We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

154
Countries delivered to

Our authors are among the
TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

19

Bayesian networks methods for traffic flow prediction

Enrique Castillo
University of Cantabria
Spain

Santos Sánchez-Cambronero and José María Menéndez
University of Castilla-La Mancha
Spain

During the last decades, "Transport Demand" and "Mobility" has been a continuously developing branch in the transport literature. This is reflected in the great amount of research papers published in scientific magazines dealing with trip matrix estimation (see (Doblas & Benítez, 2005)) and traffic assignment problems (see (Praskher & Bekhor, 2004)). Current traffic models reproduce the mobility using several data inputs, in particular prior trip matrix, link counts, etc. (see (Yang & Zhou, 1998)) which are data only from a subset of the problem variables, and its size will depend on the available budget for the study being carried on. Among the problems faced for solving these models we can emphasize the high number of possible solutions which is usually solved by choosing the solution where the model results best fits real data. Nevertheless a model does not have to reproduce only the real data, but also must reproduce accurately all the variables. To this end, the aim of this paper consists of presenting two Bayesian network models for traffic estimation, trying to bring a new tool to the transportation field. The first one deals with the problem of link flows, trip matrix estimation and traffic counting location and in the second one we propose a Bayesian network for route flow estimation (and hence link flows and $\mathcal{OD}$ flows) using data from plate scanning technique together with a model for optimal plate scanning device location. Since a Gaussian Bayesian network is used, these models allow us to update the predictions from a small subset of real data and probability intervals or regions are obtained to get an idea of the associated uncertainties. In addition dealing with data from the plate scanning approach we improve the under-specification level of the traffic flow estimation problem.

1. Some background on Bayesian network models

A Bayesian network (see, (Castillo et al., 1999)) is a pair $(\mathcal{G}, \mathcal{P})$, where $\mathcal{G}$ is a directed acyclic graph (DAG) defined on a set of nodes $\mathbf{X}$ (the random variables), and $\mathcal{P} = \{p(x_1|\pi_1), \ldots, p(x_n|\pi_n)\}$ is a set of n conditional probability densities (CPD), one for each variable, and Π_i is the set of parents of node X_i in $\mathcal{G}$. The set P defines the associated joint probability density of all nodes as

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

The graph $\mathcal{G}$ contains all the qualitative information about the relationships among the variables, no matter which probability values are assigned to them[1]. Complementary, the probabilities in $\mathcal{P}$ contain the quantitative information, i.e., they complement the qualitative properties revealed by the graphical structure.

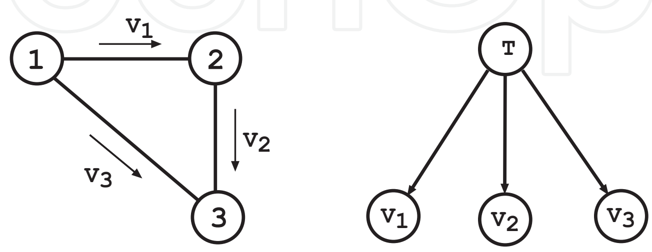


Fig. 1. A traffic network and its associated Bayesian network.

As an example of how a traffic network can be represented by means of $\mathcal{G}$ let us consider the simple traffic network in Figure 1. Assume that we have only the $\mathcal{OD}$ pair $(1,3)$ and two routes $\{(1,2),(3)\}$. Then, it is clear that the link flows v_1, v_2 and v_3 depend on the $\mathcal{OD}$ flow t, leading to the Bayesian network in the right part of Figure 1, where the arrows go from parents to sons. Note that link flow v_a has the t $\mathcal{OD}$ flow as a parent, and the t $\mathcal{OD}$ flow has v_a as son, if link a is contained in at least one path of such a $\mathcal{OD}$ pair.

Since a normal model is going to be used, the particular case of Gaussian Bayesian networks is presented next. A Bayesian network $(\mathcal{G},\mathcal{P})$ is said to be a Gaussian Bayesian network (see (Castillo et al., 1997a;b)) if and only if the joint probability distribution (JPD) associated with its variables X is a multivariate normal distribution, $N(\mu,\Sigma)$, i.e., with joint probability density function:

$$f(x) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left\{-1/2(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \tag{2}$$

where μ is the n-dimensional mean vector, Σ is the $n \times n$ covariance matrix, $|\Sigma|$ is the determinant of Σ, and μ^T denotes the transpose of μ.

In transportation problems, when some variables are observed, one needs to consider the other variables conditioned on the observations, and then the remaining variables change expected values and covariances. The following equations permit updating the mean and the covariance matrix of the variables when some of them are observed. They illustrate the basic concepts underlying exact propagation in Gaussian network models (see (Anderson, 1984)). These updating equations are:

$$\mu_{Y|Z=z} = \mu_Y + \Sigma_{YZ}\Sigma_{ZZ}^{-1}(z-\mu_Z), \tag{3}$$

$$\Sigma_{Y|Z=z} = \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}, \tag{4}$$

[1] This allows determining which information is relevant to given variables when the knowledge of other variables becomes available. As we will see in section 2.3 this fact is the basis of the traffic counts location problem.

where **Y** and **Z** are the set of unobserved and observed variables, respectively, which allow calculating the conditional means and variances given the actual evidence, i.e. they are the updating equations for the means and variance-covariance matrix of the unobserved variables and the already deterministic values of observed variables, when the later have been observed.

Note that instead of using a single process with all the evidences, one can incorporate the evidences one by one. In this way, one avoids inverting the matrix $\mathbf{\Sigma_{ZZ}}$, which for some solvers can be a problem because of its size. Note also that the conditional mean $\mu_{Y|Z=z}$ depends on z but the conditional variance $\Sigma_{Y|Z=z}$ does not.

2. Traffic link count based method for traffic flow prediction

In this section a method for traffic flow prediction using Bayesian networks with data from traffic counts is presented (see (Castillo, Menéndez & Sánchez-Cambronero, 2008a)).

2.1 Model assumptions

In our Gaussian Bayesian network for traffic flow prediction using data from traffic counts, we make the following assumptions:

Assumption 1: The vector **T** with elements t_{ks} of $\mathcal{OD}$ flows from origin k to destination s, are multivariate normal $N(\boldsymbol{\mu}_T, \mathbf{\Sigma}_T)$ random variables with mean $\boldsymbol{\mu}_T$ and variance-covariance matrix $\mathbf{\Sigma}_T$.

At this point we have to note that the **T** random variables are correlated. In particular, at the beginning and end of vacation periods the traffic increases for all $\mathcal{OD}$ pairs and strong weather conditions reduce traffic flows in all $\mathcal{OD}$ pairs. This fact can be formulated as follows:

$$t_{ks} = \zeta_{ks} U + \eta_{ks}, \tag{5}$$

where ζ_{ks} are positive real constants, U is a normal random variable $N(\mu_U, \sigma_U^2)$, and η_{ks} are independent normal $N(0, \gamma_{ks}^2)$ random variables. The meanings of these variables are as follows:

- U : A 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 general vacation period, etc.).
- $\boldsymbol{\zeta}$: A column matrix which element ζ_{ks} measures the relative weight of the traffic flow between origin k to destination s with respect to the total traffic flow (including all $\mathcal{OD}$ pairs).
- $\boldsymbol{\eta}$: A vector of independent random variables with null mean such that its ks element measures the variability of the $\mathcal{OD}$ pair ks flow with respect to its mean.

Assumption 2: The conditional distribution of each link flow **V** given the $\mathcal{OD}$ flows is the following normal distribution

$$v_{ij}|\mathbf{T} \sim N\left(\mu_{v_{ij}} + \sum_{k,s \in \Pi_{ij}} \beta_{ijks}(t_{ks} - \mu_{t_{ks}}), \psi_{ij}^2\right),$$

where v_{ij} is the traffic flow in link l_{ij}, β_{ijks} is the regression coefficient of v_{ij} on t_{ks}, which is zero if the link l_{ij} does not belong to any path of the $\mathcal{OD}$ pair ks, ψ_{ij}^2 is its variance and Π_{ij} is the set of parents of link l_{ij}.

Note that this forces our model to satisfy the flow conservation laws. If there are no errors in measurements, that is, $\psi_{ij}^2 = 0; \forall l_{ij} \in \mathcal{A}$, where $\mathcal{A}$ is the set of links, the conservation laws hold exactly. If errors are allowed ($\psi_{ij}^2 \neq 0$) they are statistically satisfied.
Note that this regression relationship, comes from the well known flow equilibrium equation:

$$v_{ij} = \sum_{ks} t_{ks} \left(\sum_r p_r^{ks} \delta_{ijr}^{ks} \right), \tag{6}$$

where t_{ks} and are the flows of the $\mathcal{OD}$ ks, p_r^{ks} is the probability of the user travelling from k to s to choose the path r, and δ_{ijr}^{ks} is the incidence matrix, i.e., it takes value 1 if link l_{ij} belongs to path r of the $\mathcal{OD}$ pair ks, and 0, otherwise.
In fact, Equation (6) can be written as

$$v_{ij} = \sum_{ks} t_{ks} \left(\sum_r p_r^{ks} \delta_{ijr}^{ks} \right) = \sum_{ks} \beta_{ijks} \mu_{t_{ks}} + \sum_{ks} \beta_{ijks} (t_{ks} - \mu_{t_{ks}}), \tag{7}$$

and then, it becomes apparent that

$$E[v_{ij}] = \mu_{v_{ij}} \quad = \quad \sum_{ks} \beta_{ijks} \mu_{t_{ks}} \tag{8}$$

$$\beta_{ijks} \quad = \quad \sum_r p_r^{ks} \delta_{ijr}^{ks}. \tag{9}$$

Note that the p_r^{ks} depend on the intensities of the traffic flow. Thus, this model is to be assumed conditional on the p_r^{ks} values. If one desires to combine this model with traffic assignment models, one can obtain the p_r^{ks} values from the predicted $\mathcal{OD}$ pair flows and iterate until convergence (this is a particular example of a bi-level method that will be explained in chapter 2.2.2 combined with the WMV assignment model).
Therefore, we can assume that the link flows are given by

$$\mathbf{V} = \boldsymbol{\beta} \mathbf{T} + \boldsymbol{\varepsilon}, \tag{10}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ are mutually independent normal random variables, independent of de random variables in $\mathbf{T}$, and ε_ℓ has mean $E[\varepsilon_\ell]$ and variance $\psi_\ell^2; \ell = 1, 2, \ldots, n$. These variables represent the traffic flow that enters or exits the link ℓ_{ij} apart from that going from the origin to the destination node of such a link. In particular, they can be assumed to be null.

Note also that assumption 1 is reasonable because when there is a general increase or decrease of flows (the U value), this affects to all ODs, and this effect can be assumed to be proportional, and the random variables η_{ks} account for random variations over these proportional distributions of flows.
Therefore, from (5), we have

$$\mathbf{T} = (\boldsymbol{\zeta} \quad | \quad \mathbf{I}) \begin{pmatrix} U \\ -- \\ \boldsymbol{\eta}^T \end{pmatrix}$$

and the variance-covariance matrix $\boldsymbol{\Sigma}_\mathbf{T}$ of the $\mathbf{T}$ variables becomes

$$\boldsymbol{\Sigma}_\mathbf{T} = (\boldsymbol{\zeta} \quad | \quad \mathbf{I}) \, \boldsymbol{\Sigma}_{(\mathbf{U},\boldsymbol{\eta})} \begin{pmatrix} \boldsymbol{\zeta}^T \\ -- \\ \mathbf{I} \end{pmatrix} = \sigma_U^2 \boldsymbol{\zeta} \boldsymbol{\zeta}^T + \mathbf{D}_{\boldsymbol{\eta}}, \tag{11}$$

where the matrices $\mathbf{\Sigma}_{(\mathbf{U},\boldsymbol{\eta})}$ and $\mathbf{D}$ are diagonal.
Then, we have

$$\begin{pmatrix} \mathbf{T} \\ \mathbf{V} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & | & \mathbf{0} \\ - & + & - \\ \boldsymbol{\beta} & | & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{T} \\ -- \\ \boldsymbol{\varepsilon} \end{pmatrix}$$

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

$$E[(\mathbf{T},\mathbf{V})] = \begin{pmatrix} E(U)\boldsymbol{\zeta} \\ -------- \\ E(U)\boldsymbol{\beta\zeta} + E(\boldsymbol{\varepsilon}) \end{pmatrix}, \tag{12}$$

and since the variance-covariance matrix of $(\mathbf{T},\boldsymbol{\varepsilon})$ is

$$\mathbf{\Sigma}_{(\mathbf{T},\boldsymbol{\varepsilon})} = \begin{pmatrix} \mathbf{\Sigma_T} & | & \mathbf{0} \\ -- & + & -- \\ \mathbf{0} & | & \mathbf{D}_\varepsilon \end{pmatrix},$$

the variance-covariance matrix of $(\mathbf{T},\mathbf{V})$ becomes

$$\begin{aligned} \mathbf{\Sigma}_{(\mathbf{T},\mathbf{V})} &= \begin{pmatrix} \mathbf{I} & | & \mathbf{0} \\ - & + & - \\ \boldsymbol{\beta} & | & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma_T} & | & \mathbf{0} \\ -- & + & -- \\ \mathbf{0} & | & \mathbf{D}_\varepsilon \end{pmatrix} \begin{pmatrix} \mathbf{I} & | & \mathbf{0} \\ - & + & - \\ \boldsymbol{\beta} & | & \mathbf{I} \end{pmatrix}^T \\ & \begin{pmatrix} \mathbf{\Sigma_T} & | & \mathbf{\Sigma_T}\boldsymbol{\beta}^T \\ --- & + & ------- \\ \boldsymbol{\beta}\mathbf{\Sigma_T} & | & \boldsymbol{\beta}\mathbf{\Sigma_T}\boldsymbol{\beta}^T + \mathbf{D}_\varepsilon \end{pmatrix}. \end{aligned} \tag{13}$$

All these assumptions imply that the joint PDF of $(t_{12}, t_{13}, \dots t_{ks}, v_{12}, v_{23}, \dots, v_{ij})$ can be written as

$$f(t_{12}, t_{13}, \dots t_{ks}, v_{12}, v_{23}, \dots, v_{ij}) = f_{N(\mu_\mathbf{T}, \mathbf{\Sigma_T})}(t_{12}, t_{13}, \dots t_{ks}) \prod_{ks} f_{N(\mu_{v_{ij}} + \sum_{ks \in \Pi_{ij}} \beta_{ijks}(t_{ks} - \mu_{t_{ks}}), \psi_{ij}^2)}(V_{ks}), \tag{14}$$

and can be used to predict traffic flows when information from traffic counts becomes available. The idea consists of using the joint distribution of $\mathcal{OD}$ pairs and link traffic 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.
Now the most convenient graph for this problem (from our point of view), is going to be described: the $\mathcal{OD}$ flows t_{ks} should be the parents of all link flows v_{ij} used by the corresponding travelers, and the error variables should be the parents of the corresponding flows, that is, the ε_{ij} must be parents of the v_{ij}, and the η_{ks} must be parents of the t_{ks}. Finally, the $\mathbf{U}$ variable must by on top (parent) of all $\mathcal{OD}$ flows, because it gives the level of them (high, intermediate or low). This solves the problem of "parent" being 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 node (see (Sumalee, 2004))

2.2 The Bayesian network model in a bi-level approach

Up to now the Bayesian network trip matrix estimation model (BNME) has been considered as a ME model, i.e., a model able to predict the $\mathcal{OD}$ and link flows from a given probability matrix with elements p_r^{ks} or β_{ijks}. The main difference with other methods is that it gives the joint conditional distributions of all not observed variables and that makes no difference between $\mathcal{OD}$ flows and link flows, in the sense that information of any one of them gives information about the others indistinctly. In particular the marginal distributions of any flow ($\mathcal{OD}$ or link) are supplied by the method, so that not only predictions can be obtained but probability intervals or regions.

In this chapter we show that the BNME can be easily combined with some assignment method to obtain the equilibrium solution of the traffic problem and therefore obtain a more realistic solution, for example in the congested case, using a bi-level approach. In particular, the proposed model is combined with an assignment model which identifies the origin and destination of the travelers who drive on a link (see (Castillo, Menéndez & Sánchez-Cambronero, 2008b)). Among its advantages, we can emphasize that the β_{ijks} coefficients are easily calculated and the most important, the route enumeration is avoided. This method, called the "*Wardrop minimum variance (WMV) method*", is next, combined with the BNME proposed method, but first let us give a detailed explanation of it.

2.2.1 The WMV assignment model

In this section a User Equilibrium based optimization problem is presented that, given the t_{ks} $\mathcal{OD}$ flows, deals with the link ℓ_{ij} flows x_{ijks} coming from node k (origin) and going to node s (destination). The balance of all these flows particularized by origins and destinations, allows us classifying the link flows by ODs.

This important information can be used, not only to have a better knowledge of the user behavior and the traffic in the network but to make decisions, for example, when some network events take place. In addition, this method avoids the route enumeration problem which is a very important issue because including a sub routine which deals with this problem is always a thorny issue.

The problem is formulated as follows:

$$\underset{\mathbf{x}}{\text{Minimize}}\, Z = \sum_{\ell_{ij}\in A} \int_0^{\sum\limits_{k,s} x_{ijks}} c_{ij}(x)dx + \frac{\lambda}{m} \sum_{\ell_{ij}\in A} \sum_{k,s} (x_{ijks} - \mu)^2 \tag{15}$$

subject to

$$t_{ks}(\delta_{ik} - \delta_{is}) = \sum_{\ell_{ij}\in\mathcal{A}} x_{ijks} - \sum_{\ell_{ji}\in\mathcal{A}} x_{jiks} \quad \forall i; \quad \forall k,s; \quad k \neq s, \tag{16}$$

$$\mu = \frac{1}{m} \sum_{\ell_{ij}\in A} \sum_{k,s} x_{ijks}, \tag{17}$$

$$x_{ijks} \geq 0 \quad \forall i,j,k,s, \tag{18}$$

where $c_{ij}(\cdot)$ is the cost function for link ℓ_{ij}, x_{ijks} is the flow through link ℓ_{ij} with origin node k and destination node s, $\lambda > 0$ is a weighting factor, δ_{ik} are the Dirac deltas ($\delta_{ik} = 0$, if $i \neq k$, $\delta_{ii} = 1$), μ is the mean of the x_{ijks} variables, and m is its cardinal. We have also assumed that the cost on a link depends only on the flow on that link.

Note that equation (16) represents the flow balance associated with the OD-pair (k,s), for all nodes, and that the problem (15)-(18) for $\lambda = 0$ is a statement of the Beckmann et al. formulation of the Wardrop UE equilibrium problem, but stated for each OD pair. As the cost function we have selected the Bureau of Public Roads (BPR) type cost functions, because it is generally accepted and has nice regularity properties, but other alternative cost functions with the same regularity properties (increasing with flow, monotonic and continuously differentiable) can be used instead. This function is as follows:

$$c_{ij}\left(\sum_{k,s} x_{ijks}\right) = c_{ij}\left[1+\alpha_{ij}\left(\frac{\sum\limits_{k,s} x_{ijks}}{q_{ij}}\right)^{\gamma_{ij}}\right], \tag{19}$$

where for a given link ℓ_{ij}, c_{ij} is the cost associated with free flow conditions, q_{ij} is a constant measuring the flow producing congestion, and α_{ij} and γ_{ij} are constants defining how the cost increases with traffic flow. So the total flow v_{ij} through link ℓ_{ij} is:

$$v_{ij} \quad = \quad \sum_{k,s} x_{ijks}. \tag{20}$$

The problem (15)-(18) for $\lambda = 0$ becomes a pure Wardrop problem and has unique solution in terms of total link flows, but it can have infinitely many solutions in terms of x_{ijks}, though they are equivalent in terms of link costs (they have the same link costs). Note that any exchange of users between equal cost sub-paths does not alter the link flows nor the corresponding costs. So, given an optimal solution to the problem, exchanging different OD users from one sub-path to the other leads to another optimal solution with different x_{ijks} values, though the same link flows v_{ij}. To solve this problem one can choose a very small values of λ. In this case, the problem has a unique solution. Note also that since for $\lambda > 0$, (15) is strictly convex, and the system (16)-(18) is compatible and convex, the problem (15)-(18) has a unique solution, which is a global optimum.

2.2.2 Combining the BN model and the WMV equilibrium model

In this section the Bayesian network model is combined with the new WMV assignment model described in section 2.2.1 using a bi-level algorithm. The aim of proposing this assignment method instead of, for example, an standard SUE assignment model is twofold. First this method avoids the route enumeration which is a very important issue. Second, once the flows x_{ijks} are known, the β_{ijks} coefficients can be easily calculated as:

$$\beta_{ijks} = \frac{x_{ijks}}{t_{ks}}. \tag{21}$$

which is a very important data for the BNME proposed method and allow us an easily implementation of it.

Algorithm 1 (Bi-level algorithm for the BN and WMV models)**.**

INPUT. *$E[U]$, the $\boldsymbol{\zeta}$ matrix of relative weight of each OD-pair, the cost coefficients $c_{ij}, \alpha_{ij}, q_{ij}$ and γ_{ij}, $\forall l_{ij} \in A$, and the observed link flows, are the data needed by the algorithm.*

OUTPUT. *The predictions of the $\mathcal{OD}$ and link flows given the observed flows.*

Step 0: Initialization. *Initialize the $\mathcal{OD}$ flows to the initial guess for $E[\mathbf{T}]$:*

$$\mathbf{T_0} = E[\mathbf{T}] = E[U]\boldsymbol{\zeta}. \tag{22}$$

Step 1: Master problem solution. *The WMV optimization problem (15)-(18) is solved.*

Step 2: Calculate the $\boldsymbol{\beta}$ matrix. *The $\boldsymbol{\beta}$ matrix, of regression coefficients of the $\mathbf{V}$ variables given $\mathbf{T}_0$, is calculated using Equation (21).*

Step 3: Subproblem: Update the $\mathcal{OD}$ and link flow predictions using the Bayesian network. *The new OD-pair $\mathbf{T}$ and link $\mathbf{V}$ flows are predicted using equations shown in this section, which are:*

$$
\begin{aligned}
E[\mathbf{V}] &= E[U]\boldsymbol{\beta}\boldsymbol{\zeta} + E[\boldsymbol{\varepsilon}] &(23)\\
\mathbf{D}_{\boldsymbol{\eta}} &= Diag\,(vE[\mathbf{T}]) &(24)\\
\boldsymbol{\Sigma}_{\mathbf{TT}} &= \sigma_U^2\boldsymbol{\zeta}\boldsymbol{\zeta}^T + \mathbf{D}_{\boldsymbol{\eta}} &(25)\\
\boldsymbol{\Sigma}_{\mathbf{TV}} &= \boldsymbol{\Sigma}_{\mathbf{TT}}\boldsymbol{\beta}^T &(26)\\
\boldsymbol{\Sigma}_{\mathbf{VT}} &= \boldsymbol{\Sigma}_{\mathbf{TV}} &(27)\\
\boldsymbol{\Sigma}_{\mathbf{VV}} &= \boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{TT}}\boldsymbol{\beta}^T + \mathbf{D}_{\boldsymbol{\varepsilon}} &(28)\\
E[\mathbf{Y}|\mathbf{Z}=\mathbf{z}] &= E[\mathbf{Y}] + \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{z} - E[\mathbf{Z}]) &(29)\\
\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}} &= \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\Sigma}_{\mathbf{ZY}} &(30)\\
E[\mathbf{Z}|\mathbf{Z}=\mathbf{z}] &= \mathbf{z} &(31)\\
\boldsymbol{\Sigma}_{\mathbf{Z}|\mathbf{Z}=\mathbf{z}} &= \mathbf{0} &(32)\\
\mathbf{T} &= E[\,\mathbf{Y}|\,\mathbf{Z}=\mathbf{z}]|_{(\mathbf{Y},\mathbf{Z})=T} &(33)
\end{aligned}
$$

Step 4: Convergence checking. *Compute actual error by means of*

$$error = (\mathbf{T}_0 - \mathbf{T})^T(\mathbf{T}_0 - \mathbf{T}). \tag{34}$$

If the error is less than the tolerance, stop and return the values of $\mathbf{T}$ and $\mathbf{V}$. Otherwise, let $\mathbf{T}_0 = \mathbf{T}$ and continue with Step 1.

Equation (22) is the initial $\mathcal{OD}$ flow matrix calculated using the random variable $\mathbf{U}$, which gives an estimation of the global flow in the system, and the relative weight vector $\boldsymbol{\zeta}$, which gives the relative importance of the different $\mathcal{OD}$ flows. This $\mathbf{T_0}$ matrix with elements t^0_{ks}, is initially the input data for the problem (15)-(18) and is the initial guess for the $\mathcal{OD}$ matrix with which the calculations are started.
As it has been indicated, for the non-observed data ($\mathcal{OD}$ or/and link flows), one can supply a probability interval, obtained from the resulting conditional probabilities given the evidence. The relevance of the proposed method consists of using the covariance structure of all the variables involved. The importance of this information has been pointed out by (Hazelton, 2003), who shows how the indeterminacy of the system of equations relating link and $\mathcal{OD}$ flows, due to the larger number of the latter, can be compensated by the covariance structure.

2.3 Optimal counting location method using Bayesian networks

This section describes how the Bayesian network model can be also used to select the optimal number and locations of the links counters based on maximum correlation (see (Castillo, Menéndez & Sánchez-Cambronero, 2008b)). To deal with this problem, a simple procedure based on the correlation matrix is described below.

Algorithm 2 (Optimal traffic counting locations.)**.**

INPUT. *The set of target variables to be predicted (normally $\mathcal{OD}$ flows), a variance tolerance, and the initial variance-covariance matrix $\mathbf{\Sigma_{ZZ}}$, or alternatively, σ_U and the matrices $\boldsymbol{\zeta}$, $\mathbf{D}_{\boldsymbol{\eta}}$, $\mathbf{D}_{\boldsymbol{\varepsilon}}$ and $\boldsymbol{\beta}$.*

OUTPUT. *The set of variables to be observed (normally link flows).*

Step 0: Initialization. *If the initial variance-covariance matrix $\mathbf{\Sigma_{TV}}$ is not given, calculate it using (23)-(28).*

Step 1: Calculate the correlation matrix. *The correlation matrix* ***Corr*** *with elements*

$$Corr_{ab} = \frac{Cov(X_a X_b)}{\sqrt{\sigma_{x_a}\sigma_{x_b}}} \tag{35}$$

is calculated from the variance-covariance matrix $\mathbf{\Sigma_{TV}}$.

Step 2: Select the target and observable variables. *Select the target variable (normally among the OD-flows) and the observable variable (normally among the link flows), by choosing the largest absolute value of the correlations in matrix* ***Corr****. Note that a value of $Corr_{ab}$ close to 1 means that variables a and b are highly correlated. Therefore if the knowledge of a certain target variable is desired, it is more convenient to observe a variable with greater correlation coefficient because it has more information than other variables on the target variable.*

Step 3: Update the variance-covariance matrix $\mathbf{\Sigma_{TV}}$. *Use formulas (30) and (32) to update the variance-covariance matrix.*

Step 4: End of algorithm checking. *Check residual variances of the target variables and determine if they are below the given threshold. If they are, stop the process and return the list of observable variables. If there are still variables to be observed, continue with Step 1. Otherwise, stop and inform that there is no solution with the given tolerance and provide the largest correlation in order to have a solution.*

Note that equation (30), which updates the variance-covariance matrix, does not need the value of the evidence, but only the evidence variable. Thus, the algorithm can be run without knowledge of the evidences. Note also that this algorithm always ends, either with the list of optimal counting locations or with a threshold[2] value for the correlation coefficient for the problem to have a solution.

[2] Because the model determines the links to be observed, this selection is done with a given error level, therefore the quality of the results depend on it.

2.4 Example of applications: The Nguyen-Dupuis network

In this section, we illustrate the previous models using the well known Nguyen-Dupuis network. It consists of 13 nodes and 19 links, as shown in Figure 2.

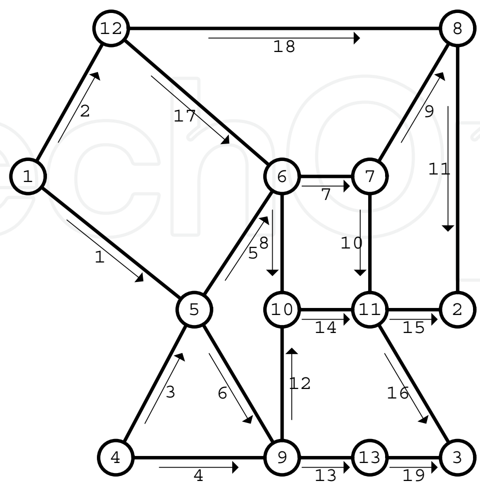


Fig. 2. The Nguyen-Dupuis network.

2.4.1 Selecting an optimal subset of links to be observed

Because the selection procedure is based on the covariance matrix of the link and $\mathcal{OD}$ flows, we need this matrix. To simplify, in this example, the matrix $\mathbf{D}_{\boldsymbol{\varepsilon}}$ is assumed to be diagonal with diagonal elements equal to 0.1 (a very small number), that is, we assume that there is practically no measurement error in the link flows. The matrix $\mathbf{D}_{\boldsymbol{\eta}}$ is also assumed to be diagonal with variances equal to the $(0.1E[t_{ks}])^2$.

The mean and standard deviation of U are assumed to be $E[U] = 100$ and $\sigma_U = 20$, respectively, and the assumed elements of the $\boldsymbol{\zeta}$ matrix are given in left part of Table 1.

$\mathcal{OD}$	ζ	Prior $\mathbf{T}_0$
1-2	0.4	40
1-3	0.8	80
4-2	0.6	60
4-3	0.2	20

link	c_{ij}	q_{ij}	α_{ij}	γ_{ij}
1 -5	7	70	1	4
1 -12	9	56	1	4
4 -5	9	56	1	4
4 -9	12	70	1	4
5 -6	3	42	1	4
5 -9	9	42	1	4
6 -7	5	70	1	4
6 -10	5	28	1	4
7 -8	5	70	1	4
7 -11	9	70	1	4

link	c_{ij}	q_{ij}	α_{ij}	γ_{ij}
8 -2	9	70	1	4
9 -10	10	56	1	4
9 -13	9	56	1	4
10-11	6	70	1	4
11-2	9	56	1	4
11-3	8	56	1	4
12-6	7	14	1	4
12-8	14	56	1	4
13-3	11	56	1	4

Table 1. Data needed for solving the example :Prior $\mathcal{OD}$ flow, $\boldsymbol{\zeta}$ matrices and link parameters.

Because we have no information about the beta matrix $\boldsymbol{\beta}$, we have used a prior $\mathcal{OD}$ trip matrix $\mathbf{T_0}$ and solved the problem (15)-(18) with the cost coefficients $c_{ij}, \alpha_{ij}, q_{ij}$ and γ_{ij} $\forall \ell \in \mathcal{A}$, shown in right part of Table 1, to obtain one. The method has been used, and Table 2 shows this initial matrix.

	Link proportions (β matrix)																		
OD	1 -5	1 -12	4 -5	4 -9	5 -6	5 -9	6 -7	6 -10	7 -8	7 -11	8 -2	9 -10	9 -13	10-11	11-2	11-3	12-6	12-8	13-3
1 -2	**0.00**	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	**0.00**	**0.00**	0.00	0.00	0.00	0.00	**1.00**	0.00
1 -3	**0.84**	0.16	0.00	0.00	0.47	0.37	0.37	0.26	0.00	0.37	0.00	**0.00**	**0.37**	0.26	0.00	0.63	0.16	**0.00**	0.37
4 -2	**0.00**	0.00	0.36	0.64	0.36	0.00	0.36	0.00	0.36	0.00	0.36	**0.64**	**0.00**	0.64	0.64	0.00	0.00	**0.00**	0.00
4 -3	**0.00**	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	**1.00**	0.00	0.00	0.00	0.00	**0.00**	1.00

Table 2. Initial $\boldsymbol{\beta}$ matrix.

With this information and considering a threshold value of 1 for the variances of the $\mathcal{OD}$ flows, one can use Algorithm 2 where the initial variance-covariance matrix $\boldsymbol{\Sigma}_{\mathbf{TV}}$ has been calculated using (25)-(28).
After solving this problem, one knows that to predict the $\mathcal{OD}$ matrix at the given quality level it is necessary to observe only the following links:

$$1-5, \quad 12-8, \quad 9-10, \quad 9-13. \tag{36}$$

It is interesting to observe the boldfaced columns associated with these links in the beta matrix in Table 2 to understand why these link flows are the most adequate to predict all the $\mathcal{OD}$ flows.

	Variances in each iteration				
Variable	0	1	2	3	Final
OD-pair 1-2	80.00	28.82	0.10	0.10	0.10
OD-pair 1-3	320.00	0.14	0.14	0.14	0.14
OD-pair 4-2	180.00	64.95	52.06	0.24	0.24
OD-pair 4-3	20.00	7.20	5.78	5.23	0.11
Link 1-5	226.48	**0.00**	**0.00**	**0.00**	**0.00**
Link 1-12	128.92	28.94	0.20	0.20	0.20
Link 4-5	23.73	8.62	6.93	0.13	0.13
Link 4-9	154.50	45.93	33.87	5.46	0.31
Link 5-6	159.32	8.68	6.98	0.16	0.16
Link 5-9	44.49	0.12	0.12	0.12	0.12
Link 6-7	117.21	8.68	6.96	0.15	0.15

	Variances in each iteration				
Variable	0	1	2	3	Final
Link 6-10	22.20	0.11	0.11	0.11	0.11
Link 7-8	23.73	8.62	6.93	0.13	0.13
Link 7-11	42.77	0.12	0.12	0.12	0.12
Link 8-2	173.30	51.48	7.08	0.23	0.23
Link 9-10	73.39	26.47	21.26	**0.00**	**0.00**
Link 9-13	112.01	7.34	5.91	5.36	**0.00**
Link 10-11	159.74	26.47	21.29	0.21	0.21
Link 11-2	73.38	26.51	21.27	0.20	0.20
Link 11-3	126.29	0.16	0.16	0.16	0.16
Link 12-6	8.20	0.10	0.10	0.10	0.10
Link 12-8	80.10	28.92	**0.00**	**0.00**	**0.00**
Link 13-3	112.01	7.35	5.92	5.36	0.20

Table 3. Variance of all variables initially and after updating the evidences in each step

Table 3 shows the variance of each variable at each stage, i.e., after the observable variables are being observed. In the Iteration 0 column, the variances when the information is not available are shown. In Iteration 1, the variances after observing the first observable variable $w_{1,5}$ are shown, and so on. It is interesting to see that the variances decrease with the knowledge of new evidences. At the end of the process all variances are very small (smaller than the selected threshold value). Note that the unobserved link flows can also be estimated with a small precision. Table 4 shows the correlation sub matrices at each iteration, together with the associated largest absolute value (boldfaced) used to choose the target and observable variable.

2.4.2 $\mathcal{OD}$ matrix estimation

Once the list of links to be observed have been obtained, one can observe them. The observed flows corresponding to these links have been simulated assuming that they are normal random variables with their corresponding means and standard deviations. The resulting flows were:

$$\hat{v}_{1,5} = 59.73; \quad \hat{v}_{12,8} = 36.12; \quad \hat{v}_{9,10} = 39.68; \quad \hat{v}_{9,13} = 49.87;$$

To estimate the $\mathcal{OD}$ flows with Algorithm 1 one needs some more data. We have assumed $E[\varepsilon] = \mathbf{0.1}$ and a tolerance value to check convergence of 0.00001. The initial guess for $\mathbf{T}$:

$$\mathbf{T_0} = E[\mathbf{T}] = E[U]\zeta. \tag{37}$$

Iteration 1				
	OD-pair			
Link	1-2	1-3	4-2	4-3
1-5	0.800	**1.000**	0.800	0.800
1-12	0.988	0.881	0.831	0.831
4-5	0.798	0.798	0.998	0.798
4-9	0.838	0.838	0.976	0.910
5-6	0.840	0.973	0.917	0.840
5-9	0.799	0.999	0.799	0.799
6-7	0.842	0.963	0.932	0.842
6-10	0.798	0.998	0.798	0.798
7-8	0.798	0.798	0.998	0.798
7-11	0.799	0.999	0.799	0.799
8-2	0.975	0.839	0.913	0.839
9-10	0.799	0.799	0.999	0.799
9-13	0.841	0.967	0.841	0.926
10-11	0.839	0.913	0.974	0.839
11-2	0.799	0.799	0.999	0.799
11-3	0.800	1.000	0.800	0.800
12-6	0.795	0.994	0.795	0.795
12-8	0.999	0.800	0.800	0.800
13-3	0.841	0.967	0.841	0.926
Target variable=OD-pair 1-3				
Observed variable=link 1-5.				

Iteration 2			
	OD-pair		
Link	1-2	4-2	4-3
1-12	0.998	0.444	0.444
4-5	0.442	0.994	0.442
4-9	0.513	0.934	0.733
5-6	0.442	0.992	0.442
6-7	0.443	0.993	0.443
7-8	0.442	0.994	0.442
8-2	0.930	0.740	0.514
9-10	0.444	0.998	0.444
9-13	0.442	0.442	0.992
10-11	0.444	0.998	0.444
11-2	0.444	0.998	0.444
12-8	**0.998**	0.444	0.444
13-3	0.442	0.442	0.992
Target variable=OD-pair 1-2			
Observed variable=link 12-8.			

Iteration 3		
	OD-pair	
Link	4-2	4-3
4-5	0.993	0.306
4-9	0.918	0.657
5-6	0.991	0.306
6-7	0.991	0.306
7-8	0.993	0.306
8-2	0.986	0.307
9-10	**0.998**	0.308
9-13	0.306	0.990
10-11	0.997	0.308
11-2	0.998	0.308
13-3	0.306	0.990
Target variable=OD-pair 4-2		
Observed variable=link 9-10.		

Iteration 4	
	OD-pair
Link	4-3
4-9	0.982
9-13	**0.989**
13-3	0.989
Target variable=OD-pair 4-3	
Observed variable=link 9-13.	

Table 4. Correlation sub matrices

The initial values of v_{ij} and x_{ijks} variables, using (37), are shown in table 5, and the resulting $\mathcal{OD}$ and link flows after convergence of the process are shown in Table 6 column 2, and the final $\boldsymbol{\beta}$ matrix is shown in Table 7.

Thus, the resulting $\mathcal{OD}$ matrix estimates and the prediction for the link flow variables are shown in Table 6. In addition, for the sake of comparison, in algorithm 1 the WMV has been replaced by a Logit SUE assignment, and the results are shown in column 4 of Table 6. Note that the results are very similar.

Link	Cost	v_{ij}	1-2	1-3	4-2	4-3
1 -5	13.0	67.3	0.0	67.3	0.0	0.0
1 -12	16.1	52.7	40.0	12.7	0.0	0.0
4 -5	9.2	21.7	0.0	0.0	21.7	0.0
4 -9	17.8	58.3	0.0	0.0	38.3	20.0
5 -6	14.9	59.3	0.0	37.5	21.7	0.0
5 -9	11.3	29.8	0.0	29.8	0.0	0.0
6 -7	6.4	51.0	0.0	29.2	21.7	0.0
6 -10	6.6	21.0	0.0	21.0	0.0	0.0
7 -8	5.0	21.7	0.0	0.0	21.7	0.0
7 -11	9.3	29.2	0.0	29.2	0.0	0.0
8 -2	14.4	61.7	40.0	0.0	21.7	0.0
9 -10	12.2	38.3	0.0	0.0	38.3	0.0
9 -13	14.6	49.8	0.0	29.8	0.0	20.0
10-11	9.1	59.3	0.0	21.0	38.3	0.0
11-2	11.0	38.3	0.0	0.0	38.3	0.0
11-3	13.2	50.2	0.0	50.2	0.0	0.0
12-6	11.8	12.7	0.0	12.7	0.0	0.0
12-8	17.6	40.0	40.0	0.0	0.0	0.0
13-3	17.9	49.8	0.0	29.8	0.0	20.0

Table 5. Link cost, link total flows, and link flows after using WMV assignment.

OD or	Link flows		
link	BN-WMV	Prior	BN-SUE
1-2	36.15	40.00	38.07
1-3	72.81	80.00	71.81
4-2	67.72	60.00	61.15
4-3	22.45	20.00	29.19
1 -5	59.73	67.27	59.73
1 -12	49.19	52.73	50.02
4 -5	28.07	21.74	34.33
4 -9	62.10	58.26	56.00
5 -6	60.48	59.25	60.87
5 -9	27.36	29.76	33.33
6 -7	52.88	50.95	52.63
6 -10	20.64	21.03	22.10
7 -8	27.92	21.74	21.94
7 -11	24.97	29.21	30.70
8 -2	64.07	61.74	58.09
9 -10	39.68	38.26	39.68
9 -13	49.87	49.76	49.87
10-11	60.28	59.28	61.80
11-2	39.80	38.26	41.14
11-3	45.45	50.24	51.36
12-6	13.04	12.73	13.87
12-8	36.12	40.00	36.12
13-3	49.82	49.76	49.63

Table 6. $\mathcal{OD}$ and link flows resulting from algorithm 1, and replacing the WMV by a Logit SUE method.

	Link proportions ($\boldsymbol{\beta}$ matrix)																		
OD	1 -5	1 -12	4 -5	4 -9	5 -6	5 -9	6 -7	6 -10	7 -8	7 -11	8 -2	9 -10	9 -13	10-11	11-2	11-3	12-6	12-8	13-3
1 -2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
1 -3	0.82	0.18	0.00	0.00	0.45	0.38	0.34	0.28	0.00	0.34	0.00	0.00	0.38	0.28	0.00	0.62	0.18	0.00	0.38
4 -2	0.00	0.00	0.41	0.59	0.41	0.00	0.41	0.00	0.41	0.00	0.41	0.59	0.00	0.59	0.59	0.00	0.00	0.00	0.00
4 -3	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 7. Final $\boldsymbol{\beta}$ matrix.

Before analyzing the previous results, one must realize that we are using two sources of information: that contained in the Bayesian network (the joint normal distribution of links and $\mathcal{OD}$ flows), and the observed link or $\mathcal{OD}$ flows. We must note that the first one can be very informative. In fact, when no observations are available it is the only one supplying information about flows, but when observations become available it can still be more informative than that contained in the observations, if the number of observed links is small.

3. Plate scanning based method for traffic prediction.

This section shows how the Bayesian network tool can be also used with data from the plate scanning technique (see (Sánchez-Cambronero et al., 2010). Therefore, first the plate scanning approach will be introduced, and then the model will be described (see (Castillo, Menéndez & Jiménez.P, 2008)).

3.1 Dealing with the information contained in the data from the plate scanning technique.

The idea of plate scanning consists of registering plate numbers and the corresponding times of the circulating vehicles when they travel on some subsets of links. This information is then used to reconstruct vehicle routes by identifying identical plate numbers at different locations and times. In order to clarify the concepts, let us consider a traffic network $(\mathcal{N}, \mathcal{A})$ where $\mathcal{N}$ is

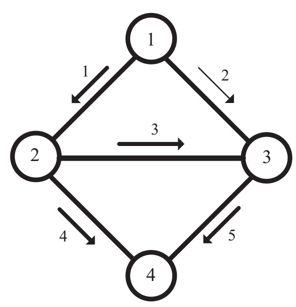


Fig. 3. The elementary example network used for illustrative purposes.

a set of nodes and $\mathcal{A}$ is a set of links. We have used the simple network in Figure 3 of 4 nodes and 5 links. Table 8 shows the 4 OD-pairs considered and the corresponding 7 paths used in this example.

$\mathcal{OD}$	path code (r)	Links		
1-4	1	1	3	5
1-4	2	1	4	
1-4	3	2	5	
2-4	4	3	5	
2-4	5	4		
1-2	6	1		
2-3	7	3		

Table 8. Set of 4 OD-pairs and 7 paths considered in the elementary example.

We assume that we have selected a nonempty subset $\mathcal{SC} \subset \mathcal{A}$ of $n_{sc} \neq 0$ links to be scanned. To illustrate, consider the scanned subset $\mathcal{SC}$ of 4 links[3]

$$\mathcal{SC} \equiv \{1, 3, 4, 5\}. \tag{38}$$

In the scanned links the plate numbers and the passing times[4] of the users are registered, i.e., the initially gathered information $\mathcal{I}$ consists of the set

$$\mathcal{I} \equiv \{(I_k, \ell_k, \tau_k);\ k = 1, 2, \ldots, m;\ \ell_k \in \mathcal{SC}\}, \tag{39}$$

[3] This subset is not arbitrary, but has been carefully selected as we will see.
[4] Passing times are used only to identify the scanned user routes.

where I_k is the identification number (plate number) of the k-th observed user, $\ell_k \in \mathcal{SC}$ is the link where the observation took place, τ_k is the corresponding pass time through link ℓ_k, and m is the number of observations.

For illustration purposes, a simple example with 28 registered items is shown in left part of Table 9, where the plate numbers of the registered cars and the corresponding links and passing times, in the format second-day-month-year, are given.

item k	# Plate I_k	Link ℓ_k	Time τ_k
1	1256 ADL	1	00001 19-12-2009
2	3789 BQP	3	00022 19-12-2009
3	7382 BCD	2	00045 19-12-2009
4	9367 CDF	1	00084 19-12-2009
5	9737 AHH	1	00123 19-12-2009
6	3789 BQP	5	00145 19-12-2009
7	7382 BCD	5	00187 19-12-2009
8	6453 DGJ	4	00245 19-12-2009
9	9737 AHH	3	00297 19-12-2009
10	9367 CDF	4	00309 19-12-2009
11	3581 AAB	1	00389 19-12-2009
12	6299 HPQ	4	00478 19-12-2009
13	9737 AHH	5	00536 19-12-2009
14	3581 AAB	3	00612 19-12-2009
15	1243 RTV	3	00834 19-12-2009
16	7215 ABC	1	00893 19-12-2009
17	8651 PPT	3	01200 19-12-2009
18	3581 AAB	5	01345 19-12-2009
19	1974 PZS	1	01356 19-12-2009
20	1256 ADL	4	01438 19-12-2009
21	2572 AZP	1	01502 19-12-2009
22	6143 BBA	3	01588 19-12-2009
23	7614 CAB	1	01670 19-12-2009
24	6143 BBA	5	01711 19-12-2009
25	1897 DEP	2	01798 19-12-2009
26	1897 DEP	5	01849 19-12-2009
27	2572 AZP	4	01903 19-12-2009
28	7614 CAB	4	01945 19-12-2009
⋮	⋮	⋮	⋮

item z	# Plate I_z	Scanned links C_{s_z}	Code s
1	1256 ADL	$\{1,4\}$	2
2	3789 BQP	$\{3,5\}$	4
3	7382 BCD	$\{5\}$	3
4	9367 CDF	$\{1,4\}$	2
5	9737 AHH	$\{1,3,5\}$	1
6	6453 DGJ	$\{4\}$	5
7	3581 AAB	$\{1,3,5\}$	1
8	4769 CCQ	$\{3\}$	7
9	2572 AZP	$\{1,4\}$	2
10	6143 BBA	$\{3,5\}$	4
11	7614 CAB	$\{1,4\}$	2
12	1897 DEP	$\{5\}$	3
13	6299 HPQ	$\{5\}$	3
14	7215 ABC	$\{1\}$	6
15	1974 PZS	$\{1\}$	6
16	1243 RTV	$\{3\}$	7
⋮	⋮	⋮	⋮

			Scanned links				
OD	r	s	1	3	4	5	$\hat{w}_s$
1-4	1	1	X	X		X	2
1-4	2	2	X		X		4
1-4	3	3				X	3
2-4	4	4		X		X	2
2-4	5	5			X		1
1-2	6	6	X				2
2-3	7	7		X			2

Table 9. Example of registered data by scanned links and the data after been processed.

Note that a single car user supplies one or more elements (I_k, ℓ_k, τ_k), in fact as many as the number of times the corresponding user passes through an scanned link. For example, the user with plate number 9737 AHH appears three times, which means he/she has been registered when passing through three scanned links (1, 3 and 5).

A cross search of plate numbers contained in the different (I_k, ℓ_k, τ_k) items of information and check of the corresponding passing times allows one determining the path or partial paths followed by the scanned users. This allows building the set

$$\{(I_z, C_{s_z}) |\ z = 1, 2, \ldots, n;\ C_{s_z} \in \mathcal{P}(\mathcal{SC})\}, \tag{40}$$

where C_{s_z} is the subset of links associated with the I_z user, which includes all links in which the user has been scanned (scanned partial path of that user), n is the number of registered users, and $\mathcal{P}(\mathcal{SC})$ is the set of parts of $\mathcal{SC}$, which contains $2^{n_{sc}}$ elements. A registered user

has an associated C_{s_z} subset only if the corresponding scanned links belong to its route. Of course, a non-registered user appears in no registered links, which corresponds to $C_{s_z} = \emptyset$. We associate with each user the subset C_{s_z} of scanned links contained in his/her route, and call a subset C_{s_z} of scanned links feasible if there exists a user which associated subset is C_{s_z}. Note that not all subsets of scanned links are feasible and each route must leads to a feasible subset. Upper right part of Table 9, shows the plate numbers and the associated scanned sub-path (C_{s_z} set of registered links for given users) of the registered users in columns two and three. For example, the user with plate number $3581AAB$ appears as registered in links 1, 3 and 5, thus leading to the set $C_{s_z} \equiv \{1, 3, 5\}$.
To obtain the feasible C sets one needs only to go through each possible path and determine which scanned links are contained in it. The two first columns of bottom part of Table 9 shows all paths, defined by the $\mathcal{OD}$ and r (the order of the path within the OD) values. The third column corresponds with the set of scanned link code s which, in this case, is the same as the route code because we have full route observability. Finally, the last columns corresponds with the scanned links and its associated sets (each indicated by an X).
An important point to note is that since all combinations of scanned links are different for all paths, and this happens because the set of links to be scanned has been adequately selected, the scanning process allows identifying the path of any scanned user. Therefore, using this information, one obtains the observed number of users $\hat{w}_s$ with associated s-values and C_s sets (see Table 9). This allows us to summarize the scanned observations as

$$\{\hat{w}_s : s \in \mathcal{S}\}, \tag{41}$$

where $\mathcal{S}$ is the set $\mathcal{S} \equiv \{1, 2, \ldots, n\}$ and n the number of different C_s sets in $\mathcal{S}$, which is the information used by the proposed model to estimate the traffic flows. Note also that standard models are unable to deal with this problem, i.e., to handle the information in the form (41).
To control this type of information, the traffic flow must be disaggregated in terms of the new variables $\hat{w}_s$, which refer to the flow registered by the scanned links in C_s. Then, one needs to write the conservation laws as follows:

$$\hat{w}_s = \sum_{r \in \mathcal{R}} \delta_{sr} f_r; \quad r \in \mathcal{R}; \quad s \in \mathcal{S}, \tag{42}$$

where f_r is the flow of route r, δ_{sr} is one if the route r contains all and only the links in C_s.

3.2 Model assumptions

In this section, the model assumptions for the BN-PLATE (see (Sánchez-Cambronero et al., 2010) model are introduced. Note that there are important differences with the model described in section 2 in which the model was built considering OD-pair and link flows, instead of route and scanned links flows, respectively.
Therefore assuming the route and subsets of scanned link 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 defined in (42) in an exact or statistical (i.e., with random errors) form. In our Gaussian Bayesian network model we make the following assumptions:

Assumption 1: The vector $\mathbf{F}$ of route flows is a multivariate normal $N(\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F)$ random variable with mean $\boldsymbol{\mu}_F$ and variance-covariance matrix $\boldsymbol{\Sigma}_F$.

For the same reason than in the BN-WMV model, it is clear that the **F** random variables are correlated. Therefore:

$$f_r = k_r U + \eta_r, \tag{43}$$

where $k_r, r = 1, \ldots, m$ are positive real constants, one for each route r, 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 : A random positive variable that measures the level of total mean flow.
- $\mathbf{K}$: A 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).
- $\boldsymbol{\eta}$: A 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.

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

$$\mathbf{W} = \boldsymbol{\Delta}\mathbf{F} + \boldsymbol{\varepsilon}, \tag{44}$$

where w_s, f_r and δ_{sr} have the same meaning than before, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ 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 got error free.

Then, following these assumptions, we have

$$\mathbf{F} = \begin{pmatrix} \mathbf{K} & | & \mathbf{I} \end{pmatrix} \begin{pmatrix} U \\ -- \\ \boldsymbol{\eta}^T \end{pmatrix} \tag{45}$$

and the variance-covariance matrix $\boldsymbol{\Sigma}_\mathbf{F}$ of the **F** variables becomes

$$\boldsymbol{\Sigma}_\mathbf{F} = \begin{pmatrix} \mathbf{K} & | & \mathbf{I} \end{pmatrix} \boldsymbol{\Sigma}_{(U,\boldsymbol{\eta})} \begin{pmatrix} \mathbf{K}^T \\ -- \\ \mathbf{I} \end{pmatrix} = \sigma_U^2 \mathbf{K}\mathbf{K}^T + \mathbf{D}_{\boldsymbol{\eta}}, \tag{46}$$

where the matrices $\boldsymbol{\Sigma}_{(U,\boldsymbol{\eta})}$ and $\mathbf{D}_{\boldsymbol{\eta}}$ are diagonal.
From (44) and (45)

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

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

$$E[(\mathbf{F}, \mathbf{W})] = \begin{pmatrix} E(U)\mathbf{K} \\ -------- \\ E(U)\boldsymbol{\Delta}\mathbf{K} + E(\boldsymbol{\varepsilon}) \end{pmatrix}, \tag{47}$$

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

$$\boldsymbol{\Sigma}_{(\mathbf{F},\mathbf{W})} = \begin{pmatrix} \boldsymbol{\Sigma}_\mathbf{F} & | & \boldsymbol{\Sigma}_\mathbf{F}\boldsymbol{\Delta}^T \\ -- & + & ---- \\ \boldsymbol{\Delta}\boldsymbol{\Sigma}_\mathbf{F} & | & \boldsymbol{\Delta}\boldsymbol{\Sigma}_\mathbf{F}\boldsymbol{\Delta}^T + \mathbf{D}_\varepsilon \end{pmatrix}. \tag{48}$$

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

3.3 Using the model to predict traffic flows

Once we have built the model, we can use its JPD (similar to the one defined in (14)) to predict route flows (and therefore $\mathcal{OD}$ and link flows) when the information becomes available. 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 **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 chapter deals with this problem in Section 3.4 providing several methods to select the best set of links to be scanned.

Step 2: Observe the plate scanning data. The plate scanning data $\hat{w}_s$ are extracted.

Step 3: Estimate the route flows. The route matrix **F** with elements f_r are estimated using the Bayesian network method, i.e., using the following formulas (see (47), (48),(3) and (4)):

$$E[\mathbf{F}] = E[U]\mathbf{K} \tag{49}$$
$$E[\mathbf{W}] = E[U]\mathbf{\Delta K} + E[\boldsymbol{\varepsilon}] \tag{50}$$
$$\mathbf{D}_{\boldsymbol{\eta}} = \text{Diag}\left(\nu E[\mathbf{F}]\right), \tag{51}$$
$$\mathbf{\Sigma_{FF}} = \sigma_U^2 \mathbf{K}\mathbf{K}^T + \mathbf{D}_{\boldsymbol{\eta}} \tag{52}$$
$$\mathbf{\Sigma_{FW}} = \mathbf{\Sigma_{FF}}\mathbf{\Delta}^T \tag{53}$$
$$\mathbf{\Sigma_{WF}} = \mathbf{\Sigma_{FW}} \tag{54}$$
$$\mathbf{\Sigma_{WW}} = \mathbf{\Delta}\mathbf{\Sigma_{FF}}\mathbf{\Delta}^T + \mathbf{D}_{\boldsymbol{\varepsilon}} \tag{55}$$
$$E[\mathbf{F}|\mathbf{W} = \mathbf{w}] = E[\mathbf{F}] + \mathbf{\Sigma_{FW}}\mathbf{\Sigma_{WW}}^{-1}(\mathbf{w} - E[\mathbf{W}]) \tag{56}$$
$$\mathbf{\Sigma}_{\mathbf{F}|\mathbf{W}=\mathbf{w}} = \mathbf{\Sigma_{FF}} - \mathbf{\Sigma_{FW}}\mathbf{\Sigma_{WW}}^{-1}\mathbf{\Sigma_{WF}} \tag{57}$$
$$E[\mathbf{W}|\mathbf{W} = \mathbf{w}] = \mathbf{w} \tag{58}$$
$$\mathbf{\Sigma}_{\mathbf{W}|\mathbf{W}=\mathbf{w}} = \mathbf{0} \tag{59}$$
$$\mathbf{F} = E[\,\mathbf{F}|\,\mathbf{W} = \mathbf{w}]|_{(\mathbf{F},\mathbf{W})=F} \tag{60}$$

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 $\mathcal{OD}$ pair flows) can be easily obtained.

3.4 The plate scanning device location problem

Due to the importance of the traffic count locations to obtain good traffic flow predictions, this section deals with the problem of determining the optimal number and allocation of plate scanning devices(see (Mínguez et al., 2010)).

3.4.1 Location rules

In real life, the true error or reliability of an estimated $\mathcal{OD}$ matrix is unknown. Based on the concept of maximal possible relative error (MPRE), (Yang & Zhou, 1998) proposed several location rules. We have derived analogous rules based on prior link and flow values and the following measure (RMSRE, root mean squared relative error):

$$\text{RMSRE} = \sqrt{\frac{1}{m}\sum_{i\in\mathcal{I}}\left(\frac{t_i^0 - t_i}{t_i^0}\right)^2}, \tag{61}$$

where[5] 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 $\mathcal{I}$. Since the prior $\mathcal{OD}$ pair flows t_i^0 are known and there are the best available information, they are used to calculate the relative error.
Given the set R of all possible routes, any of them corresponding to a unique $\mathcal{OD}$ pair, 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:

$$\text{RMSRE} = \sqrt{\frac{1}{m}\sum_{i\in I}\left(\frac{t_i^0 - \sum_{r\in R_i} f_r^0 y_r}{t_i^0}\right)^2}, \tag{62}$$

where y_r is a binary variable equal to one if route r is identified uniquely (observed) through 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 $n_{sc} = \sum_{\forall r\in R} y_r \le n_r$ 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 (62) we obtain

$$m\times\ \text{RMSRE}^2 = \sum_{i\in I}\left(1 - \sum_{r\in R_i}\frac{f_r^0}{t_i^0}y_r\right)^2, \tag{63}$$

where it can be deduced 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, (63) can be reformulated as follows

$$m\times\ \text{RMSRE}^2 = \sum_{i\in I}\left(1 - \sum_{r\in(R_i\cap\mathcal{OR})}\frac{f_r^0}{t_i^0}\right)^2 = \sum_{i\in I}\left(\sum_{i\in(R_i\cap\mathcal{UR})}\frac{f_r^0}{t_i^0}\right)^2, \tag{64}$$

[5] from now on, and for simplicity, we denoted each $\mathcal{OD}$ pair as i instead of ks and each link as a instead of ℓ_{ij}

so that routes to be observed ($y_r = 1$) should be chosen minimizing (64).
The main shortcoming of equations (63) or (64) 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:

$$\text{RMARE} = \frac{1}{m}\sum_{i\in I}\left|\frac{t_i^0 - t_i}{t_i^0}\right| = \frac{1}{m}\sum_{i\in I}\left|\frac{t_i^0 - \sum_{r\in R_i} f_r^0 y_r}{t_i^0}\right|, \tag{65}$$

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

$$\text{RMARE} = 1 - \frac{1}{m}\left(\sum_{i\in I}\sum_{r\in(R_i\cap\mathcal{OR})}\frac{f_r^0}{t_i^0}\right) = \frac{1}{m}\left(\sum_{i\in I}\sum_{r\in(R_i\cap\mathcal{UR})}\frac{f_r^0}{t_i^0}\right), \tag{66}$$

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 that this result derives in a rule that can be denominated the **Maximum Relative Route Flow** rule.
The above location rule has been derived by supposing that the prior trip distribution matrix is reasonably reliable and close to the actual true value, because the accuracy of the prior matrix has a great impact on the estimates of the true $\mathcal{OD}$ matrix. Note 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.
Since the proper identifiability of routes must be made through plate scanner devices in links, an additional rule related to links should be considered, which states that scanned links must allow us to identify uniquely the routes to be observed ($y_r = 1$) from all possible routes being considered. This rule can be denominated the **Full Identifiability of Observed Path Flows** rule.

3.4.2 Location models

The first location model to be proposed in this chapter considers full route observability, i.e. RMSRE= 0, but including budget considerations. In the transport literature, each link, 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. Then:

$$\text{M}_1 = \underset{\boldsymbol{z}}{\text{Minimize}} \sum_{a\in\mathcal{A}} \mathcal{P}_a z_a \tag{67}$$

subject to

$$\sum_{a\in\{\mathcal{A}\}}(\delta_a^r + \delta_a^{r_1})(1-\delta_a^r\delta_a^{r_1})z_a \geq 1 \left\{\begin{array}{l}\forall (r,r_1)|r<r_1\\ \sum_{a\in\mathcal{A}}\delta_a^r\delta_a^{r_1} > 0\end{array}\right. \tag{68}$$

$$\sum_{a\in\mathcal{A}} z_a\delta_a^r \geq 1; \forall r, \tag{69}$$

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, $\boldsymbol{\Delta}$ is the route incidence matrix with elements δ_a^r.

Note that constraint (68) 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 $\mathbf{\Delta}$ is different from the others) and (69) ensures that at least one link for every route is scanned (every row in the incidence matrix $\mathbf{\Delta}$ contains at least one element different from zero). Both constraints force the *maximum relative route flow* and *full identifiability of observed path flows* rules to hold. Note also that all $\mathcal{OD}$ pairs are totally covered. In addition, this model allows 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. However, budget is limited in practice, meaning that some $\mathcal{OD}$ pairs or even some routes may remain uncovered, consequently based on (66) the following model is proposed in order to observe the maximum relative route flow:

$$\mathrm{M}_2 = \underset{\mathbf{y},\,\mathbf{z}}{\text{Maximize}} \quad \sum_{\forall i \in I} \sum_{r \in R_i} \frac{f_r^0}{t_i^0} y_r \tag{70}$$

subject to

$$\sum_{a\in\{\mathcal{A}\}} (\delta_a^r + \delta_a^{r_1})(1 - \delta_a^r \delta_a^{r_1}) z_a \ge y_r \left\{ \begin{array}{l} \forall (r, r_1) | r < r_1 \\ \sum\limits_{a\in\mathcal{A}} \delta_a^r \delta_a^{r_1} > 0 \end{array} \right. \tag{71}$$

$$\sum_{a\in\mathcal{A}} z_a \delta_a^r \ge y_r; \quad \forall r, \tag{72}$$

$$\sum_{a\in\mathcal{A}} \mathcal{P}_a z_a \le \mathcal{B}, \tag{73}$$

where 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 (71) guarantees that the route r is able to be distinguished from the others if the binary variable y_r is equal to 1. Constraint (72) ensures that the route which is able to be distinguished contains at least one scanned link. Both constraints (71) and (72) 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 (71) and (72) immediately converts into inactive the constraint (69) 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 $\mathcal{B}$. 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

$$z_a = 1; \quad \forall a \in \mathcal{OL}. \tag{74}$$

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

3.5 Example of application

In this section we illustrate the proposed methods by their application to a simple example. Consider the network in Figure 4 with the routes and OD-pairs in Table 10, which shows the feasible combination of scanned links after solving the M_1 model ($\mathcal{SL} = \{1, 2, 3, 4, 7, 8\}$).
Next, the proposed method in Section 3.3 is applied.

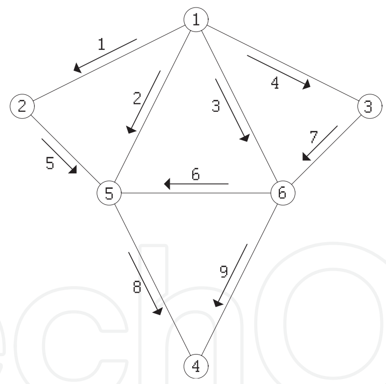


Fig. 4. The elementary example network

$\mathcal{OD}$	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	Scanned links					
(s)	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	

Table 10. Required data for the simple 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 11. The assumed mean value was $E[U] = 10$ and the value of σ_U was 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 as was explained in Section 3.1.

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 (49)-(60)

	True		Scanned links						
Route	flow	Method	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

Table 11. Route flow estimates using BN and LS approaches

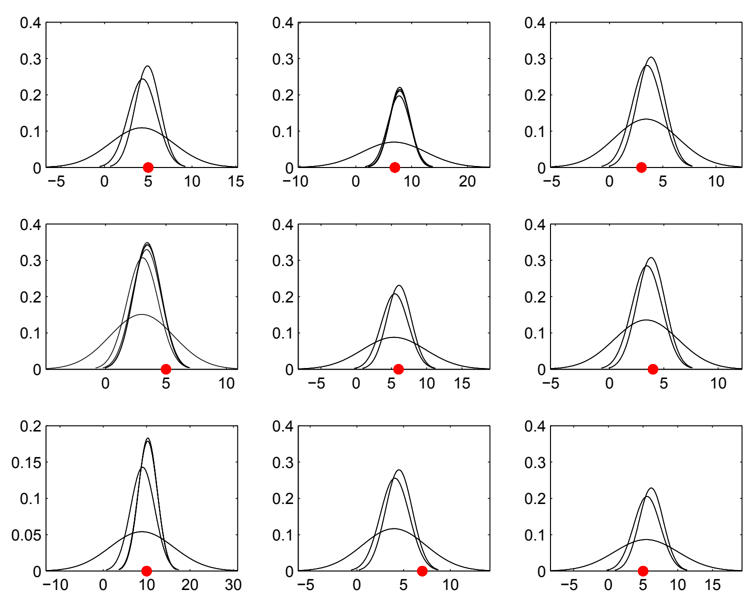


Fig. 5. Conditional distribution of the route flows

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 11. 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, we have compared the results with the standard method of Least Squares (LS) using the same data. The results appear in the second rows in Table 11. The results 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 to 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 11, 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. 5 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.

4. References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*, John Wiley and Sons, New York.

Castillo, E., Gutiérrez, J. M. & Hadi, A. S. (1997a). *Expert systems and probabilistic network models*, Springer Verlag, New York.

Castillo, E., Gutiérrez, J. M. & Hadi, A. S. (1997b). Sensitivity analysis in discrete Bayesian networks, *IEEE Transactions on Systems, Man and Cybernetics* **26(7)**: 412–423.

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

Castillo, E., Menéndez, J. M. & Sánchez-Cambronero, S. (2008b). 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. & Gómez, P. (1999). Uncertainty analyses in fault trees and Bayesian networks using form/sorm methods, *Reliability Engineering and System Safety* **65**: 29–40.

Doblas, J. & Benítez, 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.

Hazelton, M. L. (2003). Some comments on origin-destination matrix estimation, *Transportation Research* **37A**: 811–822.

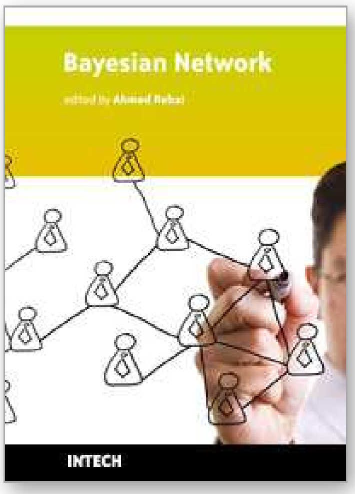
Mínguez, R., Sánchez-Cambronero, S., Castillo, E. & Jiménez, P. (2010). Optimal traffic plate scanning location for od trip matrix and route estimation in road networks, *Transportation Research B* **44**: 282–298.

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

Sánchez-Cambronero, S., Rivas, A., Gallego, I. & Menéndez, J. (2010). Predicting traffic flow in road networks using bayesian networks and data from an optimal plate scanning device location., *in* A. F. J.Filipe & B. Sharp. (eds), *Proceedings of* 2^{nd} *Internationa Conference on Agents and Artificial Intelligence.*, INSTICC-Springer., Valencia, Spain., pp. 552–559.

Sumalee, A. (2004). Optimal road user charging cordon design: a heuristic optimization approach, *Journal of Computer Aided Civil and Infrastructure Engineering* **19**: 377–392.

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



Bayesian Network
Edited by Ahmed Rebai

ISBN 978-953-307-124-4
Hard cover, 432 pages
Publisher Sciyo
Published online 18, August, 2010
Published in print edition August, 2010

Bayesian networks are a very general and powerful tool that can be used for a large number of problems involving uncertainty: reasoning, learning, planning and perception. They provide a language that supports efficient algorithms for the automatic construction of expert systems in several different contexts. The range of applications of Bayesian networks currently extends over almost all fields including engineering, biology and medicine, information and communication technologies and finance. This book is a collection of original contributions to the methodology and applications of Bayesian networks. It contains recent developments in the field and illustrates, on a sample of applications, the power of Bayesian networks in dealing the modeling of complex systems. Readers that are not familiar with this tool, but have some technical background, will find in this book all necessary theoretical and practical information on how to use and implement Bayesian networks in their own work. There is no doubt that this book constitutes a valuable resource for engineers, researchers, students and all those who are interested in discovering and experiencing the potential of this major tool of the century.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Enrique Castillo, Santos Sanchez-Cambronero and Jose Maria Menendez (2010). Bayesian Networks Methods for Traffic Flow Prediction, Bayesian Network, Ahmed Rebai (Ed.), ISBN: 978-953-307-124-4, InTech, Available from: http://www.intechopen.com/books/bayesian-network/bayesian-networks-methods-for-traffic-flow-prediction

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821