TUTORIALS IN
OPERATIONS RESEARCH
INFORMS 2012

 | ISBN 978-0-9843378-3-5
http://dx.doi.org/10.1287/educ.1120.0099

Applications of Kalman Filtering in Traffic Management and Control

Hans van Lint, Tamara Djukic
Faculty of Civil Engineering Geosciences, Department Transport and Planning, Delft University of Technology, 2600 AA Delft, The Netherlands {j.w.c.vanlint@tudelft.nl, t.djukic@tudelft.nl}

Abstract In many areas of traffic management and control the variables that are of most interest are often the ones that are most difficult to measure and estimate. Take for example vehicular density (vehicles per kilometer) and space-mean speed (kilometers per hour). A reliable real-time estimate of these quantities is critically important for real-time control of traffic networks. However, neither can be straightforwardly deduced from available sensor data. Similarly elusive quantities are origin–destination (OD) flows. These depict the amount of vehicles per hour planning to go from one place to another at a certain moment. Unless we literally know the origin and destination of *all* vehicles on a traffic network, advanced estimation techniques are required to extract OD patterns from whatever data and prior knowledge we have available. One intuitive and highly effective method to solve these types of problems (i.e., estimating a quantity x when all we have are observations y and prior knowledge about the process) is the Kalman filter, first proposed by Rudolf Kalman in 1960. In this tutorial we explain with many examples how this technique can be applied in the domain of traffic management and control to solve real-world problems. We will see that Kalman filtering is a powerful technique that works surprisingly well in many cases, but there are also clear limitations that relate to the many assumptions underlying its application.

Keywords Kalman filter; traffic management; traffic control; traffic state estimation; OD estimation; location tracking

1. Introduction: Solving Linear Problems

The underlying rationale of the celebrated Kalman [25] filter (KF) can be explained with a very simple (and mathematically loosely formulated) example. Suppose we are interested in estimating the temperature in a room. From experience we know this temperature can be modeled by means of a random walk:

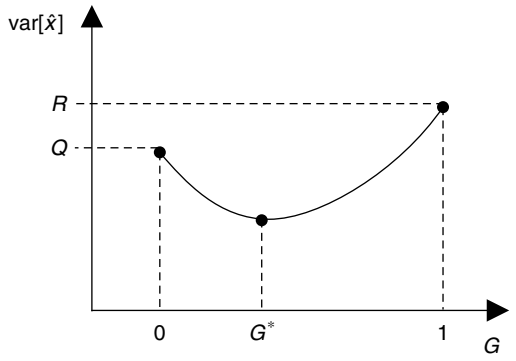
$$x_k = x_{k-1} + w, \tag{1}$$

with k depicting discrete time and w depicting some known Gaussian noise term with mean $E[w] = 0$ and variance $E[w^2] = Q$. This (process) noise term w depicts unobserved fluctuations due to physical processes in the room. Suppose we also have a thermostat that measures the temperature at one location in this room. From the specifications we learn that the thermometer observes the temperature according to

$$y_k = x_k + v, \tag{2}$$

with v depicting a second Gaussian noise term with mean $E[v] = 0$ and variance $E[v^2] = R$. The (observation) noise term v relates to characteristics of the thermometer (measurement inaccuracies, position in the room, etc.). We further assume that w and v are not correlated and do not change over time. A reasonable estimate $\hat{x}_k$ for the real room temperature in time

FIGURE 1. Determining the minimum variance weighing factor.



period k then is a weighted linear combination of the model prediction with Equation (1), which we denote $\hat{x}_k^-$, and the latest available observation y_k from our thermostat,

$$\hat{x}_k = (1-G)\hat{x}_k^- + Gy_k, \tag{3}$$

in which G represents the weighting factor. The weighted estimator in Equation (3) has the following statistical properties: $E[\hat{x}_k] = E[x_k]$, and $\text{var}[\hat{x}_k] = (1-G)^2Q + G^2R$. That implies that this estimator is unbiased, provided the assumptions on the error terms are correct. In case we would put full weight on the model estimate, we have $\text{var}[\hat{x}_k] = \text{var}[\hat{x}_k^-] = Q$. In case all weight goes to the observation, we have $\text{var}[\hat{x}_k] = \text{var}[\hat{y}_k] = R$. To determine the optimal estimator in terms of minimizing variance, we must solve $G^* = \arg\min_G \text{var}[\hat{x}_k]$, which in this case is very straightforward (see Figure 1):

$$\frac{d}{dG}\text{var}[\hat{x}_k] = -2(1-G)Q + 2GR = 0 \rightarrow G^* = \frac{Q}{Q+R}. \tag{4}$$

The optimal (minimum variance) weighting factor G^* is hence proportional to the ratio of the variance (degree of uncertainty) in our model and the variance (degree of uncertainty) in the available observations. In case the observations are unreliable ($R > Q$), more weight is put on the model prediction $\hat{x}^-$, and vice versa.

1.1. Kalman Filtering Basics

In his seminal paper in 1960, Kalman [25] rigorously showed that the minimum variance solution for *all* multivariate linear discrete dynamical systems expressed in state space form equals the elegant and intuitive estimator in Equation (3). State space systems constitute a very general class of mathematical models, which find application in many different fields, including control, automation, process technology, and transportation, to name but a few. They consist of the following equations:

$$\mathbf{x}_{k+1} = F_k\mathbf{x}_k + \mathbf{w}_k \quad \text{(process equation)}, \tag{5}$$

$$\mathbf{y}_k = H_k\mathbf{x}_k + \mathbf{v}_k \quad \text{(observation equation)}. \tag{6}$$

Clearly, Equations (1) and (2) in the thermostat example are special (univariate) cases of (5) and (6). The appeal of the Kalman filter approach lies in the fact that it provides a *recursive* solution for the optimal estimation of state variables. *Recursive* implies that we can estimate the state of a system $\mathbf{x}_k$ solely based on the last known state $\mathbf{x}_{k-1}$ and whatever new information (i.e., observations $\mathbf{y}_k$) arises. This makes the approach computationally very efficient and suitable for real-time applications. Algorithm 1 summarizes the Kalman filter equations. Before we discuss applications in the domain of traffic management and control,

we first briefly discuss a few key characteristics of the Kalman filter related to Algorithm 1 and the thermostat example above.

- First note that there are two key differences between with the filter gain G^* introduced in Equation (4) and the Kalman gain in Equation (12). The first difference is that in the Kalman filter algorithm we introduce a separate term P to depict the estimation error covariance. Examing Equation (10) it becomes clear why. Because the noise terms are independent, the estimation errors are additive. In other words, the estimation error covariance P grows proportional to the prior process noise covariance Q every time we use a new model prediction. Of course, P decreases again after we make an observation (Equation (14)). The second difference is that in Equation (4) we implicitly left out the term H_k, which in the univariate thermostat example equals 1 anyway.
- The Kalman filter gain in Equation (12) still has the same intuitive structure as the G^* in Equation (4) and can be informally understood as follows:

$$G_k = \frac{\text{uncertainty process model} \times \text{sensitivity state to observations}}{\text{uncertainty observation model} + \text{uncertainty observations}}. \tag{7}$$

The added term in the numerator ("sensitivity state," i.e., H_k) makes intuitive sense. In the limit that the observation model is constant, e.g., $H_k = C$, $C \in \mathcal{R}, \forall x_k \in [x_a, x_b]$, it clearly makes no sense to adjust the estimates of the state variables for $x_k \in [x_a, x_b]$, because changes in the observations are not related to changes in state variables over this domain.

- Algorithm 1 illustrates that the Kalman filter follows an intuitive prediction-correction scheme. In the prediction step we use our model (i.e., our knowledge on the evolution of the system state) to predict the next system state. In the correction step the observations are used to correct that prior prediction. The assumptions we make on the variance of the model and the observations govern how large or small the corrections are.
- The fact that the Kalman filter equations work for any system cast in state space form enables one to estimate state variables $\mathbf{x}$ even if we have no direct observations of these variables. As long as observations $\mathbf{y}$ are available that are (for now linearly) related to these state variables, the Kalman filter provides an efficient method to estimate the (unobserved) state variables.
- Another great feature is the fact that the Kalman filter estimates both mean and (error) covariance. This implies we have automatically error bounds around the estimated state variables, that is,

$$x_{i,k} = \hat{x}_{i,k} \pm \alpha \sqrt{P_{ii,k}}, \tag{8}$$

with, typically, $\alpha = 2$ for a nominal 95% confidence interval.

- The assumption of independent zero-mean Gaussian distributions for all noise terms makes the Kalman filter also optimal in two other respects (aside from minimizing variance). First, the Kalman filter algorithm calculates the *maximum a posteriori probability* of $\mathbf{x}_k$ given all observed data $\mathbf{y}_n$, $n = 1, \ldots, k$:

$$\hat{\mathbf{x}}_k = \arg\max_{\mathbf{x}_k} \Pr(\mathbf{x}_k \mid \mathbf{y}_1 \ldots \mathbf{y}_k).$$

Second, the solution is also optimal in terms of computing the conditional mean, that is,

$$\hat{\mathbf{x}}_k = E[\mathbf{x}_k \mid \mathbf{y}_1 \ldots \mathbf{y}_k].$$

- Although in many cases the noise parameters are assumed constant, that is, $Q_k = Q$, $\forall k$, and $R_k = R$, $\forall k$, we will see further below that these noise parameters can also be time varying, resulting in an adaptive filter.

Algorithm 1 (The Kalman filter)
Preliminaries: For the given problem we have a state space model depicted by Equations (5) and (6), in which $\mathbf{w}_k$ and $\mathbf{v}_k$ are independent, zero-mean, Gaussian noise processes of covariance matrices Q_k and R_k, respectively, with

$$E[\mathbf{w}_k\mathbf{w}_n^T] = \begin{cases} Q_k & \text{for } n=k, \\ 0 & \text{for } n\neq k; \end{cases} \qquad E[\mathbf{v}_k\mathbf{v}_n^T] = \begin{cases} R_k & \text{for } n=k, \\ 0 & \text{for } n\neq k. \end{cases}$$

Initialization:

$$\hat{\mathbf{x}}_0 = E[\mathbf{x}_0] \quad \text{and} \quad P_0 = E[\mathbf{x}_0 - E[\mathbf{x}_0]^T].$$

In case no additional information is available, P_0 is usually initialized as a matrix with a large diagonal entries, reflecting the fact that we are highly uncertain about our initial estimate of $\hat{\mathbf{x}}_0$.

For $k=1,2,\ldots$ **do**:
Predict mean and variance of state variables:

$$\hat{\mathbf{x}}_k^- = F_k\hat{\mathbf{x}}_{k-1}, \tag{9}$$

$$P_k^- = F_kP_{k-1}F_k^T + Q_{k-1}. \tag{10}$$

Predict output variables:

$$\hat{\mathbf{y}}_k^- = H_k\hat{\mathbf{x}}_k^-. \tag{11}$$

Compute the Kalman gain:

$$G_k = \frac{P_k^-H_k^T}{H_kP_k^-H_k^T + R_{k-1}}. \tag{12}$$

Update (correct) mean and covariance:

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + G_k(\mathbf{y}_k - \hat{\mathbf{y}}_k^-), \tag{13}$$

$$P_k = (1-G_kH_k)P_k^-, \tag{14}$$

in which the output errors $\mathbf{e}_k = \mathbf{y}_k - \hat{\mathbf{y}}_k^-$ computed and used in (13) are sometimes referred to as the *innovations* time series, because they reflect the amount of new information contained in the observations $\mathbf{y}_k$.
End

1.2. Tracking and Tracing a Bicycle Trip

Imagine you plan a bicycle vacation and you're interested in keeping a log of the trajectory (location and speed) along a fixed route you drive. Just for sport, you decide to base your logging on manual observations using the position of celestial bodies (the sun, moon, and stars) and landmarks. Your observations of average speed are based on these location observations but also your (physical) experiences while cycling. You plan to use these observations in the evenings to make a minimum variance estimate of your trajectory. To assess these estimates later on, you bring with you accurate global positioning system (GPS) location tracking equipment to record the actual (ground-truth) locations and average speeds. Let us further assume you drive for a maximum of 10 hours a day with an average speed of around 15 km/h and log your estimated location every hour.

This estimation problem can be expressed in state space form as follows. First consider the dynamics of location s_k and speed u_k:

$$s_k = s_{k-1} + \Delta t u_k, \tag{15}$$

$$u_k = u_{k-1} + w_k, \tag{16}$$

with Δt the discrete time period (in this case, one hour) and w_k again a (possibly time-dependent) zero-mean Gaussian noise term that models the variations in average speed you will maintain. Note that the law of motion (Equation (15)) is exact (no noise term). With $\mathbf{x}_k = [s_k, u_k]^T$, Equations (15) and (16) result in

$$\mathbf{x}_k = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \mathbf{x}_{k-1} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} w_k. \tag{17}$$

The second ingredient we require are observation equations. All four scenarios listed below are tested with (A) the observation noise covariance matrix R_k^A (which is approximately correct) and (B) covariance matrix R_k^B (which is incorrect). Both of these are listed in Table 1.

Scenario I. We use the location observations to adjust the state vector. With $y_k = s_k$, the observation equation reduces to

$$y_k = [1 \quad 0]\mathbf{x}_k + v_k^s. \tag{18}$$

Scenario II. We use the speed observations to adjust the state vector. With $y_k = u_k$, the observation equation reduces to

$$y_k = [0 \quad 1]\mathbf{x}_k + v_k^u. \tag{19}$$

Scenario III. We use both speed and location observations to adjust the state vector. With $\mathbf{y}_k = [s_k^{obs}, u_k^{obs}]^T$ and $\mathbf{v}_k = [v_k^s, v_k^u]^T$ depicting the vector of observations and a vector of associated noise terms respectively, we can write

$$\mathbf{y}_k = \mathbf{x}_k + \mathbf{v}_k. \tag{20}$$

Scenario IV. This is the same as Scenario III, now with an adaptive filtering. In the ideal world (where all our assumptions about the noise terms are correct), the observation error term $\mathbf{v}_k$ is identically distributed as the innovations time series $\mathbf{e}_k = \mathbf{y}_k - \hat{\mathbf{y}}_k^-$ computed and used in Equation (13). Before computing the Kalman gain in Equation (12), we now add the following update:

$$R_k = \gamma R_{k-1} + (1-\gamma)\mathbf{e}_k\mathbf{e}_k^T, \tag{21}$$

in which $\gamma \in [0,1]$ is a smoothing factor. The new estimate R_k from Equation (20) is then used to compute the Kalman gain.

Table 1. Assumptions and initial values.

Variable	Value	Remark
$\hat{\mathbf{x}}_0$	$[0 \quad 15]^T$	Start location and speed
P_0	$\begin{bmatrix} 10^2 & 0 \\ 0 & 10^2 \end{bmatrix}$	Initial error covariance large
Q_k	$\begin{bmatrix} 0 & 0 \\ 0 & 2^2 \end{bmatrix}, \forall k$	No error on the law of motion, an assumed variance 4 (km^2/h^2) for the speed dynamics
R_k^A	$\begin{bmatrix} 3^2 & 0 \\ 0 & 1^2 \end{bmatrix}, \forall k$	With this observation error covariance matrix, we (correctly!) assume larger errors in the location observations than in the speed observations
R_k^B	$\begin{bmatrix} 1^2 & 0 \\ 0 & 3^2 \end{bmatrix}, \forall k$	With this observation error covariance matrix, we (incorrectly) assume larger errors in the speed observations than in the location observations

All other Kalman filter-related assumptions are summarized in Table 1. Table 2 summarizes all results for the eight scenarios and also lists the data collected after day 1, including the ground-truth data. Figures 2–5 show the estimation results of the four scenarios in which the correct observation covariance (R_k^A) is used. We can learn a few important lessons from this first example.

1.3. Discussion—Why All Assumptions Matter

The first key message from this exercise is that combining observations and knowledge really pays off, under the condition that we make sensible prior assumptions (more on this below). Table 2 shows that in all A scenarios, the filtered location estimates are more accurate than the raw location observations. In other words, by combining the observations with some very basic knowledge (the law of motion in Equation (15)) and some proper assumptions about the uncertainty in both model and observations, the Kalman filter reduces the error up to a factor of 3(!) in Scenario IV-A. In this example, the best results (Scenario IV-A) are obtained when we use all available data. Note that the improvement relates mostly to the location estimates; the filter does not improve the speed estimate much when compared with the speed observations. This is because (a) the speed observations are already of good quality (much better than the location observations; root mean square error (RMSE), 0.9 km/h) and (b) we do not add any real "knowledge" about the dynamics of speed in Equation (16). The assumption is simply that speed follows a random walk, which in this case it does.

The second key message is that the (prior) assumptions we make on the observation (and process) covariance matrices matter much. Also in applying the Kalman filter, the garbage in equals garbage out principle holds. All scenarios that use the (more or less) *correct* observation noise covariance matrix R_k^A score consistently better than the very same scenarios that use R_k^B (all other inputs and parameters kept equal). Note that we refer here to the covariance *structure* (so, the relative sizes of its components) rather than the (order of) magnitude of the matrix; R_k^A and R_k^B have the same order of magnitude but differ largely in structure (the first assumes locations are more unreliable, the latter speeds). Put simply, if we put our emphasis on the wrong observations (in R_k^B on the locations instead of the speeds), the Kalman filter will not give good results and may even diverge.

Third, besides its structure, the order of magnitude of the prior covariance matrices is important. Recall from Equation (7) that the ratio between the process and observation noise determines the size of the corrections and thus the responsiveness of the filter to output errors. The smaller the observation noise is relative to the process noise, the more responsive the filter becomes. Figure 5 shows that if we adapt the observation noise to match the actual output errors with Equation (20), and our prior noise matrix is correct as in Scenario IV-A, the result is that the filter estimates follow the observations closely, particularly the more accurate speed observations as shown in Figure 5(b). An adaptive filter thus is highly responsive, which in some cases is desirable, but sometimes leads to large errors. For example, in Scenario IV-B (where the wrong observation noise matrix R_k^B is used), the adaptive filter makes things worse. The resulting filter estimates now follow the (inaccurate) location measurements very closely. This explains why Scenario IV-B performs worst of all, even worse than without filtering (to illustrate, compare s^{obs} with $\hat{s}_k$ under Scenario IV-B in Table 2).

A final observation in this experiment is that in all A scenarios the filter converges rather quickly (in a few time steps) as shown in Figures 2–4. Filter convergence essentially means that the error covariance P_k converges to constant value. As a result, also the confidence bounds (Equation (8)) converge to a constant value around the (time-varying) mean. Only in Scenario IV-A (Figure 5(b), the adaptive filter) do we see how the confidence bounds constructed with P_k become smaller as the filter continuously "learns" to better predict the observations.

TABLE 2. Estimation results of the bicycle example for all scenarios.

	Ground truth		Observations		Scenario I-A		Scenario II-A		Scenario III-A		Scenario IV-A		Scenario I-B		Scenario II-B		Scenario III-B		Scenario IV-B	
Time (h)	s^{GPS}	u^{GPS}	s^{obs}	u^{obs}	$\hat{s}_k$	$\hat{u}_k$	$\hat{s}_k$	$\hat{u}_k$	$\hat{s}_k$	$\hat{u}_k$	$\hat{s}_k$	$\hat{u}_k$	$\hat{s}_k$	$\hat{u}_k$	$\hat{s}_k$	$\hat{u}_k$	$\hat{s}_k$	$\hat{u}_k$	$\hat{s}_k$	$\hat{u}_k$
0	0.0	15.0	0.0	15.0	0.0	15.0	0.0	15.0	0.0	15.0	0.0	15.0	0.0	15.0	0.0	15.0	0.0	15.0	0.0	15.0
1	14.8	14.8	17.2	15.8	17.1	16.1	15.8	15.8	17.1	15.8	17.1	15.8	17.2	16.1	15.8	15.8	17.2	15.9	17.2	15.9
2	28.5	13.7	32.1	13.4	32.3	15.3	29.2	13.8	31.5	13.9	31.7	13.6	32.2	15.0	29.2	14.4	32.1	14.8	32.0	14.6
3	41.0	12.5	42.5	11.7	43.5	12.8	40.9	12.1	43.1	12.0	42.8	11.8	42.9	11.7	41.0	13.0	43.0	11.7	42.6	11.1
4	57.0	16.0	61.4	14.9	60.1	14.8	55.8	14.4	58.9	14.6	60.0	14.9	60.6	16.4	55.8	13.9	60.5	16.2	61.1	17.7
5	73.0	16.0	67.7	15.2	69.8	12.2	71.0	15.1	72.1	14.6	73.5	15.1	68.8	9.9	71.0	14.5	69.1	10.5	68.0	7.3
6	86.3	13.3	71.6	14.7	74.7	8.3	85.7	14.8	82.4	13.6	87.9	14.7	72.4	5.0	85.8	14.6	73.3	5.9	71.7	3.9
7	99.3	13.0	100.9	12.2	95.5	14.9	98.0	12.7	96.5	12.9	100.2	12.3	98.1	21.3	98.0	13.5	97.6	20.1	100.6	27.7
8	117.0	17.7	123.9	16.0	119.7	19.9	114.0	15.4	115.6	16.3	116.2	16.0	123.3	24.4	114.0	14.7	122.5	23.5	123.9	23.4
9	129.0	12.0	128.5	11.8	131.9	15.8	125.7	12.4	127.9	12.6	127.9	11.8	130.8	11.1	125.8	13.3	130.7	11.4	128.6	5.3
10	142.6	13.6	144.2	13.9	145.3	14.5	139.6	13.7	142.4	13.9	141.9	13.9	143.9	12.7	139.7	13.6	144.0	13.0	144.1	15.4
RMSE			5.5	0.9	4.4	2.5	1.9	1.1	2.1	0.8	1.7	0.9	5.2	4.5	1.8	1.4	4.8	4.0	5.5	6.5
CIC					90.9	63.6	100.0	81.8	90.9	81.8	72.7	9.1	36.4	54.5	100.0	90.9	36.4	45.5	9.1	9.1

Notes. All locations and speeds are in kilometers and kilometers per hour, respectively. "A" scenarios are those in which observation noise covariance matrix R_k^A is used; in "B" scenarios, R_k^B is used (see Table 1). Confidence interval coverage (CIC) (%) depicts the number of ground-truth data points falling in a 95% interval around the estimated values calculated with the estimated covariance matrix P (see Equation (8)).

FIGURE 2. Estimation results of Scenario I-A.

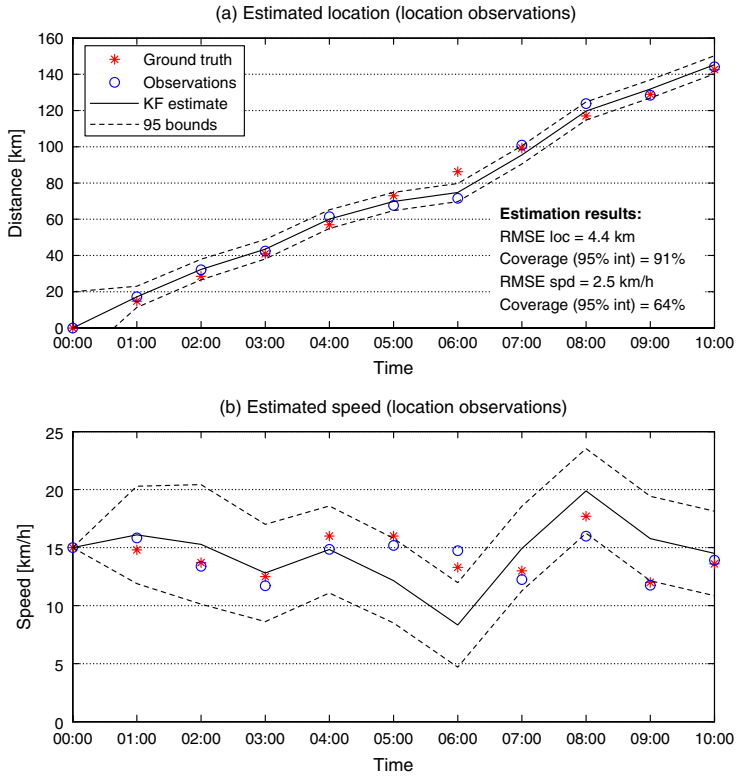


FIGURE 3. Estimation results of Scenario II-A.

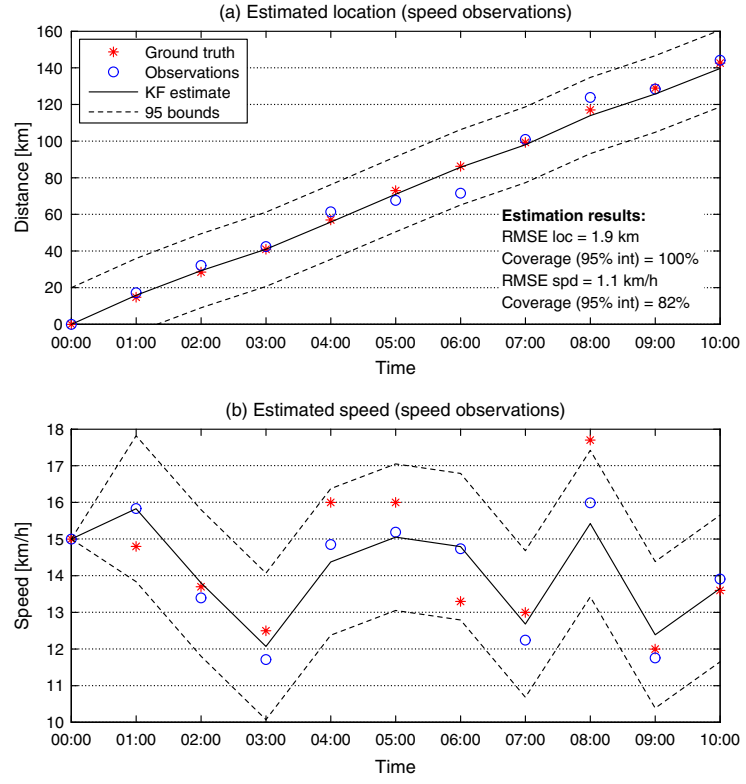


FIGURE 4. Estimation results of Scenario III-A.

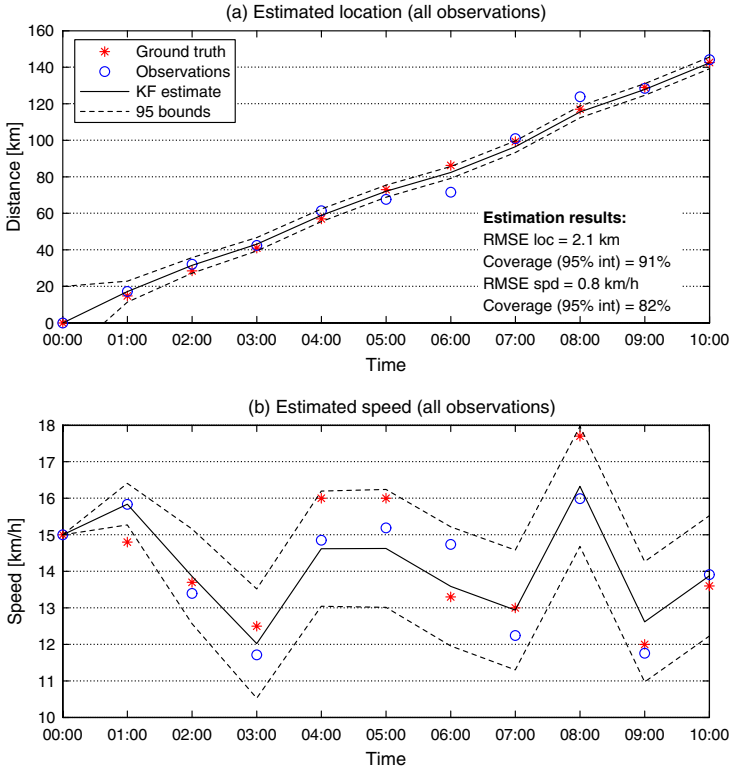
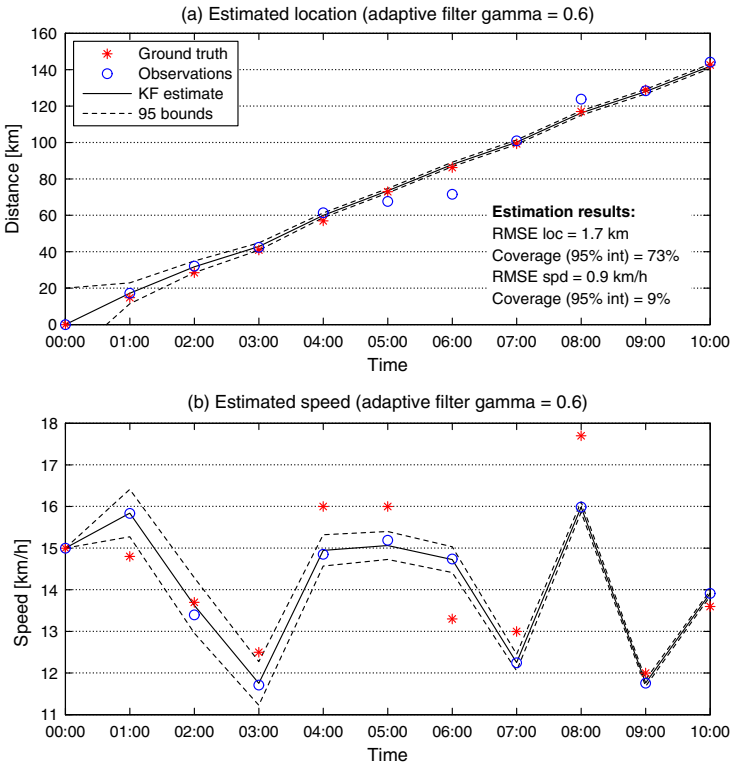


FIGURE 5. Estimation results of Scenario IV-B.



1.4. Summary and Conclusions

The Kalman filter is a highly efficient, recursive minimum variance estimator that works for all linear systems cast in state space form. However, in applying the Kalman filter many assumptions need to be made, the following most importantly.

Assumption 1. *The process and observation models can be formulated as a linear discrete state space system of equations.*

Assumption 2. *All noise terms in this system are independent, zero-mean Gaussian noise processes of known covariance matrices.*

By implication, the process and observation models are *unbiased* predictors of the state/outputs.

The validity of those assumptions largely determine whether or not the KF application is successful. In real life there are few processes for which these assumptions hold and even fewer models that are linear and that—under all circumstances—provide unbiased predictions. Nonetheless, in the remainder of this tutorial we give two other examples that illustrate how the KF may nonetheless be a very powerful tool in the field of traffic management and control. The first relates to estimating origin–destination (OD) matrices, the second to real-time traffic state estimation.

2. Complex "Linear" Problems: OD Estimation

2.1. The OD Matrix Estimation and Prediction Problem

One of the key traffic variables required for both ex post and ex ante evaluation of traffic management and policy measures is time-dependent OD demand matrices, with elements $x_{ij,k}$ depicting the number of trips going from origin i to destination j at discrete departure time interval k. For the sake of consistency, we will represent OD matrices as a single column vector $\mathbf{x}_k$ with elements x_{rk}, depicting the traffic demand from origin–destination pair $r \in \{1, \ldots, N_r\}$ departing at k.

Because OD matrices are rarely observed in practice (for that we essentially need to observe all travelers all the time), OD matrices are generally inferred from aggregate or partial information on OD flows (historical (prior) OD matrices, traffic counts, speeds, automated vehicle identification (AVI) data, floating car data, etc.). In this process, knowledge in the form of traffic (assignment) models is used, which necessarily make many assumptions related to the underlying traffic flow operations and travel behavior. Because OD estimation is such a broad research area, in this section we only scratch the surface of this problem and focus on application of the Kalman filter.

In very general terms, dynamic OD estimation approaches are designed to solve the following problem (Cascetta et al. [8]):

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} f_1(\mathbf{x}, \tilde{\mathbf{x}}) + f_2(\mathbf{y}(\mathbf{x}), \tilde{\mathbf{y}}) \tag{22}$$

subject to

$$\mathbf{x} \geq \mathbf{0}, \qquad \mathbf{y} \geq \mathbf{0}$$
$$\ldots$$

In (22), $\mathbf{x}$ denotes a vector of all time-dependent OD vectors $[\ldots, \mathbf{x}_k, \ldots]^T$ under consideration, $\tilde{\mathbf{y}}$ denotes a vector of all available observations, and f_1 and f_2 are error functions (e.g., sum of squared errors or some entropy measure). The latter term (f_2) measures the distance between predictions $\mathbf{y}(\mathbf{x})$ made with the estimated OD matrix and the actual observations $\tilde{\mathbf{y}}$. In case the term f_2 specifies the sum of squared errors, i.e., f_2: $E[(\mathbf{y}(\mathbf{x}) - \tilde{\mathbf{y}})^2]$, clearly the

Kalman filter can be used to solve it. The first term (f_1) measures the distance between the estimated OD matrix $\mathbf{x}$ and a *prior* OD matrix $\tilde{\mathbf{x}}$, which could be derived from historical data, travel surveys, or previous estimations. This term forces the estimation to converge to a solution even if there is too little information available through the measurements. Let us explain in a very simple example, outlined in Figure 6, why such prior information is needed.

In the case of Figure 6, the link flows can be explained by an, in principle, infinite number of OD matrices, which implies that the OD estimation problem is *underdetermined*. There are more degrees of freedom (unknown OD flows x_r) than independent equations (relating x_r to observations y_l) to solve these unknowns. If we assume all OD flows and observations in this small network are constant (no dynamics), the observation equation linking OD flows to link counts reads

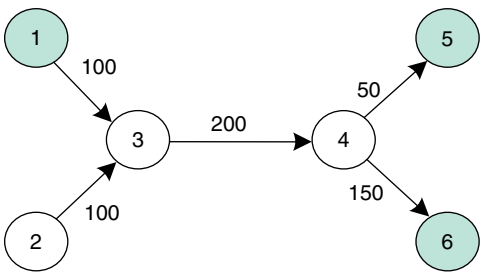
$$\mathbf{y} = A\mathbf{x}, \tag{23}$$

with A the assignment matrix with elements $a_{l,r}$ depicting the fraction of OD flow x_r contributing to the flow on link l. In the case of Figure 6, Gaussian elimination of this system leaves us with just three independent equations and four unknowns:

$$\begin{aligned}
&L1\text{: } 100 = x_{15} + x_{16} \\
&L2\text{: } 100 = \qquad\quad x_{25} + x_{26} \\
&L3\text{: } 200 = x_{15} + x_{16} + x_{25} + x_{26} \\
&L4\text{: } \quad 50 = x_{15} \qquad + x_{25} \\
&L5\text{: } 150 = \qquad x_{16} \qquad\quad + x_{26}
\end{aligned}
\xrightarrow[L4'=L4-L1]{L3'=L3-L1=L2}
\begin{aligned}
&L1\text{: } \quad 100 = x_{15} + x_{16} \\
&L2\text{: } \quad 100 = \qquad\quad x_{25} + x_{26} \\
&L4'\text{: } -50 = \qquad -x_{16} + x_{25} \\
&L5\text{: } \quad 150 = \qquad x_{16} \qquad\quad + x_{26}
\end{aligned}
\xrightarrow[L4''=-L4']{L5'=L5+L3'=L2}
\begin{aligned}
&L1\text{: } \quad 100 = x_{15} + x_{16} \\
&L2\text{: } \quad 100 = \qquad\quad x_{25} + x_{26} \\
&L4''\text{: } \quad 50 = \qquad x_{16} - x_{25}
\end{aligned}$$

A unique solution is available only in case the system is full rank, i.e., when all OD flows are observable. We will return to this "observability" problem further below. In real life there are dynamics involved in both the demand (OD flows change over time) and the traffic supply (it takes time for vehicles to travel from A to B, and even more so when the network

FIGURE 6. The key problem with OD matrix estimation: Underdeterminedness.



OD matrix A

	5	6
1	50	50
2	0	100

OD matrix B

	5	6
1	25	75
2	25	75

Note. Both matrices A and B explain the observed link flows in this simple network.

is congested). Okutani and Stephanedes [36] were among the first to formulate the time-dependent OD estimation problem in state space form such that it can be solved by a Kalman filter:

$$x_{r,\,k+1} = \sum_{p=k-q'}^{k} f_{rp} x_{rp} + w_{rk}, \tag{24}$$

$$y_{lc} = \sum_{p=1}^{k} \sum_{r=1}^{N_r} a_{lk}^{rp} x_{rp} + v_{lk}. \tag{25}$$

The process model (24) represents an autoregressive process of order q', with coefficients f_{rp}. In observation Equation (25), the assignment matrix elements a_{lk}^{rp} now depict the fraction of OD flow x_{rp} (that departed at period p) passing link l during time period k.

Although this formulation encapsulates the dynamics, there are a still a number of serious problems with this linear formulation. The first problem relates to the observation Equation (25). Particularly in congested networks, the assumption of linearity is highly debatable, because of traffic flow dynamics and queuing, and the effect of these dynamics on route choice. Moreover, in congested networks the link flows in some cases are not related to a change in traffic demand at all. This is easily understood by the following example. Consider a bottle without a bottom that is held upside down under a tap. As long as the water flow x from the tap is smaller than the capacity C of the bottleneck, the flow y in the bottleneck is determined by x. But when $x > C$, y is determined by C and not sensitive to changes in x until the bottle has emptied again. The analogy is clear: The link volumes in and downstream of congestion are determined by the capacity of bottlenecks, and not by traffic demand. The consequence is that other traffic variables (speeds, travel times) and associated observation equations are needed to solve the problem for congested networks (e.g., Frederix et al. [16]). In this tutorial we will not discuss these approaches.

The second complication is that the time-dependent OD matrix *itself* is a function of the assignment matrix. The underlying mechanism relates to departure time and route choice behavior. By and large, travelers choose departure times k (which determine x_{rk}) and routes (which determine the link fractions $a_{l,\,k}^{rp}$) such that their expected travel costs (e.g., delays) are minimized. As a consequence, $\mathbf{x}_k = \mathbf{x}(A_k)$, and $A_k = A(\mathbf{x}_k, \ldots, \mathbf{x}_{k-p'})$, with $k - p'$ the earliest time period in which vehicles departed that are still present in the network during k. A full treatment of this fixed point problem (consistency between $\mathbf{x}_k$ and A_k) is also beyond the scope of this tutorial.

The third key problem relates to the typical Kalman filter assumptions on the noise terms $w_{r,\,k}$ and $v_{l,\,k}$ in (24) and (25), respectively. Arrival processes (vehicles entering the network) are typical Poisson processes, with exponentially distributed headways. Moreover, both demand and link flows are by definition positive. Consequently, Gaussian assumptions on these noise terms make no sense, because this would allow negative OD and link flows, particularly for smaller average values.

Despite these prior considerations and cautions, many authors have proposed the Kalman filter to solve the OD estimation problem (Ashok and Ben-Akiva [2, 4], Okutani and Stephanedes [36], Chang and Wu [9], Van Der Zijpp [47]) with varying degrees of success. In the remainder of this section we will illustrate how a slightly adapted version of the Kalman filter can be used to solve the OD estimation problem. The key to this approach, which was first coined by Ashok and Ben-Akiva [2], is a redefinition of the state and observation variables to better satisfy the assumptions in the Kalman filter.

2.2. A Kalman Filter Solution Based on Deviations

In the approach of Ashok and Ben-Akiva [3], the state vector is defined as the difference between the OD flow x_{rk} and some historical ("prior") OD flow $\tilde{x}_{rk}$. In vector nation,

$$\partial \mathbf{x}_k \equiv \mathbf{x}_k - \tilde{\mathbf{x}}_k.$$

Instead of predicting the actual amounts of trips between origins and destinations as in (24), we now predict how much these trips deviate from some historical OD pattern. There are two reasons for doing this. First, this adaptation leads to a more valid process model. Suppose that a series of historical OD matrices $\tilde{\mathbf{x}}_k$ have been estimated from historical data for several previous days or months. These prior OD matrices subsume a wealth of information about the structural variations over both space and time (think of daily peak periods and the differences between tuesdays and saturdays). Instead of estimating these with a simple autoregressive model (25), we filter them out using historical data. Second, by defining the state variables as deviations from some historical time series of OD matrices, the state estimation errors are now more amenable to approximation by a normal distribution (property of the Kalman filter). In vector notation, the adjusted process equation now reads

$$\partial \mathbf{x}_{k+1} = \sum_{p=k-q'}^{k} \mathbf{f}_k^p \partial \mathbf{x}_p + \mathbf{w}_k, \tag{26}$$

where $\mathbf{f}_k^p$ is a $N_r \times N_r$ matrix with elements f_{rk}^{rp} that model the relationship between $\partial \mathbf{x}_p$ and $\partial \mathbf{x}_{k+1}$, and q' is the degree of the autoregressive process. The process noise term $\mathbf{w}_k$ depicts some known Gaussian noise term with mean $E[\mathbf{w}_k] = \mathbf{0}$ and variance $E[\mathbf{w}_k \mathbf{w}_k^T] = Q_k$, where Q_k is a $N_r \times N_r$ variance covariance matrix.

A similar transformation is applied to the observation equation that models the relationship between the state variables ∂x_k (OD deviations) and the observations $\mathbf{y}_k$ (traffic counts data):

$$\partial \mathbf{y}_k = \sum_{p=k-p'}^{k} \mathbf{a}_k^p \partial \mathbf{x}_p + \mathbf{v}_k, \tag{27}$$

with

$$\partial \mathbf{y}_k \equiv \mathbf{y}_k - \tilde{\mathbf{y}}_k.$$

In (27), $\tilde{\mathbf{y}}_k = \sum_{p=k-p'}^{k} \mathbf{a}_k^p \tilde{\mathbf{x}}_p$ represents the observations predicted with the prior OD matrix using assignment matrix $\mathbf{a}_k^p$, which is a $N_l \times N_r$ (number of links × number of OD pairs) matrix, mapping OD flows departing during intervals p to link traffic counts observed during interval k. Finally, p' is the maximum number of time intervals needed to travel between any OD pair.

Equations (26) and (27) constitute a linear state space system of equations that can be solved using the Kalman filter equations in Algorithm 1.

2.3. OD Estimation Example

We will illustrate some of the properties and limitations of this OD estimation procedure with a simple synthetic example also used by Bierlaire and Crittin [7]. The example uses the same network depicted in Figure 6, now with three OD pairs, $\{(1,5),(1,6),(2,6)\}$ (i.e., we fix OD pair $(2,5)$ to zero), and a time horizon of $N_k = 15$ time intervals of T minutes each. For the sake of simplicity, assume for each link a travel time of T minutes and an infinite capacity, which implies that $p' = 3$ (maximum number of time periods required to travel through the network). We also assume that the degree of the autoregressive process in (26) equals $q' = 2$.

Furthermore, the autoregressive process is such that $f_k^p = I$, for $p = k-3, k-2, k-1$. The variance–covariance matrices Q_k that govern the noise process $\mathbf{w}_k$ in (26) are diagonal, with the OD flows of the last time interval on the diagonal. The variance–covariance matrices R_k describing the observation noise process $\mathbf{v}_k$ in (27) are also diagonal with variance arbitrarily set to 1.

The true OD flows for OD pairs $(1,5)$, $(1,6)$, and $(2,6)$ are given in Table 3, where the unit is a number of cars per time interval (T minutes). Note that time intervals -4 to 0 are

TABLE 3. True OD demand for the example network in number of vehicles per T minutes.

Time (k)	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
OD flow 1, 5	36	12	30	12	36	30	48	12	18	36	12	12	60	42	66	42	24	42	12	18
OD flow 1, 6	36	12	30	12	36	30	48	12	18	36	12	12	60	42	66	42	24	42	12	18
OD flow 2, 6	72	24	60	24	72	60	96	24	36	72	24	24	120	84	132	84	48	84	24	36

used to warm up the simulation and to avoid starting with an empty network. The "true" assignment matrices (which incorporate how travelers choose routes and departure times) are fixed and given by

$$\mathbf{a}_k^k = \begin{pmatrix} 1/3 & 1/3 & 0 \\ 0 & 0 & 1/3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{a}_k^{k-1} = \begin{pmatrix} 1/3 & 1/3 & 0 \\ 0 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{a}_k^{k-2} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 0 & 0 \\ 0 & 1/3 & 1/3 \end{pmatrix}, \qquad \mathbf{a}_k^{k-3} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1/3 & 0 & 0 \\ 0 & 1/3 & 1/3 \end{pmatrix}.$$

In the experiment we examine the effect of the quality and availability of input data (the prior OD matrix $\tilde{\mathbf{x}}_k$ and the available traffic counts $\mathbf{y}_k$). The link counts and prior OD matrices are obtained by perturbing the true counts/OD flows listed above according to the following scenarios:

OD Scenarios 1–2. The prior OD matrices are obtained by a random perturbation of the true OD (Scenario 1), or the prior OD matrices are biased, i.e., they structurally underestimate the true OD matrix (Scenario 2).

Measurement Scenarios 1–3. The link flows resulting from the assignment of the true OD matrices have been perturbed by adding noise (sampled from independent distributions with zero mean and a coefficient of variation of 20%, 40%, or 60%, respectively, to obtain the link traffic counts data). In these three scenarios, the traffic counts are available on all links (assuming ideal detector locations on network).

Data availability scenarios. Here, we assume traffic counts are available on all links (Scenario 1); counts are available on just link 3 (Scenario 2), or on links 1 and 4 (Scenario 3).

To evaluate the quality of the final OD estimates by the Kalman filter model given in previous subsection, we use the following indicators: The normalized RMSE (NRMSE) of the estimated OD matrix $\hat{\mathbf{x}}_k$ with respect to the true OD matrix $\mathbf{x}_k$ quantifies the overall error of the estimation and penalizes large errors at a higher rate than small errors:

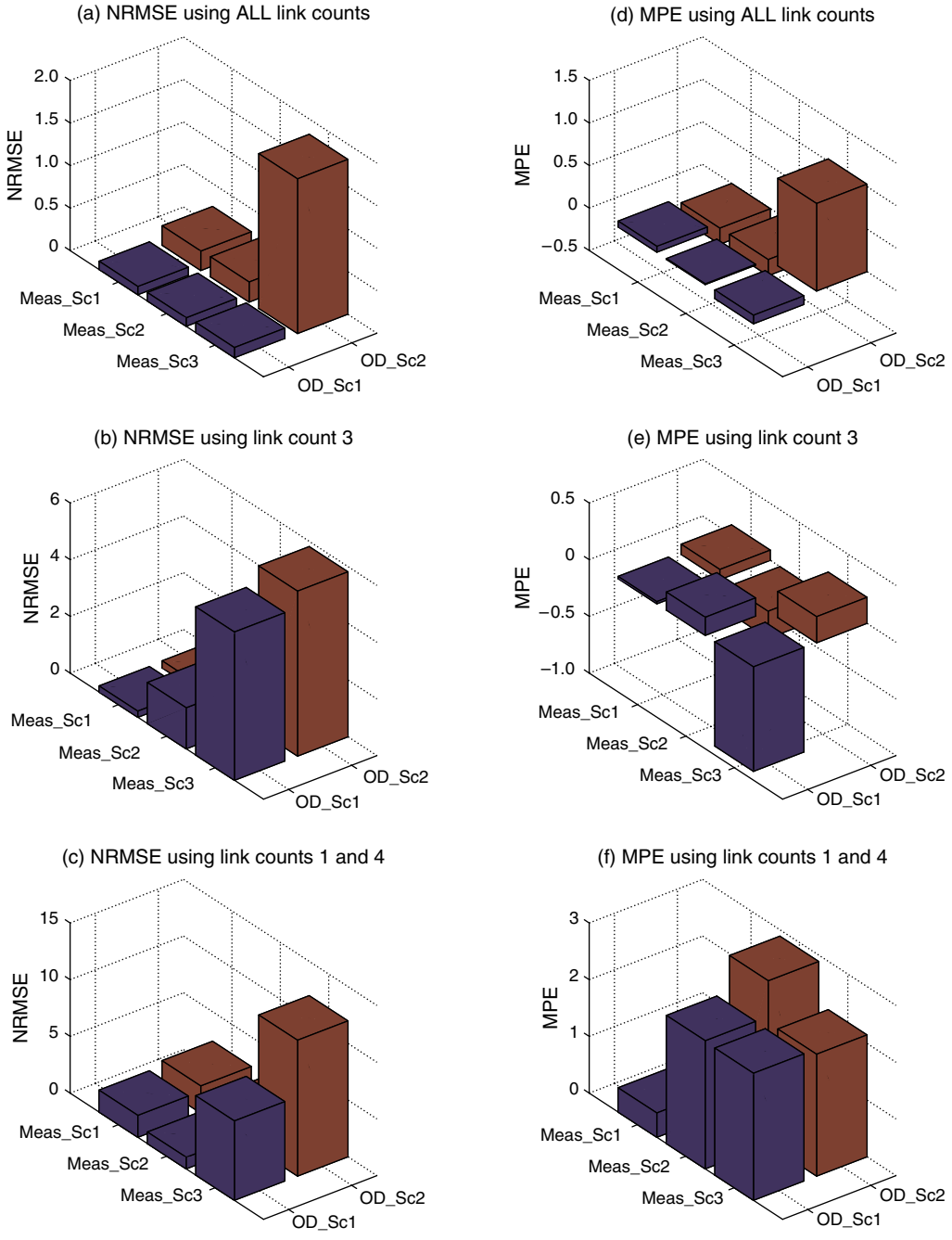
$$\text{NRMSE} = \frac{\sqrt{N_r N_k \sum_{r=1}^{N_r} \sum_{k=1}^{N_k} (\hat{x}_{rk} - x_{rk})^2}}{\sum_{r=1}^{N_r} \sum_{k=1}^{N_k} x_{rk}}. \tag{28}$$

The mean percentage error (MPE) indicates the existence of systematic under- or overestimation in the estimated OD flows:

$$\text{MPE} = \frac{1}{N_r N_k} \sum_{r=1}^{N_r} \sum_{k=1}^{N_k} \frac{\hat{x}_{rk} - x_{rk}}{x_{rk}}. \tag{29}$$

Figure 7 shows the estimation results for all scenarios considered.

FIGURE 7. Estimation results for all scenarios.



2.4. Results and Discussion

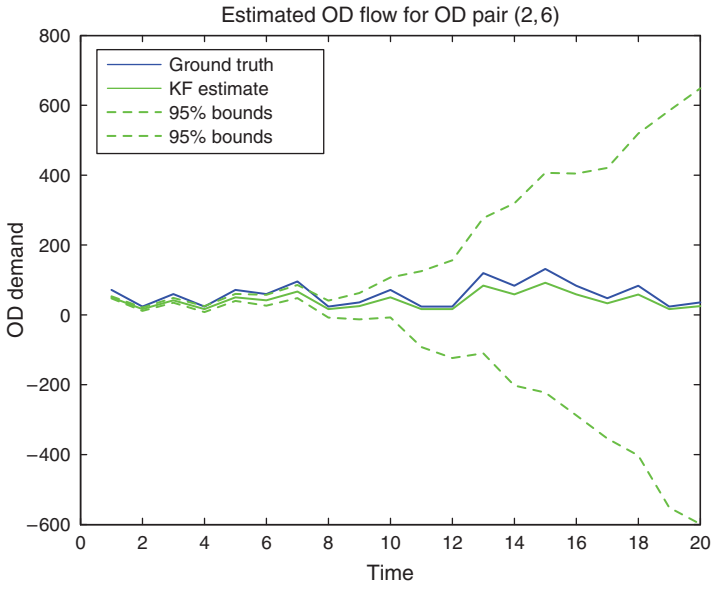
We can make the following observations from this exercise. First, Figures 7(a) and 7(d) show that the quality of the traffic counts clearly influences the reliability of OD matrix estimates, in terms of overall error (NRMSE) and bias (MPE). Put simply, more noise in the counts yields larger errors. Second, Figures 7(a) and 7(d) also indicate that structural errors in the prior OD matrix (in this case an underestimation) result in larger OD estimation errors than random errors in the prior OD. This relates to the fact that structural deviations in OD

patterns are the result of different underlying route and departure choice patterns, that is, different from the assignment matrix assumed in our case. In other words, a structural deviation in the OD matrix renders the assumption of an unbiased process Equation (26) invalid.

Finally, when comparing Figures 7(a) with 7(b) and 7(c) (or 7(d) with 7(e) and 7(f)), one can see that the three data availability scenarios yield increasingly larger errors. These errors relate directly to the observability problem introduced in Figure 6. This is obvious in the third scenario where counts are available on links 1 and 4. In this case, clearly OD pair $(2, 6)$ never gets measured, i.e., is unobservable. In case of Scenario 2 in which just counts on link 3 are available, *all* OD flows do get measured (they all pass link 3), but this does not imply that they are observable. The reader may verify this by determining the rank of the resulting system of equations along similar lines as above. How much the system is underdetermined may be easily spotted in such synthetic examples, but is much harder to spot in large real-time OD estimation problems. One tell-tale sign of under determinedness is illustrated in Figure 8. Here, the estimation of OD relation $(2, 6)$ over time is plotted with confidence intervals constructed with the error covariance matrix P_k. Now halfway through the simulation the filter starts diverging (P_k grows unboundedly).

In sum, clever use of prior information—resulting in a formulation of process and observation models using deviations—makes it possible to use the classic Kalman filter. However, although this adapted formulation validates the strong distributional assumptions made in the KF, there are additional problems that render this approach problematic for real-life large-scale OD estimation. Most importantly, in congested networks the relationship between traffic demand and the actual traffic flows (resulting from "assigning" traffic to the network) is either highly nonlinear, because of the response of travelers, or even nonexistent when in congestion flow counts are determined by capacity and not by traffic demand. In the next section we will discuss a third example of Kalman filtering that addresses the estimation of traffic variables in congested networks.

FIGURE 8. Effect of an unobservable OD matrix: The Kalman filter diverges.



3. Nonlinear Problems: Traffic State Estimation

3.1. Traffic State Estimation and Prediction

As an illustration, consider Figure 9(a), which shows a space–time contour plot of real traffic conditions (speeds) on a German freeway. Clearly visible are so-called wide moving jams that propagate against the direction of travel. Once congestion occurs, queues may spill back and block additional routes upstream. As a result, travel times increase rapidly over different routes, travelers may change their route choices, and traffic conditions may rapidly deteriorate in entire traffic networks.

Dynamic traffic management (DTM) encompasses all traffic control and information applications tailored to adjust the traffic supply (e.g., operating speed, capacity) to the prevailing network demand in real time. Examples include (adaptive) intersection control (traffic signals), freeway ramp metering, managed lanes, dynamic rerouting, and traffic information provision, to name but a few. One of the key ingredients for DTM applications is traffic state estimation and prediction. For this tutorial, we focus on one increasingly popular approach, that is, model-based traffic control as outlined in Figure 19. We will return to this scheme in the last section of this tutorial. In model-based control approaches (e.g., the RENAISSANCE system of Wang et al. [57]; the TU Delft/fileradar.nl approach of van Hinsbergen et al. [48]), macroscopic traffic flow models are used in conjunction with an extended Kalman filter (EKF) to estimate the state in a traffic network.

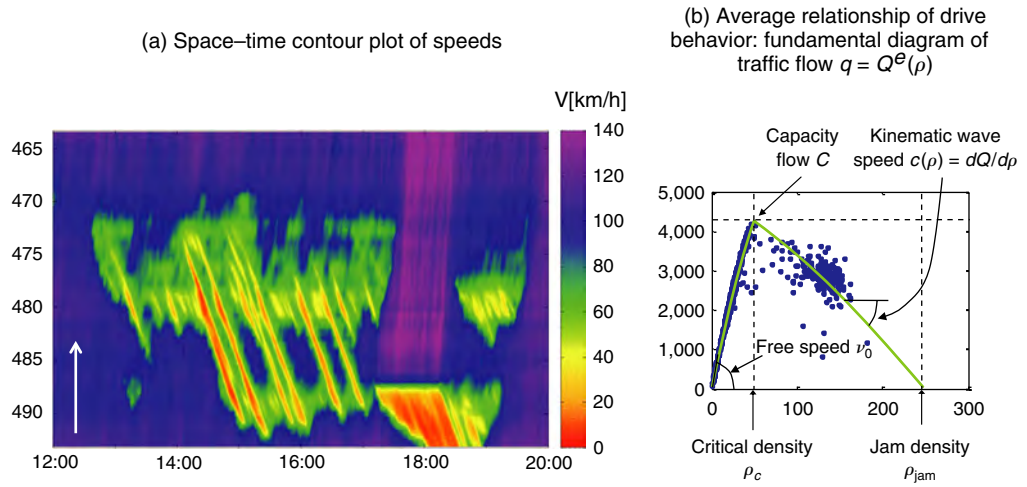
Macroscopic traffic flow models describe the traffic state in terms of flow q (the average number of vehicles passing a location per unit time), space-mean (or local harmonic mean), speed u (in kilometers per hour), and density ρ (the average number of vehicles per unit space in vehicles per kilometer), which in a homogeneous and stationary traffic stream are related to one another by

$$q = \rho u. \tag{30}$$

The simplest (process) model to predict the propagation of traffic flow over space and time (given sufficient initial and boundary conditions) is the conservation of vehicles equation

$$\frac{\partial \rho}{\partial t} + \frac{\partial q}{\partial x} = 0, \tag{31}$$

FIGURE 9. (a) Contour plot of speeds on a German freeway and (b) the fundamental diagram of traffic flow.



Source. (a) Reused with permission of Martin Treiber and Arne Kesting, http://www.traffic-states.com/. Data are from German Highway A5; Direction N; Section 493–463 km; Friday, April 20, 2001; 12:00–20:00.
Notes. (a) The arrow indicates the driving direction. (b) The dots depict actual *local* measurements.

which loosely states that the change in density over time is equal to the negative of the change in flow over space. Because we have two Equations ((30) and (31)) and three unknowns (ρ, u, and q), a third equation is required to solve this model. So-called first-order models (van Hinsbergen et al. [48]) do this by assuming an equilibrium relationship, the fundamental diagram (FD),[1] which relates the average speed u to the density ρ:

$$u = U^e(\rho). \tag{32}$$

Equation (32) is a monotonically decreasing relationship that was first described and empirically verified in the 1930s (Greenshields [18]). It essentially states that, on average, vehicles decrease their speed with a decreasing gross distance headway $s = 1/\rho$, or, vice versa, that vehicles increase their distance headway with increasing speeds. Because of (30), (32) can also be written as $q = \rho U^e(\rho)$, or as

$$q = Q^e(\rho). \tag{33}$$

The fundamental diagram $Q^e(\rho)$ usually is a convex function of ρ with a slope v_0 (equal to the free or *desired* speed) for $\rho \downarrow 0$ (i.e., for a near empty road) and a maximum flow (capacity flow C) at a critical density ρ_c, which is assumed to discriminate between freely flowing and congested traffic. Once congestion sets in, the flow decreases again, coming to a standstill (flow equals 0) at a maximum density ρ_{jam}, called the jam density. An example of the fundamental diagram is given in Figure 9(b). Substituting (33) into (31) leads to

$$\frac{\partial \rho}{\partial t} + c(\rho)\frac{\partial \rho}{\partial x} = 0, \quad c(\rho) = \frac{\partial}{\partial \rho} Q^e(\rho), \tag{34}$$

in which the derivative of the fundamental diagram $c(\rho)$ depicts the so-called kinematic wave speed, which determines the direction and the speed at which perturbations propagate over time and space. For example, in congestion (Figure 9(b), right branch) the kinematic wave speed is negative, which implies that perturbations move against the direction of travel (observe the wide moving jams in Figure 9(a)).

Figure 9(b) also shows some actual *local* data (measured with induction loops). The scatter in these data are due to a number of reasons. First, the FD is an average relationship. Because of heterogeneity in vehicle capabilities and drive behavior (trucks versus ferraris), there is a large variability around this average. Second, the FD is a steady-state relationship (no acceleration/deceleration), whereas many observations are transient and thus not on the FD. Finally, there are measurement and estimation errors. For example, by definition ρ cannot be measured locally, but has to be estimated with $\rho = q/u$ from Equation (30). The error in that estimation becomes larger as speed approaches zero, not just for numerical reasons, but also because (a) the sample size per unit time becomes smaller (in the limit of a complete standstill u can not be observed at a cross section), and (b) the assumption of homogeneous and stationary traffic conditions underpinning Equation (30) may not be valid for all measurements. We will return to these problems below.

Equations (30), (31), and (33) constitute the well-known kinematic wave model simultaneously proposed by Lighthill and Whitham [29] and Richards [41]. In a DTM/state estimation context, Equation (31) (or one of its more involved modifications) can be used as the process model, whereas the fundamental diagram (Equation (32) or (33)) may serve as an observation model for flows and speeds.

[1] First-order models do not describe acceleration/deceleration toward or from the desired equilibrium speed. In so-called higher-order traffic flow models, an additional differential equation describing the dynamics of speed is added, and there are many other ways in which first-order models have been extended with different degrees of success. For an extensive overview of traffic flow modeling, we refer the reader to Hoogendoorn and Bovy [22].

3.2. The Extended Kalman Filter

Equations (31) and (33) can be analytically solved for small examples. However, solving such models to predict traffic operations on traffic corridors or networks requires numerical approximation. Discretizing space and time into cells i of Δx_i (m) and time periods k of Δt (s), respectively, results in a *nonlinear* state space system of equations:

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \boldsymbol{\psi}_k, \mathbf{d}_k) + \boldsymbol{\eta}_k, \tag{35}$$

$$\mathbf{y}_k = \mathbf{h}(\mathbf{x}_k, \boldsymbol{\psi}_k) + \boldsymbol{\zeta}_k. \tag{36}$$

In (36), $\mathbf{x}_k$ now depicts a vector of cell densities $\boldsymbol{\rho}_k = [\ldots \rho_k^i \ldots]^T$, $\mathbf{y}_k$ is a vector of available observations (e.g., flows, *local* speeds), $\boldsymbol{\psi}_k$ is a vector of all parameters used in both process and observation models (possibly space and time specific), $\mathbf{d}_k$ is a vector of boundary conditions and disturbances (e.g., inflows, capacity constraints), and $\boldsymbol{\eta}_k$ and $\boldsymbol{\zeta}_k$ represent process and observation noise processes, respectively. An important note is that although the analytical conservation of vehicles Equation (31) is exact, the numerical approximation (35) is not. Numerical schemes (e.g., van Wageningen-Kessels et al. [53], Lebacque [26]) involve additional assumptions which introduce errors. Below we use the Godunov scheme (Lebacque [26]) to numerically approximate (36). This so-called minimum demand–supply (or maximum entropy) method uses the fundamental diagram (33) to determine the numerical fluxes $q_k^{i\to i+1}$ per time step Δt between two consecutive cells i and $i+1$ as follows:

$$q_k^{i\to i+1} = \min(D_k^i, S_k^{i+1}), \tag{37}$$

with

$$D_k^i = \begin{cases} Q^e(\rho_k^i) & \rho_k^i \le \rho_c^i, \\ C^i & \text{otherwise;} \end{cases} \qquad S_k^{i+1} = \begin{cases} C^{i+1} & \rho_k^{i+1} \le \rho_c^{i+1}, \\ Q^e(\rho_k^{i+1}) & \text{otherwise.} \end{cases}$$

Clearly, with this solution method, the choice of FD parameters (e.g., capacity C, critical density ρ_c) is important. The nonlinear system of Equations (31) and (33) can be solved for unknown state variables $\mathbf{x}_k$ by the so-called *extended Kalman filter* algorithm that is outlined in Algorithm 2. Before showing and discussing a case study and some results, let us briefly review the additional assumptions of the EKF and their consequences.

First of all, Equations (39) and (41) of the EKF assume that both process and observation models can be linearized locally (around the current state). Clearly, in case traffic is in either a free-flowing or a (heavily) congested state (i.e., in the left or right branch of the FD; see Figure 9(b)) this assumption may not be too far of, as has been confirmed by many authors (Newell [34], van Lint and Hoogendoorn [52], Treiber and Helbing [44]). However, around capacity, this assumption may be flawed and lead to corrections with the wrong sign. A direct consequence of the local approximation is that none of the convergence properties of the linear KF are guaranteed for the EKF. In practice this implies that it may work (surprisingly well)—or that it does not.

Algorithm 2 (The extended Kalman filter)
Preliminaries: For the given problem we have a state space model depicted by Equations (35) and (36), in which $\boldsymbol{\eta}_k$ and $\boldsymbol{\zeta}_k$ are independent, zero-mean, Gaussian noise processes of covariance matrices Q_k and R_k, respectively, with

$$E[\boldsymbol{\eta}_k \boldsymbol{\eta}_n^T] = \begin{cases} Q_k & \text{for } n = k, \\ 0 & \text{for } n \ne k; \end{cases} \qquad E[\boldsymbol{\zeta}_k \boldsymbol{\zeta}_n^T] = \begin{cases} R_k & \text{for } n = k, \\ 0 & \text{for } n \ne k. \end{cases}$$

Initialization:

$$\hat{\mathbf{x}}_0 = E[\mathbf{x}_0] \quad \text{and} \quad P_0 = E[\mathbf{x}_0 - E[\mathbf{x}_0]^T].$$

In case no additional information is available, P_0 is usually initialized as a matrix with a large diagonal entries, reflecting the fact that we are highly uncertain about our initial estimate of $\hat{\mathbf{x}}_0$.

For $k = 1, 2, \ldots$ **do**:
Predict mean and variance of state variables:

$$\hat{\mathbf{x}}_k^- = \mathbf{f}(\mathbf{x}_k, \boldsymbol{\psi}_k, \mathbf{d}_k), \tag{38}$$

$$P_k^- = F_{k\,|\,k-1} P_{k-1} F_{k\,|\,k-1}^T + Q_{k-1}, \tag{39}$$

in which

$$F_{k\,|\,k-1} = \left.\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\hat{\mathbf{x}}_k^-}.$$

Predict output variables:

$$\hat{\mathbf{y}}_k^- = \mathbf{h}(\mathbf{x}_k, \boldsymbol{\psi}_k). \tag{40}$$

Compute the Kalman gain:

$$G_k = \frac{P_k^- H_{k\,|\,k-1}^T}{H_{k\,|\,k-1} P_k^- H_{k\,|\,k-1}^T + R_{k-1}}, \tag{41}$$

in which

$$H_{k\,|\,k-1} = \left.\frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\hat{\mathbf{x}}_k^-}.$$

Update mean and covariance:

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + G_k(\mathbf{y}_k - \hat{\mathbf{y}}_k^-), \tag{42}$$

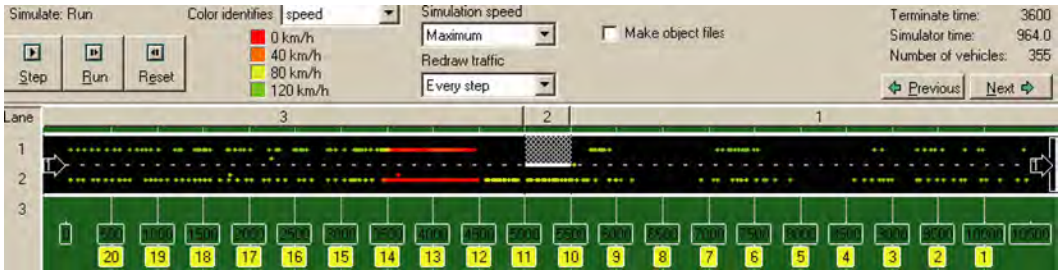
$$P_k = (1 - G_k H_{k\,|\,k-1}) P_k^-, \tag{43}$$

in which the output errors $\mathbf{e}_k = \mathbf{y}_k - \hat{\mathbf{y}}_k^-$ computed and used in (42) are again referred to as the *innovations* time series, because they reflect the amount of new information contained in the observations $\mathbf{y}_k$.
End

3.3. Freeway Traffic State Estimation—An Intuitive Example

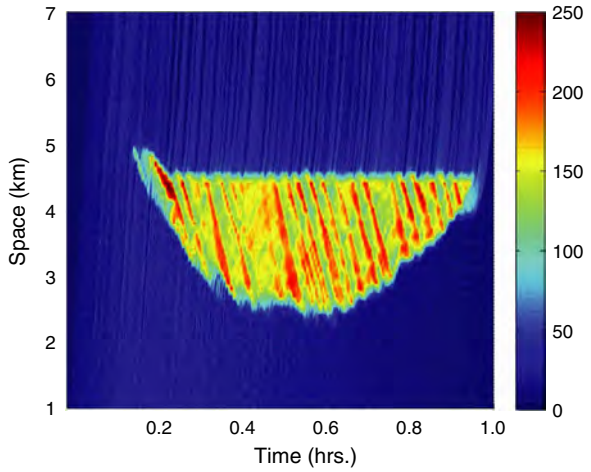
Consider a 10 km freeway stretch with no off- and on-ramps. On this freeway stretch an accident occurs after 12 minutes, which blocks one lane for an indefinite period (see Figure 10). The resulting traffic conditions are shown in Figure 11. For the sake of this example, assume

FIGURE 10. Screenshot of freeway stretch with a temporary bottleneck (e.g., an incident) coded in microsimulation model FOSIM (Vermijs and Schuurman [55]).



Note. The numbered blocks at the very bottom indicate the locations of loop detectors, which measure flow and average speed.

FIGURE 11. Ground-truth densities, in vehicles per kilometer.



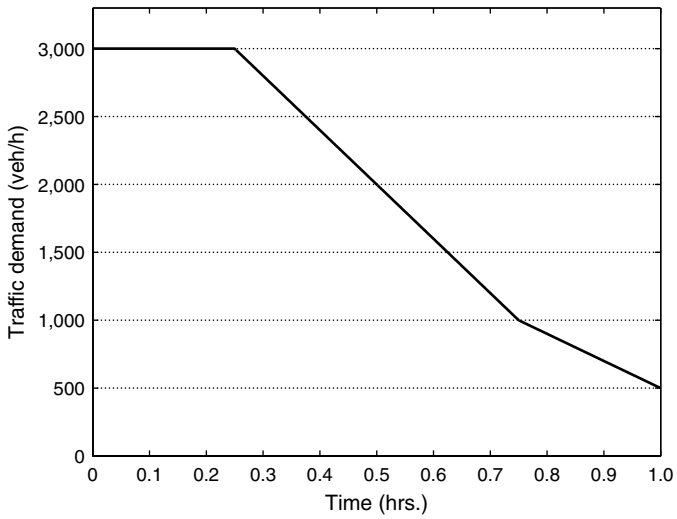
Notes. Congestion sets in because of an incident after around 12 minutes. Clearly visible are wide moving jams emanating from the bottleneck location.

that traffic demand is constant (3,000 vehicles/h) until a few minutes after the incident and then decreases because of, for example, rerouting/traffic information (Figure 12). The task of the EKF-based state estimator is essentially to reproduce the state in Figure 11 using the one-minute average speeds and flows from the detectors represented by the small boxes at the very bottom of Figure 10.

In all scenarios, flow measurements are used to reconstruct the upstream demand ($\mathbf{d}_k$), and both speed and flow measurements are used to correct the state variables using Equations (32) and (33), respectively. The EKF measurement noise covariance matrix R_k is set to match the variance of the actual measurements. We consider $3 \times 3 = 9$ scenarios:

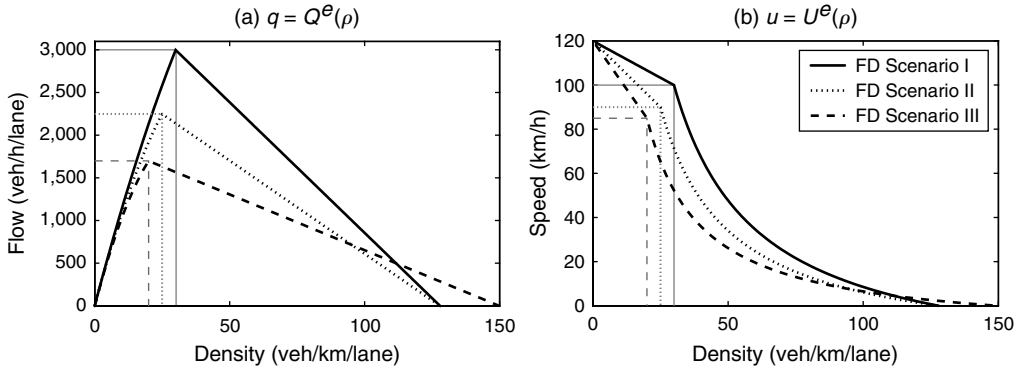
Detector availability. We consider the cases where (A) all detectors are used (spacing 500 meters); (B) one out of every two detectors is used (at location 1000, 2000, 3000, etc.), and (C) one out of six detectors is used (at location 1000, 4000, 7000, etc.; see Figure 10).

FIGURE 12. Traffic demand in the experiment.



Note. As a result of route information provision during the incident, the demand decreases sharply until the queue is fully dissolved.

FIGURE 13. Three fundamental diagram scenarios considered in the experiment.



Note. Scenario I approximately resembles the average behavior in the microsimulation; Scenarios II and III result in increasingly biased estimations.

Parameters of the fundamental diagram. We consider three cases, which are shown in Figure 13. Scenario I approximately resembles the average behavior in the Freeway Operations SIMulation (FOSIM) microsimulation (which in itself is not necessarily realistic); Scenarios II and III both underestimate capacity C, which may result in increasingly biased estimations. The parameters in the fundamental diagram, and particularly critical speed u_c and critical density ρ_c (which determine capacity flow $C = u_c\rho_c$), affect not just the quality of the observation model but also the quality of the process model.

The RMSE (in vehicles per kilometer) is used as a performance measure. Also, three relative error measures are calculated: the MPE, which indicates structural *bias*; the standard deviation of the percentage error (SPE), which is a measure for the variability around the MPE; and the mean absolute percentage error (MAPE), which provides a combined indication of the relative error (bias and variance).

3.4. Discussion—Different Performance Measures Tell Different Stories

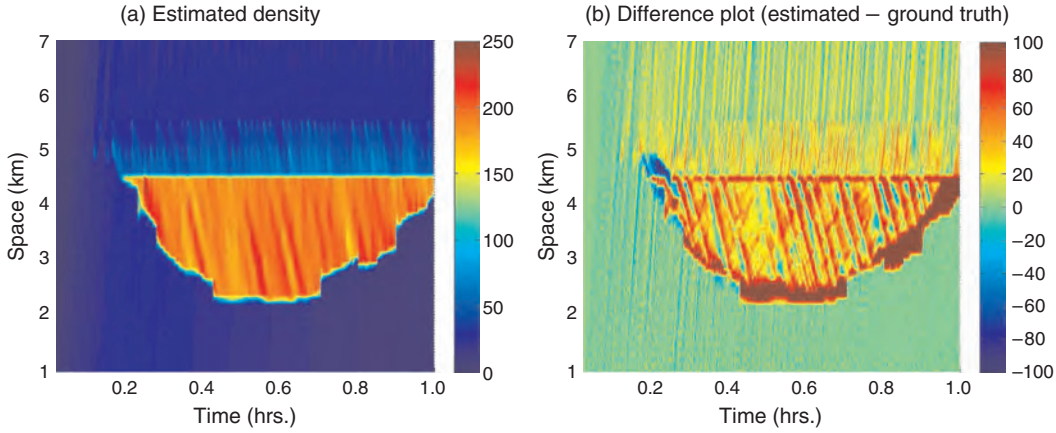
Table 4 shows the numerical results of all nine scenarios. This table should invoke some surprises. A priori one may hypothesize that the best results should occur when all available data are used in conjunction with the best-fit fundamental diagram. This, however, is not the case. There are several reasons for this, which can be illustrated with Figures 14–18, which show contour plots for the estimated densities and the estimation errors for five scenarios. We can make the following observations.

- Qualitatively, all state estimation scenarios reproduce the ground-truth congestion pattern reasonably well. The largest estimation errors occur on the space–time boundaries of

TABLE 4. Results of state estimation example.

Detector spacing	FD scenario	RMSE (veh/km)	MPE (%)	SPE (%)	MAPE (%)
A (500 m)	I	31	18	108	46
	II	26	13	100	45
	III	29	0	103	45
B (1,000 m)	I	27	13	80	39
	II	26	9	96	43
	III	27	−2	89	43
C (3,000 m)	I	40	−2	53	38
	II	23	6	75	38
	III	31	9	135	51

FIGURE 14. Contour plot of estimated densities and estimation errors for Scenario A-I (detectors at 500 m spacing, Fundamental Diagram I).



congestion, where linearization (Equations (39) and (41)) yields the largest approximation errors.

- None of the state estimation scenarios are able to accurately reconstruct the conditions *within* congestion. This is because a first-order traffic flow model is used (no acceleration/deceleration dynamics).
- Figures 14, 15, and 16 show the results for three scenarios in which all data are used with different fundamental diagrams (A-I, A-II, A-III; see also the first three rows in Table 4). Clearly, the choice of fundamental diagram matters in terms of overall bias, but it hardly has an effect in terms of the error variance. In Scenario A-I, density is predominantly overestimated; in A-II and A-III, this overestimation is compensated by larger underestimations (e.g., in Figure 16 around kilometer 5, at the onset of congestion). The curved patterns in congestion in Figures 15 and 16 do indicate that in these cases, the kinematic wave speed is underestimated.
- Figures 15, 17, and 18 show the results for three scenarios in which the *same* fundamental diagram is used but with different detector spacings (A-II, B-II, C-II). These illustrate that, in terms of performance (both bias and variance), there is no added value in using all detector observations (every 500 meters) over using just one-sixth of these data. On the contrary, the results improve slightly by using fewer observations. Too much (redundant)

FIGURE 15. Contour plot of estimated densities and estimation errors for Scenario A-II (detectors at 500 m spacing, Fundamental Diagram II).

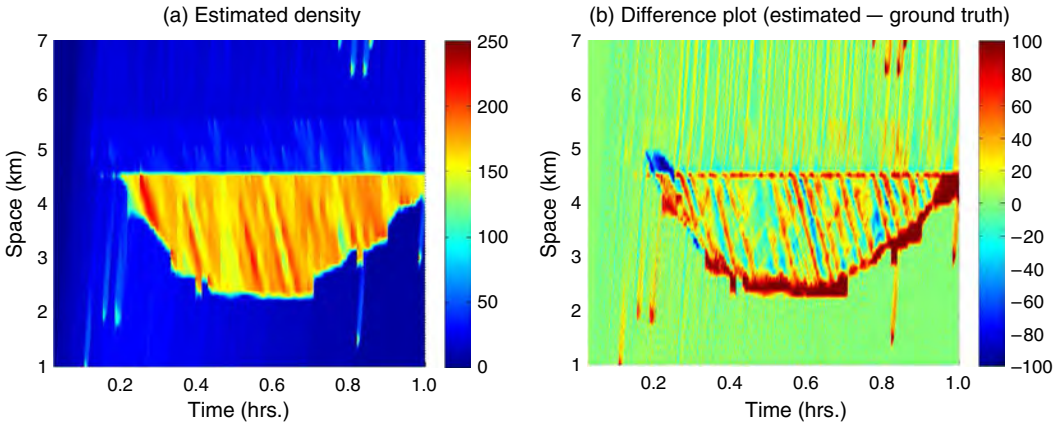


FIGURE 16. Contour plot of estimated densities and estimation errors for Scenario A-III (detectors at 500 m spacing, Fundamental Diagram III).

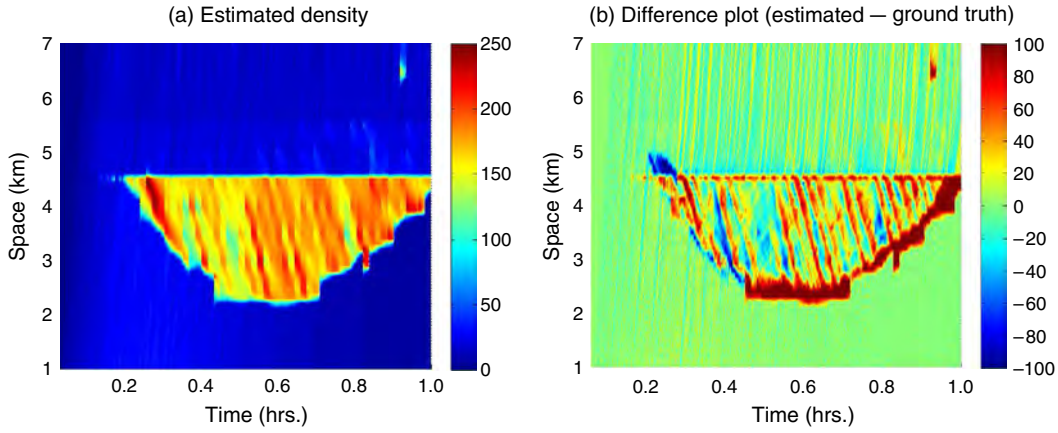


FIGURE 17. Contour plot of estimated densities and estimation errors for Scenario B-II (detectors at 1,000 m spacing, Fundamental Diagram II).

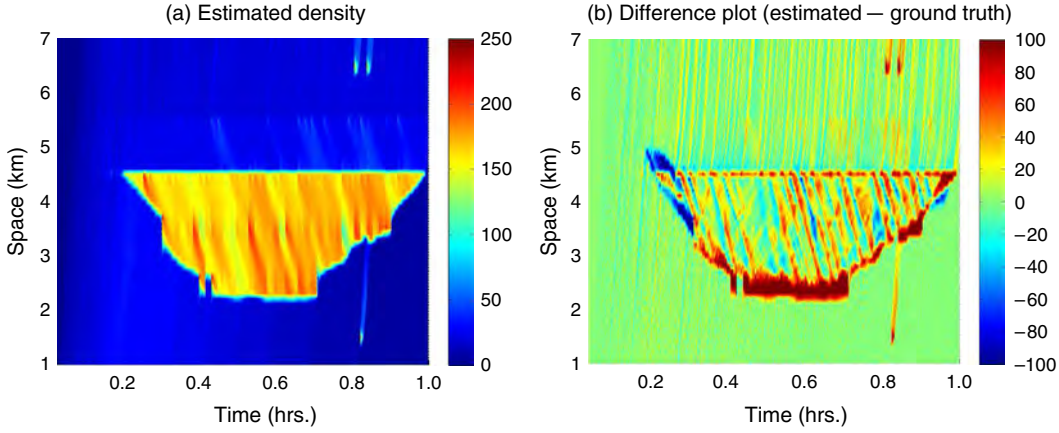
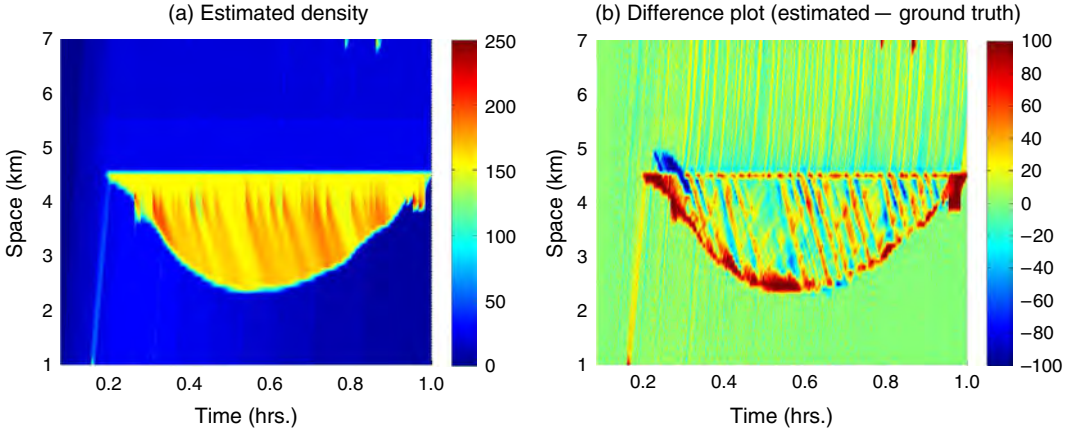


FIGURE 18. Contour plot of estimated densities and estimation errors for Scenario C-II (detectors at 3,000 m spacing, Fundamental Diagram II).



information essentially overdetermines (i.e., introduces too many constraints in) the estimation problem. More effective than using (redundant) observations is a process model that accurately implements the correlations between state variables over time and space (the kinematic wave speed in Equation (34)).

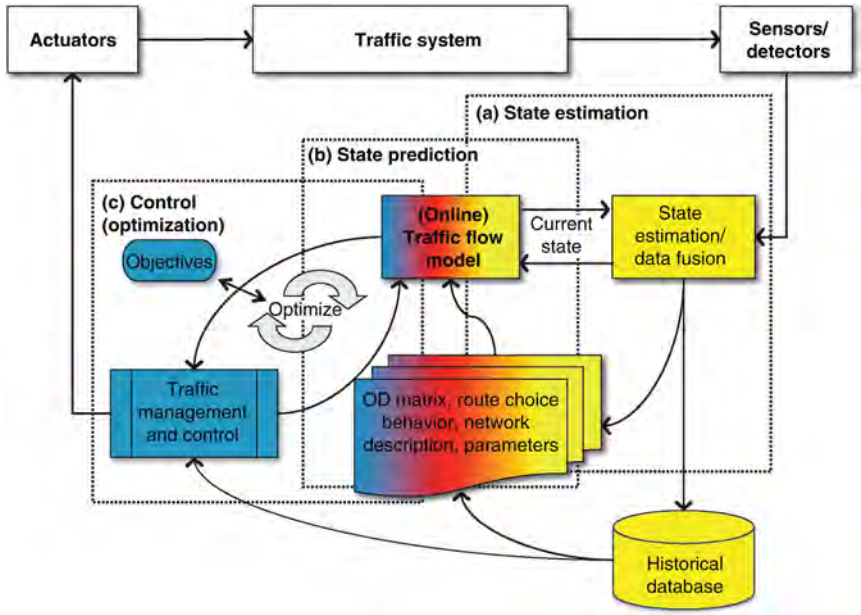
We can draw a few conclusions from these findings. First, model-based state estimation (using the EKF) can be successfully applied to reconstruct at least the first-order traffic patterns from sparse traffic data. First-order patterns relate to the spatiotemporal extent (head and tail) of congestion and the resulting average macroscopic quantities. Second, the quality of observation and process models is critically important. This particularly pertains to the fundamental diagram parameters (capacity, kinematic wave speed) that govern the spatiotemporal correlations between the state variables. Third, data redundancy does not improve estimation results—in this example, on the contrary. Finally, the EKF procedure aims to minimize the (total) estimation error covariance, a performance measure closely related to the RMSE. This exercise demonstrates that total error variance (alone) may not be a very informative performance measure. Minimizing error variance may coincide with increased bias (e.g., compare the second and third rows of Table 4). As with the OD estimation problem in the previous section, these structural errors relate to the underlying dynamics of the problem. In the OD case, a prior OD table was used to steer the estimation in the "right" direction. In traffic state estimation, this prior information comes (among other things) in the form of valid fundamental diagram parameters (which may be location and time specific).

4. Discussion and Further Applications of the (E)KF in Traffic Management and Control

In this tutorial, three relatively simple example estimation problems relevant to the field of dynamic traffic management and control have been presented in which the Kalman filter was used as the solution method. Let us first note that in literature many successful applications of the (E)KF have been reported in the context of these three application domains, for example, in location tracking (Chiang et al. [11], Bétaille and Toledo-Moreo [6], Tu and Kiang [46], Tseng and Feng [45]), traffic state estimation (Yuan et al. [62], Wang et al. [57, 58, 59], van Hinsbergen et al. [48], Li [28], Guo et al. [19], Di et al. [14], van Lint et al. [49]), and OD estimation (Barcelo et al. [5], Cho et al. [12], Zhou and Mahmassani [63], Bierlaire and Crittin [7], Michael and Rilett [31], Hu et al. [23], Van Der Zijpp [47], Cremer and Keller [13], Okutani and Stephanedes [36]). State estimation techniques are also important for the ex ante calibration and validation of the traffic models and algorithms used in these applications (Antoniou et al. [1], van Lint [50]), and for the ex post evaluation of traffic management and control applications. In that case, archived state variables are typically used to compute measures of effectiveness, such as total vehicle loss hours/delays.

In the remainder of this section we restrict ourselves to applications within the class of model-based traffic management and control, in which (traffic) state estimation plays a central role. Figure 19 outlines the components in that class of applications. The three examples in this tutorial showed that the three central assumptions in the (E)KF strongly affect the estimation quality and applicability. These assumptions are that (a) the system model is (locally approximately) linear, (b) all model and observation error terms are Gaussian noise processes, and (c) by implication, process and observation models are unbiased. In this final section we will briefly discuss approaches to relax these assumptions. The first is to use more advanced (complex) data assimilation (filtering) techniques; the second involves the use of more suitable (more linear) system models. In the final subsection we close with some concluding remarks.

FIGURE 19. Schematic representation of model-based DTM and control.



4.1. Relaxing Gaussian and Linear Assumptions: The Ensemble Kalman Filter, Unscented Kalman Filter, and Particle Filter

The objective of state estimation techniques is to estimate the posterior probability density function (pdf) of the state variables given the available observations $p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1})$ at each time step k. The (E)KF is just one solution method to do this and belongs to the more general family of recursive Bayesian estimation methods, which all follow an intuitive prediction/correction structure:

(1) Prediction (prior):

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}) = \int p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) p(\mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1})\, d\mathbf{x}_{k-1}. \tag{44}$$

(2) Update (posterior):

$$p(\mathbf{x}_k \mid \mathbf{y}_k) = \frac{p(\mathbf{y}_k \mid \mathbf{x}_k) p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1})}{\int p(\mathbf{y}_k \mid \mathbf{x}_k) p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1})\, d\mathbf{x}_k}. \tag{45}$$

In (44) and (45), respectively, $p(\mathbf{x}_k \mid \mathbf{x}_{k-1})$ represents the state transition pdf and $p(\mathbf{y}_k \mid \mathbf{x}_k)$ the observation likelihood function.

The EKF approximates the posterior (45) by analytically propagating (assumed) Gaussian random variables through a first-order linearization of the system model. In Julier et al. [24] and Wan and van der Merwe [56] a derivative-free alternative is proposed, using the unscented transformation, which is a method for calculating the statistics of a random variable that undergoes a nonlinear transformation (Julier et al. [24]). In the unscented Kalman filter (UKF), instead of computing mean and error covariance analytically (Equations (38) and (39)), a series of carefully selected samples (sigma points) $\mathcal{X}_i$ is used. These sample points completely capture ("span") the true mean and covariance of the state variables, and when propagated through the true nonlinear system (process and observation model $\mathcal{Y}_i = h(\mathcal{X}_i)$), they capture the time-dependent posterior mean and covariance accurately to the third order (Taylor series expansion) for any nonlinearity. The UKF algorithm is explained by Wan and van der Merwe [56], and many practical examples are available on the Web.

A second derivative-free alternative for the EKF is the so-called ensemble Kalman filter (EnsKF) (Evensen [15]) in which a Markov chain Monte Carlo method is applied. As in the UKF, the starting point of the EnsKF is a prior ensemble of state vectors $\mathbf{x}_i$, from which the prior mean $\hat{\mathbf{x}}_k^- = (1/N)\sum \mathbf{x}_{i,k}$ and error covariance $P_k^- = E[\mathbf{x}_{i,k} - \hat{\mathbf{x}}_k^-]$ are computed. Whereas the UKF uses a deterministic sampling strategy to obtain a new ensemble of state vectors (the sigma points) at every prediction step, the EnsKF draws this ensemble of state vectors randomly and only once in the initialization step. Resampling in the EnsKF takes place in the update/correction step, in which an observation matrix $Y_k = [\ldots, \tilde{\mathbf{y}}_{i,k}, \ldots]$, $\tilde{\mathbf{y}}_{i,k} = \mathbf{y}_{i,k} + \mathbf{v}_k$ with $\mathbf{v}_k \sim \mathrm{N}(0, R_k)$, is sampled. An excellent tutorial of the EnsKF is available in Mandel [30].

A third family of derivative-free filtering techniques for state estimation are so-called particle filters (PFs). Particle filters are classified under many names (sequential importance sampling, sequential Monte Carlo sampling) and come in many different variations. One could argue that also the EnsKF and UKF belong to the PF family (Ristic et al. [42]) as special cases. The key is that in a PF, all distributional assumptions on either prior or posterior are relaxed. In the literature we find examples of all of these alternative state estimation techniques applied to traffic management and control. Table 5 lists some references for further reading.

Table 5 also loosely summarizes the characteristics and the pros and cons of these approaches. Large-scale comparative studies between different approaches in the traffic management and control domain are rare and mostly restricted to simulated examples. For example, Hegyi et al. [20] found that UKF and EKF perform equally well on a simulated

Table 5. Characteristics and references to different filtering approaches for traffic state estimation.

	EKF	UKF	EnsKF	PF
Assumptions error terms	Gaussian	Gaussian	Gaussian	—
Characteristics	Analytical procedure	Deterministic sequential sampling	Sequential Monte Carlo sampling	Sequential Monte Carlo sampling
Pros	One-shot procedure; straightforward to implement	Derivative free; superior nonlinear estimation (compared to EKF)	Derivative free; straightforward to implement; computationally efficient; requires fewer samples than UKF for large-scale problems[a]	Theoretically superior for nonlinear estimation over all other methods
Cons	Linearization inaccurate for highly nonlinear problems; computational cost scales with state dimension	Computational cost scales with state dimension	Heuristic choice of sample size	Computational cost scales with state dimension; largest number of samples required
Examples in traffic state estimation	Yuan et al. [62], Tampere and Immers [43], Wang et al. [57, 58, 59], van Hinsbergen et al. [48], Guo et al. [19], Di et al. [14]	Hegyi et al. [20], Ngoduy [35], Mihaylova et al. [32]	Work et al. [60, 61], Chen et al. [10]	Mihaylova et al. [32, 33], Chen et al. [10]

[a]The number of samples in the UKF is deterministically chosen and is in the same order as the state dimension. The sample (ensemble) size of the EnsKF is chosen heuristically. Gillijns et al. [17], for example, showed that the EnsKF performs equally well or better than an EKF from ensemble sizes in the order of 100.

freeway corridor. Mihaylova et al. [32] made a case for particle filters (which outperform the UKF in their examples). Mihaylova et al. [33] proposed different parallelized PF configurations that improve computational speed and online applicability through parallel simulation. Van Hinsbergen et al. [48] proposed a very different approach for the same purpose. These authors showed that large-scale (i.e., network-wide) traffic state estimation is also possible with a simplified localized EKF algorithm, without loss of estimation accuracy.

4.2. Choosing (More) Appropriate System Models

4.2.1. Improved Observation Models. In the last 10 years the amount of data available from different sources has increased steeply, ranging from counts and speeds from local sensors (induction loops, radar, microwave), to travel times from AVI systems such as camera's, bluetooth "sniffers," and GPS-enabled mobile phones, to detailed samples of microscopic data from in-vehicle systems (location, lane, headway, speed, acceleration). These data differ not just in availability, accuracy, and reliability over space and time, but also in terms of their spatiotemporal semantics (van Lint and Hoogendoorn [51, 52]). Consider, for example, realized travel time over several kilometers measured with an AVI system, and local (harmonic) mean speed measured with an induction loop. The first quantity is the result from a line integral (a trajectory) over time and space (travel time equals travel distance over average journey speed), whereas the second integrates information over a cross section for a fixed time period. These data cannot be straightforwardly aligned and fused with recursive techniques such as the EKF or any of its relatives. In a number of contributions, Ou et al. [37, 38, 39, 40] demonstrate that different data assimilation techniques are required to effectively fuse such semantically different data sources. Nonetheless, these so-called data–data consistency techniques can in turn be used to remove structural bias (e.g., due to local averaging) from the available sensor data *before* these are used in recursive state estimation methods and control applications. Another range of data assimilation approaches were proposed by Herrera and Bayen [21] and Work et al. [60], in which different observation models are used to assimilate Lagrangian data (i.e., data from mobile (in-vehicle) sensors).

4.2.2. Improved Process Models. Improving data assimilation methods is one path to improving nonlinear state estimation. A second path is to improve the system models themselves, that is, to choose and formulate process and observation models more carefully such that they are better tailored for the state estimation task. One example was given in §2.2, in which the OD estimation process (and observation) models were reformulated such that the resulting system was more suitable for use with the Kalman filter. As outlined in §3.2, the use of classic traffic flow models may lead to a violation of the linearity assumption around capacity and may lead to EKF corrections with the wrong sign. Yuan et al. [62] showed that freeway state estimation using the EKF and a Lagrangian formulation of the first-order LWR model (Lighthill and Whitham [29], Richards [41]) leads to large improvements. Although still nonlinear, the Lagrangian LWR model does not suffer from mode switching around capacity (information always travels in the same direction), which leads to more accurate numerical simulation (Leclercq et al. [27], van Wageningen-Kessels et al. [54]) and effectively renders both process and observation models more linear and more suitable for use with the EKF (Yuan et al. [62]). Also, here a different formulation with the same model leads to large performance benefits.

4.2.3. Parameter Estimation and Adaptivity. A final direction in which state estimation approaches in our field can be refined and improved relates to the many parameters involved in both process and observation models. These parameters, $\boldsymbol{\psi}_k$, determine individual or average drive behavior (e.g., the fundamental diagram in Figure 9(b)) and travel behavior (e.g., in (simulation) models that are used to construct the assignment matrix in §§2.1 and 2.2), and are clearly also time-varying random variables. As shown in the freeway

state estimation example, a wrong setting of these parameters may lead to biased process and observation models. The EKF (or any of its more advanced relatives) can also be used to adapt such parameters in real time, using the following system model:

$$\boldsymbol{\psi}_k = \boldsymbol{\psi}_{k-1} + \mathbf{w}_k, \tag{46}$$

$$\mathbf{y}_k = \mathbf{h}(\boldsymbol{\psi}_k, \ldots) + \mathbf{v}_k, \tag{47}$$

in which (46) represents the (unknown) dynamics of these parameters by means of a random walk, and (46) depicts an observation model that could represent an entire traffic simulation and assignment model or any other method to relate observations to the parameters. In the work of van Lint [50], for example, this method is used for the online calibration of travel time prediction models; in the work of Antoniou et al. [1] this approach is used to adapt the parameters in a traffic simulation model. This adaptive approach to parameter estimation can also be employed in parallel with state estimation, often referred to as dual estimation (e.g., in Wang et al. [58], van Lint et al. [49], Hegyi et al. [20]). An important note to make is that the dynamics of drive or travel behavioral parameters are different (slower over time and more coarse over space) than the dynamics of state variables such as speed and density.

4.3. Concluding Remarks

Despite the large and fast-growing body of knowledge available in the literature, the implementation of model-based traffic management and control (Figure 19) in actual traffic management centers is still in its infancy. We can put forward several tentative reasons for this. One relates to the difficulty of porting the models and data assimilation techniques from research environments to actual traffic management practice. Traffic and transport are very broad domains in which many disciplines meet, including traffic engineering, behavioral science, psychology, operations research, applied mathematics, computer science, and control engineering, to name a few. Each of these have their own terminology and worldview, which makes it difficult to convince road operators and politicians of the need for knowledge-based traffic management and control. A related point is that in many countries and cities around the world, (car) mobility is predominantly viewed and governed as a (necessarily) self-organising process, in which too-rigid management and control is not considered appropriate beyond the control required for safety and for efficient traffic operations under free-flowing or mildly congested network traffic. This view is reflected also in the steep increase in research and development of cooperative vehicle-based technologies in the European Union, the United States, and Japan, which, besides improved safety, enable high-speed platooning, automated lane keeping, and assisted merging. Clearly, these technologies offer great potential for traffic management and control in the future, not just to make traffic more safe and efficient, but also more sustainable.

The problem with this strong focus on "bottom-up" technologies and methods is that self-organisation (of drive and route choice behavior) no longer leads to efficient traffic operations in heavily congested networks, because of spill back, capacity drop (which makes congested traffic up to 30% less efficient than free-flow traffic), and suboptimal (user optimal) route choice. In many metropolitan areas in the world (e.g., Beijing, Paris, Cairo), traffic gridlock occurs on a daily basis. Such gridlocks cannot be solved by local (vehicle) intelligence only. These vehicles must drive on an intelligent infrastructure that is managed by intelligent network operators. In case of very dense network traffic conditions, *system optimal* coordinated and integrated traffic management and control is required. For that to happen, a system-oriented perspective on (car) mobility must be put higher on the political agenda.

Acknowledgments

The authors acknowledge the 2012 *TutORials* volume editor, Pitu Mirchandani, for his invitation and collaboration on this chapter, Thomas Schreiter for his contribution to some of the underlying exercises, and Serge Hoogendoorn for his collaboration on the underlying lecture slides from which this chapter was drafted.

References

[1] C. Antoniou, M. Ben-Akiva, and H. N. Koutsopoulos. On-line calibration of traffic prediction models. *Proceedings of the 84th Annual Meeting Transportation Research.* National Academies Press, Washington, DC, 2005.

[2] K. Ashok and M. Ben-Akiva. Matrix estimation and prediction for real-time traffic management systems. C. F. Daganzo, ed. *Transportation and Traffic Theory: Proceedings of the 12th ISTTT.* Elsevier, New York, 465–484, 1993.

[3] K. Ashok and M. E. Ben-Akiva. Alternative approaches for real-time estimation and prediction of time-dependent origin-destination flows. *Transportation Science* 34(1):21–36, 2000.

[4] K. Ashok and M. E. Ben-Akiva. Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows. *Transportation Science* 36(2):184–198, 2002.

[5] J. Barcelo, L. Montero, L. Marquos, and C. Carmona. Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. *Transportation Research Record: Journal of the Transportation Research Board* (2175):19–27, 2010.

[6] D. Bétaille and R. Toledo-Moreo. Creating enhanced maps for lane-level vehicle navigation. *IEEE Transactions on Intelligent Transportation Systems* 11(4):786–798, 2010.

[7] M. Bierlaire and F. Crittin. An efficient algorithm for real-time estimation and prediction of dynamic OD tables. *Operations Research* 52(1):116–127, 2004.

[8] E. Cascetta, D. Inaudi, and G. Marquis. Dynamic estimators of origin-destination matrices using traffic counts. *Transportation Science* 27(4):363–373, 1993.

[9] G.-L. Chang and J. Wu. Recursive estimation of time-varying origin-destination flows from traffic counts in freeway corridors. *Transportation Research Part B: Methodological* 28(2):141–160, 1994.

[10] H. Chen, H. A. Rakha, and S. Sadek. Real-time freeway traffic state prediction: A particle filter approach. *The 14th International IEEE Conference on Intelligent Transportation Systems.* IEEE, New York, 626–631, 2011.

[11] C.-T. Chiang, P.-H. Tseng, and K.-T. Feng. Hybrid unified Kalman tracking algorithms for heterogeneous wireless location systems. *IEEE Transactions on Vehicular Technology* 61(2):702–715, 2012.

[12] H.-J. Cho, Y.-J. Jou, and C.-L. Lan. Time dependent origin-destination estimation from traffic count without prior information. *Networks and Spatial Economics* 9(2):145–170, 2009.

[13] M. Cremer and H. Keller. A new class of dynamic methods for the identification of origin-destination flows. *Transportation Research Part B: Methodological* 21(2):117–132, 1987.

[14] X. Di, H. X. Liu, and G. A. Davis. Hybrid extended Kalman filtering approach for traffic density estimation along signalized arterials. *Transportation Research Record* (2188):165–173, 2010.

[15] G. Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics* 53:343–367, 2003.

[16] R. Frederix, F. Viti, R. Corthout, and C. M. J. Tampre. New gradient approximation method for dynamic origin-destination matrix estimation on congested networks. *Transportation Research Record: Journal of the Transportation Research Board* (2263):19–25, 2012.

[17] S. Gillijns, O. Barrero Mendoza, B. L. R. De Moor J. Chandrasekar, D. S. Bernstein, and A. Ridley. What is the ensemble Kalman filter and how well does it work? *Proceedings of the 2006 American Control Conference, Minneapolis*, 4448–4453, 2006.

[18] B. D. Greenshields. The photographic method of studying traffic behaviour. *Proceedings of the 13th Annual Meeting of the Highway Research Board, Washington, DC*, 382–399, 1933.

[19] J. Guo, J. Xia, and B. L. Smith. Kalman filter approach to speed estimation using single loop detector measurements under congested conditions. *Journal of Transportation Engineering* 135(12):927–934, 2009.

[20] A. Hegyi, D. Girimonte, R. Babuska, and B. De Schutter. A comparison of filter configurations for freeway traffic state estimation. *The 9th International IEEE Conference on Intelligent Transportation Systems.* IEEE, New York, 1029–1034, 2006.

[21] J. C. Herrera and A. M. Bayen. Incorporation of Lagrangian measurements in freeway traffic state estimation. *Transportation Research Part B: Methodological* 44(4):460–481, 2010.

[22] S. P. Hoogendoorn and P. H. L. Bovy. State-of-the-art of vehicular traffic flow modeling. *Proceedings of the Institution of Mechanical Engineers, Part 1: Journal of Systems and Control Engineering* 215(1):283–303, 2001.

[23] S.-R. Hu, S. M. Madanat, J. V. Krogmeier, and S. Peeta. Estimation of dynamic assignment matrices and OD demands using adaptive Kalman filtering. *Intelligent Transportation Systems Journal* 6(3):281–300, 2001.

[24] S. Julier, J. Uhlmann, and H. Durrant-White. A new method for nonlinear transformation of means and covariances in filters and estimators. *IEEE Transactions on Automatic Control* 45(3):477–482, 2000.

[25] R. E. Kalman. A new approach to linear filtering and prediction problems. *ASME Basic Engineering Journal* 82(Series D):35–45, 1960.

[26] J. P. Lebacque. The Gudunov scheme and what it means for first order traffic flow models. J. B. Lesort, ed. *Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Lyon, France*, 647–677, 1996.

[27] L. Leclercq, J. Laval, and E. Chevallier. The Lagrangian coordinates and what it means for first order traffic flow models. R. E. Allsop, M. G. H. Bell, B. G. Heydecker, eds. *Proceedings of the 17th International Symposium on Transportation and Traffic Theory*. Elsevier, London, 735–753, 2007.

[28] B. Li. A non-gaussian Kalman filter with application to the estimation of vehicular speed. *Technometrics* 51(2):162–172, 2009.

[29] M. Lighthill and G. Whitham. On kinematic waves. II. A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences* 229(1178):317–345, 1955.

[30] J. Mandel. A brief tutorial on the ensemble Kalman filter. E-print, arXiv, http://arxiv.org/abs/0901.3725v1, 2009.

[31] P. D. Michael and L. R. Rilett. Real-time OD estimation using automatic vehicle identification and traffic count data. *Computer-Aided Civil and Infrastructure Engineering* 17(1):7–21, 2002.

[32] L. Mihaylova, R. Boel, and A. Hegyi. Freeway traffic estimation within particle filtering framework. *Automatica* 43(2):290–300, 2007.

[33] L. Mihaylova, A. Hegyi, A. Gning, and R. K. C Boel. Parallelized particle and Gaussian sum particle filters for large-scale freeway traffic systems. *IEEE Transactions on Intelligent Transportation Systems* 13(1):36–48, 2012.

[34] G. F. Newell. A simplified car-following theory: A lower order model. *Transportation Research Part B* 36(3):195–205, 2002.

[35] D. Ngoduy. Low-rank unscented Kalman filter for freeway traffic estimation problems. *Transportation Research Record* (2260):113–122, 2011.

[36] I. Okutani and Y. J. Stephanedes. Dynamic prediction of traffic volume through Kalman filtering theory. *Transportation Research Part B: Methodological* 18(1):1–11, 1984.

[37] Q. Ou, H. van Lint, and S. P. Hoogendoorn. Traffic data fusion—Fusing heterogeneous and unreliable data from traffic sensors. R. Babuska, F. C. A. Groen, eds. *Interactive Collaborative Information Systems*, Vol. 281. Springer, Berlin, 511–545, 2010.

[38] Q. Ou, J. W. C. van Lint, and S. P. Hoogendoorn. Piecewise inverse speed correction by using individual travel times. *Transportation Research Record* (2049):92–102, 2008.

[39] Q. Ou, J. W. C. van Lint, and S. P. Hoogendoorn. An integrated algorithm for fusing travel times, local speed and flow. *13th International Conference on Information Fusion, Edinburgh, UK*, 1–8, 2010.

[40] Q. Ou, J. W. C. van Lint, and S. P. Hoogendoorn. Fusing heterogeneous and unreliable data from traffic sensors. *Studies in Computational Intelligence* 281:511–545, 2010.

[41] P. I. Richards. Shock waves on the highway. *Operations Research* 4(1):42–51, 1956.

[42] B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House Publishers, Boston, 2004.

[43] C. Tampere and B. Immers. An extended Kalman filter application for traffic state estimation using CTM with implicit mode switching and dynamic parameters. *The 11th International IEEE Conference on Intelligent Transportation Systems.* IEEE, New York, 209–216, 2007.

[44] M. Treiber and D. Helbing. Reconstructing the spatio-temporal traffic dynamics from stationary detector data. *Cooper@tive Transport@tion Dyn@mics* 1(3):1–24, 2002.

[45] P.-H. Tseng and K.-T. Feng. Hybrid network/satellite-based location estimation and tracking systems for wireless networks. *IEEE Transactions on Vehicular Technology* 58(9):5174–5189, 2009.

[46] P.-J. Tu and J.-F. Kiang. Estimation on location, velocity, and acceleration with high precision for collision avoidance. *IEEE Transactions on Intelligent Transportation Systems* 11(2):374–379, 2010.

[47] N. Van Der Zijpp. Dynamic origin-destination matrix estimation from traffic counts and automated vehicle identification data. *Transportation Research Record: Journal of the Transportation Research Board* 1607(1):87–94, 1997.

[48] C. P. I. J. van Hinsbergen, T. Schreiter, F. S. Zuurbier, J. W. C. van Lint, and H. J. van Zuylen. Localized extended Kalman filter for scalable real-time traffic state estimation. *IEEE Transactions on Intelligent Transportation Systems* 13(1):385–394, 2012.

[49] H. van Lint, A. Hegyi, and S. P. Hoogendoorn. Dual EKF state and parameter estimation in multi-class first-order traffic flow models. *17th International Federation of Automatic Control World Congress (IFAC), Seoul, South Korea*, 14078-14084, 2008.

[50] J. W. C. van Lint. Online learning solutions for freeway travel time prediction. *IEEE Transactions on Intelligent Transportation Systems* 9(1):38–47, 2008.

[51] J. W. C. van Lint and S. P. Hoogendoorn. The technical and economic benefits of data fusion for real-time monitoring of freeway traffic: Preliminary results and implications of a study with simulated data. *Proceedings of the 11th World Conference on Transport Research, University of California, Berkeley*, 1–15, 2007.

[52] J. W. C. van Lint and S. P. Hoogendoorn. A robust and efficient method for fusing heterogeneous data from traffic sensors on freeways. *Computer-Aided Civil and Infrastructure Engineering* 25(8):596–612, 2010.

[53] F. L. M. van Wageningen-Kessels, J. W. C. van Lint, S. P. Hoogendoorn, and C. Vuik. Numerical time stepping schemes for macroscopic traffic flow models. H. J. van Zuylen and A. J. van Binsbergen, eds. *Proceedings of the 10th TRAIL-Congress.* TRAIL Research School, Delft, The Netherlands, 1–10, 2008.

[54] F. L. M. van Wageningen-Kessels, J. W. C. van Lint, S. P. Hoogendoorn, and C. Vuik. Lagrangian formulation of a multi-class kinematic wave model. *Transportation Research Record: Journal of the Transportation Research Board* (2188):29–36, 2010.

[55] R. G. M. M. Vermijs and H. Schuurman. Evaluating capacity of freeway weaving sections and on-ramps using the microscopic simulation model FOSIM. R. Akçelik, ed. *Proceedings of the Second International Symposium on Highway Capacity*, Vol. 2. Australian Road Research Board, Sydney, Australia, 651–670, 1994.

[56] E. A. Wan and R. van der Merwe. The unscented Kalman filter for nonlinear estimation. *IEEE Symposium on Adaptive Systems for Signal Processing, Communications and Control.* IEEE, New York, 153–158, 2000.

[57] Y. Wang, M. Papageorgiou, and A. Messmer. RENAISSANCE—A unified macroscopic model-based approach to real-time freeway network traffic surveillance. *Transportation Research Part C: Emerging Technologies* 14(3):190–212, 2006.

[58] Y. Wang, M. Papageorgiou, and A. Messmer. Real-time freeway traffic state estimation based on extended Kalman filter: Adaptive capabilities and real data testing. *Transportation Research Part A: Policy and Practice* 42(10):1340–1358, 2008.

[59] Y. Wang, M. Papageorgiou, A. Messmer, P. Coppola, A. Tzimitsi, and A. Nuzzolo. An adaptive freeway traffic state estimator. *Automatica* 45(1):10–24, 2009.

[60] D. B. Work, O.-P. Tossavainen, Q. Jacobson, and A. M. Bayen. Lagrangian sensing: Traffic estimation with mobile devices. *2009 American Control Conference, St. Louis*, 1536–1543, 2009.

[61] D. B. Work, O.-P. Tossavainen, S. Blandin, A. M. Bayen, T. Iwuchukwu, and K. Tracton. An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices. *The 47th IEEE Conference on Decision and Control*. IEEE, New York, 5062–5068, 2008.

[62] Y. Yuan, J. W. C. van Lint, R. E. Wilson, F. Van Wageningen-Kessels, and S. P. Hoogendoorn. Real-time Lagrangian traffic state estimator for freeways. *IEEE Transactions on Intelligent Transportation Systems* 13(1):59–70, 2012.

[63] X. Zhou and H. S. Mahmassani. A structural state space model for real-time traffic origin-destination demand estimation and prediction in a day-to-day learning framework. *Transportation Research Part B: Methodological* 41(8):823–840, 2007.