Improved Kalman Filtering Approach for Estimating Origin-Destination Matrices for Freeway Corridors

NANNE J. VAN DER ZIJPP AND RUDI HAMERSLAG

The estimation of origin-destination (OD) matrices for freeway corridors by using inner-link induction loop data is examined. A trip generation model is used, and various parameter optimization and statistics-based methods are examined to estimate the split parameters in the model. A Kalman-based method that uses the model-predicted link-flow variances and covariances while processing the measurements is described. A simple but effective solution to the problem of initializing the Kalman filter and imposing the natural constraints to the estimates is presented. The resulting method is tested on both simulated and observed data and is compared with other methods such as least squares and constrained optimization, showing that the Kalman-based method leads to the best results.

Vehicle movement estimates are generally summarized in origin-destination (OD) tables. These tables contain the number of trips for each combination of origin and destination. For a freeway system origins correspond with on-ramps (entrances), whereas destinations relate to off-ramps (exits). Dynamically updated OD tables are required for various strategies aimed at optimal usage of existing freeway capacity. Examples of such strategies are ramp metering, route guidance, and incident management. Often induction loop data are the only continuously updated source of information, producing the number of observed vehicles per time slice. Induction loops generate an abundance of traffic counts. To be able to analytically calculate an OD table within a time slice, however, additional techniques are necessary. A first example of such a technique is the use of a traffic model that defines explicit relationships between OD flows. A second example is the use of an a priori trip table. The distance to this a priori trip table, according to some criterion, is minimized by using traffic counts as a boundary condition. Examples of these approaches can be found in Cascetta and Nguyen (*1*), Hamerslag and Immers (*2*), Bell (*3*), Hendrickson and McNeil (*4*), and van Zuylen and Willumsen (*5*).

Although the use of such techniques when applied to aggregated data sets can be well defended, it is questionable whether the inherent assumptions of the above-mentioned techniques are valid when applied to subnetworks like intersections or freeway corridors. First, these subnetworks contain neither real origins nor real destinations. Second, because of low aggregation levels, stochastic influences are likely to be dominant.

Therefore in this paper a class of OD estimators that works with a weaker assumption, the assumption of constant split ratios, is studied. According to this assumption for each entrance the *fractions* of traffic destined for a certain exit can be assumed to be changing slowly or even remain constant. This assumption changes the underspecific problem into an overspecified problem.

The split ratio approach was first introduced by Cremer and Keller (*6*), who used a recursive formula to estimate the unknown split proportions. Since then various techniques have been used to estimate the split proportions. First, the correlation procedure was proposed by Cremer (*7*). This procedure is equivalent to the least-squares method. Later the method was improved by Cremer and Keller (*8*), who used constrained optimization. Simultaneously Kalman filtering was applied to this problem by both Cremer and Keller (*8*) and Nihan and Davis (*9*). Finally maximum likelihood approaches have been employed by Nihan and Davis (*10*) and Bell et al. (*11*). A combination of split ratio and modeling approaches can be found in Keller and Ploss (*12*), whereas Bell (*13*) added the problem of platoon dispersion.

The problem statement used in this paper will show many similarities to the problem statements used in the above-mentioned work. Three new elements are added, however. First, the measurement vector contains not only exit volumes but can also contain inner-link volumes. Second, the split parameters are interpreted as split probabilities rather than fixed fractions of entering volumes in a trip generation model. The third addition is the incorporation of a time shift in the problem. Entrance volumes and measurements from all locations are processed simultaneously. Therefore each measurement must be processed with a delay for all measurements to refer to one set of split parameters.

The main problem with the processing of inner-link volumes is the strong measurement dependency due to redundancy. To adequately describe the properties of the system and its measurements the trip generation model presented by van der Zijpp and Hamerslag (*14*) is used. This model distinguishes between split probabilities and split proportions, an idea already used by Davis and Nihan (*15*) and Davis (*16*) for static OD estimation. The trip generation model describes not only how split probabilities change through time but also the choice of destination as a random choice process and which noise is involved when monitoring entering traffic and inner-link volumes.

For prediction purposes the split probabilities have more significance than the split proportions. The OD estimation problem is therefore converted into the estimation of the split *probabilities* in the trip generation model. For this purpose least squares, constrained optimization, maximum likelihood, and Kalman filtering have been considered. Each method is described in terms of the variables used in the problem statement, and when necessary computational aspects are discussed. From these methods only the Kalman filter approach and the maximum likelihood approach allow the specification of dependency between measurements. From

Faculty of Civil Engineering, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, The Netherlands.

these two the Kalman filter has been selected because it is the only method for which tractable expressions for the dependent measurement case could be derived.

Several problems, however, hinder the straightforward application of a Kalman filter to the problem of estimating the split probabilities from induction loop data. First, since the Kalman filter is a recursive method, a set of initial conditions needs to be available. Second, the measurement properties need to be defined, since noise occurs because of differences between split probabilities and split proportions and inaccuracies in induction loop observations. Finally there are several equality and inequality constraints that apply to the split probabilities. For each entrance split probabilities must not only add up to 1 but each individual split probability must also be nonnegative and less than 1. Depending on how these problems are solved one can expect a Kalman filter to do better or worse. The results presented by Cremer and Keller (*8*), for example, suggest that constrained optimization gives better results than Kalman filtering at the cost of high computation times.

The section Improved Kalman Filtering Approach describes a solution to each of these problems, resulting in an improved Kalman-based method. The method is tested against constrained optimization and least squares by using both simulated and empirical data. The test results are included in the last section.

PROBLEM STATEMENT

For the problems treated in this paper route choice is supposed to play no role, although this is not really a constraint of the methods under consideration; see for example Davis (*16*). All implemented methods take nonzero travel times into account. The problem of determining the delays is treated at the end of this section. For simplicity of notation the travel times are not mentioned in the equations.

Notation

The definitions of the terms used in the equations are as follows:

- $q(t)$ = vector of length m whose elements $q_i(t)$ are the observed volumes at entrance i that are processed during interval t.
- $y(t)$ = vector of length p whose elements $y_h(t)$ are the counted volumes at location h that are processed during interval t.
- $B(t)$ = $m \times n$ matrix whose elements $b_{ij}(t)$ are the proportion of trips leaving i destined for j. Let $b_{i:}(t)$ represent row i of $B(t)$, that is, the split parameters associated with entrance i. Then $b(t) = [b'_{1:}(t)\ b'_{2:}(t) \ldots b'_{m:}(t)]'$ is defined as a vector of length $m \times n$ that contains the elements of $B(t)$ row by row.
- $F(t)$ = $m \times n$ matrix whose elements $f_{ij}(t)$ give the flow from i to j. Let $f_{i:}(t)$ define row i of $F(t)$. Then $f(t) = [f'_{1:}(t)\ f'_{2:}(t) \ldots f'_{m:}(t)]'$ is a vector of length $m \times n$ that contains the elements of $F(t)$ row by row.

Trip Generation Model

The problem is to estimate the unknown parameters $B(t)$. Referring to van der Zijpp and Hamerslag (*14*), we argue that $b_{ij}(t)$ should be considered the *probability* that a vehicle will leave the network at exit j given the fact that it originated from entrance i. Such a probability does not really exist, but for our purposes drivers selecting randomly their destination upon entrance onto the network is an acceptable model of the system.

Working with split probabilities rather than split proportions has two advantages. First, the assumption that $b(t)$ is a slowly moving process can be better defended here since the randomly triggered difference between the split proportions and split probabilities is eliminated. Second, some useful properties of the measurements can be derived, such as variances and covariances given a set of split probabilities. Hence from now on we use the following definition for the split parameters:

$$b_{ij} \triangleq P[\text{exit at } j | \text{enter at } i] \tag{1}$$

By definition the following constraints apply to the split parameters:

$$0 \leq b_{ij}(t) \leq 1 \qquad i = 1 \ldots m,\ j = 1 \ldots n \tag{2}$$

$$\sum_{j=1}^{n} b_{ij}(t) = 1 \qquad i = 1 \ldots m \tag{3}$$

Like in Nihan and Davis (*9*) we refer to these constraints as the natural constraints. The split parameters are assumed to vary slowly over time, driven by a zero mean drift parameter $w(t)$:

$$b(t) = b(t - 1) + w(t) \tag{4}$$

Another aspect we would like to consider is that all volumes are observed with noise because of inaccuracy of the induction loop observations. Introduce $q^*(t)$ and $y^*(t)$ as the vectors of real input- and inner-link volumes, whereas $q(t)$ and $y(t)$ are the measured values. All noise components are considered to be independent and zero mean and have variances σ_q^2 or σ_y^2. Therefore,

$$q(t) = q^*(t) + r(t)$$

$$E[r(t)] = 0, \quad E[r(t)r(t)'] = \sigma_q^2 I \tag{5}$$

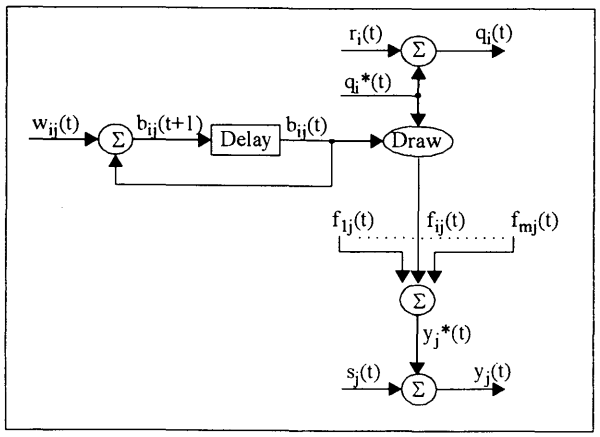
and

$$y(t) = y^*(t) + s(t)$$

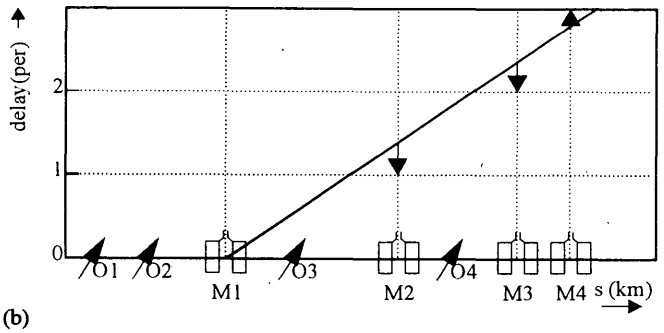
$$E[s(t)] = 0, \quad E[s(t)s(t)'] = \sigma_y^2 I \tag{6}$$

Often the on-ramps are not monitored directly and one must calculate these entrance volumes by taking the difference of two consecutive inner-link volumes. Experiments have shown that neglecting noise in the entrance volume vector seriously affects the quality of the estimate, especially when the Kalman filter was applied. One reason for this is that the Kalman filter uses the entrance volume vector as a boundary condition. Therefore errors in the entrance volumes are subscribed to measurement noise. This causes a strong dependency among the elements of this noise vector.

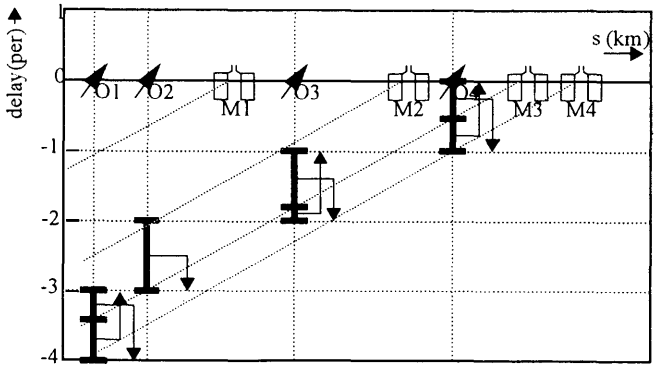
The above assumptions define the trip generation model that was presented earlier by van der Zijpp and Hamerslag (*14*). This model is summarized in Figure 1(*a*), which shows a system in which $b(t + 1)$ is obtained from $b(t)$ and drift variable $w(t)$, which



(a)



(b)



(c)

FIGURE 1 Modeling assumptions: (*a*) trip generation model, (*b*) selecting simultaneously processed measurements, and (*c*) matching entrance volumes with measurements.

will be considered a random input. The variables $b(t)$ are used as probabilities in a drawing process. For each entrance $q_i^*(t)$ experiments are done. The observed entrance volume vector $q(t)$ is obtained by taking the sum of $q^*(t)$ and $r(t)$, as described in Equation 5. The results of the drawing processes are merged into a link volume vector $y^*(t)$, to which (see Equation 6) a noise vector $s(t)$ must be added to obtain the measured values $y(t)$.

Measurements are available as traffic counts, which are obtained from induction loops. By definition these counts are linear combinations of flows. Since the route choice issue is neglected here, an OD flow is either totally or not at all contained in a measurement. Therefore,

$$y^*(t) = U'f(t) \tag{7}$$

with U denoting an $mn \times p$ matrix whose elements can either be one or zero, indicating that a flow does or does not contribute to the measurement. Note that this matrix does not depend on the time period. The transpose was used solely to keep conformity with literature. By using Equations 1 and 5 the following approximation for the flows can be derived:

$$f_{ij}(t) \approx q_i(t)b_{ij}(t) \tag{8}$$

Substituting this approximation in Equation 7 and combining this with Equation 6 allows us to calculate an $mn \times p$ matrix $H(t)$ with

$$y(t) = H'(t)b(t) + v(t) \tag{9}$$

This equation will later be referred to as the *measurement equation*. The vector $v(t)$ accounts for all measurement errors and the effects of the random selection process described in the trip generation model. The properties of this measurement error are discussed in the section Improved Kalman Filtering Approach.

Calculating Correct Time Delay

The measurements are processed with a time delay to let all measurements refer to the same set of split parameters and entrance volumes. However the time axis is divided into intervals, and the average travel times between entrances and measurement locations generally do not match the length of the intervals. To minimize errors a two-step process was followed.

The first step involves the selection of the measurements that will be processed simultaneously. To apply the natural constraint (Equation 3) to the estimate of $b(t)$, for each entrance this estimate must represent the splits during only one interval. To optimally fulfill this condition the relative travel times between the measurement locations are calculated and rounded to an integer number of intervals. This is illustrated in Figure 1(*b*), which shows the delay (in periods) as a function of the distance s (in kilometers). The locations of origins O1 through O4 and measurement locations M1 through M4 are indicated on the x-axis. The gradient of the line corresponds to the average speed. In the experiments described at the end of this paper this average could be derived directly from the input data, because measurements are carried out with double induction loops that monitor both intensity and speed.

The second step involves the selection of the corresponding entrance volumes. Since the average travel times do not exactly equal an integer number of periods, entrance volumes from at least two periods are assumed to contribute to a measurement. Therefore a weighted sum of the entrance volumes should be substituted in Equation 8. The weight factors can be determined from Figure 1(*c*). They correspond to the length of the vertical intervals in Figure 1(*c*). The arrows join the weight factors with the corresponding delay. By taking a weighted sum of two entrance volumes the optimal approximation of the entering volume during a certain period is obtained. However the entering traffic cannot be

assumed to be evenly spread over time, and the travel times are not exactly known. Therefore it is inevitable that an error is introduced in the entering volume observation. For this reason the noise variable $r(t)$ was included in the trip generation model.

ESTIMATION OF SPLIT PARAMETERS

When we use the trip generation model described in the previous section, the problem of estimating the OD matrix is reduced to estimating the split parameters in the trip generation model. In this section five existing methods are described. Since these methods have been described in other contributions, this paper provides only a brief summary showing how the methods can be applied to problems in which the measurements contain inner-link volumes instead of exiting volumes only. The methods being considered here are the least-squares method, inequality-constrained least-squares method, constrained optimization, maximum likelihood, and Kalman filtering method. The first three methods can be classified as parameter optimization methods, whereas the other two are statistics-based methods.

Least-Squares Method

The least-squares method is aimed at solving the following problem:

$$\min_{\hat{b}(t)} \sum_{k=1}^{t} \|y(k) - H'(k)\hat{b}(t)\|^2 \tag{10}$$

By expanding this expression and setting the derivatives to $\hat{b}(t)$ to zero, the least-squares estimate can be calculated by

$$\hat{b}(t) = \left[\sum_{k=1}^{t} H(k)H'(k)\right]^{-1}\left[\sum_{k=1}^{t} H(k)y(k)\right] \tag{11}$$

A unique solution is guaranteed if mn independent columns can be found in the matrices $H(1) \ldots H(t)$. From Equation 11 it can be seen that it is possible to employ the least-squares method by using a constant amount of storage space by the following algorithm:

$$\hat{b}(t) = HH_{tot}^{-1}(t)HY_{tot}(t)$$

$$HH_{tot}(t) = HH_{tot}(t-1) + H(t)H'(t)$$

$$HY_{tot}(t) = HY_{tot}(t-1) + H(t)y(t) \tag{12}$$

By introducing a discounting factor the method can be adapted to track a time-varying $b(t)$. This transforms the problem into

$$\min_{\hat{b}(t)} \sum_{k=1}^{t} \lambda^{t-k}\|y(k) - H'(k)\hat{b}(t)\|^2 \tag{13}$$

Again putting the derivatives to zero gives a minimum:

$$\hat{b}(t) = \left[\sum_{k=1}^{t} \lambda^{t-k}H(k)H'(k)\right]^{-1}\left[\sum_{k=1}^{t} \lambda^{t-k}H(k)y(k)\right] \tag{14}$$

This gives rise to the following algorithm:

$$\hat{b}(t) = HH_{tot}^{-1}(t)HY_{tot}(t)$$

$$HH_{tot}(t) = \lambda HH_{tot}(t-1) + H(t)H'(t)$$

$$HY_{tot}(t) = \lambda HY_{tot}(t-1) + H(t)y(t) \tag{15}$$

Inequality-Constrained Least-Squares Method

Equation 15 does not guarantee that the natural inequality constraint (Equation 2) is met. Imposing this condition would therefore improve the solution. On the other hand this would convert the problem from an unconstrained minimization into an inequality-constrained minimization problem:

$$\min_{\hat{b}(t)} \sum_{k=1}^{t} \|y(k) - H'(k)\hat{b}(t)\|^2 \tag{16}$$

subject to

$$\hat{b}_l(t) \geq 0 \qquad l = 1 \ldots mn \tag{17}$$

This problem consumes much more computation time than the unconstrained problem. When solved by an interior steepest-descent method the computation times tend to be high because of the ill-conditioned matrix HH_{tot}. Not only does this hinder the testing of the method (an average test run would take 2 hr for the cases described in this paper) but also in case of real-time applications the duration of the computation could easily exceed the time available.

Therefore a less time-consuming algorithm is needed. For the time being the best results are obtained with an iterative algorithm that employs conjugate search directions that are projected on the feasible region when necessary. Moreover the searches are restricted to the feasible region, and the search direction is reset to steepest descent after each truncated search or change of active constraints. Calculation times are approximately 10 times longer than those by the straightforward matrix inversion method that could be used for the nonconstrained case. This suffices for problems of the size studied in this paper.

Constrained Optimization

If both the inequality and equality constraints (Equations 2 and 3) are imposed, an even better solution should be obtained. The satisfaction of the equality constraint (Equation 3) can be guaranteed by substituting the following in Equation 10:

$$\hat{b}(t) = b^0 + Gb^1(t) \tag{18}$$

with b^0 satisfying the equality constraints in Equation 3 and G being a $mn \times m(n-1)$ matrix chosen in such a way that $Gb^1(t)$ does not disturb the satisfaction of the equality constraints for all $b^1(t)$. Although many combinations of b^0 and G satisfy the nec-

essary conditions, for practical reasons we use

$$b^0 = \begin{bmatrix} 0 \\ \cdots \\ 0 \\ 1 \\ \cdots \\ \cdots \\ \cdots \\ 0 \\ \cdots \\ 0 \\ 1 \end{bmatrix}, \quad G = \begin{bmatrix} & I & & & \\ -1/n & -1/n \cdots & -1/n & & \\ & & \cdots & & \\ & & & \cdots & \\ & & & & \cdots \\ & & & & I \\ & & & & -1/n \quad -1/n \cdots \quad -1/n \end{bmatrix} \tag{19}$$

This substitution transforms the problem in Equation 10 into

$$\min_{b^1(t)} \sum_{k=1}^{t} \|y(k) - H'(k)(b^0 + Gb^1(t))\|^2 \tag{20}$$

subject to

$$b^1_l(t) \geq 0 \qquad l = 1 \ldots m(n-1)$$

$$\sum_{j=1}^{n-1} b^1_{(i-1)(n-1)+j}(t) \leq 1 \qquad i = 1 \ldots m \tag{21}$$

Solving this inequality-constrained problem and substituting the resulting $b^1(t)$ in Equation 18 gives a solution for $\hat{b}(t)$ that satisfies all required conditions. For solving the problem in Equation 20 the algorithms of the inequality-constrained problem can be used, although the projection of the search direction on the feasible region requires more computation time because of the nonorthogonal inequality constraints.

Maximum Likelihood

The previous methods can all be considered parameter optimization methods. They are designed to minimize the distance between measured and predicted values. Apart from these methods we distinguish the statistics-based methods. These methods are defined in terms of the probability distributions related to the unknown parameters $b(t)$. The most common statistics-based method is the maximum likelihood (ML) technique. When applied to the problem of determining the split parameters in the trip generation model the ML solution would be defined by

$$\max_{\hat{b}(t)} P[y(1) \ldots y(t)|\hat{b}(t)] \tag{22}$$

Calculation of the ML solution normally requires the derivation of a probability distribution from the system shown in Figure 1(*a*), which is not tractable. Nihan and Davis (*10*) presented an ML approach that did not require this derivation by using the EM algorithm proposed by Dempster et al. (*17*). However this was done for the simplified system in which $b(t)$ was constant rather than slowly varying and in which no noise on the entrance volume observations was present. Moreover the resulting algorithm was nonrecursive.

Another ML approach has been presented by Bell et al. (*11*). This approach is fully disaggregate but is computationally too demanding to be useful in practice. So although ML estimators have desirable properties no ML estimator that suits our needs is available.

Kalman Filtering

Another statistics-based estimation technique is the Kalman filter. The Kalman filter is a widely applied method for parameter estimation in dynamic systems. Before a Kalman filter can be used two equations must be supplied: the state equation and the measurement equation. The state equation describes how the unknown parameters evolve through time. The measurement equation describes the relation between the unknown parameters and the measurement. In both equations it is possible to specify uncertainty by way of noise terms. In our case the state parameters represent the split probabilities. These parameters are assumed to change only slowly over time. Therefore we use the following state equation:

$$b(t) = b(t-1) + w(t) \tag{23}$$

The measurement equation describes the way the state parameters are observed. In this case we can use Equation 9:

$$y(t) = H'(t)b(t) + v(t) \tag{24}$$

In these equations $w(t)$ and $v(t)$ must be zero mean noise processes with known covariance matrices;

$$E[w(t)w(p)'] = Q_t\delta_{tp}$$

$$E[v(t)v(p)'] = R_t\delta_{tp} \tag{25}$$

with δ_{tp} equal to 1 if t equals p and zero otherwise. On the basis of these equations and a knowledge of the covariance matrices a widely used estimation technique has been derived: the Kalman filter. A description of this technique can be found in many textbooks [see for example Anderson and Moore (*18*) and Catlin (*19*)]. The Kalman filter equations for the problems in Equation 23, 24, and 25 are:

$$\hat{b}(t) = \hat{b}(t-1) + K_t[y(t) - H'(t)\hat{b}(t-1)]$$

$$K_t = \textstyle\sum_{t-1} H(t)[H(t)'\sum_{t-1} H(t) + R_t]^{-1}$$

$$\textstyle\sum_t = \sum_{t-1} - \sum_{t-1} H(t)[H(t)'\sum_{t-1} H(t) + R_t]^{-1} H'(t)\sum_{t-1} + Q_t \tag{26}$$

These equations define a recursion that should be started with an initial estimate $\hat{b}(0)$ and an initial covariance matrix Σ_0. Given the assumptions in Equations 23 through 25, the Kalman filter leads to the minimum variance linear estimator; that is, the estimate is a linear function of the measurements $y(1) \ldots y(t)$, and the filter implicitly finds the matrix A and vector c that solve the following problem:

$$\min_{A,c} E[\|b(t) - A.[y(1) \ldots y(t)] - c\|^2] \tag{27}$$

Moreover this estimate can be shown to be unbiased. If besides earlier assumptions the noise terms and the initial state have Gaussian distributions the Kalman filter can be shown to produce unbiased estimates that have minimum variance over all estimators [see Anderson and Moore (*18*)].

The advantages of the Kalman filtering method are the computational efficiency of the method and the possibility of process-

ing interdependent measurements. Other advantages are that the calculations can be done recursively and that together with the estimate for the split matrix a variance-covariance matrix is calculated. This matrix gives an indication of the reliability of the estimate.

IMPROVED KALMAN FILTERING APPROACH

Despite the nice theoretical properties of the method several problems hinder the straightforward application of a Kalman filter. The first problem is that no initial values $\hat{b}(0)$ and Σ_0 are available. Some experimenting shows that the problem of initializing the filter cannot be seen apart from a second problem: how to impose the natural inequality constraints in Equation 2. It seems natural to specify very large diagonal values of Σ_0, since this expresses a lack of information about $b(0)$ and results in forgetting the initial value $\hat{b}(0)$ as quickly as possible. On the other hand the initial variance is bounded above since the split parameters are bounded to the interval [0,1]. Also specifying large initial variances results in many violations of the inequality constraints during the start-up phase of the filter. The problem of dealing with these inequality constraints has already been treated by Nihan and Davis (*9*), who proposed several constraining algorithms. This paper shows that a much simpler and effective way of dealing with both initial conditions and inequality constraints is possible.

Another problem is the lack of information about the noise covariance matrices Q_t and R_t in Equation 25. The results produced by the Kalman filter strongly depend on these matrices. Therefore a good approximation of these matrices should improve the estimate. In this section the measurement noise covariance matrix is derived from the trip generation model shown in Figure 1(*a*). This derivation produces the matrix R_t as a function of the split probability $b(t)$. This is an approximation since only an estimate of $b(t)$ is available. The last problem treated in this section is the use of the natural equality constraints (Equation 3). In Nihan and Davis (*9*) a normalization procedure is used to impose these constraints. In this paper the natural equality constraint is imposed via the perfect measurement concept [see Anderson and Moore (*18*)]. The consequences for the method are discussed.

Initial Conditions and Inequality Constraints

The Kalman filter described in the previous section has one commonly recognized interpretation, that is, that of a linear minimum variance estimator. However the Kalman filter can also be interpreted as an example of Bayesian estimation [see Catlin (*19*)]. As shown by Maher (*20*), assuming a Gaussian a priori distribution of the state vector and performing a Bayesian update with a measurement that has a Gaussian distribution (conditionally to the state vector) leads to a Gaussian a posteriori distribution. The equations derived for the scalar measurement case in Maher (*20*) can be shown to match the Kalman filter measurement update equations. In van der Zijpp and Hamerslag (*21*) the results are generalized to nonconstant state parameters and nonscalar measurements. A central role in this derivation is played by Bayes rule:

$$p[b(t)|y(1)\ldots y(t)] = \frac{p[y(t)|b(t), y(1)\ldots y(t-1)]\, p[b(t)|y(1)\ldots y(t-1)]}{p[y(t)|y(1)\ldots y(t-1)]} \quad (28)$$

The validity of Bayes rule follows from the definition of conditional probability. According to Bayes rule the a posteriori distribution can be derived from the a priori distribution and the likelihood function of the measurement vector. Figure 2(*a*) illustrates the principle of Bayesian updating for a scalar Gaussian random variable and a scalar measurement. The a posteriori distribution is obtained by multiplying the a priori density and the likelihood function and normalizing the result.

Inequality Constraints

Since natural inequality constraints bound the split probabilities $b(t)$ to an mn-dimensional hypercube [0,1], the a priori probability function should be zero outside this hypercube. One way in which this can be achieved is by multiplying the a priori probability

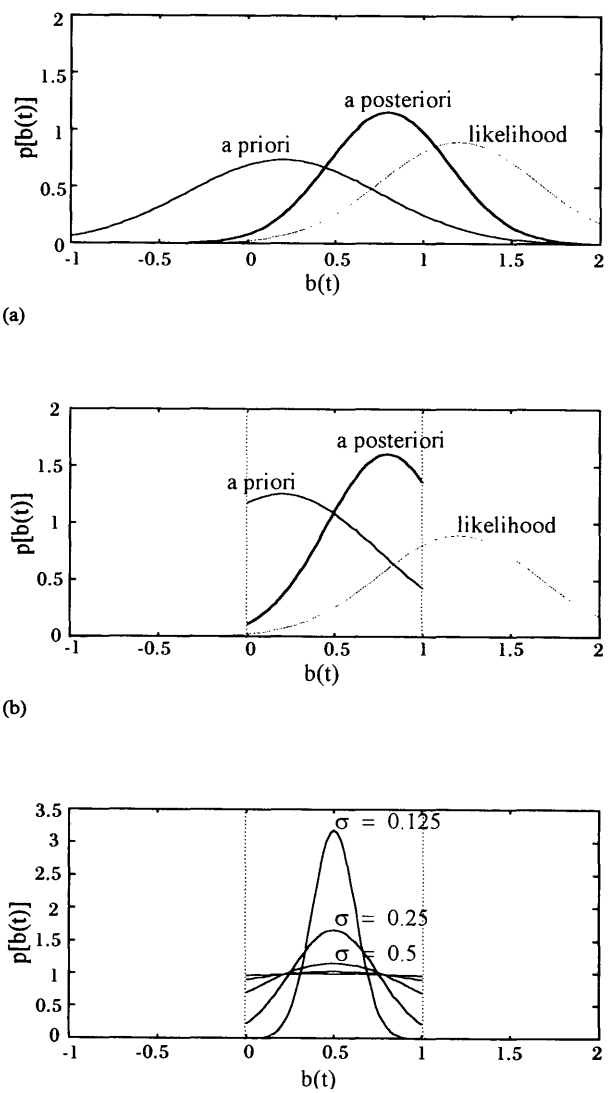


FIGURE 2 Bayesian updating: (*a*) Bayesian update, (*b*) Bayesian update, truncated a priori distribution, and (*c*) uniform distribution approached by truncated normal distribution.

function with an indicator function $I[0,1]()$. This function equals 1 if all elements of $b(t)$ satisfy the inequality constraints and is zero elsewhere. This leaves the shape of the distribution of $b(t)$ intact on the hypercube [0,1], whereas it defines a zero probability elsewhere. To ensure that the new function integrates to unity the a priori function should be multiplied by a factor $F()$. $F()$ can be expressed as a function of $\hat{b}(t-1)$ and Σ_{t-1}. In this way we get a truncated MVN distribution.

$$p[b(t)|y(1)\ldots y(t-1)] = \frac{F\left(\hat{b}(t-1), \sum_{t-1}\right)}{\sqrt{|\sum_{t-1} + Q_t|}\cdot(\sqrt{2\pi})^{mn}} \times \exp -\frac{1}{2}[b(t) - \hat{b}(t-1)]' \times \left(\sum_{t-1} + Q_t\right)^{-1} [b(t) - \hat{b}(t-1)] \times I_{[0,1]}[b(t)] \quad (29)$$

If we check how this assumption affects the derivation of an a posteriori distribution we conclude that when Equation 29 is used Equation 28 must be multiplied by the indicator function $I()$ and by a factor $F()$, and also the value of the normalizing constant $p[y(t)|y(1)\ldots y(t-1)]$ will be different.

However this operation does not affect the shape of the a posteriori distribution within the hypercube [0,1]. Therefore the a posteriori distribution will still be defined by a truncated MVN distribution, characterized by some $\hat{b}(t)$ and Σ_t. Moreover the recursion that determines $\hat{b}(t)$ and Σ_t from $\hat{b}(t-1)$ and Σ_{t-1} has not been changed. Therefore the Kalman filter equations can be used without modification, despite the presence of inequality constraints. Because of the modified circumstances, the Kalman filter results need another interpretation. The variables $\hat{b}(t)$ and Σ_t still characterize the probability distribution but can no longer be used as mean and variance [see Figure 2(*b*)]. Therefore the filtered results need some postprocessing before a point estimate can be presented. A first option, calculation of the true mean, requires the evaluation of an integral for which no analytical solution exists. Numerical integration is no option either because $b(t)$ is a high-dimensional vector. The second-best option is finding the maximum a posteriori (MAP) estimator for $b(t)$ [see Beck and Arnold (*22*)]. This can be found by maximizing the a posteriori density of $b(t)$:

$$\min_{b(t)} (b(t) - \hat{b}(t))'\Sigma_t^{-1}[b(t) - \hat{b}(t)],$$

$$0 \le b_i(t) \le 1,\ i = 1, 2 \ldots mn \quad (30)$$

To find the minimum solution the methods for constrained optimization can be used. These methods were described in a previous section. A potential drawback of this approach is the increase in computation time. When computation time is a bottleneck one can option for a suboptimal postprocessing method.

Initialization of Filter

In the foregoing we used the principle of Bayesian updating to derive a version of the Kalman filter that incorporates inequality constraints. As it will turn out, simultaneously we find a solution to the problem of initializing the Kalman filter. A common way of initializing a Bayesian filter when no a priori information is available is to use a uniform distribution. This expresses that, on the basis of the a priori information, every solution is equally likely. Working with the indicator function enables us to define an initial distribution that is arbitrarily close to the uniform distribution simply by defining Σ_0 as a diagonal matrix with very large diagonal elements. Figure 2(*c*) illustrates how a truncated Gaussian distribution approaches a uniform distribution if the variance increases.

Derivation of Measurement Noise Properties

The estimate obtained from a Kalman filter strongly depends on the assumed variance-covariance matrix for the measurement noise. Therefore in van der Zijpp and Hamerslag (*14*) such a matrix was derived on the basis of the trip generation model shown in Figure 1(*a*), which shows a system that is clearly different from the one described by the measurement Equation 24, since the measurements are numbers of successful experiments rather than linear combinations of the unknown parameters. However in terms of the expected value and the variance there is no difference between both systems. Therefore as far as the Kalman filter is concerned, we can treat the measurements from Figure 1(*a*) as if they were obtained from a linear system, as long as a covariance matrix R_t for the noise vector $v(t)$ is supplied.

A starting point for the derivation of such a matrix is the conditional distribution of the flows, given the entrance volume, $q_i^*(t)$, which is defined by a multinomial distribution:

$$P[f_{i1}(t)\ldots f_{in}(t)|q_i^*(t)] = \frac{q_i^*(t)!}{\prod_{j=1}^{n} f_{ij}(t)!} \prod_{j=1}^{n} b_{ij}(t)^{f_{ij}(t)} \quad (31)$$

By combining this with Equation 5 it can be shown that the following equations define the covariance matrix for the measurement $y(t) = U'f(t)$ as a function of the split vector $b(t)$.

$$R_t = \text{cov}[y(t), y(t)] = U'\text{cov}[f(t), f(t)]U \quad (32)$$

with

$$\text{cov}[f_{ij}(t), f_{hk}(t)] = q_i(t)b_{ij}(t)\delta_{ih}\delta_{jk} + [\sigma_q^2 - q_i(t)]b_{ij}(t)b_{hk}(t)\delta_{ih} \quad (33)$$

Since the exact value of the split vector is unknown, the estimate of the split vector is used instead. The covariance matrix is therefore only an approximation to the true matrix.

Equality Constraints

Another way of improving the Kalman filter estimate is by imposing the natural equality constraints (Equation 3). For the purpose of imposing the natural equality constraints Niham and Davis (*10*) proposed a normalization procedure. Since that procedure was meant to act separately from the active parameter estimation

method, it does not take full advantage of the possibilities of Kalman filtering.

Because the natural equality constraints are just another linear combination of the unknown split parameters, these constraints can be imposed as measurements to the Kalman filter. These kinds of measurements are referred to as *perfect observations*, because no noise on these observations is present. In matrix notation,

$$e = F'b(t),\ e = \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix},\ F' = \begin{bmatrix} 1 \cdots 1 & & & \\ & 1 \cdots 1 & & \\ & & \cdots & \\ & & & 1 \cdots 1 \end{bmatrix} \tag{34}$$

Anderson and Moore (*18*) show two ways to deal with these kinds of observations. The first way is to reduce the order of the filter by an order m (m denotes the number of entrances). This can be done by a change of coordinate basis, similar to the one used while calculating the solution to the constrained optimization problem. The second way is to proceed as with any measurement by using a zero matrix for the measurement noise matrix. In this case a recursion similar to Equation 26 is valid. For ease of implementation the latter method was used in the study described in this paper.

Define $\hat{b}^+(t)$ and Σ_t^+ as the updated estimate and variance-covariance matrix after performing a measurement update by Equation 34. Now $\hat{b}^+(t)$ and Σ_t^+ are obtained via:

$$\hat{b}^+(t) = \hat{b}(t) + K_t^+[e - F'\hat{b}(t)]$$

$$K_t^+ = \sum_t F\left[F'\sum_t F\right]^{-1}$$

$$\sum_t^+ = \sum_t - \sum_t F\left[F'\sum_t F\right]^{-1} F'\sum_t \tag{35}$$

These update equations lead to a singular variance-covariance matrix Σ_t^+. However $\hat{b}^+(t)$ and Σ_t^+ still define the density function of $b(t)$ on the domain in which $b(t)$ satisfies the natural equality constraints. Outside this domain the density function is zero. As a result Equation 30 transforms into:

$$\min_{b(t)} [b(t) - \hat{b}^+(t)]'\sum_t^{+\text{pinv}}[b(t) - \hat{b}^+(t)],$$

$$0 \le b_{(i-1)n+j}(t) \le 1,\ i = 1 \ldots m,\ j = 1 \ldots n$$

$$\sum_{j=1}^{n} b_{(i-1)n+j}(t) = 1,\ i = 1 \ldots m \tag{36}$$

where pinv is defined as the pseudo-inverse operator [see also Anderson and Moore (*18*)].

EXPERIMENTS

Experiments were carried out with both simulated and real data. The advantage of using simulated data is that the original matrix is available to evaluate the different methods. However these experiments give only limited insight into whether a method would work in practice. Therefore a second series of experiments was done by using minute-by-minute induction loop data from the Amsterdam beltway.

First Experiment, Simulated Data

The simulated data have been obtained by programming the trip generation model shown in Figure 1(*a*). The on-ramp volumes were generated by using a Poisson random generator. The split probabilities were obtained by taking a weighted sum of two extreme split vectors:

$$B(t) = \alpha(t)B_1 + [1 - \alpha(t)]B_2,$$

$$\alpha(t) = \frac{1}{2}[1 + \cos(2\pi t/T)],\ T = 144 \tag{37}$$

The network used consists of four entrances and four exits [Figure 3(*a*)].

As an evaluation criterion the square root of the mean squared error (RMSE) of the split parameters was used, that is,

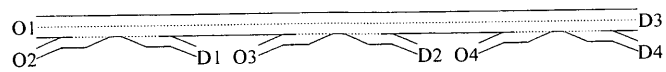
$$\text{RMSE} = \sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[\hat{b}_{ij}(t) - b_{ij}(t)]^2} \tag{38}$$

All methods described in this paper were tested. To make a fair comparison all methods were optimized for parameters that reflect the rate of change in the dynamic OD. The results from the previous section were used to determine the noise error covariance matrix required by the Kalman filter and to the natural constraints. For this experiment the Kalman-based method produced the best results; this was followed by constrained optimization, inequality-constrained least squares, and ordinary least squares [see Figure 3(*b*)]. The results are also presented in scatter diagrams [see Figure 3(*c*)]. These diagrams show for a number of periods the estimated split values plotted against the real values.

Second Experiment, Empirical Data

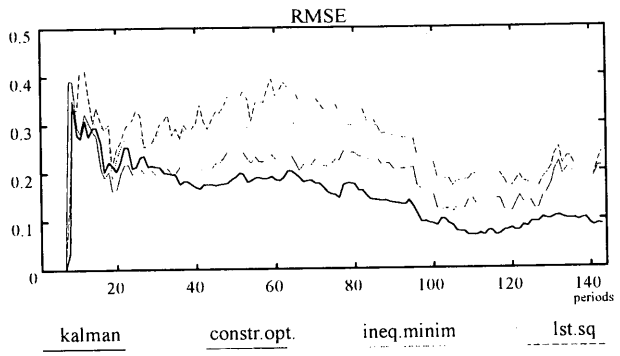
The second series of experiments was done by using induction loop data from the Amsterdam beltway. For this experiment one direction of an 11-km freeway corridor was selected. This corridor has five entrances and five exits and is equipped with 19 detector stations. All data were aggregated to periods of 5 min. Again various methods were compared. This time only the diagonal elements of the variance covariance matrix prescribed by Equation 32 were used while applying the Kalman filter.

For this experiment observed trip matrices were not available. Therefore the evaluation criterion in Equation 38 could not be used. Instead the flow-predicting capabilities for a set L of reference locations were used. Set L is a set of n_L reference locations. It contains induction loops on locations for which the volumes are expected to be sensitive to the split parameters, for example, between off-ramps and on-ramps. To prevent data from being used at the same time to calculate and evaluate $\hat{b}(t)$, the volumes were predicted by multiplying 5-min-old split parameter estimates by

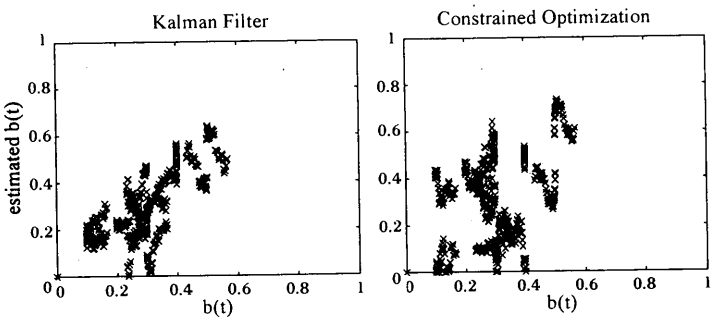


Origins		Destinations							
	vehicles /period	*B1*				*B2*			
		D1	D2	D3	D4	D1	D2	D3	D4
O1	666	.2	.1	.3	.4	.3	.2	.2	.3
O2	500	.3	.1	.3	.3	.4	.2	.2	.2
O3	333		.3	.4	.3		.4	.4	.2
O4	333			.5	.5			.6	.4

(a)



(b)



(c)

FIGURE 3 Simulation results: (*a*) simulation setup, (*b*) simulation results, and (*c*) true versus estimated $b(t)$, $t > 100$.

With the evaluation criterion in Equation 39 it was not possible to prove major differences in performance between Kalman filtering, constrained optimization, and inequality-constrained least squares. Only the nonconstrained least-squares method clearly gave results worse than those obtained by the other methods. After the evening rush hour the RMSE for all methods increased, probably because of suddenly changing OD patterns. When data from other days were evaluated RMSE plots with similar patterns appeared. This indicates that it might be useful to use a historic data base in which the permitted rate of change or even the direction of the changes in OD patterns are stored. Of all of the evaluated methods the Kalman filter seems the most suitable one for use in working with such a data base.

Although RMSE values do not differ significantly, comparing the split proportions estimated by different methods shows significant differences in estimated value; see for example Figure 4(*b*), which shows estimated splits for both the Kalman filter and the constrained least-squares methods.

To decide which of the two sets of parameters is more likely to correspond to the observed volumes, a second measure of effectiveness is introduced: the value of the likelihood function of the observations $y(t)$. Again $\hat{b}(t)$ is replaced by $\hat{b}(t-1)$ to prevent the use of observed volumes for estimation and evaluation purposes at the same time. The resulting likelihood is defined by

$$p[y(t)|\hat{b}(t)] \approx \frac{1}{(2\pi)^{p/2}\sqrt{|R_{t-1}|}} \exp -\frac{1}{2} \times [y(t) - H'(t)\hat{b}(t-1)]' \times R_{t-1}^{-1}[y(t) - H'(t)\hat{b}(t-1)] \quad (40)$$

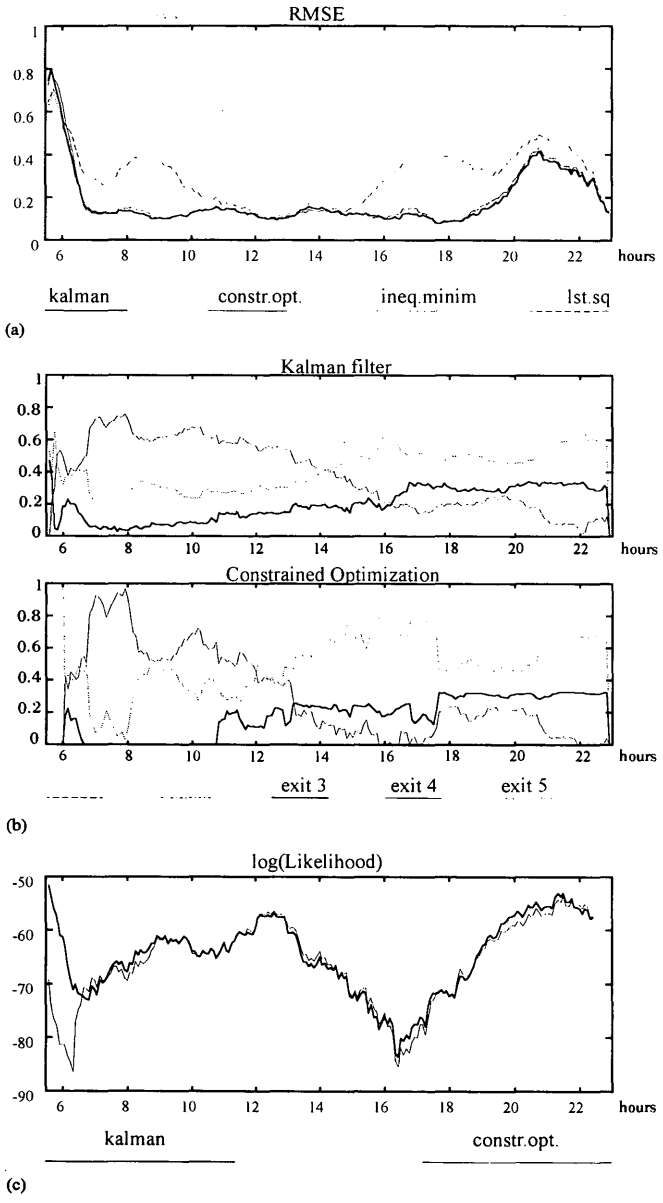


FIGURE 4 Empirical results: (*a*) moving average of RMSE, (*b*) estimated splits, entrance 4, and (*c*) moving average of log (likelihood).

In Figure 4(*c*) the moving average of the logarithm of this likelihood is displayed. Figure 4(*c*) shows that a test of the hypothesis by using a likelihood ratio would generally favor the Kalman filter-generated solution.

CONCLUSIONS

The problem of estimating dynamic OD matrices was converted to the problem of estimating split parameters in a trip generation model. A Kalman-based method was compared with other methods like least squares and constrained optimization.

A new way of initializing the Kalman filter and of imposing the natural inequality and equality constraints was derived from theory. A measurement noise covariance matrix that was derived from the trip generation model was used.

The resulting method was programmed and tested. Tests with simulated data indicate that the Kalman-based filter method performs better than the other methods. Tests with real data indicate that results can be improved by using a Kalman filter combined with a data base in which optimal tuning parameters for the filter are stored.

REFERENCES

1. Cascetta, E., and S. Nguyen. A Unified Framework for Estimating or Updating Origin/Destination Matrices from Traffic Counts. *Transportation Research B*, Vol. 22B, No. 6, 1988, pp. 437–455.
2. Hamerslag, R., and B. H. Immers. Estimation of Trip Matrices: Shortcomings and Possibilities for Improvement. In *Transportation Research Record 1203*, TRB, National Research Council, Washington, D.C., 1988, pp. 27–39.
3. Bell, M. G. H. Variances and Covariances for Origin-Destination Flows When Estimated by Log-Linear Models. *Transportation Research-B*, Vol. 19B, No. 6, 1985, pp. 497–507.
4. Hendrickson, C., and S. McNeil. Estimation of Origin/Destination Matrices with Constrained Regression. In *Transportation Research Record 976*, TRB, National Research Council, Washington, D.C., 1984.
5. van Zuylen, H. J., and L. G. Willumsen. The Most Likely Trip Matrix Estimated from Traffic Counts. *Transportation Research-B*, Vol. 14B, 1980, pp. 281–293.
6. Cremer, M., and H. Keller. Dynamic Identification of Flows from Traffic Counts at Complex Intersections. *Proc., Eighth International Symposium on Transportation and Traffic Theory*, 1981.
7. Cremer, M. Determining the Time-Dependent Trip Distribution in a Complex Intersection for Traffic Responsive Control. *IFAC Control in Transportation Systems.* Baden-Baden, Germany, 1983.
8. Cremer, M., and H. Keller. A New Class of Dynamic Methods for the Identification of Origin-Destination Flows. *Transportation Research-B*, Vol. 21B, No. 2, 1987, pp. 117–132.
9. Nihan, N. L., and G. A. Davis. Recursive Estimation of Origin-Destination Matrices from Input/Output Counts. *Transportation Research-B*, Vol. 21B, No. 2, 1987, pp. 149–163.
10. Nihan, N. L., and G. A. Davis. Application of Prediction-Error Minimization and Maximum Likelihood to Estimate Intersection O-D Matrices from Traffic Counts. *Transportation Science*, Vol. 23, No. 2, May 1989.
11. Bell, M. G. H., D. Inaudi, J. Lange, and M. Maher. Techniques for the Dynamic Estimation of O-D Matrices in Traffic Networks. *Proc., Drive Conference*, February 4 to 6, 1991.
12. Keller, H., and G. Ploss. Real-Time Identification of O-D Network Flows from Counts for Urban Traffic Control. *Proc., 10th Symposium on Traffic Theory*, 1987.
13. Bell, M. G. H. The Real Time Estimation of Origin-Destination Flows in the Presence of Platoon Dispersion. *Transportation Research-B*, Vol. 25B, 1991, pp. 115–125.
14. van der Zijpp, N. J., and R. Hamerslag. The Real Time Estimation of Origin-Destination Matrices for Freeway Corridors. *Proc., 26th ISATA Conference*, 1993.
15. Davis, G. A., and N. L. Nihan. A Stochastic Process Approach to the Estimation of Origin Destination Parameters from Time-Series of Traffic Counts. Presented at 70th Annual Meeting of the Transportation Research Board, Washington, D.C., 1991.
16. Davis, G. A. *A Stochastic Dynamic Model of Traffic Generation and Its Application to the Maximum Likelihood Estimation of Origin-Destination Parameters.* Ph.D. thesis. University of Washington, 1989.
17. Dempster, A. P., N. M. Laird, and D. B. Dubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Vol. 39, Series B, 1977, pp. 1–38.
18. Anderson, B. D. O., and J. B. Moore. *Optimal Filtering.* Prentice-Hall, Incorporated, Englewood Cliffs, N.J., 1979.
19. Catlin, D. E. Estimation, Control, and the Discrete Kalman Filter. *Applied Mathematical Sciences*, 71. Springer-Verlag, 1989.

20. Maher, M. J. Inferences on Trip Matrices from Observations on Link Volumes: A Bayesian Statistical Approach. *Transportation Research B*, Vol. 17B, No. 6, 1983, pp. 435–447.
21. van der Zijpp, N. J., and R. Hamerslag. A Bayesian Approach to Estimate Origin-Destination Matrices for Freeway Corridors. Presented at the Universities Transport Study Group 1994 Conference, 1994.
22. Beck, J. V., and K. J. Arnold. *Parameter Estimation in Engineering Science*. John Wiley & Sons, Incorporated, New York, 1977.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.