PREDICTING TRAFFIC FLOW IN ROAD NETWORKS
Using Bayesian Networks with Data from an Optimal Plate Scanning Device Location

S. Sánchez-Cambronero, A. Rivas, I. Gallego and J. M. Menéndez
Department of Civil Engineering, University of Castilla-La Mancha, 13071 Ciudad Real, Spain

Keywords: Traffic data updating, Gaussian Bayesian networks, Probability intervals, Plate scanning, RMARE.

Abstract: This paper deals with the problem of predicting route flows (and hence, Origin-Destination (OD) pair and link flows) and updating these predictions when plate scanned information becomes available. To this end, a normal Bayesian network is built which is able to deal with the joint distribution of route and link flows and the flows associated with all possible combinations of scanned link flows and associated random errors. The Bayesian network provides the joint density of route flows conditioned on the observations, which allow us not only the independent or joint predictions of route flows, but also probability intervals or regions to be obtained. A procedure is also given to select the subset of links to be observed in an optimal way. An example of application illustrate the proposed methodology and shows its practical applicability and performance.

1 INTRODUCTION

As is well known, in traffic problems several types of flows can be considered, such us route, OD-pair, link and node flows, but other types are also possible including disaggregated versions of flows, as OD-pair flows through a given link or node and with a given origin and/or destination, etc. In this paper, we aim at estimating route flows, because once they are estimated, other traffic flows such as OD-pair and link flows can be immediately calculated from the corresponding incidence matrices. We also deal with another type of flow, which corresponds to flows passing only through a given subset of links. This is the natural type of flow for the plate scanning technique.

In the traffic literature, several problems related to these types of traffic flows were studied, such as trip matrix estimation (ME) or traffic assignment problems. The assignment model has as inputs the OD pair flows and produces as outputs the probabilities of a traveler to select the different routes of an OD pair (see (Castillo et al., 2008c) for deterministic models and (Praskher and Bekhor, 2004) for stochastic models). On the contrary, the ME problem has as inputs these probabilities and link flows and produces as outputs the OD pair flows. However, these two problems are closely connected and some inconsistencies can occur in the traffic estimation if they are solved separately. To solve this problem, they can be combined into one problem, in which the trip matrix flow estimation and the traffic assignment problems become a single one. Several techniques have been proposed to solve this combined problem, which include, the well known bi-level approaches (see (Conejo et al., 2006)).

In the ME problem one tries to estimate origin-destinations (OD) trip matrices based on some observed link flows. Unfortunately, this is an under-specified problem, due to the fact that the number of paths between OD pairs is normally much larger than the number of observed links (i.e. the incidence matrix is not full rank), and there are infinitely many solutions satisfying the conservation laws. In order to have a unique solution, one has to provide more information by means of a prior OD trip matrix, which can come from many different sources, including an out-of-date, subjectively-guessed, or obtained by an alternative method, etc. Based on the observations (OD or link flows) and this prior OD matrix, the OD matrix can be estimated by many different methods, such as the least squares and the generalized least squares methods ((Doblas and Benitez, 2005)), the entropy or information based methods ((Van Zuylen and Willumsen, 1980)), and statistical based methods (see (Cascetta and Nguyen, 1988), for classical approaches, and (Tebaldi and West, 1996) for the Bayesian approaches). Some other works used

Sánchez-Cambronero S., Rivas A., Gallego I. and M. Menéndez J. (2010).
PREDICTING TRAFFIC FLOWIN ROAD NETWORKS - Using Bayesian Networks with Data from an Optimal Plate Scanning Device Location.
In *Proceedings of the 2nd International Conference on Agents and Artificial Intelligence - Artificial Intelligence*, pages 552-559
DOI: 10.5220/0002717705520559

Bayesian networks such as (Sun et al., 2006) and (Castillo et al., 2008b).

In this paper we have used the plate scanning technique in order to deal with the problems exposed above: the under-specification on one hand, and the inconsistences between OD matrix estimation and the assignment problems on the other (see (Castillo et al., 2008a)). The idea is to register the plate numbers of the circulating vehicles together with the corresponding times at some subsets of links and use this information to reconstruct vehicle routes. The plate scanning approach to traffic flow estimation and reconstruction has been revealed as a very promising alternative to other existing methods based on link flows or traffic surveys, as was done in other standard methods, because it provides much richer information about traffic flows than simply observing (counting) link flows (see (Watling, 1994) or (Castillo et al., 2008a)).

The new contribution of this paper consists of presenting a Bayesian network to estimate route flow based on plate scanning. It combines two recently developed techniques, Bayesian network and plate scanning, to predict traffic flows. Using both thechniques, the random dependence structure of traffic flows including not only OD-pair and link flows but route flows and flows associated with subsets of links are provided. In addition, a procedure is also given to select the subset of links to be observed in an optimal way subject to a given budget.

The paper is organized as follows. In section 2 the problem of selecting an optimal subset of links to be scanned is dealt with. Section 3 introduces Bayesian networks and describes the proposed model for route flow estimation. In Section 4 a example of applications is used to illustrate the effectiveness of the proposed model and clarify some of its implementation details. Finally, Section 5 provides some conclusions.

2 THE PLATE SCANNING DEVICE LOCATION PROBLEM

In real life, the true error or reliability of an estimated $\mathcal{OD}$ matrix is unknown. (Yang et al., 1991) proposed the concept of maximal possible relative error (MPRE), which represents the maximum possible relative deviation of the estimated $\mathcal{OD}$ matrix from the true one. Based on this concept (Yang and Zhou, 1998) proposed several location rules. In this paper, since the scanner location problem is of different nature to the counting location problem based on link flows, we derive an analogous formulation based on prior link and flow values and the following measure (RMSRE, root mean squared relative error):

$$\text{RMSRE} = \sqrt{\frac{1}{m}\sum_{i\in I}\left(\frac{t_i^0 - t_i}{t_i^0}\right)^2}, \qquad (1)$$

where t_i^0 and t_i are the prior and estimated flow of $\mathcal{OD}$-pair i, respectively, and m is the number of $\mathcal{OD}$-pairs belonging to the set $\mathcal{I}$. Note that we propose this alternative formulation because our model uses prior information and we also assume that the real network flows will be similar to those given by the prior approach, therefore our models try to reproduce through an estimation method the prior $\mathcal{OD}$ pair flows as exactly as possible, when other information is not available. Since the prior $\mathcal{OD}$ pair flows t_i^0 are known, they are used to calculate the relative error.

Given the set R of all possible routes, if R_i is the set of routes belonging to $\mathcal{OD}$-pair i, we have $t_i^0 = \sum_{r\in R_i} f_r^0$, and then the RMSRE can be expressed as:

$$\text{RMSRE} = \sqrt{\frac{1}{m}\sum_{i\in I}\left(\frac{t_i^0 - \sum_{r\in R_i} f_r^0 y_r}{t_i^0}\right)^2}, \qquad (2)$$

where y_r is a binary variable equal to one if route r is identified uniquely (observed) by the scanned links, and zero otherwise. Note that the minimum possible RMSRE-value corresponds to $y_r = 1;\ \forall r \in R$, where $t_i = t_i^0$ and RMSRE=0. However, if $n_{\text{sc}} = \sum_{r\in R} y_r \le n_r$ then RMSRE>0, and then, one interesting question is: how do we select the routes to be observed so that the RMSRE is minimized? From (2) we obtain

$$m\times\ \text{RMSRE}^2 = \sum_{i\in I}\left(1 - \sum_{r\in R_i}\frac{f_r^0}{t_i^0} y_r\right)^2, \qquad (3)$$

where it can be concluded that the bigger the value of $\sum_{r\in R_i}\frac{f_r^0}{t_i^0}y_r$ the lower the RMSRE. If the set of routes is partitioned into observed ($\mathcal{OR}$) and unobserved ($\mathcal{UR}$) routes associated with $y_r = 1$ or $y_r = 0$, respectively, (3) can be reformulated as follows

$$\begin{aligned} m\times\ \text{RMSRE}^2 &= \sum_{i\in I}\left(1 - \sum_{r\in(R_i\cap\mathcal{OR})}\frac{f_r^0}{t_i^0}\right)^2 \\ &= \sum_{i\in I}\left(\sum_{i\in(R_i\cap\mathcal{UR})}\frac{f_r^0}{t_i^0}\right)^2, \end{aligned} \qquad (4)$$

so that routes to be observed ($y_r = 1$) should be chosen minimizing (4).

The main shortcoming of equations (3) or (4) is their quadratic character which makes the RMSRE minimization problem to be nonlinear. Alternatively, the following RMARE (root mean absolute value relative error) based on the mean absolute relative error norm can be defined:

$$\text{RMARE} = \frac{1}{m}\sum_{i\in I}\left|\frac{t_i^0 - t_i}{t_i^0}\right| = \frac{1}{m}\sum_{i\in I}\left|\frac{t_i^0 - \sum_{r\in R_i} f_r^0 y_r}{t_i^0}\right|, \quad (5)$$

and since the numerator is always positive for error free scanners ($0 \le \sum_{r\in R_i} f_r^0 y_r \le t_i^0$; $\forall i \in I$), the absolute value can be dropped, so that the RMARE as a function of the observed and unobserved routes is

$$\begin{aligned}\text{RMARE} &= 1 - \frac{1}{m}\left(\sum_{i\in I}\sum_{r\in(R_i\cap\mathcal{OR})}\frac{f_r^0}{t_i^0}\right)\\ &= \frac{1}{m}\left(\sum_{i\in I}\sum_{r\in(R_i\cap\mathcal{UR})}\frac{f_r^0}{t_i^0}\right),\end{aligned} \quad (6)$$

which implies that minimizing the RMARE is equivalent to minimizing the sum of relative route flows of unobserved routes, or equivalently, maximize the sum of relative route flows of observed routes.

Note also that even though the knowledge of prior $\mathcal{OD}$ pair flows could be difficult in practical cases, the aim of the proposed formulation is determining which $\mathcal{OD}$ flows are more important than others in order to prioritize their real knowledge. In fact the prior $\mathcal{OD}$ matrix is only used as a weighting factor for O-D pairs flows. Alternatively, the MPRE criterion proposed by (Yang et al., 1991) could be used for those cases where a prior O-D matrix is unavailable. Note that existing methods such us the one proposed by (Yang and Zhou, 1998) and according to their maximal flow-interception rule, also use a flow pattern associated with a prior O-D matrix.

The first location model to be proposed in this paper considers full route observability, i.e. RMSRE= 0, but including budget considerations. In the transport literature, each link, denoted by a, is considered independently of the number of lanes it has. Obviously, when trying to scan plate numbers links with higher number of lanes are more expensive (usually the number of scanning devices is bigger):

$$M_1 = \underset{z}{\text{Minimize}} \sum_{a\in\mathcal{A}} \mathcal{P}_a z_a \quad (7)$$

subject to

$$\sum_{a\in\{\mathcal{A}\}}(\lambda_a^r + \lambda_a^{r_1})(1 - \lambda_a^r\lambda_a^{r_1})z_a \ge 1 \left\{\begin{array}{l}\forall (r,r_1)|r<r_1\\ \sum_{a\in\mathcal{A}}\lambda_a^r\lambda_a^{r_1} > 0\end{array}\right. \quad (8)$$

$$\sum_{a\in\mathcal{A}} z_a\lambda_a^r \ge 1; \forall r, \quad (9)$$

where z_a is a binary variable taking value 1 if the link a is scanned, and 0, otherwise, r and r_1 are paths, Λ is the route incidence matrix with elements λ_a^r.

Note that constraint (8) guarantees that the selected subset of scanned links is able to distinguish the users of any given pair of paths r and r_1 based on their scanned plate numbers, i.e. there exists at least one scanned link which is in path r and not in path r_1 or vice-versa. In addition, constraint (9) ensures that any route or path contains at least one scanned link, and therefore information, not only of all $\mathcal{OD}$ pairs but all the routes, becomes available. $\mathcal{P}_a$ is the cost of plate scanning link a. Note that constraint (8) forces to select the scanned links so that every route is uniquely defined by a given set of scanned links (every row in the incidence matrix Λ is different from the others) and (9) ensures that at least one link for every route is scanned (every row in the incidence matrix Λ contains at least one element different from zero). Both constraints (8) and (9) force to observe the maximum relative route flow and provide the full identifiability of observed path flows. Note also that all $\mathcal{OD}$ pairs are totally covered. In addition, this model allows us the estimation of the required budget resources $\mathcal{B}^* = \sum_{a\in\mathcal{A}} \mathcal{P}_a z_a^*$ for covering all $\mathcal{OD}$ pairs in the network which obviously must be the minimum for full identifiability of routes. However, budget is limited in practice, meaning that some $\mathcal{OD}$ pairs or even some routes may remain uncovered, consequently based on (6) the following model is proposed in order to observe the maximum relative route flow:

$$M_2 = \underset{\mathbf{y},z}{\text{Maximize}} \sum_{\forall i\in I}\sum_{r\in R_i}\frac{f_r^0}{t_i^0}y_r \quad (10)$$

subject to

$$\sum_{a\in\{\mathcal{A}\}}(\lambda_a^r + \lambda_a^{r_1})(1 - \lambda_a^r\lambda_a^{r_1})z_a \ge y_r \left\{\begin{array}{l}\forall (r,r_1)|r<r_1\\ \sum_{a\in\mathcal{A}}\lambda_a^r\lambda_a^{r_1} > 0\end{array}\right. \quad (11)$$

$$\sum_{a\in\mathcal{A}} z_a\delta_a^r \ge y_r; \quad \forall r, \quad (12)$$

$$\sum_{a\in\mathcal{A}} \mathcal{P}_a z_a \le \mathcal{B}, \quad (13)$$

where f_r^0 and t_i^0 are the route and $\mathcal{OD}$-pair flows, respectively, of a prior $\mathcal{OD}$ matrix, y_r is a binary variable equal to 1 if route r can be distinguished from others and 0 otherwise, z_a is a binary variable which is 1 if link a is scanned and 0 otherwise, and $\mathcal{B}$ is the available budget.

Constraint (11) guarantees that the route r is able to be distinguished from the others if the binary variable y_r is equal to 1. Constraint (12) ensures that the route which is able to be distinguished contains at least one scanned link. Both constraints (11) and (12) ensure that all routes such that $y_r = 1$ can be uniquely identified using the scanned links. It is worthwhile mentioning that using y_r instead of 1 in the right hand side of constraints (11) and (12) immediately converts into inactive the constraint (9) for those routes the flow of which are not fully identified.

Note that the full identifiability of observed path flows is included in the optimization itself and it will be ensured or not depending on the available budget $\mathcal{B}$, i.e. depending on whether or not constraint (13) becomes active. For example, if the available budget equals the optimal value of the objective function given by model M_1 ($\mathcal{B} = \mathcal{B}^*$), model M_2 provides full $\mathcal{OD}$ coverage. Note also that previous models can be easily modified in order to include some practical considerations as for example the fact that some detectors are already installed and additional budget is available. To do that one only need to include the following constraint to models M_1 or M_2

$$z_a = 1; \quad \forall a \in \mathcal{OL}. \tag{14}$$

where $\mathcal{OL}$ is the set of already observed links (links with scanning devices already installed).

3 THE PROPOSED MODEL FOR TRAFFIC PREDICTION

Bayesian network models have been used frequently to solve a wide range of practical problems (see, for example, (Castillo et al., 1995), (Bouckaert et al., 1996), (Castillo et al., 1996), (Castillo et al., 1999), or (Castillo and Kjaerulff, 2003)). In this section we have used the Bayesian network tool to built a model for traffic prediction using data from plate scanning devices. In addition a detailed description and justification of its main assumptions are presented.

3.1 Bayesian Networks

A Bayesian network is a pair $(\mathcal{G}, \mathcal{P})$, where $\mathcal{G}$ is a directed acyclic graph of a set of nodes $\mathbf{X}$, which are the random variables, and a set $\mathcal{P} = \{p(x_1|\pi_1), \ldots, p(x_n|\pi_n)\}$ of n conditional probability densities, where Π_i is the set of parents of node X_i in $\mathcal{G}$. The graph $\mathcal{G}$ contains qualitative information about the relationships among the variables, and $\mathcal{P}$ contains the quantitative information and defines the associated joint probability density of all nodes as

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|\pi_i). \tag{15}$$

Bayesian networks are very useful to represent the statistical relationship among multivariate random variables. In particular, (Sun et al., 2006), (Castillo et al., 2008b; Castillo et al., 2008c) apply Bayesian networks to traffic flow problems.

Bayesian networks have a high practical interest because: (a) the conditional independence relations among the $\mathbf{X}$ variables can be inferred directly from the graph $\mathcal{G}$, which is relevant to given variables when the knowledge of other variables become available, and (b) the updating of probabilities can be very easily done when new variables become known.

In this paper we use Gaussian Bayesian networks (GBN), that is, Bayesian networks such that their joint probability distributions of all their variables are multivariate normal $N(\mu, \Sigma)$ distributions. This assumption is very common in the transport literature

3.2 Updating Information in GBN after having Evidences

When one works with Gaussian Bayesian networks, it is possible to introduce the observed value of several variables of the network an computing the probability distribution of the rest of variables.

Let Y and Z be two sets of random variables representing unobserved and observed variables, respectively, and having a multivariate Gaussian distribution with mean vector and covariance matrix given by

$$\mu = \begin{pmatrix} \mu_Y \\ \mu_Z \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix},$$

respectively, where μ_Y and Σ_{YY} are the mean vector and covariance matrix of Y, μ_Z and Σ_{ZZ} are the mean vector and covariance matrix of Z, and Σ_{YZ} is the covariance of Y and Z. Then the conditional probability distribution (CPD) of Y given $Z = z$ (the evidence) is multivariate Gaussian with mean vector $\mu_{Y|Z=z}$ and covariance matrix $\Sigma_{Y|Z=z}$ that are given by

$$\mu_{Y|Z=z} = \mu_Y + \Sigma_{YZ}\Sigma_{ZZ}^{-1}(z - \mu_Z), \tag{16}$$

$$\Sigma_{Y|Z=z} = \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}. \tag{17}$$

Note that the conditional mean $\mu_{Y|Z=z}$ depends on z but the conditional variance $\Sigma_{Y|Z=z}$ does not. Therefore equations (16) and (17) suggest a procedure to calculate the means and variances of any subset of variables $Y \subset X$, given a set of evidential nodes $Z \subset X$ whose values are $Z = z$.

3.3 Model Assumptions

Assuming the route flows are multivariate random variables, we build a Gaussian Bayesian network using the special characteristics of traffic flow variables. To this end, we consider the route flows as parents and the subsets of scanned link flows as children and reproduce the conservation law constraints in an exact or statistical (i.e., with random errors) form. In our Gaussian Bayesian network model we make the following assumptions:

Assumption 1. It is clear that the **F** of route flows random variables are correlated. For example, during peak commuting periods traffic increases for all routes and strong weather conditions reduce traffic flows in all routes. In order to represent these correlations and obtain the associated variance-covariance matrix, we make the following assumption:

$$F_r = k_r U + \eta_r, \tag{18}$$

where $k_r, r = 1,\ldots,m$ are positive real constants, U is a normal random variable $N(\mu_U, \sigma_U^2)$, and η_r are independent normal $N(0,\gamma_r^2)$ random variables. The meanings of these variables are as follows:

- U: Random positive variable that measures the level of total mean flow. This means that flow varies randomly and deterministically in situations similar to those being analyzed (weekend period, labor day, beginning or end of a holiday period, etc.).
- **K**: Column matrix whose element k_r measures the relative weight of the route r flow with respect to the total traffic flow (including all routes). It measures the importance or level of traffic flow associated with route r (the larger the value of k_r, the larger the flow traffic in route r).
- η : Vector of independent random variables with null mean such that its r element measures the variability of the route r flow with respect to its mean.

Then, we have

$$\mathbf{F} = \left(\mathbf{K} \quad | \quad \mathbf{I} \right) \begin{pmatrix} U \\ -- \\ \eta^T \end{pmatrix} \tag{19}$$

and the variance-covariance matrix $\Sigma_{\mathbf{F}}$ of **F** is

$$\Sigma_{\mathbf{F}} = \left(\mathbf{K} \quad | \quad \mathbf{I} \right) \Sigma_{(U,\eta)} \begin{pmatrix} \mathbf{K}^T \\ -- \\ \mathbf{I} \end{pmatrix} \tag{20}$$

$$= \sigma_U^2 \mathbf{K}\mathbf{K}^T + \mathbf{D}\eta, \tag{21}$$

where the matrices $\Sigma_{(U,\eta)}$ and $\mathbf{D}\eta$ are diagonal.

Assumption 2. The flows associated with the combinations of scanned link flows and counted link flows can be written as

$$\mathbf{W} = \Delta\mathbf{F} + \varepsilon, \tag{22}$$

where the W_s variables represent the traffic flow associated with each feasible combination of scanned links, which is related to the route flows; δ_{sr} (element of matrix Δ) is 1 if route r contains all and only the links associated with W_s, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ are mutually independent normal random variables, independent of the random variables in **F**, and ε_s has mean $E(\varepsilon_s)$ and variance $\psi_s^2; s = 1,2,\ldots,n$. The ε_s represents the error in the corresponding subset of scanned links. In particular, they can be assumed to be null i.e. the plate data is assumed error free.

Then, we have

$$\begin{pmatrix} \mathbf{F} \\ \mathbf{W} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & | & \mathbf{0} \\ - & + & - \\ \Delta & | & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{F} \\ -- \\ \varepsilon \end{pmatrix},$$

which implies that the mean $E[(\mathbf{F},\mathbf{W})]$ is

$$E[(\mathbf{F},\mathbf{W})] = \begin{pmatrix} E(U)\mathbf{K} \\ -------- \\ E(U)\Delta\mathbf{K} + E(\varepsilon) \end{pmatrix}, \tag{23}$$

and the variance-covariance matrix of $(\mathbf{F},\mathbf{W})$ is

$$\Sigma_{(\mathbf{F},\mathbf{W})} = \begin{pmatrix} \Sigma_{\mathbf{F}} & | & \Sigma_{\mathbf{F}}\Delta^T \\ -- & + & ---- \\ \Delta\Sigma_{\mathbf{F}} & | & \Delta\Sigma_{\mathbf{F}}\Delta^T + \mathbf{D}_\varepsilon \end{pmatrix}. \tag{24}$$

All these assumptions imply that the joint PDF of $(F_1, F_2, \ldots, F_m, W_1, W_2, \ldots, W_n)$ can be written as

$$f(f_1, f_2, \ldots, f_m, w_1, w_2, \ldots, w_n) = f_{N(\mu_{\mathbf{F}},\Sigma_{\mathbf{F}})}(f_1, f_2, \ldots, f_m) \prod_{s=1}^{n} f_{N(\mu_s + \sum_{r\in\Pi_s} \Delta_{sr}(f_r - \mu_{F_r}), \psi_s^2)}(w_s) \ . \tag{25}$$

To complete our Bayesian network model we need to define the graph. Any probability distribution can be represented by a directed graph. The only problem, to build the Bayesian network graph, is the number of links required, that can be large if the order of nodes is not adequately chosen.

In this paper we give what we think is the most convenient graph (see Fig. 2): the route flows F_r are the parents of all link flow combinations W_s used by the corresponding travelers, and the error variables are the parents of the corresponding flows, that is, the ε_s are the parents of the W_s, and the η_r are the parents of the F_r. Finally, the U variable is on top (parent) of all route flows, because it gives the level of them (high, intermediate or low). This solves the problem of "parent" being not well defined, without the need for recursion - in general graphs, one could seemingly have a "deadlock" situation in which it is not clear what node is the parent of which other nodes.

In this paper we consider the simplest version of the proposed model, which considers only the route flows, and the scanned link flow combinations. Therefore, a further requires that a model with all variables must be built i.e. including the mean and variance matrix of the all variables (U, $\eta_r; r = 1,2,\ldots,m$ and $\varepsilon_s; s = 1,2,\ldots,n$).

3.4 Traffic Prediction

Once we have built the model, we can use its JPD (25) to predict route and link traffic flows when some information becomes available. The idea consists of using the joint distribution of routes flows conditioned on the available information. In fact, since the remaining variables (those not known) are random, the most informative item we can get is its conditional joint distribution, and this is what the Bayesian network methodology supplies. In this section we propose an step by step method to implement the plate scanning-Bayesian network model:

Step 0: Initialization Step. Assume an initial $\mathbf{K}$ matrix (for example, obtained from solving a SUE problem for a given out-of-date prior OD-pair flow data), the values of $E[U]$ and σ_U, and the matrices $\mathbf{D}_\varepsilon$ and $\mathbf{D}_\eta$.

Step 1: Select the Set of Links to be Scanned. The set of links to be scanned must be selected. This paper deals with this problem in Section 2 providing several methods to select the best set of links to be scanned.

Step 2: Observe the Plate Scanning Data. The plate scanning data w_s are extracted.

Step 3: Estimate the Route Flows. The route matrix $\mathbf{F}$ with elements f_r are estimated using the Bayesian network method, i.e., using the formulas (see (16) and (17)):

$$E[\mathbf{F}] = E[U]\mathbf{K} \quad (26)$$

$$E[\mathbf{W}] = E[U]\Delta\mathbf{K} + E[\varepsilon] \quad (27)$$

$$\mathbf{D}_\eta = \text{Diag}\left(\nu E[\mathbf{F}]\right), \quad (28)$$

$$\Sigma_{\mathbf{FF}} = \sigma_U^2 \mathbf{K}\mathbf{K}^T + \mathbf{D}_\eta \quad (29)$$

$$\Sigma_{\mathbf{FW}} = \Sigma_{\mathbf{FF}}\Delta^T \quad (30)$$

$$\Sigma_{\mathbf{WF}} = \Sigma_{\mathbf{FW}} \quad (31)$$

$$\Sigma_{\mathbf{WW}} = \Delta\Sigma_{\mathbf{FF}}\Delta^T + \mathbf{D}_\varepsilon \quad (32)$$

$$E[\mathbf{F}|\mathbf{W}=\mathbf{w}] = E[\mathbf{F}] + \Sigma_{\mathbf{FW}}\Sigma_{\mathbf{WW}}^{-1}(\mathbf{w} - E[\mathbf{W}]) \quad (33)$$

$$\Sigma_{\mathbf{F}|\mathbf{W}=\mathbf{w}} = \Sigma_{\mathbf{FF}} - \Sigma_{\mathbf{FW}}\Sigma_{\mathbf{WW}}^{-1}\Sigma_{\mathbf{WF}} \quad (34)$$

$$E[\mathbf{W}|\mathbf{W}=\mathbf{w}] = \mathbf{w} \quad (35)$$

$$\Sigma_{\mathbf{W}|\mathbf{W}=\mathbf{w}} = \mathbf{0} \quad (36)$$

$$\mathbf{F} = E[\mathbf{F}|\mathbf{W}=\mathbf{w}]|_{(\mathbf{F},\mathbf{W})=F} \quad (37)$$

where ν is the coefficient of variation selected for the η variables, and we note that $\mathbf{F}$ and $\mathbf{W}$ are the unobserved and observed components, respectively.

Step 4. Obtain the F Vector. Return the the f_r route flows as the result of the model. Note that from $\mathbf{F}$ vector, the rest of traffic flows (link flows and OD pair flows) can be easily obtained.

4 EXAMPLE OF APPLICATIONS

In this section we illustrate the proposed methods by their application to a simple example. We assume that plate scanning traffic data have no errors.

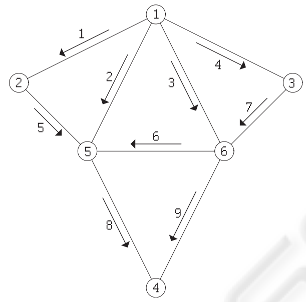

Figure 1: The elementary example network.

Table 1: Required data for the simple example.

OD	path code (r)	Links
1-4	1	1 5 8
1-4	2	2 8
1-4	3	3 9
1-4	4	3 6 8
1-4	5	4 7 9
1-4	6	4 7 6 8
2-4	7	5 8
2-4	8	7 6 8
3-4	9	7 9

set code (s)	Scanned links 1	2	3	4	7	8
1	X					X
2		X				X
3			X			
4			X			X
5				X	X	
6				X	X	X
7						X
8					X	X
9					X	

Consider the network in Fig. 1 with the routes and OD-pairs in Table 1, which shows the feasible combination of scanned links after solving the M_1 model.

As described in Section 3, the graph of the associated Bayesian network is shown in Fig. 2 for $\mathcal{SL} = \{1,2,3,4,7,8\}$. Note that the route node F_r has as parents only node U and η_r, and any flow from plate scanning data node W_s has its associated routes as parents, i.e., those routes with all and only all the corresponding subset of scanned links (see Table 1). Next, the proposed method in Section 3 is applied.

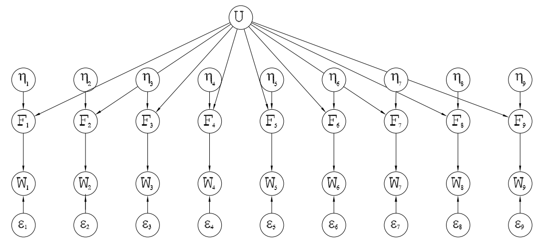

Figure 2: BN associated with the example.

Step 0: Initialization Step. To have a reference flow, we have considered that the true route flows are those shown in the second column of Table 2. The assumed mean value was $E[U] = 10$ and the value of σ_U is 8. The initial matrix $\mathbf{K}$ is obtained by multiplying each true route flow by an independent random uniform $U(0.4,1.3)/10$ number. The $\mathbf{D}_\varepsilon$ is assumed diagonal matrix, the diagonal of which are almost null (0.000001) because we have assumed error free in the plate scanning process. $\mathbf{D}_\eta$ is also a diagonal matrix which values are associated with a variation coefficient of 0.4.

Step 1: Select the Set of Links to be Scanned. The set of links to be scanned have been selected using the M_2 model for different available budget, i.e. using the necessary budget for the devices needed to be installed in the following links:

$$\mathcal{SL} \equiv \{1,2,3,4,7,8\};\ \mathcal{SL} \equiv \{1,4,5,7,9\};$$
$$\mathcal{SL} \equiv \{1,4,7,9\};\ \mathcal{SL} \equiv \{4,7,9\};$$
$$\mathcal{SL} \equiv \{1,5\};\ \mathcal{SL} \equiv \{2\}.$$

Step 2: Observe the Plate Scanning Data. The plate scanning data W_s is obtained by scanning the selected links (a detailed explanation of how this can be done appears in (Castillo et al., 2008a)).

Step 3: Estimate the Route Flows. The route flows $\mathbf{F}$ with elements f_r are estimated using the Bayesian network method and the plate scanning data, i.e., using the formulas (26)-(37)

Table 2: Route flow estimates using BN and LS approaches.

Route	True flow	Method	Scanned links 0	1	2	3	4	5	6
1	5.00	BN	4.26	4.35	5.00	4.91	5.00	5.00	5.00
		LS	4.26	4.26	5.00	4.26	5.00	5.00	5.00
2	7.00	BN	6.84	7.00	7.76	7.89	7.91	7.85	7.00
		LS	6.84	7.00	6.84	6.84	6.84	6.84	7.00
3	3.00	BN	3.45	3.52	3.91	3.00	3.00	3.00	3.00
		LS	3.45	3.45	3.45	3.00	3.00	3.00	3.00
4	5.00	BN	3.00	3.07	3.41	3.46	3.47	3.45	5.00
		LS	3.00	3.00	3.00	3.00	3.00	3.00	5.00
5	6.00	BN	5.36	5.47	6.08	6.00	6.00	6.00	6.00
		LS	5.36	5.36	5.36	6.00	6.00	6.00	6.00
6	4.00	BN	3.37	3.45	3.82	4.00	4.00	4.00	4.00
		LS	3.38	3.38	3.38	4.00	4.00	4.00	4.00
7	10.00	BN	8.90	9.08	10.00	10.25	10.28	10.00	10.00
		LS	8.90	8.90	10.00	8.90	8.90	10.00	10.00
8	7.00	BN	3.97	4.06	4.50	7.00	7.00	7.00	7.00
		LS	3.97	3.97	3.97	7.00	7.00	7.00	7.00
9	5.00	BN	5.45	5.57	6.18	5.00	5.00	5.00	5.00
		LS	5.45	5.45	5.45	5.00	5.00	5.00	5.00

The method has been repeated for different subsets of scanned links shown in step 2 of the process. The resulting predicted route flows are shown in Table 2. The first rows correspond to the route predictions using the proposed model. With the aim of illustrating the improvement resulting from the plate scanning technique using Bayesian networks when compared with the standard method of Least Squares (LS), for example (see (Castillo et al., 2008a)), we have implemented this model using the same data. The results appear in the second rows in Table 2. A comparison of the results obtained from both methods confirm that the plate scanning method using Bayesian networks outperforms the standard method of Least Squares for several reasons:

- The BN tool provides the random dependence among all variables. This fact allows us improve the route flow predictions even though when we have no scanned link belonged to this particular route. Note that using the LS approach the prediction is the prior flow (the fourth column in Table 2, i.e with 0 scanned links in the network)
- The BN tool provides not only the variable prediction but also the probability intervals for these predictions using the JPD function. Fig. 3 shows the conditional distributions of the route flows the different items of accumulated evidence. From left to right and from top to bottom $F_1, F_2 \ldots$ predictions are shown. In each subgraph the dot represents the real route flow in order to analyze the predictions.

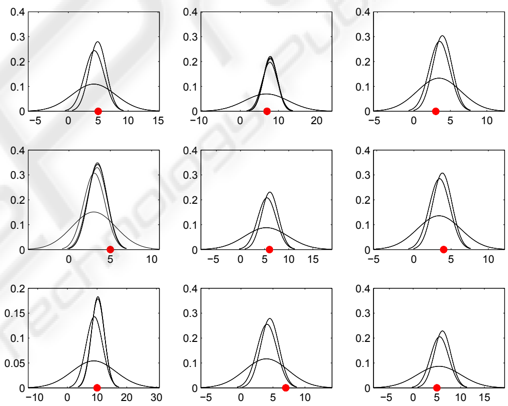

Figure 3: Conditional distribution of the route flows.

It is necessary to point out that the proposed models have been applied to real size networks as for example the city of Cuenca (Spain) but the results cannot be showed for space problems. The network consists of 672 links, 232 nodes, 139 OD-pairs and 528 routes. In this network, 100 scanned links are sufficient for full observability using M_1 proposed model.

5 CONCLUSIONS

The main conclusions that can be drawn from this paper are the following:

1. Bayesian networks are very natural tools for reproducing the random dependence structure of traffic flows including not only OD-pair and link flows but route flows and flows associated with subsets of links. Therefore, the combination of

Bayesian networks and scanned link flows seems to be a very good and practical tool to predict traffic flows. The example in this paper illustrate the improvement of this combination when combined with other methods and shows that it outperforms other alternatives.

2. The updating techniques for Bayesian networks allow us obtaining the distribution of route flows conditioned by the observed flows, accounting for all the information available (evidences).
3. The knowledge of plate scanned observations modify substantially the means and reduces the variance of the route flows leading to more precise predictions, which improve with increasing number of scanned links and can be exact for an adequate selection of the set of scanned and counted links i.e. using the M_1 proposed model.
4. Several models have been presented for an adequate selection and location of plate scanning devices including budget constraints together with the consideration of already existing devices. In addition, they allow us improving the route flow estimations.
5. Errors in scanned links can produce important alterations of the parameters estimates, because several users of several routes can be confounded if they are not been observed in some links. This aspect is not the focus of this paper and its full treatment will be dealt with in an outgoing work. In any case one approach for solving this problem is treated in (Castillo et al., 2008a).

REFERENCES

Bouckaert, R., Castillo, E., and Gutiérrez, J. M. (1996). A modified simulation scheme for inference in Bayesian networks. *International Journal of Approximate Reasoning*, 14(1):55–80.

Cascetta, E. and Nguyen, S. (1988). A unified framework for estimating or updating origin/destination matrices from traffic counts. *Transportation Research*, 22B:437–455.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1995). Parametric structure of probabilities in bayesian networks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU 95)*, pages 89–98. Symbolic and Quantitative Approaches to Reasoning and Uncertainty.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1996). A new method for efficient symbolic propagation in discrete bayesian networks. *Networks*, 28(1):31–43.

Castillo, E. and Kjaerulff, U. (2003). Sensivity analysis in Gaussian Bayesian networks using a symbolic-numerical technique. *Reliability Engineering and System Safety*, 79/2:139–148.

Castillo, E., Menéndez, J. M., and Jiménez.P (2008a). Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations. *Transportation Research B*, 42:455–481.

Castillo, E., Menéndez, J. M., and Sánchez-Cambronero, S. (2008b). Predicting traffic flow using Bayesian networks. *Transportation Research B*, 42:482–509.

Castillo, E., Menéndez, J. M., and Sánchez-Cambronero, S. (2008c). Traffic estimation and optimal counting location without path enumeration using Bayesian networks. *Computer Aided Civil and Infraestructure Engineering*, 23:189–207.

Castillo, E., Sarabia, J. M., Solares, C., and Gómez, P. (1999). Uncertainty analyses in fault trees and Bayesian networks using form/sorm methods. *Reliability Engineering and System Safety*, 65:29–40.

Conejo, A., Castillo, E., Mínguez, R., and García-Bertrand, R. (2006). *Decomposition Techniques in Mathematical Programming. Engineering and Science Applications*. Springer, Berlin.

Doblas, J. and Benitez, F. G. (2005). An approach to estimating and updating origin-destination matrices based upon traffic counts preserving the prior structure of a survey matrix. *Transportation Research*, 39B:565–591.

Praskher, J. N. and Bekhor, S. (2004). Route choice models used in the stochastic user equilibrium problem: a review. *Transportation Reviews*, 24:437–463.

Sun, S. L., Zhang, C. S., and Yu, G. Q. (2006). A Bayesian network approach to traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):124–132.

Tebaldi, C. and West, M. (1996). Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association*, 93:557–576.

Van Zuylen, H. J. and Willumsen, L. (1980). Network tomography: The most likely trip matrix estimated from traffic-counts. *Transportation Research B*, 14:291–293.

Watling, D. (1994). Maximum likelihood estimation of an origindestination matrix from a partial registration plate survey. *Transportation Research*, 28B:289–314.

Yang, H., Iidia, Y., and Sasaki, T. (1991). An analysis of the realiability of an origin-destination trip matrix estimated from traffic counts. *Transportation Research*, 25B:351–363.

Yang, H. and Zhou, J. (1998). Optimal traffic counting locations for origin-destination matrix extimation. *Transportation Research*, 32B:109–126.