Hindawi
Journal of Advanced Transportation
Volume 2017, Article ID 5730648, 11 pages
https://doi.org/10.1155/2017/5730648

WILEY | Hindawi

Research Article

Macroscopic Traffic State Estimation: Understanding Traffic Sensing Data-Based Estimation Errors

Paul B. C. van Erp, Victor L. Knoop, and Serge P. Hoogendoorn

Department of Transport & Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, 2628 CN Delft, Netherlands

Correspondence should be addressed to Paul B. C. van Erp; p.b.c.vanerp@tudelft.nl

Received 24 May 2017; Accepted 1 October 2017; Published 1 November 2017

Academic Editor: Martin Trépanier

Copyright © 2017 Paul B. C. van Erp et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Traffic state estimation is a crucial element in traffic management systems and in providing traffic information to road users. In this article, we evaluate traffic sensing data-based estimation error characteristics in macroscopic traffic state estimation. We consider two types of sensing data, that is, loop-detector data and probe speed data. These data are used to estimate the mean speed in a discrete space-time mesh. We assume that there are no errors in the sensing data. This allows us to study the errors resulting from the differences in characteristics between the sensing data and desired estimate together with the incomplete description of the relation between the two. The aim of the study is to evaluate the dependency of this estimation error on the traffic conditions and sensing data characteristics. For this purpose, we use microscopic traffic simulation, where we compare the estimates with the ground truth using Edie's definitions. The study exposes a relation between the error distribution characteristics and traffic conditions. Furthermore, we find that it is important to account for the correlation between individual probe data-based estimation errors. Knowledge related to these estimation errors contributes to making better use of the available sensing data in traffic state estimation.

1. Introduction

Traffic state estimation is an important element in traffic management applications and traffic information services. The traffic state can be described on different levels [1]. The microscopic traffic state describes the traffic on an individual vehicle level, thus using the time and space headways and individual vehicle speed. The macroscopic traffic state describes the traffic flow conditions using the mean speed, density, and flow.

In traffic state estimation, different types of information can be used. Traffic sensing data is collected via different kind of sensors, for example, loop-detectors [2, 3] and mobile phones (probes) [4, 5]. In addition to sensing data, information related to the traffic dynamics is captured in the form of traffic flow models, for example, the LWR model [6, 7]. Traffic flow models are based on physical laws and historical data. These information types, thus sensing data and traffic flow models, allow us to estimate the traffic state. The difference between the true and estimated traffic states is denoted as the estimation error. In this research, the focus is put on these estimation errors.

It is valuable to have information related to the estimation errors for applications of the related estimates. For instance, in traffic state estimation, different types of information, for example, traffic sensing data-based estimates and traffic flow model-based prediction, can be fused. Examples of such applications are [3–5, 8]. These applications all consider (amongst others) a variant of the Kalman Filter (KF) [9] for information fusion. KFs assume Gaussian distributed errors with a to-be-defined (co)variance. Alternatively, a Particle Filter (PF) [10] can be used for information fusion, in which we are free to define any type of expected error distribution. In the provided references related to the KF, the researchers all assume constant error variances related to the data-based, that is, loop-detector, or probe data-based estimates. They thus do not assume any dependency of the error characteristics on features such as the traffic conditions or varying sensing data characteristics. In addition to information fusion, information related to estimation or

prediction errors can also have a more direct value for road users. For instance, in travel time prediction, a probability function can be provided instead of a single expected value [11]. This may be valuable for routing advise, as the travel time variability may negatively influence the attractiveness of certain routes.

In this research, the following way of thinking is considered. If we have knowledge related to the relation between estimation error characteristics and potentially observable features, we may improve our understanding of the error characteristics related to an estimate. Examples of potentially observable features are traffic conditions and sensing data characteristics. Improved understanding of the estimation error characteristics is valuable in the applications discussed above, for example, traffic state estimation using a variant of the KL [3, 4, 8].

The objective of this research is to expose the dependency of traffic sensing data-based estimation errors on traffic conditions and sensing data characteristics. We define traffic sensing data-based estimation as the estimation of a desired output based on traffic sensing data. Both the sensing data as the desired output have specific characteristics, for example, type of variable and spatial/temporal characteristics. If these differ, we have to make assumptions to describe the relation between the two. In this research, we estimate the mean speed in discrete time and space; that is, time is discretized in periods and space in cells (road segments), similar to [3, 4, 8, 12, 13]. The properties of the traffic sensing data-type we consider, that is, loop-detector data and probe speed data, are based on existing research [4, 14].

This research focuses on specific combinations of traffic sensing data and estimation output. The findings can be used for applications which consider similar data-types and estimation output. However, more generally, we opt to show that the estimation error characteristics can depend on the traffic conditions and (varying) sensing data characteristics. Any application that requires defining the estimation error characteristics, for example, information fusion using a KF, can take this into account. However, depending on the specific application, this may require extra research.

In this article, we first describe the (macroscopic) traffic conditions within a discrete space-time mesh. Next, we discuss traffic sensing data-based traffic state estimation and our focus related to this topic. After describing the conducted experiments, we present and discuss the results. Finally, the conclusions of this research are presented.

2. Variables Used to Describe the Traffic Conditions

The traffic conditions can be described as a function of space x and time t. For computational reasons, it is valuable to consider the traffic conditions in discretized space and time [15]. To discretize the space x, the road stretch is subdivided into I cells, where i and Δ_i, respectively, denote the cell number and length of cell i. The number of lanes within i is given by λ_i. Furthermore, time t is discretized in time periods with duration T, which are denoted by $p = 1, \ldots, P$. Each combination of i and p corresponds to a discrete area in the space-time domain. In this discrete space-time mesh, the macroscopic traffic variables mean speed, flow, and density are, respectively, denoted by $u(i, p)$, $q(i, p)$, and $\rho(i, p)$.

In the literature, different methodologies are proposed to calculate the macroscopic variables in a discrete space-time mesh based on the microscopic variables. For instance, Wang and Papageorgiou [3] propose calculating each variable independently based on the downstream (flow) and end-of-period (mean speed and density) conditions. Alternatively, Edie [16] proposed a generalized formulation of the macroscopic variables. In this research, we follow Edie's formulation as it considers the conditions over the entire space-time area instead of only the end-of-period and downstream boundary conditions.

In traffic state estimation in a discrete space-time mesh, homogeneous conditions (defined as constant over space) and stationary conditions (defined as constant over time) are often assumed [8]. Different vehicle classes (e.g., passenger cars and trucks) can coexist in homogeneous and stationary conditions, namely, if the conditions within these classes are homogeneous and stationary.

Assumptions related to homogeneity and stationarity can be important when applying a traffic flow model. For instance, the Cell Transmission Model (CTM) [12, 13] assumes a constant cell outflow during the entire period p. It is, however, also important in sensing data-based estimation. In nonhomogeneous conditions, a loop-detector placed at the upstream cell boundary may observe different traffic conditions than one placed at the downstream boundary. If the conditions are homogeneous, both loop-detectors observe the same conditions and a loop-detector placed at any location within the cell is representative for the conditions in the entire cell. Furthermore, the variation in individual vehicle speeds v can increase when the traffic conditions are nonhomogeneous and nonstationary. This can lead to a larger estimation uncertainty when estimating the traffic conditions based on individual probe speeds.

In reality, traffic is nonhomogeneous and nonstationary [17]. Such traffic conditions can still be expressed in the macroscopic traffic flow variables, but these variables may be incapable of uniquely describing the traffic conditions. For instance, in terms of the (traditional) macroscopic variables, an area in which vehicles are decelerating due to downstream congestion (jam inflow) and in which vehicles are accelerating when leaving a jam may be the same, while in reality the conditions differ.

To capture the conditions that are nonhomogeneous or nonstationary, extra traffic variables can be used. In this research, we add a single extra traffic variable related to the nonhomogeneity of traffic, that is, the change in speed over space. Although it is possible to add more variables, adding this single variable suffices for the analysis conducted in our experiments.

3. Sensing Data-Based Mean Speed Estimation

The explanation of sensing data-based mean speed estimation is split into three parts. First, we discuss the traffic sensing data considered in this research. Second, the estimation

approach to obtain the mean speed from the sensing data is presented. And third, we discuss how the estimation error distribution can be described.

3.1. Traffic Sensing Data Characteristics. Seo [18] states that we can regard traffic data collection as a special case of traffic state estimation. In the procedure to obtain traffic data from raw sensor signals, exogenous assumptions are required. In this research, the starting point is traffic sensing data. We assume that these data do not contain errors. This assumption allows us to study the errors induced due to differences between the sensing data characteristics and desired estimate characteristics and incomplete description of the relation between the two.

We consider two types of traffic sensing data, that is, loop-detector data and probe speed data. The characteristics of the loop-detector data are based on the loop-detector data available in Netherlands, that is, lane-specific one-minute aggregated (time-mean) speeds u_l^T and flows q_l [14]. Following [3], the loop-detectors are located at the downstream boundary of discrete road segments. In line with [4, 5], we consider instantaneous individual vehicle speeds from probe vehicles, that is, v_n, where n describes the vehicle's ID. It is assumed that the probes are observed at the end of each period. These data can be collected from GPS-enabled mobile phones and navigation systems.

3.2. Estimation Approach. The desired estimation output has specific characteristics, that is, variable type and spatial/temporal characteristics. In this research, we estimate the mean speed u for a cell (discrete road segment) i and period p, that is, $u(i, p)$. This desired output is estimated based on the two traffic sensing data-types discussed above.

The traffic sensing data (partly) and desired output differ in terms of variable type and spatial/temporal characteristics. Therefore, we have to define models to estimate $u(i, p)$ based on the sensing data. The models used in this research are taken from prior research, that is, [4, 14].

The loop-detector data and probe speed data-based estimates are, respectively, denoted as $\widehat{u}_{\text{ld}}$ and $\widehat{u}_{\text{probe}}$. Based on the loop-detector data, the speed is estimated by taking the weighted harmonic mean of the lane-specific speeds [14]. Here, we consider the loop-detector data which relates to the cell and period for which u is estimated.

$$\widehat{u}_{\text{ld}} = \frac{\sum_{l=1}^{\lambda} q_l}{\sum_{l=1}^{\lambda} \left(q_l / u_l^T\right)}. \tag{1}$$

We consider the mean v_n of the j number of vehicles observed in a specific cell and period as the probe data-based u estimate, that is, $\widehat{u}_{\text{probe}}$ [4].

$$\widehat{u}_{\text{probe}} = \frac{1}{j} \sum_{n=1}^{j} v_n. \tag{2}$$

3.3. Estimation Error Distribution. The traffic sensing data-based estimates may differ from the true $u(i, p)$. We denote this difference as the sensing data-based estimation error. The characteristics of these errors may be described using the error distribution.

We describe the estimation error distribution using four statistics, namely, the mean, variance, skewness, and kurtosis. The mean, variance, skewness, and kurtosis, respectively, relate to the first, second, third, and fourth standardized moments of a distribution [19]. The skewness addresses the symmetry of a distribution and the kurtosis provides information related to the peakedness or alternatively the "fat tails" of a distribution [19]. For perfect Gaussian distribution, the skewness and kurtosis are, respectively, equal to 0 and 3. By means of the Jarque-Bera (JB) test [20], we can test for normality. The null hypothesis of the JB test is normality.

4. Experimental Set-Up

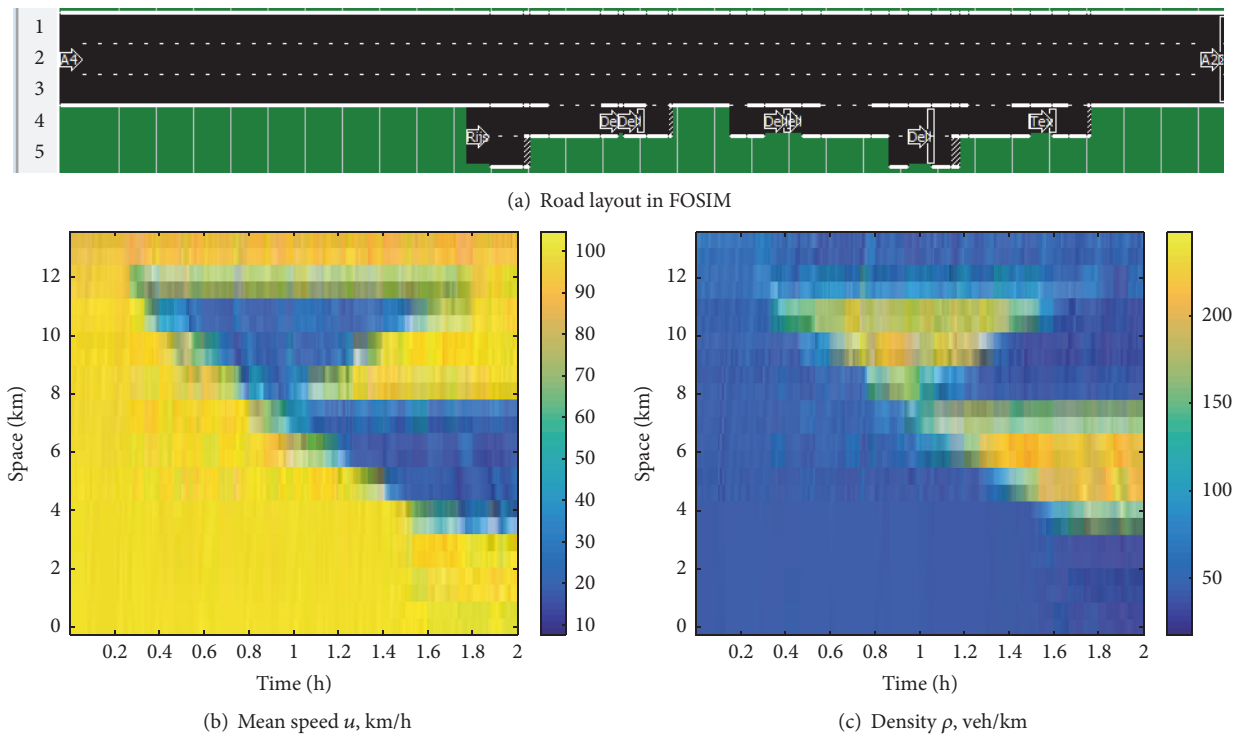
The objective of the experiments is to expose the characteristics of the data-based mean speed estimation errors. In this section, we discuss the data used in this research and explain the conducted experiments.

4.1. Data Collection. In this research, we consider synthesized data collected using the microscopic simulation program FOSIM. The microscopic models and calibration used in FOSIM are described in [21]. Furthermore, it is validated for Dutch freeways [22, 23]. FOSIM allows us to retrieve trajectory data for each individual vehicle. The trajectory data are used for two purposes. Firstly, we construct the traffic sensing data, that is, loop-detector data and probe speed data, with the characteristics described in the previous section. Secondly, we construct the ground truth, as will be explained below. Combined, these allow us to obtain the traffic sensing data-based estimation errors and evaluate their characteristics.

Real trajectory datasets are scarce and are often limited in terms of spatial and temporal coverage. For instance, the NGSIM [24] trajectory dataset covers a study area of approximately 640 m for a 45-minute period. FOSIM allows us to simulate traffic for a much larger spatial and temporal coverage.

We consider a schematized road stretch of the Dutch A13 freeway from the Hague to Rotterdam which has a speed limit of 100 km/h in our experiments. The length of the road stretch is 13,878 m, with five on-ramps and four off-ramps. The road is discretized in 24 cells with lengths ranging from 520 to 770 m. We consider a two-hour time domain which is discretized into periods of 15 seconds: $T = 15/3600$ h. This discretization is based on the approach followed by [3]. The road layout and the traffic conditions in terms of u and ρ are shown in Figure 1. Within this space-time domain, two standing queues are observed.

4.2. Ground Truth. The ground truth is important to describe the true traffic conditions in a discrete area in the space-time domain and determine the estimation errors by comparing the data-based estimates with the ground truth. As explained before, we describe the macroscopic traffic conditions by four variables, namely, the mean speed, density, flow, and



(a) Road layout in FOSIM

(b) Mean speed u, km/h

(c) Density ρ, veh/km

FIGURE 1: The schematized road stretch in FOSIM (a) and the true traffic conditions ((b) and (c)) in terms of the mean speed u and density ρ.

change in speed over space. The former three variables are determined using [16] generalized formulation.

$$\rho = \frac{\sum_n r_n}{\Delta T},$$
$$q = \frac{\sum_n d_n}{\Delta T}, \quad (3)$$
$$u = \frac{q}{\rho},$$

where the time spent and distance traveled by vehicle n within the space-time area are, respectively, denoted by r_n and d_n.

The change in speed over space is obtained by performing an OLS regression over the end-of-period individual vehicle speeds. For this purpose, we consider the following equation: $v_n = \theta_0 + \theta_1 x_n$. The change in speed over space is captured in θ_1.

Given the ground truth u and data-based estimate $\widehat{u}$, the estimation error η can be determined:

$$\eta = u - \widehat{u}. \quad (4)$$

4.3. Evaluation of Estimation Error Characteristics. In this research, we evaluate the dependency of loop-detector and probe data-based estimation error distributions on the traffic conditions. Here, the probe data-based estimates are based on individual probe observations. In reality, we may have multiple probe observations for a given combination of cell and period. Therefore, we will also consider the dependency of the probe data-based estimation error characteristics on the observed fraction of the traffic flow. This sensing dataset characteristic is denoted by the number of probe observations.

4.3.1. Dependency on Traffic Conditions. The traffic conditions can be described using the traffic flow variables. The estimation errors may be directly explained based on the traffic flow variables using, for instance, linear regression. Problems with this approach are that the explanatory variables are correlated (as described by fundamental diagrams, e.g., [25]) and nonlinear relations may exist between the explanatory and dependent variables. Instead of considering the traffic flow variables as independent variables, we want to identify the different traffic conditions in the considered dataset based on the combination of these variables. For this purpose, the true traffic conditions are grouped into a discrete number of clusters using the K-means clustering algorithm. Next, the estimation error characteristics of the observations assigned to each cluster can be described. This approach allows us to get an insight into the dependency of the estimation errors on the traffic conditions.

Using the K-means clustering algorithm [26], each area in time and space is assigned to one of the defined K number of clusters based on the traffic conditions. These conditions are described by three variables, namely, the mean speed u, density ρ, and change in individual vehicle speeds over space $\partial v/\partial x$. The vector $z^{(o)}$ contains these three variables

for observation o, where o relates to a single area in space-time domain. The macroscopic traffic variable flow q is left out because u and ρ combined contain this information, that is, $q = \rho u$, and in contrast to q are able to uniquely describe the homogeneous and stationary traffic conditions.

K-Means clustering follows an iterative procedure to minimize a cost function J_{cost}. This iterative procedure is susceptible to local optima. This means that the algorithm can converge to different solutions. To overcome this problem, the algorithm is applied 10 times and the solution with the lowest cost J_{cost} is selected.

We consider the cost function given by (5). This function considers the mean squared difference between the individual observations $z^{(o)}$ and centroid of the assigned cluster $\mu_{c^{(o)}}$ for the total number of observations m. Here, the cluster centroid $\mu_{c^{(o)}}$ is defined as the mean values for each variable of all observations assigned to cluster c.

$$J_{\text{cost}} = \frac{1}{m}\sum_{o=1}^{m}\left\|z^{(o)} - \mu_{c^{(o)}}\right\|^2. \tag{5}$$

Before applying the clustering algorithm, the explanatory variables (features) are scaled. If feature scaling is not applied, it is possible that features with larger absolute difference will dominate the clustering. Therefore, for each feature, the z-scores are considered. The z-score of an observation is equal to the difference between the observation and the mean of all observations divided by the standard deviation of all observations.

To find the optimal K, we plot J_{cost} as a function of K. Increasing K will decrease the cost, since a more refined clustering is possible. However, this plot allows us to visually compare benefits in terms of J_{cost} of adding more clusters. The optimal K is selected by searching for a kink, which is referred to as the elbow, in the plot. Up till this point, adding a cluster yields a relatively large benefit and thus a decrease in J_{cost}, while the added value of increasing the number of clusters is limited. Therefore, the selected number of clusters is at the location of the elbow. This selection procedure is subjective as we have to define what is a kink and what is not. Therefore, it is also important to interpret the cluster characteristics and see if they make sense. As will be shown in the results, the selection procedure works well in our application.

After each observation has been assigned to a cluster, the estimation error distribution characteristics can be determined per cluster. We are specifically interested in the differences between clusters. Here, the cluster characteristics, as described by the cluster centroids, can be compared with the error characteristics, as described by the error statistics.

We assume that the loop-detectors are located at the downstream boundary of each road segment. Based on these data, a single u estimate is obtained for each combination of i and p, $\{i, p\}$. Furthermore, in this part of the research, we consider every possible individual probe data-based speed estimate. The number of vehicles on road segment i in period p is defined as $N(i, p)$. Therefore, for each combination of $\{i, p\}$, one loop-detector and $N(i, p)$ probe data-based estimates are obtained.

4.3.2. Dependency on Observed Fraction of the Traffic Flow. The probe data-based estimate (see (2)) and thus the probe data-based estimation error depend on the number of probe observations j. We are interested in the effect of j on estimation error characteristics. Here, we focus on the estimation error variance, as this is an important feature in traffic state estimation methodologies, which apply a KF and were discussed in Introduction, that is, [3–5, 8].

Models to Explain Probe Data-Based Estimation Error Variance. We will compare two models that describe the influence of the number of probe observations j on the estimation error variance. The difference between these models relates to the expected correlation between estimation errors of individual probe observations.

The first model assumes that individual probe data-based estimates have the same error variance and the estimation errors are not correlated. In contrast, the second model does assume that there is a correlation between the individual estimation errors. The rationale behind the second model is as follows. The mean speed is dependent on the speeds of all vehicles within the considered area in space-time. In this case, the difference between an individual vehicle speed v_n of vehicle n and the mean speed u, that is, the estimation error, is expected to be correlated with the difference between v_m and u, that is, the estimation error based on v of vehicle m. For example, if we have two observations, the difference with respect to the mean (error) of the two observations has a correlation of minus one.

The assumptions discussed above are used to analytically derive the two models. For both models, it is assumed that v are Gaussian-distributed with mean u and variance σ_v^2. Given this distribution, we can say that the estimation error distribution based on single probe observations is a zero-mean Gaussian distribution with variance σ_v^2. This variance is constant for a given area in the space-time domain but can differ between areas, for example, due to traffic conditions. In the first model, each observation is seen as an independent observation with no relation to other observations; thus $E[\varepsilon_n\varepsilon_m] = 0$ for $n \neq m$, where ε_n denotes the difference between u and v_n, that is, the estimation error when estimating u based on the individual speed v of vehicle n. In the second model, we take into account the fact that the estimation errors based on different observations are (negatively) correlated. Each probe vehicle is given an equal probability of being observed, which means that the observation sample is taken from a random draw. Based on this assumption, we say that the expected covariance between two different probe observations, for example, of vehicles n and m, is constant; thus $E[\varepsilon_n\varepsilon_m] = c$ for $n \neq m$.

The difference between the two models is the assumption of the size of the sample for which we are interested in the mean. No correlation corresponds to the assumption that the observations are drawn from an infinite sample, while a finite sample yields results in a (negative) correlation between errors. Therefore, the two models will be denoted as Assumed Infinite Sample (AIS) and Assumed Finite Sample (AFS) models in the remainder of this article. Based on the

assumptions stated above, the analytical derivations result in the AIS model and AFS model.

$$E\left[\eta^2\right]_{\text{AIS}} = \frac{1}{j}\sigma_v^2, \tag{6}$$

$$E\left[\eta^2\right]_{\text{AFS}} = \frac{N-j}{j\,(N-1)}\sigma_v^2. \tag{7}$$

The derivations of these models are shown in the Appendix.

True Probe Data-Based Estimation Error Variance. The two models that are proposed to describe the probe data-based estimation error variance are dependent on the observed number of vehicles j, total number of vehicles N, and variance of individual vehicle speeds σ_v^2. To evaluate and compare the fit of the two models, we want to compare the model estimates with the true error variance. The true error variance is approximated using a Monte Carlo Experiment (MCE) and is therefore denoted as MCE. For each area in space and time, thus combination of i and p, we require a MCE-based error variance approximation for each potential value of j, that is, σ_{ipj}^2. To obtain σ_{ipj}^2, the following procedure is followed for each potential value of j:

(1) Random draw of j observations from $N(i, p)$ probes

(2) Calculating estimation error for each set using $\eta = u - 1/j \sum_{n=1}^{j} v_n$

(3) Repeating steps (1) and (2) 500 times

(4) Calculating error variance σ_{ipj}^2 of the set of 500 estimation errors

The MCE-based error variance approximation, that is, σ_{ipj}^2, is defined as the error variance observed in the data. Note that this is still an approximation based on the 500 sample sets that are randomly drawn. If the experiment is performed again with new random draws, the error variances may slightly differ. It is expected that the MCE becomes more accurate when the number of random draws increases. For this paper, we consider the provided MCE over 500 runs as the representative ground truth.

Comparison of the Models. We want to compare the error variance observed in the data, that is, the MCE approximation σ_{ipj}^2, with the two model-based estimates of the error variance, that is, $\widehat{\sigma}_{ipj}^2$. To get an initial insight into the differences in the fit of the two models, two discrete areas in the space-time domain are selected. These are the representative areas; thus combinations of i and p, for the clusters with the lowest and highest ρ, which are related to N, are selected. To obtain a representative area, we consider the cost function given by (5). The contribution of individual areas can be computed using $z^{(o)}$ and $\mu_{c(o)}$. Per cluster, we select the area, that is, observation o, which has the smallest squared difference with the cluster centroid, that is, $\mu_{c(o)}$, as the representative area.

To get an overall insight, for both models, the Mean Absolute Percentage Error (MAPE) is calculated. The following equation is used to calculate the MAPE:

$$\text{MAPE} = \frac{1}{P}\sum_{p=1}^{P}\frac{1}{I}\sum_{i=1}^{I}\frac{1}{N\,(i,p)}\sum_{j=1}^{N(i,p)}\frac{\left|\sigma_{ipj}^2 - \widehat{\sigma}_{ipj}^2\right|}{\sigma_{ipj}^2} \times 100\%. \tag{8}$$

We will consider the overall MAPE together with the MAPE for bins of the penetration rate with a size of 10%. By considering the MAPE for different bins, we are able to discuss the effect of the penetration rate on the accuracy of the two model-based estimates of the error variance. The choice for a bin size of 10% is not crucial for this discussion. For both the overall fit and the fit per penetration rate bin, the penetration rate is capped at 90%. This limit is imposed to overcome the problem that the Percentage Error (PE), and thus the MAPE, goes to infinity when $j = N$. In this case, σ_{ipj}^2 will be equal to zero.

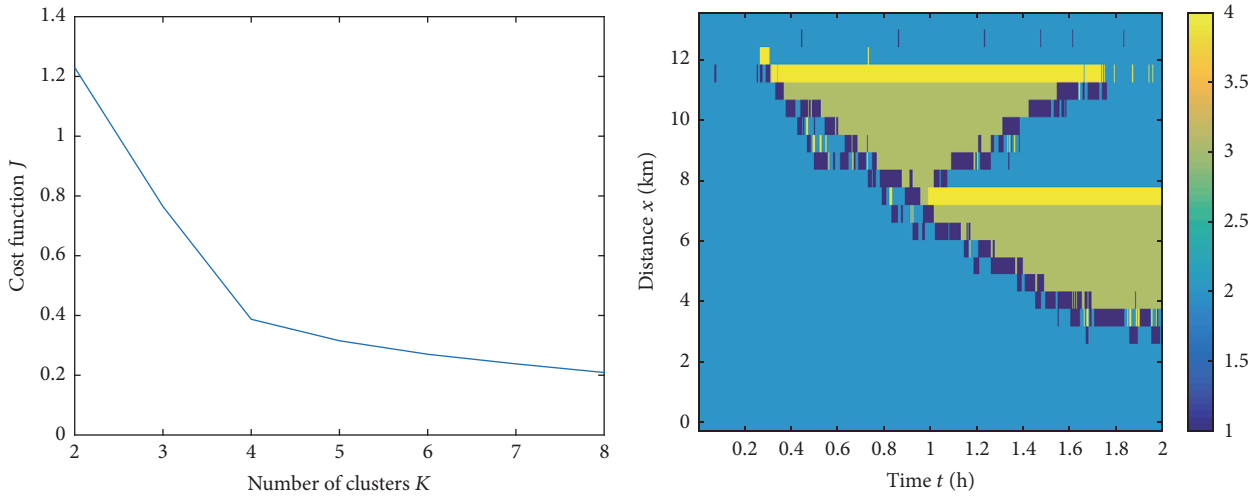
5. Results and Findings

The results and findings of the experiments presented in the previous section are discussed here. We first consider the loop-detector and probe data-based estimation error dependency on the traffic conditions. Next, we zoom in into the error variance of the probe data-based estimates for different numbers of probe observations and traffic conditions.

5.1. Dependency on Traffic Conditions. The relation between K and the cost is shown in Figure 2(a). The elbow is observed for $K = 4$, which is therefore selected as the optimal K. In Figure 2(b), the cluster classification is shown in the space-time domain. Furthermore, the cluster centroids are provided in Table 1. Based on these centroids and the comparison between Figures 1 and 2(b), the clusters are interpreted. For interpretation based on ρ and q, it is important to know that we are considering a three-lane road stretch, that is, $\lambda = 3$. In the cluster interpretation, $\partial v/\partial x$ plays an important role. A value close to zero corresponds to (near) homogeneous conditions, which can be either free flow or congestion based on the other variables. Throughout the remainder of this paper, we will refer to the free flow and congested space-time areas as homogeneous traffic conditions. A negative value means that the speed decreases when moving downstream. This means that vehicles are (or have to) decelerating, which can correspond to the inflow of a congested area or jam. Vice versa, a positive value can correspond to jam outflow.

The clustering of areas in space and time based on the traffic conditions allows us to evaluate the sensing data-based estimation error dependency on traffic conditions. For each cluster, the loop-detector and probe data-based estimation error characteristics are described. By means of the mean, variance, skewness, and kurtosis of the independent errors, we gain insight into the distribution of these errors. Furthermore, we test for normality, and thus whether the errors are Gaussian-distributed, using the JB test. The results are given in Table 1.

The first observation is that, for each error distribution, the null hypothesis of Gaussian-distributed error is rejected;



(a) Cost versus number of clusters

(b) Space-time plot clusters $K = 4$

FIGURE 2: Cost function dependent on the number of clusters and visualization of the cluster in the space-time domain for the optimal number of clusters $K = 4$.

TABLE 1: Interpretation of the four clusters based on the cluster centroids and the error distribution characteristics for probe and loop-detector data-based estimates per cluster.

		Clusters			
		1	2	3	4
		Jam inflow	Free flow	Congestion	Jam outflow
Cluster centroids	u [km/h]	58.5	96.0	23.8	47.7
	ρ [veh/km]	87.1	51.3	187.4	108.9
	q [veh/h]	4596	4878	4369	5049
	$\partial v/\partial x$ [km/(km h)]	−0.108	−0.001	−0.002	0.096
	Num. obs.	601	8424	1889	630
Probe data error char.	Mean	0.55	−0.04	0.01	0.05
	Variance	784.0	87.9	48.8	455.9
	Skewness	0.06	0.72	−1.48	−0.19
	Kurtosis	2.99	6.91	17.00	2.67
	JB test	1	1	1	1
	Num. obs.	31187	248461	213184	41982
Loop-detector data error char.	Mean	15.17	−0.08	−0.40	−31.23
	Variance	115.2	20.1	15.4	107.5
	Skewness	0.51	1.76	−0.17	1.03
	Kurtosis	3.43	19.50	8.75	5.25
	JB test	1	1	1	1
	Num. obs.	601	8424	1889	630

thus JB test = 1. The nonhomogeneous conditions, that is, jam inflow and outflow, yield challenges for data-based traffic state estimation in a discrete estimation mesh. Loop-detector data-based estimates can be biased in these conditions. This is caused by the combination of loop-detector location and change in speed over space. Following existing literature, we placed the loop-detectors at the downstream boundaries on the cells. If the speed decreases over space, as is the case for jam inflow, the mean speed is underestimated using loop-detector data. Vice versa, if the speed increases over space, as is the case for jam outflow, the mean speed is overestimated using loop-detector data. Challenges also arise for probe data-based estimation, but these are a result of other effects. Caused by variation in speed over space in addition to the variation in speed due to vehicle and driver heterogeneity, the total variance of the individual vehicle speeds increases with respect to homogeneous conditions. This yields an increased probe data-based estimation uncertainty. For the homogeneous traffic conditions, we make a distinction between free flow and congested conditions. For probe data-based estimates, the uncertainty is larger in free flow with respect to congestion. This is caused by the larger

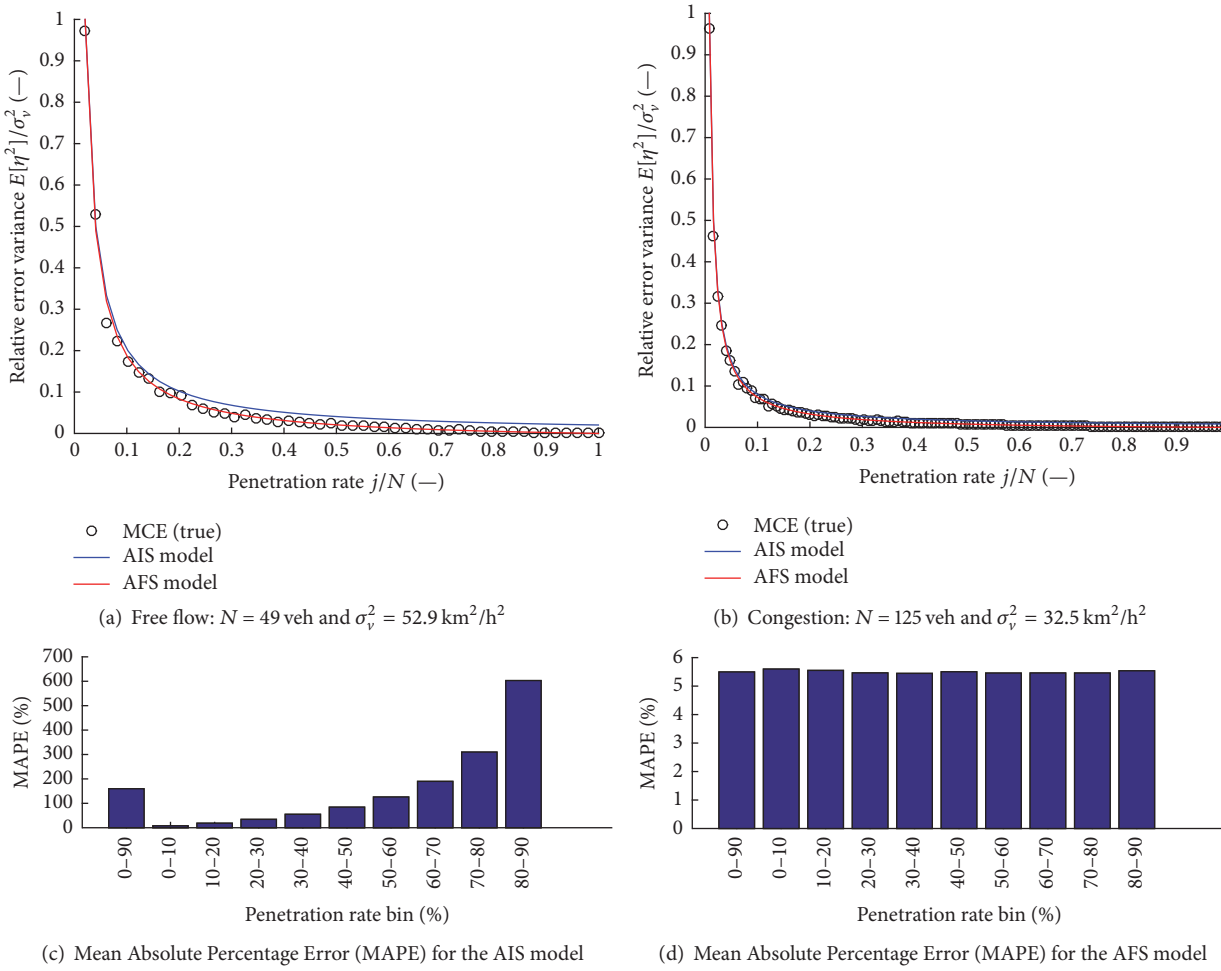


Figure 3: Visualization of the AIS model and AFS model fit and shape for an individual representative of free flow and congested area in the space-time domain. Furthermore, the total model fit, in terms of MAPE, is given for both models.

speed variation due to vehicle and driver heterogeneity in free flow conditions. A similar relation is visible for loop-detector data-based estimates; however, here the relative difference between free flow and congested conditions is smaller.

A direct comparison between the accuracies of loop-detector and probe data-based estimations in free flow and congested conditions should not be made based on the results depicted in Table 1. These probe data-based estimates are based on individual vehicle speed data of a single probe, while we may observe more than one probe. For this reason, we will evaluate how the estimation error is affected when more than one probe observation is available below.

5.2. Dependency on Observed Fraction of the Traffic Flow. In this section, we want to evaluate the performances of the two proposed models, that is, the AIS model and AFS model, to describe the probe data-based estimation error variance. Figures 3(a) and 3(b) show representative areas in the space-time domain for free flow and congested conditions, respectively: $\{i, p\} = \{18, 76\}$ and $\{i, p\} = \{20, 205\}$. Keep in mind that these figures relate to a single area in the space-time domain. Considering a single area can help to understand the difference between the two models and how these relate to the true estimation error variance. In line with the findings from the previous section, the estimation error variance for an individual observation is larger in free flow than in congestion. Furthermore, we can see that N is smaller in free flow than in congestion. In the figures, it can be observed that the MCE sometimes (slightly) increases with the number of observations, while this is not expected for the true error variance. This can be explained by the limited set of random draws, that is, 500, used to obtain the MCE-based error variance.

The AIS model has the same shape based on j but differs based on penetration rate. The relative accuracy of the AIS model decreases when j increases. The reason may be that the AIS model fails to capture the important (negative) correlation between estimation errors. The AFS model does include these effects, resulting in a higher accuracy. It is clear that the AFS model is dependent on the penetration rate j/N, as for $j = N$ the estimated error variance is zero. Furthermore, the effect of having more observations, thus when j increases, becomes also clear when comparing the two figures. In Figure 3(b), the AFS model decreases more rapidly

as a function of j/N than in Figure 3(a). This can be explained by the fact that in the congested case j is larger than in the free flow case for a given penetration rate j/N. As the MCE-based error variance follows this line, we may say that this effect seems to explain the true error variance. If N increases, the AIS model seems to become more accurate. This makes sense as the AFS model approaches the AIS model for larger values of N and they are the same if $N \to \infty$. If one considers the AIS model for a given penetration rate, the probe data-based estimate thus becomes more accurate in terms of the point estimate and the error variance description.

Up to now, we considered two observations to gain an insight into the two models and MCE-based error variance. However, we are interested in the overall fit and whether the discussed characteristics hold for the entire sample. The MAPE is considered to describe the overall and penetration rate bin-specific fit. These are depicted in Figures 3(c) and 3(d) for, respectively, AIS model and AFS model.

Figures 3(c) and 3(d) show that the AFS model has a much better fit than the AIS model. The AIS model especially has problems describing the estimation error variance at higher penetration rates. At these penetration rates, the negative correlation between estimation errors of individual probe observation becomes a more important factor. As the AIS model misses this factor while it is included by the AFS model, the AIS model is outperformed by the AFS model. The MAPE of the AFS model is relatively constant over the penetration rate bins.

6. Conclusions and Discussion

In this research, the traffic sensing data-based estimation error characteristics are evaluated by means of experiments. We focus on two combinations of a desired estimation output: type of traffic sensing data and estimation approach. Here, we estimate the mean speed in a discrete space-time mesh based on loop-detector data and probe speed data.

In the experiments, we observe a relation between the estimation error characteristics and traffic conditions and sensing data characteristics. We find that the extent to which traffic conditions are nonhomogeneous negatively influences the estimation error characteristics. For instance, for both data-types, we observe larger error variances in jam inflow and outflow than in free flow and congested conditions. Also, our loop-detector data-based estimates are biased in nonhomogeneous traffic conditions. Furthermore, we show that it is valuable to take into account the correlation between estimation errors to describe the probe data-based estimation error variance based on multiple probe observations.

Our experiments are conducted with the microscopic simulation program FOSIM. The results are thus influenced by the use of this simulation environment. However, we expect that the findings in this paper are also applicable to real-world applications as our explanations relate to real-world situations. As an example, we can consider the effects of the level on nonhomogeneity of the traffic conditions on the estimation errors in terms of the mean and variance. We try to explain these effects based on the nonhomogeneity itself. As we observe nonhomogeneous traffic conditions both in simulated environments and in the real world, we expect that this explanation is also valid in real-world applications. Furthermore, we think that the notion that the estimation errors can be explained based on certain features is of general importance. Even if one deals with different circumstances, for example, other data-types or desired estimation output, it is valuable to consider this notion and try to explain the estimation errors.

Knowledge related to estimation errors is valuable when combining (fusing) different types of information or when weighing different alternatives. Examples of the latter are control decisions within a dynamic traffic management system or routing decisions of road users. This knowledge can lead to improved performance for these different applications without requiring additional (expensive) sensing data. Implementing knowledge related to estimation errors only marginally adds to the computation cost and does not require development of new, complex methodologies. However, before this can be put into practice, research is required to expose the added value for performance of applications in which the estimates are used as an input, for example, traffic state estimation or control.

Appendix

Derivation the AIS Model and AFS Model

The individual vehicle speeds within a cell and period are assumed to be Gaussian-distributed with mean u and variance σ_v^2:

$$v_n = u + \varepsilon_n, \quad \varepsilon_n \sim N\left(0, \sigma_v^2\right). \tag{A.1}$$

The mean speed u is equal to the estimated mean speed $\widehat{u}$ plus the estimation error η:

$$u = \widehat{u} + \eta. \tag{A.2}$$

We can estimate u based on j number of individual vehicle speed observations:

$$\widehat{u} = \frac{1}{j}\sum_{n=1}^{j} v_n = u + \frac{1}{j}\sum_{n=1}^{j} \varepsilon_n. \tag{A.3}$$

The estimation error variance, that is, $E[\eta^2]$, is given by $E[(\widehat{u} - E[\widehat{u}])^2]$; thus

$$E\left[\eta^2\right] = E\left[\left(\widehat{u} - E\left[\widehat{u}\right]\right)^2\right] \tag{A.4}$$

$$= E\left[\left(u - u + \frac{1}{j}\sum_{n=1}^{j} \varepsilon_n\right)^2\right] \tag{A.5}$$

$$= E\left[\left(\frac{1}{j}\sum_{n=1}^{j} \varepsilon_n\right)^2\right]. \tag{A.6}$$

Following the reasoning discussed in Section 4.3.2, we can make different assumptions related to (co)variance of the

individual differences (errors) between v_n and u. This results in the two models, that is, Assumed Infinite Sample (AIS) and Assumed Finite Sample (AFS) models. Based on (A.1), for both models, $E[\varepsilon_n \varepsilon_n] = \sigma_v^2$. However, the AIS model assumes that $E[\varepsilon_n \varepsilon_m] = 0$ for $n \neq m$, while AFS model assumes that $E[\varepsilon_n \varepsilon_m] = c$ for $n \neq m$.

Continuing from (A.6), for the AIS model, the error variance becomes

$$E\left[\eta^2\right]_{\text{AIS}} = \frac{1}{j^2} E\left[\sum_{n=1}^{j} \varepsilon_n^2\right] = \frac{1}{j^2} j\sigma_v^2 = \frac{1}{j}\sigma_v^2, \tag{A.7}$$

which thus yields (6).

Continuing from (A.6), for the AFS model, the error variance becomes

$$E\left[\eta^2\right]_{\text{AFS}} = \frac{1}{j^2} E\left[\left(\sum_{n=1}^{j} \varepsilon_n\right)^2\right]. \tag{A.8}$$

In contrast to the AIS model, $E[(\sum_{n=1}^{j} \varepsilon_n)^2]$ does not simplify to $E[\sum_{n=1}^{j} \varepsilon_n^2]$. Instead, the $(j^2 - j)$ number of terms of $E[\varepsilon_n \varepsilon_m]$ where $n \neq m$is still of importance. Therefore, we obtain

$$E\left[\eta^2\right]_{\text{AFS}} = \frac{1}{j^2}\left(j\sigma_v^2 + \left(j^2 - j\right)c\right) = \frac{1}{j}\sigma_v^2 + \frac{j-1}{j}c. \tag{A.9}$$

The last step is to find c. For this purpose, we say that the error variance is equal to zero when we observe all vehicles; that is, $j = N$. This yields

$$\begin{aligned} \frac{1}{N}\sigma_v^2 + \frac{N-1}{N}c &= 0, \\ (N-1)\,c &= -\sigma_v^2, \\ c &= -\frac{1}{N-1}\sigma_v^2. \end{aligned} \tag{A.10}$$

Next, we combine the above relations:

$$\begin{aligned} E\left[\eta^2\right]_{\text{AFS}} &= \frac{1}{j}\sigma_v^2 - \frac{j-1}{j}\frac{1}{N-1}\sigma_v^2 \\ &= \frac{N-1}{j\,(N-1)}\sigma_v^2 - \frac{j-1}{j\,(N-1)}\sigma_v^2 \\ &= \frac{N-j}{j\,(N-1)}\sigma_v^2, \end{aligned} \tag{A.11}$$

which yields (7).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors acknowledge the Netherlands Organisation for Scientific Research (NWO) for providing the funding used to perform this research.

References

[1] A. D. May, *Traffic Flow Fundamentals*, Prentice Hall, 1990.

[2] S. Hoogendoorn, R. Landman, J. Van Kooten, and M. Schreuder, "Integrated Network Management Amsterdam: Control approach and test results," in *Proceedings of the 2013 16th International IEEE Conference on Intelligent Transportation Systems: Intelligent Transportation Systems for All Modes, ITSC 2013*, pp. 474–479, The Hague, Netherlands, October 2013.

[3] Y. Wang and M. Papageorgiou, "Real-time freeway traffic state estimation based on extended Kalman filter: a general approach," *Transportation Research Part B: Methodological*, vol. 39, no. 2, pp. 141–167, 2005.

[4] C. Nanthawichit, T. Nakatsuji, and H. Suzuki, "Application of probe-vehicle data for real time traffic state estimation and short term travel time prediction on a freeway," *Transportation Research Record*, vol. 58900, no. 1855, pp. 49–59, 2003.

[5] J. C. Herrera and A. M. Bayen, "Incorporation of Lagrangian measurements in freeway traffic state estimation," *Transportation Research Part B: Methodological*, vol. 44, no. 4, pp. 460–481, 2010.

[6] M. J. Lighthill and G. B. Whitham, "On kinematic waves. II. A theory of traffic flow on long crowded roads," *Proceedings of the Royal Society A Mathematical, Physical and Engineering Sciences*, vol. 229, pp. 317–345, 1955.

[7] P. I. Richards, "Shock waves on the highway," *Operations Research*, vol. 4, no. 1, pp. 42–51, 1956.

[8] 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*, vol. 13, no. 1, pp. 385–394, 2012.

[9] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, pp. 35–45, 1960.

[10] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[11] D. Woodard, G. Nogin, P. Koch, D. Racz, M. Goldszmidt, and E. Horvitz, "Predicting travel time reliability using mobile phone GPS data," *Transportation Research Part C: Emerging Technologies*, vol. 75, pp. 30–44, 2017.

[12] C. F. Daganzo, "The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transportation Research Part B: Methodological*, vol. 28, no. 4, pp. 269–287, 1994.

[13] C. F. Daganzo, "The cell transmission model, part II: network traffic," *Transportation Research Part B: Methodological*, vol. 29, no. 2, pp. 79–93, 1995.

[14] V. Knoop and S. Hoogendoorn, "Empirics of a generalized macroscopic fundamental diagram for urban freeways," *Transportation Research Record*, no. 2391, pp. 133–141, 2013.

[15] T. Bellemans, B. De Schutter, and B. De Moor, "Models for traffic control," *Journal A*, vol. 430, no. 3-4, pp. 13–22, 2002.

[16] L. C. Edie, "Discussion of traffic stream measurements and definitions," in *Proceedings of the 2nd Int. Symp. On the Theory of Traffic Flow*, OECD, Paris, France, 1965.

[17] L. H. Immers and S. Looghe, "Traffic Flow Theory," Tech. Rep., Katholieke Universiteit Leuven, 2002.

[18] T. Seo, *Traffic Estimation with Vehicles Observing Other Vehicles [Ph.D. thesis]*, Tokyo Institute of Technology, 2015.

[19] J. B. Ramsey, H. J. Newton, and J. L. Harvill, "Moments and the shape of histograms," in *The Elements of Statistics: With Applications to Economics and the Social Sciences*, chapter 4, pp. 77–119, Duxbury/Thomson Learning, 2004.

[20] C. M. Jarque and A. K. Bera, "A test for normality of observations and regression residuals," *International Statistical Review*, vol. 55, no. 2, pp. 163–172, 1987.

[21] T. Dijker and P. Knoppers, "FOSIM 5.1 Gebruikershandleiding (Users Manual)," Tech. Rep., Technische Universiteit Delft, 2006.

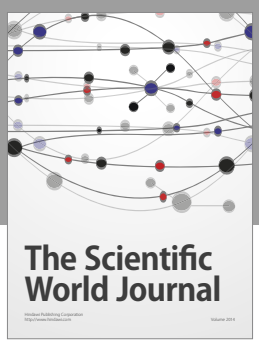
[22] M. M. Minderhoud and K. Kirwan, "Validatie FOSIM voor asymmetrische weekvakken - CAPWEEF fase 1," Tech. Rep., Laboratorium voor Verkeerskunde, Faculteit Civiele Techniek en Geowetenschappen, Technische Universiteit Delft, Delft, Netherlands, 2001.

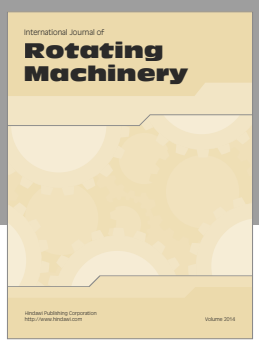
[23] N. Henkens, W. Mieras, and D. Bonnema, "Validatie FOSIM," Tech. Rep., Sweco, De Bilt, Netherlands, 2017.

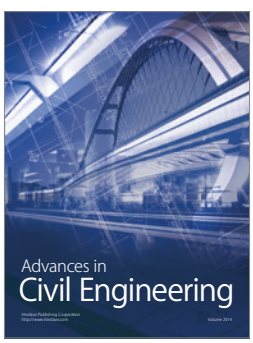
[24] J. Colyar and J. Halkias, 2007, https://www.fhwa.dot.gov/publications/research/operations/07030/index.cfm.

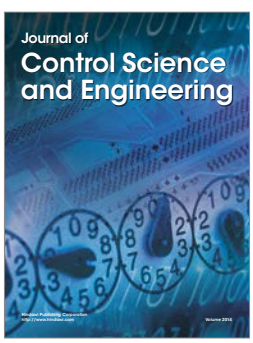
[25] S. Smulders, "Control of freeway traffic flow by variable speed signs," *Transportation Research Part B: Methodological*, vol. 24, no. 2, pp. 111–132, 1990.

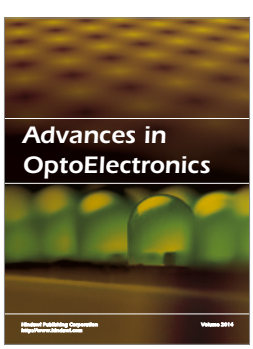
[26] P.-N. Tan, M. Steinbach, and V. Kumar, "Cluster analysis: basic concepts and algorithms," in *Introduction to Data Mining*, chapter 8, pp. 487–568, 2006.

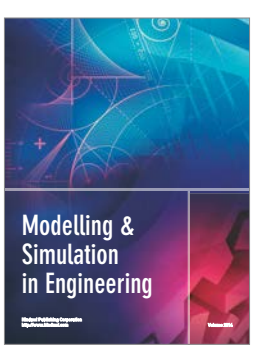


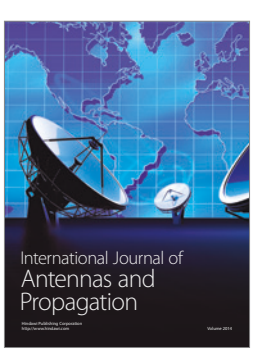


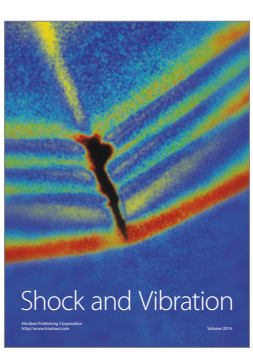


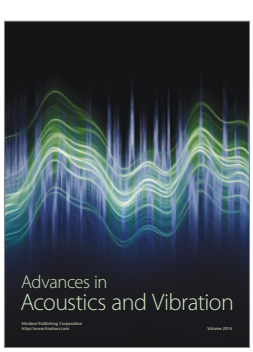












Hindawi

Submit your manuscripts at
https://www.hindawi.com