Research Article

A Bayesian Network Model for Origin-Destination Matrices Estimation Using Prior and Some Observed Link Flows

Lin Cheng, Senlai Zhu, Zhaoming Chu, and Jingxu Cheng

School of Transportation, Southeast University, Nanjing 210096, China

Correspondence should be addressed to Senlai Zhu; zhusenlai@163.com

Received 1 November 2013; Accepted 11 March 2014; Published 13 April 2014

Academic Editor: Wuhong Wang

Copyright © 2014 Lin Cheng 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.

This paper presents a Bayesian network model for estimating origin-destination matrices. Most existing Bayesian methods adopt prior OD matrices, which are always troublesome to be obtained. Since transportation systems normally have stored large amounts of historical link flows, a Bayesian network model using these prior link flows is proposed. Based on some observed link flows, the estimation results are updated. Under normal distribution assumption, the proposed Bayesian network model considers the level of total traffic flow, the variability of link flows, and the violation of the traffic flow conservation law. Both the point estimation and the corresponding probability intervals can be provided by this model. To solve the Bayesian network model, a specific procedure which can avoid matrix inversion is proposed. Finally, a numerical example is given to illustrate the proposed Bayesian network method. The results show that the proposed method has a high accuracy and practical applicability.

1. Introduction

Information about the traffic demand, which commonly means the origin-destination (OD) matrices, has been traditionally used by transportation planning agencies to evaluate the impact of various strategic transportation plans. And the real-time OD matrices are essential for real-time traffic applications, especially in the intelligent transportation system (ITS), such as the real-time route guidance via a dynamic traffic assignment or the evaluation of various ITS deployment alternatives [1, 2].

Various methods have been proposed to estimate OD matrices by using aggregate data such as OD demand counts and/or a set of traffic counts observed on the links. Using information derived from traffic counts is very attractive because they are cheap, easy, and immediate data. However, based on these data, we cannot obtain a unique OD matrix because the number of OD pairs is much larger than the number of links in large-scale transportation networks and there are infinite solutions satisfying the conservation law.

In order to have a unique solution which must be close to the actual one, one has to give more information. Normally, people use a prior OD matrix which can be obtained by many different methods, such as an old out-of-date or subjectively guessed OD matrix. These methods for estimating OD matrices can be classified as (1) least squares [3–9] and generalized least squares [10–12] methods, (2) entropy or information based methods [13, 14], and (3) statistical based methods.

Providing variability information of the traffic flow estimation is the most important advantage of the statistical methods. Normally, other methods give only the particular values of the OD and link flows, while statistical methods could also provide the corresponding probability intervals. The statistical methods can be categorized as follows. (1) Classical methods [15–17]: the traffic flows are assumed multivariate random variables given some parametric families, such as Poisson, Gamma, and multivariate normal. Then, the problem reduces to estimating the parameters and becomes a standard statistical problem. (2) Bayesian methods [18–21]: these methods also consider parametric families of distributions, but the parameters are considered as random variables themselves. Particularly, among the Bayesian methods, using Bayesian network [22–24] can easily know the relationship of all the variables (link flows and OD flows) and then simplify the calculation.

Whether using prior information (historical information or experience) or not is the main difference between Bayesian methods and classical statistical methods. In the Bayesian
methods, based on some prior information, the prior distribution of some parameters or variables can be determined. Then by updating the sample information (observed information), we can derive the posterior distribution, which is the fundamental inferential tool of the Bayesian methods.

Generally, the quality of the prior information can affect the accuracy of the estimation when using the Bayesian methods. The prior information used by almost all existing Bayesian methods for estimating OD matrices is a prior OD matrix. However, it is difficult to guarantee the accuracy of the prior OD matrix, which is outdated or subjectively guessed. Moreover, it is even impossible to get a prior OD matrix in some cases, especially in a newly developed city.

In reality, there are usually large amounts of historical link flow data stored in the cities’ transportation system data base. Compared with a prior OD matrix, prior (historical) link flows are more accurate as they were obtained by traffic detectors or manual investigation. Therefore, in this paper, in order to estimate OD matrices, we propose a Bayesian network (BN) method using prior link flows and a set of new observed link flows. Based on these prior link flows, we can derive the prior distribution of link flows and OD flows. Then, by updating a set of observed link flows, we can modify the means and reduce the variances of the remaining variables. Using these updated means and variances, we can obtain the posterior distribution of all the variables. Based on the posterior distribution, both the point estimation and the corresponding probability intervals can be provided.

Note that the level of total traffic flow varies randomly and deterministically in similar situations (vacation, peak hour, special weather conditions, etc.) [20, 24]. So the proposed BN model also considers the level of total traffic flow, which is very useful for many real-time traffic applications. In addition, the BN model also considers the variability of link flows and the violation of the conservation law.

The rest of the paper is organized as follows. Section 2 briefly introduces Bayesian network and Gaussian Bayesian networks. In Section 3, the proposed BN model for estimating OD matrices and its main assumptions are described. In Section 4, using the Bayesian network model, a specific procedure for estimating OD matrices is proposed. In Section 5, a numerical example is provided to illustrate the proposed model and clarify some of its implementation details. Finally, some conclusions are provided in Section 6.

2. Bayesian Network and Gaussian Bayesian Network

In this section, we briefly review the Bayesian network and Gaussian Bayesian network, which are the basic tools of this paper.

Definition 1 (Bayesian network). A Bayesian network is a pair \((G, P)\), where \(G\) is a directed acyclic graph (DAG) defined on a set of nodes \(X, P = \{ p(x_1 | \pi_1), \ldots, p(x_n | \pi_n) \}\) is a set of \(n\) conditional probability densities (CPDs), and \(\pi_i\) is the set of parents of node \(X_i\) in \(G\). The set \(P\) defines the associated joint probability density (JPD) as

\[
p(X) = \prod_{i=1}^{n} p(x_i | \pi_i).
\] (1)

The graph \(G\) contains all the qualitative information about the relationships among the variables. As a supplement, the probabilities in \(P\) quantify the qualitative information in graph \(G\).

In Bayesian networks, the factorization of JPD implied by (1) is normally very simple and the conditional independence relations among variables can be inferred directly from the graph \(G\), which makes the evidence propagation easy. Due to these advantages, Bayesian network models have been used widely to solve a large variety of practical problems [25, 26].

Bayesian networks can be applied to many distributions. For the sake of illustration, we consider the important and particular case of Gaussian Bayesian networks, in which the traffic flows distribution is supposed to be a normal distribution. A normal distribution for traffic flows is reasonable, because these random variables are the sum of a great number of independent Bernoulli experiments in which the users decide where to travel and which routes to choose. In the literature, Gaussian Bayesian networks have been used frequently [24, 27].

Definition 2 (Gaussian Bayesian network). A Bayesian network \((G, P)\) is said to be a Gaussian Bayesian network if and only if the joint probability distribution (JPD) associated with its variables \(X\) is a multivariate normal distribution, \(N(\mu, \Sigma)\), that is, with joint probability density function:

\[
f(X) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2}(X - \mu)^T \Sigma^{-1} (X - \mu) \right\},
\] (2)

where \(\mu\) is the mean vector, \(\Sigma\) is the \(n \times n\) covariance matrix, \(|\Sigma|\) is the determinant of \(\Sigma\), and \(\mu^T\) is the transpose of \(\mu\).

The JPD of the variables in a Gaussian Bayesian network can be specified as in (1) by the product of a set of CPDs, whose joint probability density function is

\[
f(x_i | \pi_i) \sim N \left( \mu_i + \sum_{j=1}^{i-1} \beta_{ij} (x_j - \mu_j), \psi_i \right).
\] (3)

where \(\beta_{ij}\) is the regression coefficient of \(X_j\) in the regression of \(X_i\) on its parents \(\pi_i\).

And the conditional variance of \(X_i\) is

\[
\psi_i = \Sigma_i - \Sigma_{i\pi} \Sigma_{\pi i}^{-1} \Sigma_{\pi i}.
\] (4)

where \(\Sigma_i\) is the unconditional variance of \(X_i\), \(\Sigma_{\pi i}\) is the covariance matrix between \(X_i\) and the variables in \(\pi_i\), and \(\Sigma_{\pi i}\) is the covariance matrix of \(\pi_i\).

3. Proposed Bayesian Network Model

Since Bayesian network has so many advantages as introduced in Section 2, in this section we propose a Bayesian
network (BN) model to reproduce the probabilistic structure of link and OD flows.

3.1. Model Assumptions. Assuming we have some prior (historical) link flows, in order to give the prior distribution of the link flows, we make the following assumptions.

Assumption 3. The link flows are given by

\[ V = KU + \eta. \]  

(5)

Assumption 4. The variable \( U \) is a normal random variable with mean \( \mu_U \) and variance-covariance matrix \( \sigma_U^2 \), where \( U \) is a normal random variable and measures the level of total mean flow. It reflects that traffic flows vary randomly and deterministically in similar situations (vacation, peak hour, special weather conditions, etc.). \( K \) is a vector, whose elements measure the relative weights of link flows with respect to the total traffic flow; \( \eta \) is a vector of independent normal random variables with zero mean; and \( \eta_a \) measures the discrepancy of the flow of link \( a \) with respect to its mean.

Note that traffic flows vary randomly and deterministically in similar situations (vacation, peak hour, special weather conditions, etc.) [20, 24]. Assumption 3 can take this into account. The distribution of \( U \) varies with the situation. Based on prior link flows and considering the similar situation, the distribution of \( U \) and the initial vector \( K \) are determined. Then we can easily derive the prior distribution of link flows, which will be shown later. Assumption 4 is a normal assumption, which is also adopted in Maher [18], Hazelton [20], Castillo et al. [24], and so forth.

To give the prior distribution of OD flows, we first consider the well-known conservation law equation:

\[ V_a = \sum_k \left( \sum_i T_i P_{ak} \delta_{ik} \right) = \sum_k T_k \left( \sum_i P_{ik} \delta_{ik} \right), \]  

(6)

where \( T_i \) and \( V_a \) are the flows of OD pair \( i \) and link \( a \), respectively. \( \delta_{ik} \) is the incidence element; that is, it takes value 1 if link \( a \) belongs to route \( k \) of OD pair \( i \) and 0, otherwise. \( p_{ik} \) is the proportion of users from OD pair \( i \) choosing route \( k \). In this paper, the route choice proportions are defined by a logit model as follows:

\[ p_{ik} = \frac{\exp(-\theta c_{ik})}{\sum_{l \in R} \exp(-\theta c_{il})}. \]  

(7)

where \( \theta \) is a parameter measuring travelers’ sensitivity to the cost difference between routes; \( c_{ik} \) is the cost associated with route \( k \) of OD pair \( i \).

Equation (6) can be written as

\[ V_a = \sum_i \left( \sum_k P_{ik} \delta_{ik} \right) T_i. \]  

(8)

Set \( d_{ai} = \sum_k p_{ik} \delta_{ik} \) to represent the proportions of users from OD pair \( i \) choosing link \( a \). Then (8) can be rewritten in the form of matrix as

\[ V = DT. \]  

(9)

3.2. The Complete Model. Based on Definitions 1 and 2, in order to complete our BN model, we need to define an associated graph \( G \). For example, consider the simple network shown in the left of Figure 1, which has 2 nodes, 2 links, and 1 OD pair 1-2. The right of Figure 1 shows the associated Bayesian network. The link flow node \( v_a \) has as parents the corresponding node \( U \) and \( \eta_a \). The OD flow node \( t_i \) has as parents the corresponding node \( U \) and \( \eta_a \).

Then according to Assumptions 3 and 4, we get the variance-covariance matrix of \( V \):

\[ \Sigma_V = (K 1) \Sigma_{U\eta} \left( K^T \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right) = \sigma_U^2 K K^T + D_\eta, \]  

(13)

where \( \Sigma_{U\eta} \) and \( D_\eta \) are diagonal matrixes. \( D_\eta \) is the variance-covariance matrix of \( \eta \).

Based on Assumption 5, we get

\[ \begin{pmatrix} V \\ T \end{pmatrix} = \begin{pmatrix} K & 1 \\ 0 & \beta \end{pmatrix} \begin{pmatrix} V \\ \varepsilon \end{pmatrix}. \]  

(14)

Then, the mean \( E[(V, T)] \) is

\[ E[(V, T)] = \begin{pmatrix} E(U) K \\ E(U) \beta K + E(\varepsilon) \end{pmatrix}. \]  

(15)
And the variance-covariance matrix of \((V, T)\) is
\[
\Sigma_{(V,T)} = \begin{pmatrix}
\Sigma_V & \Sigma_V \beta^T \\
\beta \Sigma_V & \beta \Sigma_V \beta^T + D_\varepsilon
\end{pmatrix},
\]
(16)
where \(D_\varepsilon\) is the variance matrix of \(\varepsilon\).

In summary, all random variables involved in our model are related by the linear expression:
\[
\begin{pmatrix}
U \\
\eta \\
V \\
T
\end{pmatrix} = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
K & 0 & 1 & 0 \\
\beta K & \beta & 1 & 1
\end{pmatrix}
\begin{pmatrix}
U \\
\eta \\
V \\
T
\end{pmatrix}.
\]
(17)

The mean \(E[U, \eta, \varepsilon, V, T]\) is
\[
\begin{pmatrix}
E(U) \\
E(\eta) \\
E(\varepsilon) \\
E(V) \\
E(T)
\end{pmatrix} = \begin{pmatrix}
E(U) \\
0 \\
E(\varepsilon) \\
E(U) K \\
E(U) \beta K + E(\varepsilon)
\end{pmatrix}.
\]
(18)

And the variance-covariance matrix \(\Sigma_{(U, \eta, \varepsilon, V, T)}\) is
\[
\begin{pmatrix}
\sigma_u^2 & 0 & 0 & \sigma_u^2 K^T & \sigma_u^2 K^T \beta^T \\
0 & D_\eta & 0 & D_\eta & D_\eta \beta^T \\
0 & 0 & D_\varepsilon & 0 & D_\varepsilon \\
\sigma_u^2 K & D_\eta & 0 & \sigma_u^2 KK^T + D_\eta & \sigma_u^2 KK^T + D_\eta \beta^T \\
\sigma_u^2 \beta K & D_\eta \beta & D_\varepsilon & \sigma_u^2 \beta K \beta + D_\varepsilon & \sigma_u^2 \beta K \beta + D_\varepsilon + D_\varepsilon
\end{pmatrix}.
\]
(19)

Then, the prior distribution (joint probability density function) of all the variables can be given as
\[
f(V_1, V_2, \ldots, V_m, T_1, T_2, \ldots, T_n) = f_{N(\mu_x, \Sigma)}(V_1, V_2, \ldots, V_m) \prod_{i=1}^n \prod_{i \neq \pi_i} T_i = \prod_{a=1}^m p(V_a | \pi_a)
\]
(20)

### 4. Estimating OD Matrices Using the Proposed BN Model

In this section, using the proposed BN model, we describe how to estimate the OD matrices when some new observed link flows are available.

Since we have obtained the prior distribution of all the variables, we can use the following equations to update the mean and the covariance matrix of the variables [23, 24] based on some observed variables. Note that one only needs to consider the unobserved variables conditioned on the observed variables and then update the expected values and covariance of the remaining variables. These equations are
\[
\mu_{YZ|z} = \mu_Y + \Sigma_{YZ} \Sigma_{ZZ}^{-1} (z - \mu_Z),
\]
(21)
\[
\Sigma_{YZ|z} = \Sigma_{YY} - \Sigma_{YZ} \Sigma_{ZZ}^{-1} \Sigma_{ZY},
\]
(22)
\[
\mu_{Z|z} = z,
\]
(23)
\[
\Sigma_{Z|z} = 0,
\]
(24)
where \(Y\) and \(Z\) are the sets of unobserved and observed variables, respectively; \(\mu_Y\) and \(\Sigma_{YY}\) are the mean vector and covariance matrix of \(Y\); \(\mu_Z\) and \(\Sigma_{ZZ}\) are the mean vector and covariance matrix of \(Z\); and \(\Sigma_{YZ}\) is the covariance matrix of \(Y\) and \(Z\).

Given a set of evidential nodes \(Z\) whose values are known to be \(Z = z\), by (21) and (22), we can derive the mean vector and covariance matrix of the unobserved nodes in \(Y\). Thus, the conditional distribution of \(Y\) can be obtained. Equations (23) and (24) state that the expected values of the observed variables coincide with their observed values and their variances and covariances are null. In order to simplify the calculation, we can use an incremental method, that is, updating evidence from \(Z\) one by one. Thus, we do not need to calculate the matrix inverse operation, because the matrix degenerates to a scalar. In this case, \(\Sigma_{YZ}\) is a column vector and \(\Sigma_{ZZ}\) is a scalar (i.e., \(\Sigma_{ZZ}^{-1} = 1/\sigma_{ZZ}\)).

If we want to give the point estimation as well as the corresponding probability intervals, we can solve the following maximum posterior distribution problem to get the point estimation, whose results are normally the conditional means:
\[
\text{Maximize } f_{N(\mu_x, \Sigma)}(t_1, t_2, \ldots, t_n) \prod_{a=1}^m p(V_a | \pi_a) |_Z
\]
(25)
where \(Z\) is the set of the observed variables, including those observed link flows and/or OD flows.

In summary, the specific procedure for estimating OD matrices and those unobserved link flows is given as follows.

**Step 0.** Initialize the model. According to Assumptions 3 and 4, based on prior (historical) link flows, we can determine the distribution of \(U\) and the initial matrix \(K\). Then we can obtain the initial link flows \(V = (V_1, V_2, \ldots, V_m) = K(E(U))\). Thus the initial route choice proportion \(p_{ik}\) is calculated as follows:
\[
h_a(V_a) = h_a^0 \left(1 + \alpha_a \left(\frac{V_a}{C_a}\right)^{\beta_a}\right),
\]
(26)
\[
\delta_{ik} = \sum_{a \in A} h_a(V_a) \delta_{ik}^a,
\]
(27)
\[
p_{ik} = \exp\left(-\theta_{ik}\right) \sum_{k \in R} \exp\left(-\theta_{ik}\right),
\]
(28)
where (26) is the link cost function, where \(h_a^0\) is the cost associated with free flow conditions, \(C_a\) is the link capacity, and
\[ \alpha_a \text{ and } \beta_a \text{ are constants defining how the cost increases with traffic flow; (27) is the route cost function; (28) calculates the route choice proportion defined in (7).} \]

**Step 1.** Solve the BN model. According to model assumptions, using the initial route choice proportion matrix \( P \), we can get the prior distribution of traffic flows (prior means and variances) using the following formulas:

\[
\beta = \left[ (P \delta)^T P \delta \right]^{-1} (P \delta)^T, \\
E[V] = KE[U], \\
E[T] = \beta E[V] + E[\epsilon], \\
\Delta_\eta = \text{Diag}(v E[V]), \\
\Sigma_{VV} = \sigma_\eta^2 \mathbf{K} \mathbf{K}^T + \Delta_\eta, \\
\Sigma_{VT} = \Sigma_{VV} \beta^T, \\
\Sigma_{TT} = \beta \Sigma_{VV} \beta^T + \Delta_\eta, \\
\]

where (29) is for calculating the regression coefficient matrix given in (12). Equations (30) and (31) are for calculating the means of \( V \) and \( T \). Equation (32) defines the diagonal variance matrix of \( \eta \); that is, \( \text{Var}(\eta_i) = (E(\nu_i) \times v)^2 \), where \( v \) is the coefficient of variation. Equations (33) to (36) define the variance-covariance matrix in (19).

**Step 2.** Update the observed link flows, using the formulas:

\[
E[YZ | z] = E[Y] + \Sigma_{YZ} \Sigma_{ZZ}^{-1} (z - E[Z]), \\
\Sigma_{YZ} = \Sigma_{YY} - \Sigma_{ZV} \Sigma_{ZZ}^{-1} \Sigma_{ZY}, \\
E[Z | Z = z] = z, \\
\Sigma_{Z|Z = z} = 0, \\
T = E[Y | Z = z] |_{Y \in \mathbf{T}}, \\
\]

where (37) and (38) are for updating the means and variance-covariance matrix of the unobserved variables, where \( Y \) and \( Z \) refer to the unobserved and observed components of \( (T, V) \), respectively. Equations (39) and (40) state that the expected values of the observed variables are their observed values and their variances and covariances are zero, as given in (23) and (24). Equation (41) takes the conditional means as point estimation for the OD and link flows, as the results of the maximum posterior distribution problem given in (25).

**Step 3.** Calculate the new route choice proportions. Since link flows are updated in Step 2, the route choice proportions also need to be updated. Given that the matrix \( T \) is obtained by (41), the new route choice proportion \( p_{ik}^* \) is calculated using the following expressions:

\[
V_a = \sum_{i \in A} \sum_{k \in K} \Delta_{ik}^{\alpha} \left( V_{ai} \right), \\
h_a(V_a) = h_{\alpha} \left( 1 + \alpha \left( \frac{V_{ai}}{C_{ai}} \right)^{\beta} \right), \\
\epsilon^i_k = \sum_{a \in A} h_{\alpha} (V_{ai}) \theta_{ak}, \\
\]

where (42) is the conservation law equation given in (6).

**Step 4.** Test convergence. If \( \Sigma_{ik}(p_{ik} - p_{ik}^*)^2 < \xi \), where \( \xi \) is a small number to control convergence of the process, then stop the process and return the OD flow \( T_i \), the link flow \( V_a \), and the route choice proportion \( p_{ik}^* \). Otherwise, continue with Step 5.

**Step 5.** Update route choice proportions and the matrix \( K \), using the expressions:

\[
p_{ik} = \rho p_{ik}^* + (1 - \rho) p_{ik} \quad \forall i, k, \\
K = \frac{V}{E(U)},
\]

and go to Step 1, where (46) is for updating the route choice proportion matrix, where \( \rho, 0 < \rho < 1, \) is a relaxation factor; (47) is for updating the matrix \( K \). The values of variables \( V \) are obtained by (42).

## 5. Example: The Nguyen-Dupuis Network

In this section, we illustrate the proposed methods using the well-known Nguyen-Dupuis network, shown in Figure 2. It consists of 13 nodes, 19 links, and 4 OD pairs: 1-2, 1-3, 4-2, and 4-3.
6 Discrete Dynamics in Nature and Society

Table 1: Network parameters of the Nguyen-Dupuis network.

<table>
<thead>
<tr>
<th>Link</th>
<th>$h_a$</th>
<th>$C_{a}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>900</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>700</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>700</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>900</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>800</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>600</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>900</td>
</tr>
<tr>
<td>8</td>
<td>13</td>
<td>500</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>300</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>400</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>700</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>700</td>
</tr>
<tr>
<td>13</td>
<td>9</td>
<td>600</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>700</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>700</td>
</tr>
<tr>
<td>16</td>
<td>8</td>
<td>700</td>
</tr>
<tr>
<td>17</td>
<td>7</td>
<td>300</td>
</tr>
<tr>
<td>18</td>
<td>15</td>
<td>700</td>
</tr>
<tr>
<td>19</td>
<td>11</td>
<td>700</td>
</tr>
</tbody>
</table>

The network data are shown in Table 1 and the associated parameters in (26) are assumed to be $\alpha_a = 0.15$, $\beta_a = 4$ for any link.

The assumed true OD matrices, which are used later for testing the quality of the estimation, are shown in Table 4 under the heading “True flow.” The true link flows are obtained by solving the multinomial logit assign model with parameter $\theta = 1.0$ for the stochastic loading.

The prior information is assumed as follows: the expected value of the level of total traffic flow $U$ and its standard deviation are $E(U) = 50$ and $\sigma(U) = 10$, respectively and the initial matrix $K$ is

$$K = (1.5, 0.9, 0.9, 0.7, 1.7, 0.8, 1.8, 0.1, 0.8, 1.0, 1.5, 0.3, 1.2, 0.3, 0.5, 0.8, 0.1, 0.7, 1.2)^T.$$ (48)

The observed link flows are assumed to be $V_5 = 82.57$, $V_7 = 87.38$, $V_{10} = 48.07$, $V_{13} = 58.66$, and $V_{18} = 37.12$. And it is supposed that they are known in this order. Since they are observed, their values are equal to the true link flows (as shown in Table 4).

Step 0-Step 1. Initialize and give the prior distribution of all the variables.

Based on the prior information, we can get the prior distribution of the traffic flows. The prior mean and variances are shown in the second columns of Tables 2 and 3, respectively. To simplify the calculation, in this example, the expectation and variance-covariance of $\epsilon$ are assumed to be null (i.e., there is no uncertainty in the conservation law). In addition, to obtain the variance-covariance matrix $D_{\epsilon\epsilon}$ we have selected $\nu = 0.1$ in (32).

Step 2-Step 5. Give the posterior distribution by updating the observed link flows one by one.

Table 2: Point estimation of traffic flows after updating observed link flows one by one.

<table>
<thead>
<tr>
<th>Item</th>
<th>0</th>
<th>5</th>
<th>7</th>
<th>10</th>
<th>13</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>37.16</td>
<td>39.02</td>
<td>39.05</td>
<td>39.86</td>
<td>40.11</td>
<td>39.88</td>
</tr>
<tr>
<td>$T_2$</td>
<td>82.88</td>
<td>77.95</td>
<td>78.32</td>
<td>78.63</td>
<td>80.51</td>
<td>81.22</td>
</tr>
<tr>
<td>$T_3$</td>
<td>68.37</td>
<td>59.64</td>
<td>59.94</td>
<td>59.23</td>
<td>60.85</td>
<td>59.95</td>
</tr>
<tr>
<td>$T_4$</td>
<td>12.68</td>
<td>22.48</td>
<td>22.29</td>
<td>23.28</td>
<td>19.54</td>
<td>19.06</td>
</tr>
<tr>
<td>$V_1$</td>
<td>75.00</td>
<td>74.64</td>
<td>74.92</td>
<td>75.26</td>
<td>76.98</td>
<td>77.55</td>
</tr>
<tr>
<td>$V_2$</td>
<td>45.00</td>
<td>42.36</td>
<td>42.48</td>
<td>43.28</td>
<td>43.69</td>
<td>43.61</td>
</tr>
<tr>
<td>$V_3$</td>
<td>45.00</td>
<td>46.30</td>
<td>46.32</td>
<td>45.73</td>
<td>46.22</td>
<td>46.03</td>
</tr>
<tr>
<td>$V_4$</td>
<td>35.00</td>
<td>35.82</td>
<td>35.92</td>
<td>36.78</td>
<td>34.14</td>
<td>32.80</td>
</tr>
<tr>
<td>$V_5$</td>
<td>85.00</td>
<td>82.57</td>
<td>82.57</td>
<td>82.57</td>
<td>82.57</td>
<td>82.57</td>
</tr>
<tr>
<td>$V_6$</td>
<td>40.00</td>
<td>38.37</td>
<td>38.62</td>
<td>38.37</td>
<td>40.69</td>
<td>41.17</td>
</tr>
<tr>
<td>$V_7$</td>
<td>90.00</td>
<td>87.24</td>
<td>87.38</td>
<td>87.38</td>
<td>87.38</td>
<td>87.38</td>
</tr>
<tr>
<td>$V_8$</td>
<td>5.00</td>
<td>1.43</td>
<td>1.45</td>
<td>1.55</td>
<td>1.51</td>
<td>1.54</td>
</tr>
<tr>
<td>$V_9$</td>
<td>40.00</td>
<td>39.38</td>
<td>39.40</td>
<td>39.31</td>
<td>39.31</td>
<td>39.31</td>
</tr>
<tr>
<td>$V_{10}$</td>
<td>50.00</td>
<td>47.86</td>
<td>47.92</td>
<td>48.07</td>
<td>48.07</td>
<td>48.07</td>
</tr>
<tr>
<td>$V_{11}$</td>
<td>75.00</td>
<td>75.64</td>
<td>75.73</td>
<td>76.24</td>
<td>76.58</td>
<td>76.43</td>
</tr>
<tr>
<td>$V_{12}$</td>
<td>15.00</td>
<td>15.14</td>
<td>15.42</td>
<td>15.44</td>
<td>16.20</td>
<td>15.43</td>
</tr>
<tr>
<td>$V_{13}$</td>
<td>60.00</td>
<td>59.05</td>
<td>59.12</td>
<td>59.71</td>
<td>58.66</td>
<td>58.66</td>
</tr>
<tr>
<td>$V_{14}$</td>
<td>15.00</td>
<td>16.57</td>
<td>16.87</td>
<td>16.99</td>
<td>17.71</td>
<td>16.84</td>
</tr>
<tr>
<td>$V_{15}$</td>
<td>25.00</td>
<td>23.06</td>
<td>23.31</td>
<td>22.89</td>
<td>24.42</td>
<td>23.57</td>
</tr>
<tr>
<td>$V_{16}$</td>
<td>40.00</td>
<td>41.37</td>
<td>41.49</td>
<td>42.22</td>
<td>41.34</td>
<td>41.34</td>
</tr>
<tr>
<td>$V_{17}$</td>
<td>5.00</td>
<td>6.10</td>
<td>6.16</td>
<td>6.29</td>
<td>6.40</td>
<td>6.51</td>
</tr>
<tr>
<td>$V_{18}$</td>
<td>35.00</td>
<td>36.26</td>
<td>36.32</td>
<td>36.40</td>
<td>37.29</td>
<td>37.12</td>
</tr>
<tr>
<td>$V_{19}$</td>
<td>60.00</td>
<td>59.05</td>
<td>59.12</td>
<td>59.71</td>
<td>58.66</td>
<td>58.66</td>
</tr>
</tbody>
</table>

Table 2 shows how the means of the traffic flows changed after updating the observed link flows one by one. After each update, the point estimation of link flows $\{V_1, V_2, \ldots, V_{19}\}$ and OD flows $\{T_1, T_2, \ldots, T_4\}$ is provided. It can be seen that once $V_7$ and $V_{10}$ become known and updated, the point estimation of $V_5$ in Table 2 remains unchanged and its variance in Table 3 becomes zero (boldfaced in the table). Because, due to the flow conservation in node 6, once $V_7$ and $V_{10}$ become known, $V_6$ becomes known. Similarly, $V_{19}$ becomes known once $V_{13}$ is given; $V_{18}$ becomes known once $V_{19}$ is given; $V_{17}$ becomes known once $V_{18}$ and $V_{19}$ are given; $V_{15}$ becomes known once $V_{17}$ is given; $V_{14}$ becomes known once $V_{15}$, $V_{16}$, and $V_{18}$ are given. In other words, due to the conservation law, the observed link flows are known in this order: $V_5 = 82.57$, $V_7 = 87.38$, $V_{10} = 48.07$, $V_{13} = 58.66$, $V_{19} = 58.66$, $V_{16} = 41.24$, $V_{18} = 37.12$, $V_{11} = 76.43$, $V_{15} = 23.57$, and $V_{14} = 16.84$.

Table 3 shows how the variances of the traffic flows changed after updating the observed link flows one by one. Note that after some link flows (including the observed link flows and those derived from the observed link flows and the conservation laws) are known, their means remain constant and their variances become null (boldfaced in the tables). And normally the variances of the unknown variables (OD flows and those unobserved link flows) decrease with each update. Note that the smaller the variances, the higher the precision of the estimation. So after a series of update, the estimation becomes more accurate. This derives a method to
determine how many links and what links need to be observed when estimating traffic flows by the Bayesian network model, that is, the network sensor location problem (NLSI) [28]. Note that the variances updating equation (22) has no relevance with the values of the observed link flows. So we can solve NLSI without observing any link. First, by the Bayesian network model, we can get the prior distribution of all the variables shown in (19). Next, we can take the link which can reduce the variance of the OD flows maximally by updating as the first observed link. Then we update the variances of traffic flows, determine the second observed link, and iterate until the variances decrease to meet the requirement of the estimation precision or until the budget exceeds the constraint.

By the updated means (point estimation) and variances, we can obtain the posterior distribution of OD flows and those unobserved link flows. Figure 3 illustrates how the marginal densities of OD flows and those unobserved link flows evolve from their initial form to their final form (boldfaced) by updating the observed link flows one by one. It can be seen that the variances of the unknown variables are normally decreasing with each update. The proposed method can provide a control of the conservation law as well. The final forms (boldfaced) in Figure 3 are the posterior densities of OD flows and those unobserved link flows. These posterior densities supply complete statistical information about the unknown variables. By these posterior densities, we can provide the point estimation as well as the corresponding probability intervals.

In order to test the quality of the estimation, Table 4 compares the true flow and the point estimation of the proposed BN model. Because $V_5$, $V_7$, $V_9$, $V_{10}$, $V_{11}$, $V_{13}$, $V_{14}$, $V_{15}$, $V_{16}$, $V_{18}$, and $V_{19}$ are observed, their values are equal to the true flow. And studying the OD pairs and those unobserved links (boldfaced in the table), it can be seen that the estimation and true flow are basically the same. The relative errors are all very small and the maximum relative error value of the OD flows estimation is only 4.70%. This illustrates that the proposed BN model has a high accuracy.

### 6. Conclusions

In this paper, we use a Bayesian network model to estimate origin-destination matrices based on prior link flows and a set of observed link flows. Normally, large amounts of historical link flows are stored in the cities’ transportation system. Compared with an outdated or subjectively guessed prior OD matrix, prior link flows are more accurate as they are obtained by traffic detectors or manual investigation. The proposed

<table>
<thead>
<tr>
<th>Item</th>
<th>True flow</th>
<th>Proposed method</th>
<th>Relative error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>40</td>
<td>39.88</td>
<td>0.30</td>
</tr>
<tr>
<td>$T_2$</td>
<td>80</td>
<td>81.22</td>
<td>1.53</td>
</tr>
<tr>
<td>$T_3$</td>
<td>60</td>
<td>59.95</td>
<td>0.08</td>
</tr>
<tr>
<td>$T_4$</td>
<td>20</td>
<td>19.06</td>
<td>4.70</td>
</tr>
<tr>
<td>$V_1$</td>
<td>76.64</td>
<td>77.55</td>
<td>1.19</td>
</tr>
<tr>
<td>$V_2$</td>
<td>43.36</td>
<td>43.61</td>
<td>0.58</td>
</tr>
<tr>
<td>$V_3$</td>
<td>46.13</td>
<td>46.03</td>
<td>0.22</td>
</tr>
<tr>
<td>$V_4$</td>
<td>33.87</td>
<td>32.80</td>
<td>3.16</td>
</tr>
<tr>
<td>$V_5$</td>
<td>82.57</td>
<td>82.57</td>
<td>0</td>
</tr>
<tr>
<td>$V_6$</td>
<td>40.21</td>
<td>41.17</td>
<td>2.39</td>
</tr>
<tr>
<td>$V_7$</td>
<td>87.38</td>
<td>87.38</td>
<td>0</td>
</tr>
<tr>
<td>$V_8$</td>
<td>1.42</td>
<td>1.54</td>
<td>8.45</td>
</tr>
<tr>
<td>$V_9$</td>
<td>39.31</td>
<td>39.31</td>
<td>0</td>
</tr>
<tr>
<td>$V_{10}$</td>
<td>48.07</td>
<td>48.07</td>
<td>0</td>
</tr>
<tr>
<td>$V_{11}$</td>
<td>76.43</td>
<td>76.43</td>
<td>0</td>
</tr>
<tr>
<td>$V_{12}$</td>
<td>15.41</td>
<td>15.43</td>
<td>0.13</td>
</tr>
<tr>
<td>$V_{13}$</td>
<td>58.66</td>
<td>58.66</td>
<td>0</td>
</tr>
<tr>
<td>$V_{14}$</td>
<td>16.84</td>
<td>16.84</td>
<td>0</td>
</tr>
<tr>
<td>$V_{15}$</td>
<td>23.57</td>
<td>23.57</td>
<td>0</td>
</tr>
<tr>
<td>$V_{16}$</td>
<td>41.34</td>
<td>41.34</td>
<td>0</td>
</tr>
<tr>
<td>$V_{17}$</td>
<td>6.24</td>
<td>6.51</td>
<td>4.33</td>
</tr>
<tr>
<td>$V_{18}$</td>
<td>37.12</td>
<td>37.12</td>
<td>0</td>
</tr>
<tr>
<td>$V_{19}$</td>
<td>58.66</td>
<td>58.66</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>0</th>
<th>5</th>
<th>7</th>
<th>10</th>
<th>13</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>65.81</td>
<td>25.98</td>
<td>20.88</td>
<td>16.94</td>
<td>14.35</td>
<td>2.93</td>
</tr>
<tr>
<td>$T_2$</td>
<td>294.56</td>
<td>54.98</td>
<td>29.91</td>
<td>20.46</td>
<td>15.66</td>
<td>15.66</td>
</tr>
<tr>
<td>$T_3$</td>
<td>203.36</td>
<td>22.72</td>
<td>7.59</td>
<td>5.41</td>
<td>5.40</td>
<td>3.30</td>
</tr>
<tr>
<td>$T_4$</td>
<td>18.56</td>
<td>16.98</td>
<td>15.59</td>
<td>14.10</td>
<td>3.24</td>
<td>3.16</td>
</tr>
<tr>
<td>$V_1$</td>
<td>281.25</td>
<td>100.33</td>
<td>81.13</td>
<td>69.99</td>
<td>67.48</td>
<td>65.57</td>
</tr>
<tr>
<td>$V_2$</td>
<td>101.25</td>
<td>32.31</td>
<td>26.08</td>
<td>23.15</td>
<td>21.73</td>
<td>20.73</td>
</tr>
<tr>
<td>$V_3$</td>
<td>101.25</td>
<td>38.61</td>
<td>31.01</td>
<td>25.85</td>
<td>24.33</td>
<td>23.10</td>
</tr>
<tr>
<td>$V_4$</td>
<td>61.25</td>
<td>23.11</td>
<td>18.65</td>
<td>16.72</td>
<td>13.27</td>
<td>11.73</td>
</tr>
<tr>
<td>$V_5$</td>
<td>361.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$V_6$</td>
<td>80.00</td>
<td>26.51</td>
<td>21.55</td>
<td>18.19</td>
<td>18.85</td>
<td>18.48</td>
</tr>
<tr>
<td>$V_7$</td>
<td>405.00</td>
<td>137.07</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$V_8$</td>
<td>1.25</td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>$V_9$</td>
<td>80.00</td>
<td>27.93</td>
<td>22.44</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$V_{10}$</td>
<td>125.00</td>
<td>41.25</td>
<td>33.19</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$V_{11}$</td>
<td>281.25</td>
<td>103.05</td>
<td>82.88</td>
<td>71.83</td>
<td>66.79</td>
<td>0</td>
</tr>
<tr>
<td>$V_{12}$</td>
<td>11.25</td>
<td>4.13</td>
<td>3.44</td>
<td>2.95</td>
<td>2.99</td>
<td>2.60</td>
</tr>
<tr>
<td>$V_{13}$</td>
<td>180.00</td>
<td>62.81</td>
<td>50.52</td>
<td>44.05</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$V_{14}$</td>
<td>11.25</td>
<td>4.94</td>
<td>4.11</td>
<td>3.57</td>
<td>3.57</td>
<td>0</td>
</tr>
<tr>
<td>$V_{15}$</td>
<td>31.25</td>
<td>9.58</td>
<td>7.85</td>
<td>6.47</td>
<td>6.79</td>
<td>0</td>
</tr>
<tr>
<td>$V_{16}$</td>
<td>80.00</td>
<td>30.82</td>
<td>24.88</td>
<td>22.03</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$V_{17}$</td>
<td>1.25</td>
<td>0.67</td>
<td>0.55</td>
<td>0.50</td>
<td>0.47</td>
<td>0.46</td>
</tr>
<tr>
<td>$V_{18}$</td>
<td>61.25</td>
<td>23.68</td>
<td>19.07</td>
<td>16.91</td>
<td>15.83</td>
<td>0</td>
</tr>
<tr>
<td>$V_{19}$</td>
<td>180.00</td>
<td>62.81</td>
<td>50.52</td>
<td>44.05</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Bayesian network model can make use of these historical link flows and also consider the level of total traffic flow, which is really useful for many real-time traffic applications, especially in the ITS.

Using the Bayesian network model and updating the observed variables (including the observed link flows and those derived from the observed link flows and the conservation law) can modify the means and reduce the variances of the remaining variables. These updated means and variances allow us to obtain the posterior distribution of the unobserved variables based on those observed. Thus, the methods can provide not only point estimation but also the corresponding probability intervals. In addition, an incremental procedure is developed for solving the Bayesian network model without the intensive computation of matrix inversion, which can make this model apply easily in large-scale networks.

Moreover, in this paper, a normal distribution for traffic flows is assumed. It is reasonable because these random variables are the sum of a great number of independent Bernoulli experiments in which the users decide where to travel and which routes to choose. For future research, it is worthwhile to relax the normal distribution assumption.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.
Acknowledgments

This research is supported by the National Natural Science Foundation of China (no. 51778110 and no. 51378119) and Graduate Innovation Project of Jiangsu Province (no. CXZZ12_0113). Comments provided by anonymous referees are much appreciated.

References


