Aust. N. Z. J. Stat. **49**(3), 2007, 221–239 doi: 10.1111/j.1467-842X.2007.00477.x

ELICITING A DIRECTED ACYCLIC GRAPH FOR A MULTIVARIATE TIME SERIES OF VEHICLE COUNTS IN A TRAFFIC NETWORK

CATRIONA M. QUEEN[1]*, BEN J. WRIGHT[1] AND CASPER J. ALBERS[1]

The Open University

Summary

The problem of modelling multivariate time series of vehicle counts in traffic networks is considered. It is proposed to use a model called the linear multiregression dynamic model (LMDM). The LMDM is a multivariate Bayesian dynamic model which uses any conditional independence and causal structure across the time series to break down the complex multivariate model into simpler univariate dynamic linear models.

The conditional independence and causal structure in the time series can be represented by a directed acyclic graph (DAG). The DAG not only gives a useful pictorial representation of the multivariate structure, but it is also used to build the LMDM. Therefore, eliciting a DAG which gives a realistic representation of the series is a crucial part of the modelling process. A DAG is elicited for the multivariate time series of hourly vehicle counts at the junction of three major roads in the UK. A flow diagram is introduced to give a pictorial representation of the possible vehicle routes through the network. It is shown how this flow diagram, together with a map of the network, can suggest a DAG for the time series suitable for use with an LMDM.

Key words: conditional independence; dynamic linear model; linear multiregression dynamic model; model elicitation; traffic modelling.

1. Introduction

This paper considers the problem of modelling multivariate time series of vehicle counts in traffic networks. This can be a difficult problem. The model needs to be complex enough to accommodate the multivariate structure of the time series, but it also needs to be simple enough to work in real time if the model is to be of practical use. The model also needs to be easily adaptable to cope with any changes which may occur in the network caused by external events, such as roadworks or bad weather. In this respect, it is also helpful if the model is interpretable.

Several modelling approaches have been used for vehicle counts in traffic networks. These include historical, data-based algorithms, classical time series methods, neural networks and Bayesian dynamic linear models (see, for example, Smith & Demetsky, 1997; Dougherty & Cobbett, 1997; Kirby, Watson & Dougherty, 1997; Whittaker, Garside & Lindvelt, 1997; Tebaldi, West & Karr, 2002). Here a Bayesian dynamic linear modelling approach is used for the problem. Both the Bayesian models of Whittaker *et al.* (1997) and Tebaldi *et al.* (2002) use vehicle counts over one-minute intervals. However, the data considered in this paper are hourly and have quite different modelling requirements. Firstly, hourly

*Author to whom correspondence should be addressed.

[1]Department of Statistics, The Open University, Milton Keynes, MK7 6AA, UK.
e-mail: C.Queen@open.ac.uk

Acknowledgments. We would like to thank the referees for their very useful comments on an earlier version of this paper.

data are far less noisy so that brief periods of unusual activity (such as caused by congestion) are largely smoothed away. Secondly, during an hour, each vehicle may be counted at several different data collection points. The data in this paper therefore require a different model.

A particular type of multivariate Bayesian dynamic linear model (DLM) (West & Harrison, 1997) will be used called a Linear Multiregression Dynamic Model or LMDM (Queen & Smith, 1993). An LMDM represents any heuristic conditional independence relationships and causal drive within a multivariate time series by a directed acyclic graph (DAG) (see, for example, Cowell *et al.*, 1999). This DAG not only gives a useful pictorial representation of the multivariate structure of the time series, but is also used in the LMDM to decompose a complex multivariate time series model into simpler workable components. Thus the LMDM can accommodate the multivariate structure of the time series as represented by the DAG, and yet is also computationally simple. The model is adaptable and any external information which may affect the network can be integrated into the model through intervention (see West & Harrison, 1997). In addition, an LMDM can also often be defined so that its parameters are interpretable.

The elicitation of a DAG which accurately represents the structure of the series is a crucial part of the LMDM modelling process and this paper focuses on this important elicitation problem. A DAG will be elicited for a particular traffic network at the junction of three major roads — the M25, A2 and A282 — east of London, UK. Although just a single network is considered here, the principles of the elicitation methods presented can be applied to any network.

The traffic data are in the form of hourly counts of vehicles passing over induction loops in the road surface at a number of data collection sites. A diagram of the network showing the layout of the data collection sites is given in Figure 1. Each site is identified by a number and white arrows on the diagram indicate the direction of traffic flow on each part of the network. The network is such that traffic flows into the network, through a number of data collection sites, and then out of the network. During normal conditions it will only take a few minutes for a vehicle to traverse the network.

The structure of the paper is as follows. In the next section a brief overview of the LMDM is given and the type of DAG suitable for an LMDM is considered. In Section 3 a diagram is introduced, called the flow diagram, which gives a pictorial representation of possible vehicle routes through the network. It is shown, in Section 4, how this flow diagram, together with Figure 1, can be used to elicit a suitable DAG for an LMDM. The model associated with the elicited DAG is implemented in Section 5 over a 12 week period. Changes in a traffic network can occur from time to time and Section 6 describes how these changes can be accommodated by the DAG and the LMDM. Finally, Section 7 contains some concluding remarks.

2. The Linear Multiregression Dynamic Model

This section provides a brief (non-technical) overview of the LMDM. For a full account of the model, see Queen & Smith (1993).

Consider a multivariate time series $\boldsymbol{Y}_t = (Y_t(1), \ldots, Y_t(n))^\top$. Suppose that there is a conditional independence and causal structure defined across the series, so that, at each time

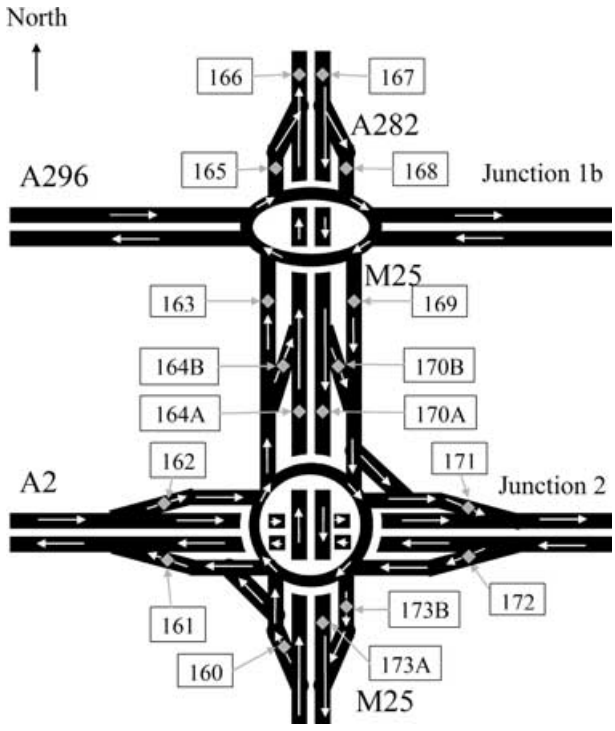

Figure 1. Diagram showing the layout of data collection sites used around the M25/A2/A282 junction. The grey diamonds are the data collection sites, each of which is numbered. The white arrows indicate the direction of traffic flow on each part of the network.

$t = 1, 2, \ldots,$

$$Y_t(i) \amalg \{\{Y_t(1), \ldots, Y_t(i-1)\} \backslash \text{pa}(Y_t(i))\} \,|\text{pa}(Y_t(i)) \qquad \text{for } i = 2, \ldots n$$

which reads "$Y_t(i)$ is independent of $\{Y_t(1), \ldots, Y_t(i-1)\}\backslash\text{pa}(Y_t(i))$ given $\text{pa}(Y_t(i))$" (using the notation that "\" reads "excluding"), where $\text{pa}(Y_t(i)) \subseteq \{Y_t(1), \ldots, Y_t(i-1)\}$. Each variable in the set $\text{pa}(Y_t(i))$ is called a parent of $Y_t(i)$ and $Y_t(i)$ is known as a child of each variable in the set $\text{pa}(Y_t(i))$. A DAG represents these conditional independence relationships within the system pictorially, where each $Y_t(i)$ is represented by a node on the graph and there is a directed arc to $Y_t(i)$ from each of its parents. For example, Figure 2 shows a DAG for four time series at time t, where $\text{pa}(Y_t(2)) = \emptyset$, $\text{pa}(Y_t(3)) = \{Y_t(1), Y_t(2)\}$ and $\text{pa}(Y_t(4)) = \{Y_t(3)\}$.

As $Y_t(i)$ is independent of $\{Y_t(1), \ldots, Y_t(i-1)\}\backslash\text{pa}\,(Y_t(i))$ given $\text{pa}(Y_t(i))$, a forecasting model for $Y_t(i)$ need only depend on $\text{pa}(Y_t(i))$, rather than all the series at time t. An LMDM uses this idea and models the multivariate time series by n separate univariate models – for $Y_t(1)$ and $Y_t(i) \,|\, \text{pa}(Y_t(i))$, $i = 2, \ldots, n$. For each $Y_t(i)$ with parents $\text{pa}(Y_t(i))$, the (conditional) univariate model is simply a regression DLM with $\text{pa}(Y_t(i))$ as linear regressors. For those series without parents, any suitable univariate DLM may be used. As long as the parameters for each (conditional) univariate model are mutually independent *a priori*, they can be updated separately. Forecasts for $Y_t(1)$ and $Y_t(i) \,|\, \text{pa}(Y_t(i))$, $i = 2, \ldots, n$, can also be found separately. The marginal forecast distributions for $Y_t(i)$, $i = 2, \ldots, n$, may not

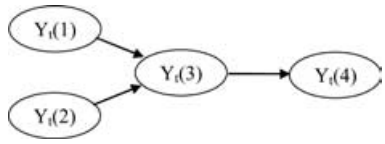

Figure 2. Directed acyclic graph representing four time series at time t.

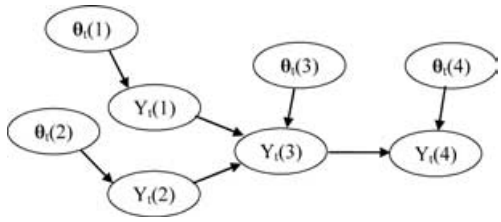

Figure 3. Directed acyclic graph from Figure 2 with independent model parameters included.

be mathematically simple. However, the moments of the marginal forecast distributions can be easily calculated (see Queen, Wright & Albers, 2007).

The joint one-step ahead forecast distribution of $\boldsymbol{Y}_t$ can be expressed as the product of $Y_t(1)$'s forecast distribution and the individual univariate conditional forecast distributions, $Y_t(i) \mid \text{pa}(Y_t(i)), i = 2, \ldots, n$. Even though regression is linear so that each of the univariate forecast distributions for $Y_t(1)$ and $Y_t(i) \mid \text{pa}(Y_t(i)), i = 2, \ldots, n$, is Gaussian, these models can yield highly non-Gaussian joint forecast distributions. As such, they are analogous to non time series graphical models in that although the sub-problems can be fairly simple to work with (in this case univariate DLMs), the joint distribution can be highly complex (Cowell *et al.*, 1999).

To illustrate an LMDM, consider again the time series represented by the DAG in Figure 2. Let $\boldsymbol{\theta}_t(i)$ be the model parameters for $Y_t(i), i = 1, \ldots, 4$. The parameter $\boldsymbol{\theta}_t(i)$ can be considered as another parent of $Y_t(i)$ on the DAG. Further, the LMDM assumes independent (Gaussian) priors for $\boldsymbol{\theta}_t(1), \ldots, \boldsymbol{\theta}_t(4)$. The DAG, including the model parameters, is given in Figure 3. As $Y_t(1)$ and $Y_t(2)$ are both without parents, each of these series can be modelled separately using any suitable univariate DLMs. Both $Y_t(3)$ and $Y_t(4)$ have parents and so these would both be modelled by (separate) univariate regression DLMs with the two regressors $Y_t(1)$ and $Y_t(2)$ for $Y_t(3)$'s model and the single regressor $Y_t(3)$ for $Y_t(4)$'s model.

The forecast distributions for $Y_t(1)$, $Y_t(2)$, and the conditional forecast distributions for $Y_t(3) \mid (Y_t(1), Y_t(2))$ and $Y_t(4) | Y_t(3)$ are all Gaussian, each calculated separately using the priors for $\boldsymbol{\theta}_t(1)$, $\boldsymbol{\theta}_t(2)$, $\boldsymbol{\theta}_t(3)$ and $\boldsymbol{\theta}_t(4)$, respectively. The marginal forecast distributions for $Y_t(3)$ and $Y_t(4)$ may not be mathematically simple and marginal moments usually need to be calculated. After observing $y_t(i)$, the (Gaussian) posterior distributions for each $\boldsymbol{\theta}_t(i) | y_t(i)$ can be calculated. Because of the DAG structure, the parameters $\boldsymbol{\theta}_t(1)$, $\boldsymbol{\theta}_t(2)$, $\boldsymbol{\theta}_t(3)$, $\boldsymbol{\theta}_t(4)$ remain independent *a posteriori*. The (Gaussian) prior distribution for each $\boldsymbol{\theta}_{t+1}(i)$ is then calculated from the posterior for $\boldsymbol{\theta}_t(i) | y_t(i)$, and so the process continues.

It is important to note that whereas two DAGs may exhibit the same conditional independence statements, they can yield quite different LMDMs. For example, consider the two DAGs in Figure 4. In both DAGs $A_t \amalg C_t \mid B_t$. However, they would yield quite different LMDMs. For LMDMs, the DAG needs to represent the conditional independence structure

Figure 4. Two directed acyclic graphs both representing $A_t \amalg C_t \mid B_t$, but yielding different Linear Multiregression Dynamic Models.

related to causality, so that (following Wermuth & Lauritzen, 1990) variables which are hypothesized to be causally linked should be connected by a directed arc following the direction of causation.

3. How does traffic pass through the network?

To investigate the relationships between the time series of vehicle counts, consider first how traffic flows through the sites in the network. Eliciting conditional independence relationships which are consistent with the direction of traffic flow will help to establish the conditional independence structure related to causality.

It is useful to make the simplifying assumption that drivers will behave rationally and follow the most direct route through the network. For example, in Figure 1 suppose that a vehicle is entering the network southbound on the A282 and wishes to exit southbound on the M25. Then it is assumed that the vehicle will take the most direct route (continuing on the M25 at both Junctions 1b and 2, passing sites 167, then 170A and finally 173A) and not use a more indirect route (such as leaving and then rejoining the M25 at Junction 2, passing sites 167, then 170B and finally 173B). Although some drivers may behave irrationally in this way, it is unlikely that such behaviour is common.

By considering the layout of sites (Figure 1), it is possible to draw a diagram representing how vehicles pass through the sites in the network. Represent each site by an oval and draw an arrow leading from one site to another if there is a direct route from one to the other; see Figure 5. This will be called the flow diagram. Note that flows into and out of the network itself are also shown in the flow diagram.

The flow diagram helps to give some insight into the relationships between the time series of vehicle counts. For example, knowing that vehicles at site 167 can go to sites 168, 170A or 170B only, means that these time series will be highly correlated and that the sum of

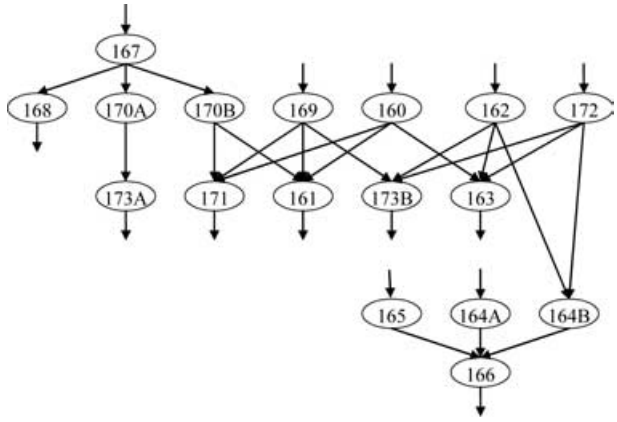

Figure 5. Flow diagram for the network.

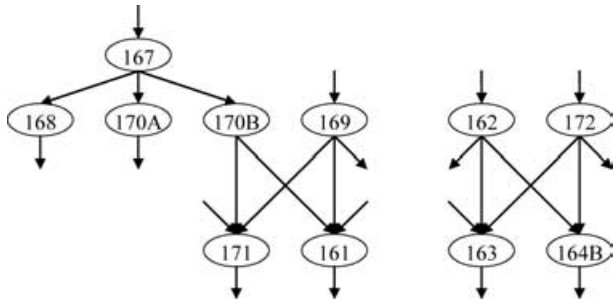

Figure 6. Flow diagram for the network with missing and trivial sites removed.

vehicles at sites 168, 170A and 170B at time t should be the number of vehicles counted at site 167 at time t (approximately, accounting for vehicles which are between sites at the end of the hour and measurement error in the vehicle detectors). It also means that the time series at 167 is hypothesized to be causally linked to the series at sites 168, 170A and 170B.

Unfortunately due to faulty data collection equipment, no data were collected at some sites. The missing data could be estimated using Markov chain Monte Carlo techniques (see Whitlock & Queen, 2000). However, to allow evaluation of the model, only time series for which data are observed will be considered here. The sites for which no data are available are 160, 166, 173A and 173B. Figure 6 shows a new flow diagram with these sites removed. When site 166 is removed, sites 164A and 165 become disconnected from the rest of the network. As this paper aims to examine the multivariate nature of the traffic network, these two sites shall be dropped from the model here for simplicity. Notice also that the network is now subdivided into two separate subnetworks.

4. Eliciting a DAG for the network

Consider the flow diagram in Figure 6, which shows how vehicles flow through the (observed) sites in the network. This flow diagram, together with the diagram of the traffic network in Figure 1, will be used to heuristically elicit a DAG for the time series which can be used for an LMDM. To do this it is helpful to form the sites into three groups: {167, 168, 170A, 170B}, {170B, 169, 171, 161} and {162, 172, 163, 164B}.

For each site s, denote the vehicle count at time t by $Y_t(s)$. Following Vaughan (2001), it is natural to make the assumption that hourly traffic counts follow a Poisson distribution with a different rate for each of the 24 hours in a day. Therefore, for hour t and site s, $Y_t(s) \sim \text{Po}(\mu_t(s))$, where $\text{Po}(\mu_t(s))$ denotes a Poisson distribution with mean $\mu_t(s)$. In what follows the following results will be used:

- For any two variables X_1 and X_2 such that $X_1 \sim \text{Po}(\mu_1)$ and $X_2 \sim \text{Po}(\mu_2)$, the distribution of $(X_1|X_1 + X_2)$ is binomial so that $(X_1 \mid X_1 + X_2) \sim \text{Bi}(X_1 + X_2, \mu_1/(\mu_1 + \mu_2))$, where $\text{Bi}(n, p)$ denotes a binomial distribution of sample size n with parameter p.
- For $\text{Bi}(n, p)$ (n large), $\text{Bi}(n, p) \approx N(np, np(1 - p))$.
- For $\text{Po}(\mu)$ (μ large), $\text{Po}(\mu) \approx N(\mu, \mu)$.

4.1. DAG for sites 167, 168, 170A and 170B

As traffic only flows from site 167 to the other three, $Y_t(167)$ should be (approximately) equal to the sum of $Y_t(168)$, $Y_t(170\text{A})$ and $Y_t(170\text{B})$. It is therefore possible to define the conditional distribution

$$Y_t(168) \,|\, Y_t(167) \sim \text{Bi}\left(Y_t(167), \frac{\mu_t(168)}{\mu_t(167)}\right)$$

with similar conditional distributions for $Y_t(170\text{A})|Y_t(167)$ and $Y_t(170\text{B})|Y_t(167)$. This could be represented by a DAG with $Y_t(168)$, $Y_t(170\text{A})$ and $Y_t(170\text{B})$ as children of $Y_t(167)$. As hourly vehicle counts are typically large, these conditional binomial distributions can be approximated by conditional normal distributions — for example,

$$Y_t(168)|Y_t(167) \approx N\left(Y_t(167)\left(\frac{\mu_t(168)}{\mu_t(167)}\right), Y_t(167)\left(\frac{\mu_t(168)}{\mu_t(167)}\right)\left(1-\frac{\mu_t(168)}{\mu_t(167)}\right)\right).$$

Then the observation equation in an LMDM for the (conditional) univariate model for $Y_t(168)$, for example, would be of the form:

$$Y_t(168) = Y_t(167)\left(\frac{\mu_t(168)}{\mu_t(167)}\right) + v_t(168), \qquad v_t(168) \sim N(0, V_t(168)).$$

However, the parameters for the three distributions for $Y_t(168)$, $Y_t(170\text{A})$ and $Y_t(170\text{B})$, conditional on $Y_t(167)$, are not independent of one another. Consequently the parameters for each univariate model in an LMDM could not be considered mutually independent. To ensure independent model parameters, the DAG needs to be elicited in a slightly different way.

Consider $Y_t(167)$. From Figure 1, at Junction 1b, a proportion of the vehicles making up $Y_t(167)$ will continue southbound onto the M25 to become $Y_t(170\text{A}) + Y_t(170\text{B})$, and the rest will leave to become $Y_t(168)$. Of the vehicles making up $Y_t(170\text{A}) + Y_t(170\text{B})$, a proportion will leave the M25 at Junction 2 to become $Y_t(170\text{B})$ and the rest will continue on the M25 to become $Y_t(170\text{A})$. Thus there are two alternative conditional distributions:

$$\begin{aligned}(Y_t(170\text{A}) + Y_t(170\text{B}))|Y_t(167) &\sim \text{Bi}\,(Y_t(167), \alpha_t)\\ Y_t(170\text{B})|(Y_t(170\text{A}) + Y_t(170\text{B})) &\sim \text{Bi}\,(Y_t(170\text{A}) + Y_t(170\text{B}), \beta_t)\end{aligned}$$

where $\alpha_t = (\mu_t(170\text{A}) + \mu_t(170\text{B}))/\mu_t(167)$ and $\beta_t = \mu_t(170\text{B})/(\mu_t(170\text{A}) + \mu_t(170\text{B}))$. Both parameters α_t and β_t are interpretable:

α_t = proportion of vehicles at 167 continuing south on to the M25 at Junction 1b
β_t = proportion of those vehicles continuing south on the M25 after Junction 1b that leave the M25 at Junction 2

These conditional distributions can be represented by the DAG in Figure 7. Independence of parameters is now a reasonable assumption because there is no structural reason to believe otherwise. Here both $Y_t(168)$ and $Y_t(170\text{A})$ are logical functions of their parents and are known once their parents are known. Following the terminology of WinBUGS software (http://www.mrc-bsu.cam.ac.uk/bugs/), these will be called logical variables and denoted on the DAG by a double oval. Note that $Y_t(170\text{A}) + Y_t(170\text{B})$ and/or $Y_t(170\text{B})$ could have been chosen as the logical variables instead. The series $Y_t(168)$ and $Y_t(170\text{A})$ were chosen simply because site 170B leads to other parts of the network.

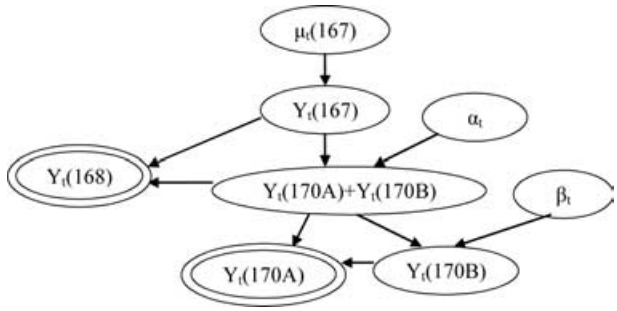

Figure 7. Directed acyclic graph representing $Y_t(167)$, $Y_t(168)$, $Y_t(170A)$ and $Y_t(170B)$, together with the model parameters.

Approximating the Poisson distribution for $Y_t(167)$ and the conditional binomial distributions to normality, the observation equations in an LMDM for the DAG in Figure 7 are of the following forms:

$$\begin{aligned}
Y_t(167) &= \mu_t(167) + v_t(167), & v_t(167) &\sim N(0, V_t(167)) \\
Y_t(170A) + Y_t(170B) &= Y_t(167)\alpha_t + v_t(170A + 170B), & & \\
& & v_t(170A + 170B) &\sim N(0, V_t(170A + 170B)) \\
Y_t(170B) &= (Y_t(170A) + Y_t(170B))\,\beta_t + v_t(170B), & v_t(170B) &\sim N(0, V_t(170B)) \\
\text{and} \quad Y_t(168) &= Y_t(167) - (Y_t(170A) + Y_t(170B)) & & \\
Y_t(170A) &= (Y_t(170A) + Y_t(170B)) - Y_t(170B). & &
\end{aligned}$$

Note that although the observation equations are algebraically the same for each time t, the actual parameters ($\mu_t(167)$, α_t and β_t) will vary depending on which hour of the day t is. For example, if time t is the hour 1:00–2:00 a.m. when the roads are very quiet, $\mu_t(167)$ will be quite small, but if time t is the hour 8:00–9:00 a.m. during the morning rush hour, $\mu_t(167)$ will be much larger.

4.2. DAG for sites 170B, 169, 171 and 161

From Figure 6, vehicles at 170B flow to both sites 171 and 161, and the same is true for vehicles at site 169. Vehicles from the unobserved site 160 also flow to both sites 171 and 161. At first sight it seems reasonable to draw a DAG with $Y_t(171)$ and $Y_t(161)$ as children of both $Y_t(170B)$ and $Y_t(169)$. However, the parameters for the conditional distributions $Y_t(171)|(Y_t(170B), Y_t(169))$ and $Y_t(161)|(Y_t(170B), Y_t(169))$ are not independent. Thus $Y_t(171)$ and $Y_t(161)$ cannot be modelled as separate children of $Y_t(170B)$ and $Y_t(169)$ using an LMDM. Instead, following the methods of Subsection 4.1, a DAG can be elicited using the conditional distributions $(Y_t(161) + Y_t(171))|(Y_t(170B), Y_t(169))$ and $Y_t(161)|(Y_t(161) + Y_t(171))$. The latter of these conditional distributions is straightforward:

$$Y_t(161)|(Y_t(171) + Y_t(161)) \sim \text{Bi}(Y_t(171) + Y_t(161), \delta_t),$$

where $\delta_t = \mu_t(161)/(\mu_t(171) + \mu_t(161))$, the proportion of those vehicles joining the A2 that travel west bound. However, the conditional distribution $(Y_t(161) + Y_t(171))|(Y_t(170B), Y_t(169))$ requires more thought.

Consider the sum $Y_t(171) + Y_t(161)$. The vehicles making up this sum come from three sources: 170B, 169 and the unobserved site 160. All vehicles at 170B flow to either 171 or 161, whereas only a proportion of the vehicles at 169 do. Write

$$Y_t(171) + Y_t(161) = X_t(u) + Y_t(170\text{B}) + X_t(169) \tag{1}$$

where $X_t(u) =$ vehicles at 171 or 161 at time t inherited from the unobserved site
$X_t(169) =$ vehicles at 171 or 161 at time t inherited from site 169

Model $X_t(x) \sim \text{Po}(\mu_t(X(x)))$. Then

$$X_t(u) \sim \text{Po}(\mu_t(X(u)) \quad \text{and} \quad X_t(169)|Y_t(169) \sim \text{Bi}(Y_t(169), \gamma_t)$$

where $\gamma_t = \mu_t(X(169))/\mu_t(169)$, the proportion of those vehicles travelling south from Junction 1b that join the A2. Approximate the Poisson distributions for $X_t(u)$ and $Y_t(170\text{B})$, and the conditional binomial distribution for $X_t(169)|Y_t(169)$ to normality. Then, using equation (1), the conditional distribution for $(Y_t(171) + Y_t(161))|(Y_t(170\text{B}), Y_t(169))$ is approximately normal with mean $\mu_t(X(u)) + Y_t(170\text{B}) + Y_t(169)\gamma_t$.

A DAG representing the conditional distributions defined here for this part of the network is shown in Figure 8. An LMDM using this DAG has the two regressors $Y_t(170\text{B})$ and $Y_t(169)$ in the DLM for $Y_t(171) + Y_t(161)$, with the regression parameter for $Y_t(170\text{B})$ set to 1. Unfortunately, $Y_t(170\text{B})$ and $Y_t(169)$ are highly correlated leading to problems with collinearity. In conventional regression, collinearity is often dealt with by dropping one of the correlated regressors. However, here it is desirable to retain the information contained in both parents. Let mean169 and sd169 be the mean and standard deviation of hourly counts of vehicles at site 169 over some period in the past, and let mean170B and sd170B be the corresponding values at site 170B. Define two new orthogonal variables

$$Z_t(1) = U_t + V_t, \qquad Z_t(2) = U_t - V_t,$$

with

$$U_t = \frac{Y_t(169) - \text{mean169}}{\text{sd169}} \quad \text{and} \quad V_t = \frac{Y_t(170\text{B}) - \text{mean170B}}{\text{sd170B}}.$$

Without any loss of information, $Z_t(1)$ and $Z_t(2)$ can be considered as independent regressors in the DLM for $Y_t(171) + Y_t(161)$, instead of the correlated variables $Y_t(170\text{B})$ and $Y_t(169)$.

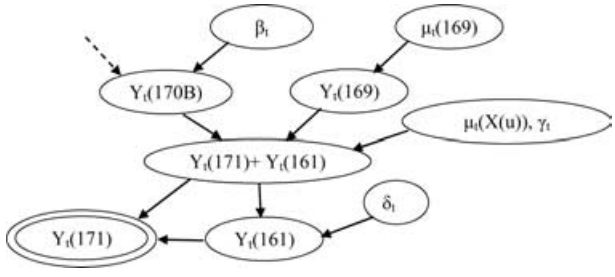

Figure 8. Directed acyclic graph representing $Y_t(170\text{B})$, $Y_t(169)$, $Y_t(171)$ and $Y_t(161)$, together with the model parameters.

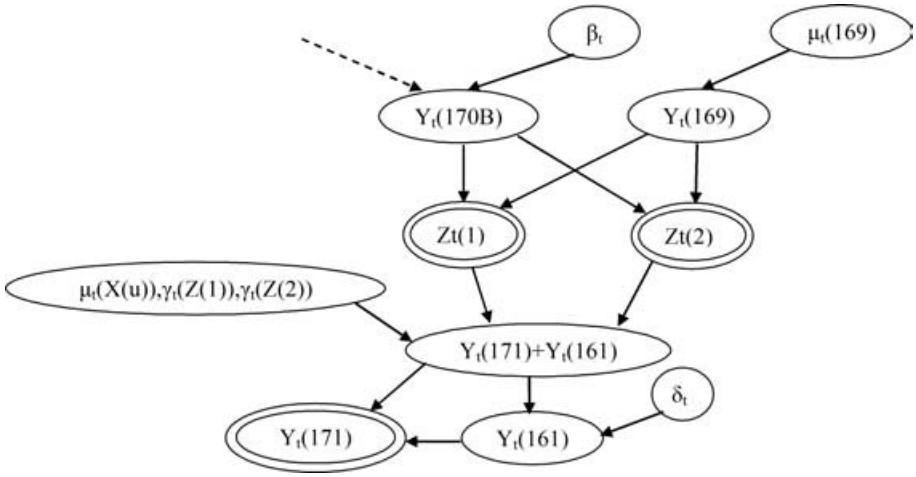

Figure 9. Final elicited directed acyclic graph representing $Y_t(170B)$, $Y_t(169)$, $Y_t(171)$ and $Y_t(161)$, together with the model parameters.

Introduce $Z_t(1)$ and $Z_t(2)$ into the DAG as logical children of $Y_t(170B)$ and $Y_t(169)$, and parents of $Y_t(171) + Y_t(161)$. Then $(Y_t(171) + Y_t(161))|(Z_t(1), Z_t(2))$ is approximately normal with mean

$$\mathrm{E}(Y_t(171) + Y_t(161)|(Z_t(1), Z_t(2))) = \mu_t(X(u)) + Z_t(1)\gamma_t(Z(1)) + Z_t(2)\gamma_t(Z(2)) \quad (2)$$

for parameters $\gamma_t(Z(1))$ and $\gamma_t(Z(2))$. Unfortunately, $\gamma_t(Z(1))$ and $\gamma_t(Z(2))$ are not easily interpretable. However, using $Z_t(1)$ and $Z_t(2)$ does allow us to build a DAG which still respects the conditional independence structure of the series and produces an LMDM which is still computationally simple. The final elicited DAG for sites 170B, 169, 171 and 161 is given in Figure 9.

Approximating the Poisson distribution for $Y_t(169)$ and the conditional binomial distribution for $Y_t(161)|(Y_t(171) + Y_t(161))$ to normality, an LMDM representing the DAG in Figure 9 then has the following observation equations for $Y_t(169)$, $Y_t(171) + Y_t(161)$ and $Y_t(161)$:

$$\begin{aligned}
Y_t(169) &= \mu_t(169) + v_t(169), & v_t(169) &\sim N(0, V_t(169)) \\
Y_t(171) + Y_t(161) &= \mu_t(X(u)) + Z_t(1)\gamma_t(Z(1)) + Z_t(2)\gamma_t(Z(2)) + v_t(171 + 161), \\
& & v_t(171 + 161) &\sim N(0, V_t(171 + 161)) \\
Y_t(161) &= (Y_t(171) + Y_t(161))\,\delta_t + v_t(161), & v_t(161) &\sim N(0, V_t(161)) \\
\text{and} \quad Y_t(171) &= (Y_t(171) + Y_t(161)) - Y_t(161).
\end{aligned}$$

4.3. DAG for sites 162, 172, 163 and 164B

The flow diagram for this group of sites is almost identical in structure to that for the four sites in Subsection 4.2. As was the case there, it is not possible to simply draw a DAG for an LMDM with $Y_t(163)$ and $Y_t(164B)$ as children of $Y_t(162)$ and $Y_t(172)$, because the parameters of the conditional distributions $Y_t(163)|(Y_t(162), Y_t(172))$

and $Y_t(164B)|(Y_t(162), Y_t(172))$ would not be independent. Instead, following the methods of Subsections 4.1 and 4.2, a DAG can be elicited using the conditional distributions $(Y_t(163) + Y_t(164B))|(Y_t(162), Y_t(172))$ and $Y_t(163)|(Y_t(163) + Y_t(164B))$. The DAG and associated LMDM for this part of the model can be elicited in exactly the same way as was described for the second group of sites in Subsection 4.2. To avoid repetitiveness, the details will be omitted here, except to mention that (as with the four sites in Subsection 4.2) in order to avoid problems of collinearity, two new orthogonal variables are introduced into the model. These are

$$Z_t(3) = U_t^* + V_t^*, \qquad Z_t(4) = U_t^* - V_t^*$$

with

$$U_t^* = \frac{Y_t(162) - \text{mean162}}{\text{sd162}} \quad \text{and} \quad V_t^* = \frac{Y_t(172) - \text{mean172}}{\text{sd172}}$$

where mean162 and sd162 are the mean and standard deviation of hourly counts of vehicles at site 162 over some period in the past, and mean172 and sd172 are the corresponding values at site 172.

The final DAG for sites 162, 172, 163 and 164B is given in Figure 10 and the associated observation equations for an LMDM are as follows:

$$\begin{aligned}
Y_t(162) &= \mu_t(162) + v_t(162), & v_t(162) &\sim N(0, V_t(162)) \\
Y_t(172) &= \mu_t(172) + v_t(172), & v_t(172) &\sim N(0, V_t(172)) \\
Y_t(163) + Y_t(164B) &= \mu_t(X'(u)) + Z_t(3)\varepsilon_t(Z(3)) + Z_t(4)\varepsilon_t(Z(4)) + v_t(163 + 164B), \\
& & v_t(163 + 164B) &\sim N(0, V_t(163 + 164B)) \\
Y_t(164B) &= (Y_t(163) + Y_t(164B))\,\zeta_t + v_t(164B), & v_t(164B) &\sim N(0, V_t(164B)) \\
\text{and} \quad Y_t(163) &= (Y_t(163) + Y_t(164B)) - Y_t(164B).
\end{aligned}$$

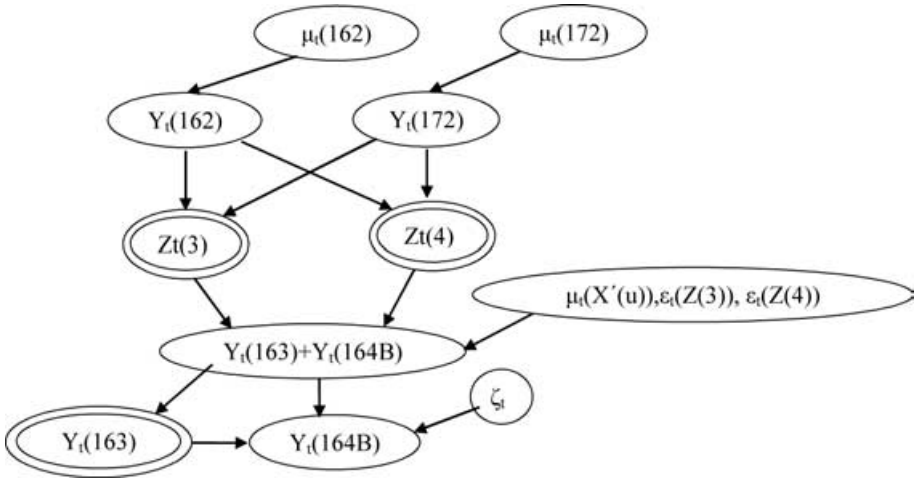

Figure 10. Final elicited directed acyclic graph representing $Y_t(162)$, $Y_t(172)$, $Y_t(163)$ and $Y_t(164B)$, together with the model parameters.

The parameter $\mu_t(X'(u))$ represents the vehicles at site 163 inherited from the unobserved site 160. The parameter $\mu_t(X(u))$ (from equation 2) represents the vehicles at sites 161 or 171 inherited from the same unobserved site. So the parameters $\mu_t(X'(u))$ and $\mu_t(X(u))$ are correlated. However, as $Y_t(160)$ is unobserved, for practical purposes knowing one of these parameters will not actually tell us very much about the other. So, for simplicity it will be assumed that $\mu_t(X'(u))$ and $\mu_t(X(u))$ are independent.

5. Model performance

The focus of this paper is the elicitation of a suitable DAG representing the traffic network, and its associated LMDM. Using the resulting LMDM for forecasting is an interesting problem in its own right and is addressed in detail in another paper (in preparation). It is, however, useful to illustrate the model's performance over a short period when the series are 'well-behaved', so that forecasting is straightforward.

The time series of vehicle counts exhibit different daily patterns for different days of the week. For example, the general pattern of hourly counts observed on Sundays is quite different to that observed on Mondays. These different daily patterns can be easily accommodated in the model with a little extra work. However, for simplicity, in this illustration only data for Tuesday, Wednesday and Thursday each week are considered, as the daily patterns are similar for these days.

The LMDM defined in Section 4 was applied to data for Tuesday-Thursday over a 14 week period. As genuine expert priors were not available, the first two weeks of data were used to form initial priors for the model parameters. The one-step ahead forecasts were then calculated for the remaining 12 weeks.

Vehicle count series often exhibit unexpected changes in behaviour, due to events such as road accidents or adverse weather conditions. As a consequence, outliers can occur in the series, producing unexpected large forecast errors. This is illustrated in Figure 11. The graphs on the left in Figure 11 show the one-step ahead forecasts (solid line) and ± 2 forecast standard deviations (dotted lines) obtained using the LMDM, together with the actual values observed (dots) for series $Y_t(167)$ and $Y_t(170\text{A}) + Y_t(170\text{B})$ during week eight. The graphs on the right are the one-step ahead forecast errors with ± 2 forecast standard deviation error bars for the same series over the same time period. For both series, an unusually large negative forecast error occurs at time 560 (07.00-08.00 on Thursday, week 8), followed by an unusually large positive forecast error. This pattern of forecast errors is consistent with a slowdown in traffic flow, for example due to a temporary block in the road following a crash, followed by an increase in traffic flow as the problem is resolved and delayed cars move through the network. Such patterns are not uncommon in traffic networks. The median squared error (MedianSE) has therefore been used when assessing model forecast performance, as this measure is more robust to possible outliers such as these.

Recall that vehicles at site 167 move to sites 170A, 170B and 168. It is therefore no surprise that $Y_t(167)$ and $Y_t(170\text{A}) + Y_t(170\text{B})$ exhibit similar patterns at times 560 and 561. Figure 12 shows the one-step ahead forecast and error graphs for series $Y_t(168)$ and $Y_t(163)$. As might be expected, the same pattern can also be seen for $Y_t(168)$ at times 560 and 561. However, the same pattern is not evident for series $Y_t(163)$, as site 163 is counting vehicles travelling in the opposite direction.

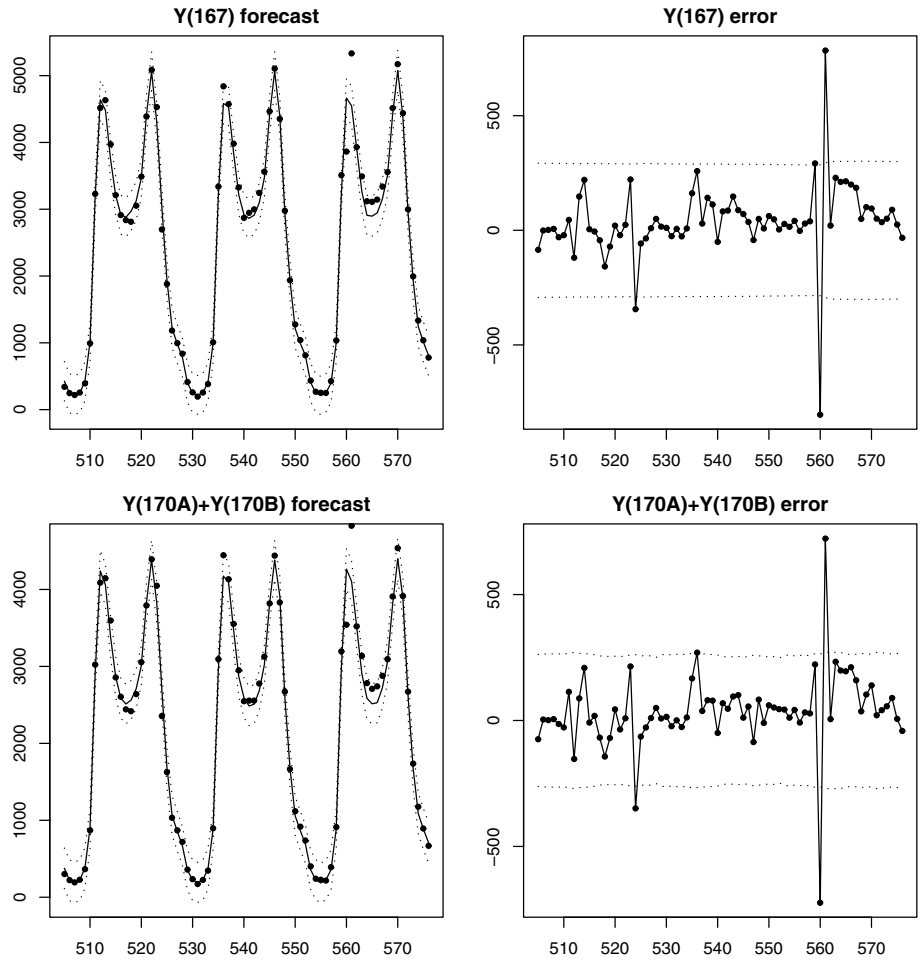

Figure 11. The graphs on the left show the one-step ahead forecasts (solid line) and ±2 forecast standard deviations (dotted lines) obtained using the Linear Multiregression Dynamic Model, together with the actual values observed (dots) for series $Y_t(167)$ and $Y_t(170A) + Y_t(170B)$ during week eight. The graphs on the right are the one-step ahead forecast errors with ±2 forecast standard deviation error bars for the same series over the same time period.

As with any Bayesian DLM, intervention can be used in the LMDM to accommodate unexpected changes in the behaviour of series, and thus improve forecast performance. Although the forecast performance was poor for all three of the series $Y_t(167)$, $Y_t(170A) + Y_t(170B)$ and $Y_t(168)$ at times 560 and 561, the LMDM exploits the causality between the series so that intervention is only required in the single univariate DLM for $Y_t(167)$. This was done as follows. The observation $y_{560}(167)$ was unexpected and so was treated simply as an outlier. During the following time period ($t = 561$), as the road blockage clears and vehicles start moving, the delayed vehicles (from the previous hour) are expected to pass site 167, in addition to the vehicles that arrive during hour $t = 561$. The expected number of cars delayed from hour 560 is $f_{560}(167) - y_{560}(167)$, where $f_{560}(167)$ is the one-step forecast

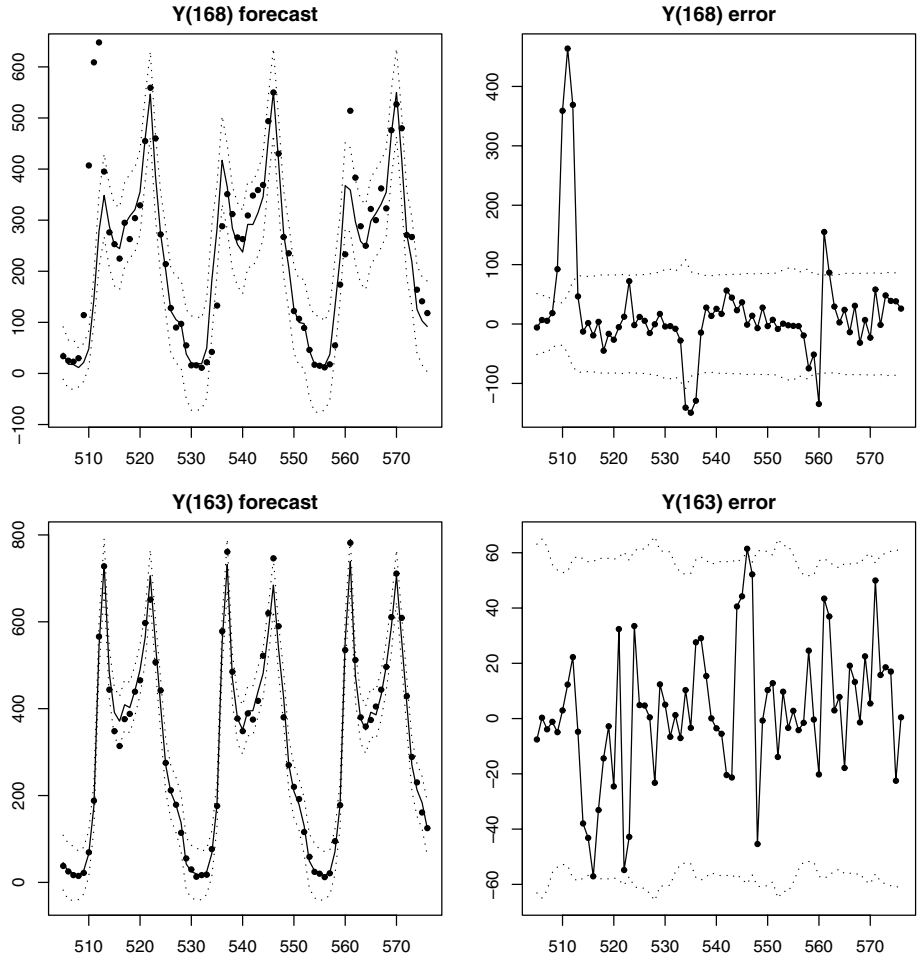

Figure 12. The graphs on the left show the one-step ahead forecasts (solid line) and ±2 forecast standard deviations (dotted lines) obtained using the Linear Multiregression Dynamic Model, together with the actual values observed (dots) for series $Y_t(168)$ and $Y_t(163)$ during week eight. The graphs on the right are the one-step ahead forecast errors with ±2 forecast standard deviation error bars for the same series over the same time period.

for $y_{560}(167)$. Thus the observation equation is adjusted at hour 561 so that

$$Y_{561}(167) = \mu_{561}(167) + f_{560}(167) - y_{560}(167) + v_{561}(167), \quad v_{561}(167) \sim N(0, V_{561}(167)).$$

The resulting plots of one-step forecasts and forecast errors during week eight, after this intervention was used, are shown in Figure 13 for series $Y_t(167)$ and $Y_t(170\text{A}) + Y_t(170\text{B})$. Following intervention in the model for $Y_{561}(167)$, there is now a good forecast for $Y_{561}(167)$. There is also a good forecast for $Y_{561}(170\text{A}) + Y_{561}(170\text{B})$, due to the fact that $Y_t(167)$ is a parent of $Y_t(170\text{A}) + Y_t(170\text{B})$. There is a similar improvement in the forecast for $Y_{561}(168)$ (not shown), whereas the forecast for $Y_{561}(163)$, for example, is practically unaffected by the change in $Y_{561}(167)$'s model.

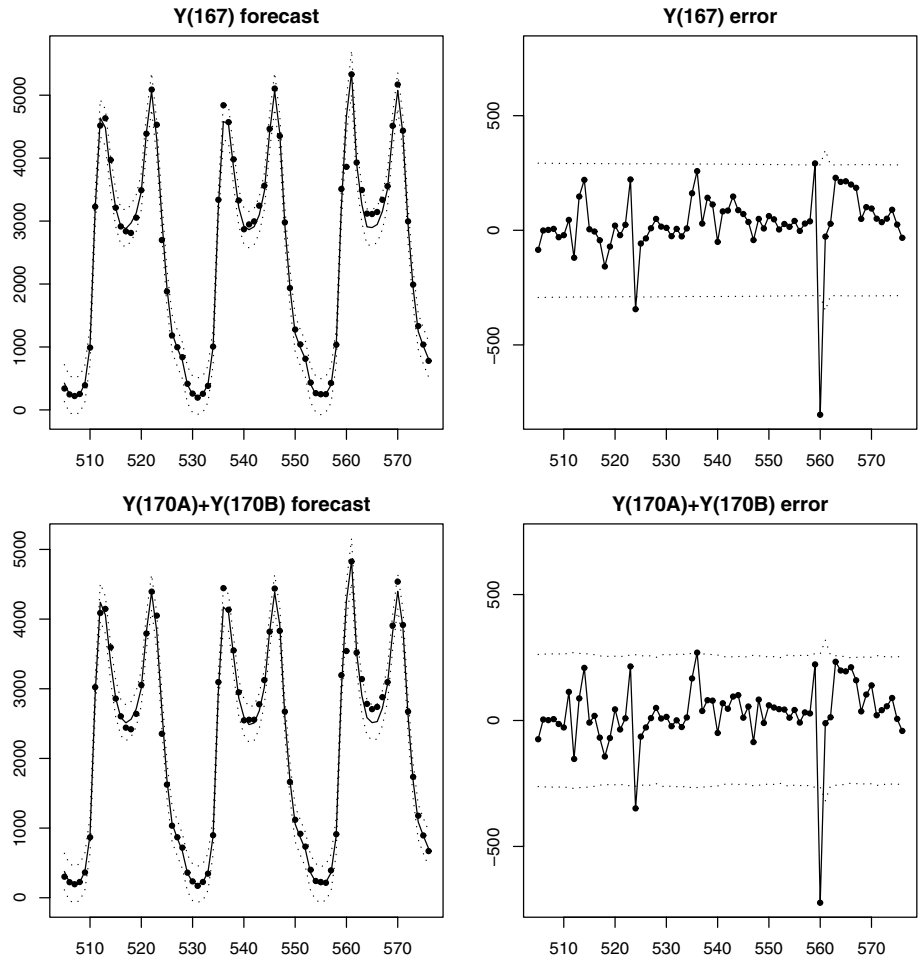

Figure 13. The graphs on the left show the one-step ahead forecasts (solid line) and ±2 forecast standard deviations (dotted lines) obtained using the Linear Multiregression Dynamic Model, together with the actual values observed (dots) for series $Y_t(167)$ and $Y_t(170A) + Y_t(170B)$, after intervention was used for $Y_t(167)$ at times 560 and 561. The graphs on the right are the one-step ahead forecast errors with ±2 forecast standard deviation error bars for the same series over the same time period.

Table 1 shows the MedianSE for each series using the LMDM, both with and without intervention. The improvement in forecast performance when using intervention is clearly seen. It is interesting to note that while the forecast for $Y_t(170A) + Y_t(170B)$ improved after intervening in $Y_{561}(167)$'s model, the forecasts for the individual series $Y_t(170B)$ did not.

The natural model to compare the LMDM's performance with is a standard multivariate DLM. However, unfortunately, this model cannot be used for these series because the series are so highly correlated that the computation of the DLM updating algorithm breaks down (the inverse of an estimate of the observation covariance matrix cannot be calculated since the determinant is too small). Instead, the model performance of univariate DLMs were used for comparison because univariate DLMs are similar to the LMDM in the level of simplicity in implementation. The MedianSE for the one-step forecasts for univariate DLMs

TABLE 1

MedianSEs for each series using the LMDM, both with and without intervention, and using univariate DLMs without any intervention

Series	LMDM (no intervention)	LMDM (intervention)	Univariate DLMs (no intervention)
$Y(167)$	4915	4834	4915
$Y(170A) + Y(170B)$	4734	4637	
$Y(168)$	125	125	132
$Y(170A)$	2272	2225	2298
$Y(170B)$	821	823	835
$Y(169)$	202	202	202
$Y(161) + Y(171)$	7984	7984	
$Y(161)$	927	927	796
$Y(171)$	3971	3971	3973
$Y(162)$	812	812	812
$Y(172)$	1088	1088	1088
$Y(163) + Y(164B)$	1599	1599	
$Y(163)$	174	174	175
$Y(164B)$	1019	1019	981

are given in the final column of Table 1. The values of the MedianSE for the LMDM (without intervention) are lower for all series except $Y_t(161)$ and $Y_t(164B)$. The models for both of these series depend on the unobserved series $Y_t(160)$, which perhaps explains the LMDM's poor performance for these series.

When intervention is required (as it frequently is for traffic networks), the LMDM has a clear advantage over univariate (or indeed, standard multivariate) DLMs. As was seen in Figures 11 and 12, when there is an unexpected large forecast error in one series, then this is invariably accompanied by an unexpected large forecast error in its children's series and further down the DAG. The LMDM only requires intervention in the parent's model, whereas univariate or standard multivariate DLMs require intervention not only in the parent's model, but also in all of its children's models. This is not only more work, but it can also be hard to formulate the required intervention in the children's models. For example, if $Y_t(167)$ was estimated to increase by an extra 500, say, at hour t, then it is not immediately obvious how $Y_t(171)$, for example, might change as a consequence. In large networks, intervening for so many series would simply be infeasible in real time.

6. Accommodating changes in the network

Changes in the network can occur for a variety of reasons and for various lengths of time. For example, an accident may cause a short term temporary diversion; roadworks may cause a longer term temporary diversion; or the road layout may be altered permanently. The DAG representing the time series of vehicle counts will need to be altered to accommodate any such changes in the network. Luckily, because of the structure of the LMDM, much of the posterior information for parameters in the original DAG can be carried forward into the new DAG. Additionally, it is also possible for the posteriors for the original parameters to help form priors for any new parameters. The following example will illustrate how this might be done.

6.1. Example: blocked road at site 170B

Suppose that the road at site 170B is temporarily blocked from time t due to roadworks, for example. Suppose further that vehicles travelling southbound who wish to leave the M25 at Junction 2 are diverted via Junction 1B and sites 168 and 169.

Figure 14 shows the flow diagram which reflects the new possible routes through the network. Using exactly the same methods as in Section 4, this flow diagram and the diagram of the traffic network (Figure 1) can be used to elicit a DAG for the new network. A suitable new DAG is given in Figure 15 for the first part of the flow diagram. Sites 162, 172, 163 and 164B are unaffected by the network change, and so their DAG would remain as in Figure 10.

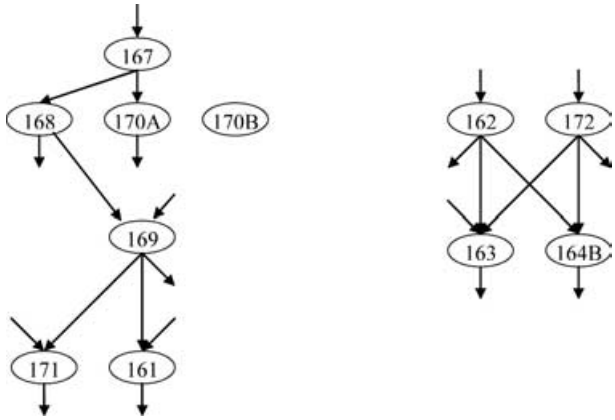

Figure 14. Flow diagram for the network when the road is blocked at site 170B.

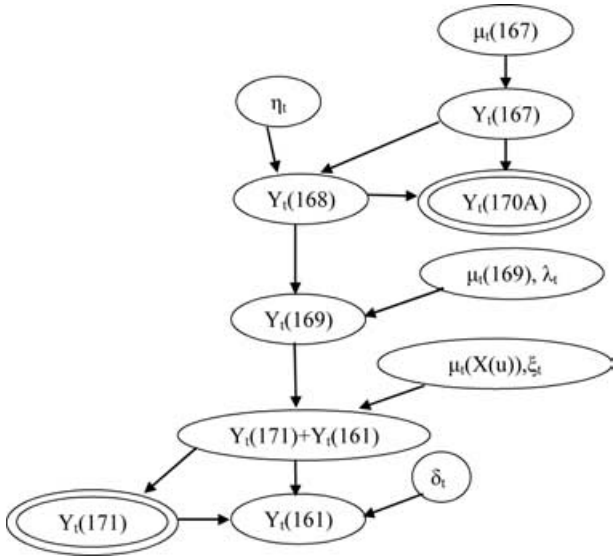

Figure 15. New directed acyclic graph for the first part of the flow diagram when the road is blocked at site 170B.

The new observation equations are as follows.

$$
\begin{aligned}
Y_t(167) &= \mu_t(167) + v_t(167), && v_t(167) \sim N(0, V_t(167)) \\
Y_t(168) &= \eta_t Y_t(167) + v_t(168), && v_t(168) \sim N(0, V_t(168)) \\
Y_t(169) &= \mu_t(169) + \lambda_t Y_t(168) + v_t(169), && v_t(169) \sim N(0, V_t(169)) \\
Y_t(171) + Y_t(161) &= \mu_t(X(u)) + \xi_t Y_t(169) + v_t(171+161), \\
& && v_t(171+161) \sim N(0, V_t(171+161)) \\
Y_t(161) &= \delta_t\,(Y_t(171) + Y_t(161)) + v_t(161), && v_t(161) \sim N(0, V_t(161)) \\
\text{and} \quad Y_t(170\text{A}) &= Y_t(167) - Y_t(168) \\
Y_t(171) &= (Y_t(171) + Y_t(161)) - Y_t(161)
\end{aligned}
$$

There are three new parameters here and each is easily interpretable: η_t is the proportion of those vehicles at 167 leaving the M25 southbound at Junctions 1b or 2, λ_t is the proportion of those vehicles leaving southbound at Junction 1b that are following the diversion to Junction 2, and ξ_t is the proportion of those vehicles travelling south from Junction 1b (which includes diverted traffic) that join the A2.

Notice that several of the parameters ($\mu_t(167)$, $\mu_t(169)$, $\mu_t(X(u))$, δ_t) remain in the model and so their posteriors carry through to form priors at time t under the new model. It is possible to elicit priors for the new parameters η_t and λ_t as follows. Calculate the prior means at time t for the original parameters α_t and β_t and the one step ahead forecasts for $Y_t(170\text{B})$ and $Y_t(168)$ assuming no change in the DAG and model (i.e. assuming the DAG in Figure 7 still holds). Denote these by $\hat{\alpha}_t$, $\hat{\beta}_t$, $\hat{Y}_t(170\text{B})$ and $\hat{Y}_t(168)$. Denoting the first $t-1$ observations by $\underline{y}^{t-1}$, prior mean estimates of the new parameters η_t and λ_t can be obtained using

$$
\mathrm{E}(\eta_t|\underline{y}^{t-1}) = 1 - \hat{\alpha}_t + \hat{\alpha}_t\hat{\beta}_t \quad \text{and} \quad \mathrm{E}(\lambda_t|\underline{y}^{t-1}) = \frac{\hat{Y}_t(170\text{B})}{\hat{Y}_t(170\text{B}) + \hat{Y}_t(168)}.
$$

The most obvious prior estimate for ξ_t is of the form

$$
\mathrm{E}(\xi_t|\underline{y}^{t-1}) = \frac{\hat{Y}_t(170\text{B}) + \hat{Y}_t(169)\gamma_t}{\hat{Y}_t(170\text{B}) + \hat{Y}_t(169)}
$$

where γ_t was defined in Section 4.2 as the proportion of vehicles at 169 flowing to 171 or 161, and $\hat{Y}_t(169)$ is the one step ahead forecast for $Y_t(169)$ assuming no change in the DAG and model. However, the parameter γ_t was not used in $Y_t(171) + Y_t(161)$'s model because of the collinearity problem, so posterior information for γ_t is not available from the LMDM. A prior which is more vague therefore needs to be placed on ξ_t.

Once the temporary blockage at 170B is removed at time $t+k$, the original DAG (Figure 7 and 9) and LMDM for the network are again used. Posterior information at time $t+k-1$ forms priors at time $t+k$ for those parameters used in both models ($\mu_{t+k}(167)$, $\mu_{t+k}(169)$, $\mu_{t+k}(X(u))$, δ_{t+k}), whereas the posteriors at time $t-1$ can be used to form priors for α_{t+k}, β_{t+k}, $\gamma_{t+k}(Z(1))$ and $\gamma_{t+k}(Z(2))$.

7. Discussion

This paper has primarily focused on the crucial first stage in implementing the LMDM — namely, the elicitation of a DAG and associated simple model accommodating the multivariate

structure of the series. A traffic network can potentially have a huge number of counting sites which would make eliciting the DAG by hand impractical. There is therefore a need for developing an automated procedure based on the techniques presented in this paper.

Although logical variables in the network have a deterministic relationship with their parents, in practice these variables may not actually be deterministic due to measurement error in the vehicle detectors. Measurement error will only be a problem if the observed values of the logical variables are not matching up with the expected values for these series. This would be a particular problem if poor forecasting of the logical variables is then feeding through to any children of logical variables. Any logical variable with poor forecasts should therefore account for measurement error in its model. However, if the forecasts are reasonable, there is little to be gained by including measurement error in a logical variable's model.

Using the model for forecasting is an interesting problem in itself and there are several important issues which have not been covered in this paper. Firstly, the model needs to accommodate different patterns of traffic volumes for different days of the week. Secondly, intervention required in traffic networks is not always as simple as the intervention illustrated in this paper and often needs more sophisticated intervention techniques. And finally, for practical purposes, the model requires an automatic monitoring system which can detect, and react to, both short and long-term changes in traffic flow.

References

COWELL, R.G., DAWID, A.P., LAURITZEN, S.L. & SPIEGELHALTER, D.J. (1999). *Probabilistic Networks and Expert Systems.* New York: Springer Verlag.

DOUGHERTY, M.S. & COBBETT, M.R. (1997). Should we use neural networks or statistical models for short-term motorway traffic forecasting? *Internat. J. Forecasting* **13**, 21–31.

KIRBY, H.R., WATSON, S.M. & DOUGHERTY, M.S. (1997). Short-term inter-urban traffic forecasts using neural networks. *Internat. J. Forecasting* **13**, 43–50.

QUEEN, C.M. & SMITH, J.Q. (1993). Multiregression dynamic models. *J. Royal Statist. Soc. Ser. B* **55**, 849–870.

QUEEN, C.M., WRIGHT, B.J. & ALBERS, C.J. (2007). Forecast covariances in the linear multiregression dynamic model. *J. Forecasting* (forthcoming).

SMITH, B.L. & DEMETSKY, M.J. (1997). Traffic flow forecasting: comparison of modelling approaches. *J. Transportation Engineering* **123**, 261–266.

TEBALDI, C., WEST, M. & KARR, A.F. (2002). Statistical analyses of freeway traffic flows. *J. Forecasting* **21**, 39–68.

VAUGHAN, R.J. (2001). The distribution of traffic volumes. *J. Transportation Science*, **70**, 97–110.

WERMUTH, N. & LAURITZEN, S.L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *J. Royal Statist. Soc. Ser. B* **52**, 21–50.

WEST, M. & HARRISON, P.J. (1997). *Bayesian Forecasting and Dynamic Models.* 2nd edn. New York: Springer Verlag.

WHITLOCK, M.E. & QUEEN, C.M. (2000). Modelling a traffic network with missing data. *J. Forecasting* **19**, 561–574.

WHITTAKER, J., GARSIDE, S. & LINDVELT, K. (1997). Tracking and predicting a network traffic process. *Internat. J. Forecasting* **13**, 51–61.