction horizon becomes longer, and as the boundaries vary more rapidly in time.

The case study considered traffic on a motorway, whereas the ambition is to apply the EKF-CTM online model also to urban traffic. Future work will focus on the differences between urban and motorway traffic and how to overcome these in the EKF-CTM model. The proven capability of tracking rapid changes in parameter values (here: capacity) is encouraging, as it is expected that urban traffic will be dominated (among others) by time-varying turning fractions and signal plans (capacities at intersections).

Another issue for future work is the prediction of boundary conditions. It is well-known that time series are better predictable if one can isolate a systematic (periodic) component and the process to be predicted is stationary’. One way of doing so is to track deviations from an average pattern, which is difficult for unobserved boundaries (e.g. unmeasured inflows). Future work will focus on methods for using the online model to establish average patterns from ex-post estimation of unobserved boundaries, and using these for better predictions (self-learning properties).

Finally, future work will address the challenge of incorporating floating car data and data from vehicle re-identification systems. The main problem to be solved is that such data contains information not only on the most recent time increment (e.g. latest minute), but on a much longer history of states (e.g. 5 – 20 minutes). This is not trivially compatible with the prediction – correction cycle of the Kalman filter. However, a solution to this issue is increasingly important as future traffic monitoring systems for urban networks will rely more and more on this type of data sources.

ACKNOWLEDGEMENT

The research reported in this paper was financially supported by the Institute for the Promotion of Innovation by Science and Technology in Flanders IWT (post-doctoral fellowship no. IWT OZM 050315). Their support is herewith acknowledged. We also wish to thank Transport & Mobility Leuven, who provided financial support during the early stages of the research.

REFERENCES


Figure 1: a) space-time plot of density (microsimulation reference); b) space-time plot of density (EKF-CTM online estimation); c) EKF-CTM online capacity estimation; d) travel time in reference and online model; for the reference, distinction is made between instantaneous and experienced (trajectory) travel time.
Figure 2: Space-time plots of density; a) microsimulation reference; b) EKF-CTM on-line estimation (ex-post); c) EKF-CTM 15-minute prediction, without knowledge of the incident duration; d) EKF-CTM 15-minute prediction, incident duration known; e) EKF-CTM 30-minute prediction, incident duration unknown; f) EKF-CTM 30-minute prediction, incident duration known
Figure 3: Travel time estimates using the EKF-CTM online traffic model