Eliciting a DAG for a multivariate time series of vehicle counts in a traffic network.

Catriona M Queen (The Open University)[1]

Ben J Wright (The Open University)

Summary

In this paper we elicit a directed acyclic graph (DAG) 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 suitable DAG which represents the conditional independence structure across the time series. We discuss how the DAG can be used to define a linear multiregression dynamic model for the multivariate time series, so that each individual series is simply modelled by a univariate dynamic linear model.

Keywords: Model elicitation; traffic network; DAG; conditional independence; multivariate time series.

1 Introduction.

Modelling multivariate time series of vehicle counts in a traffic network 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 traffic networks. These include historical, data-based algorithms, classical time series methods and neural networks (see, for example, Stephanedes et al (1981), Ahmed and Cook, (1979), and Dougherty and Cobbett (1997)). However, non of these modelling approaches have managed to satisfy all the model requirements given above.

We take a different approach and use a particular type of multivariate Bayesian dynamic model (West and Harrison, 1997) called a Linear Multiregression Dynamic Model

[1]Address for correspondence: Department of Statistics, Faculty of Mathematics and Computing, The Open University, Milton Keynes, MK7 6AA, UK. Tel: (+44) 01908 659585 Fax: (+44) 01908 652140 Email: C.Queen@open.ac.uk

or LMDM (Queen and 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, Smith (1990)). 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 easily integrated into the model through intervention (see West and 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. Using the LMDM for forecasting in a traffic network with a given DAG is an interesting problem in itself and is considered elsewhere (see, for example, Wright (2005)). We shall elicit a DAG for a particular traffic network at the junction of three major roads — the M25, A2 and A282 — east of London, UK. Although we are considering just a single network here, the elicitation methods presented are quite general and the principles can be applied to any network.

The traffic data are in the form of hourly counts of vehicles passing each of a number of data collection sites. Each site is identified by a number and their geographical layout is illustrated in the diagram in Figure 1. 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. Changes in a traffic network can occur from time to time and Section 5 describes how these changes can be accommodated by the DAG and the LMDM. Finally, Section 6 contains some concluding remarks.

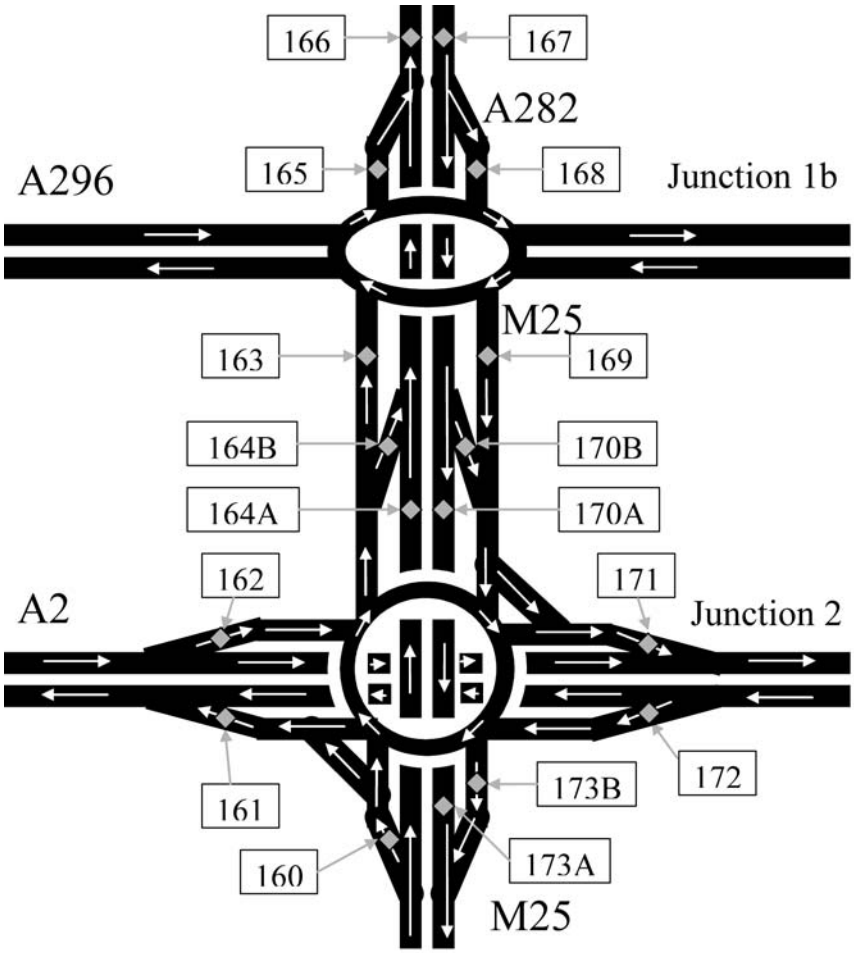

Figure 1: Locations of data collection sites used around the M25/A2/A282 junction. The grey diamonds are the data collection sites, each of which is numbered. White arrows indicate the direction of traffic flow.

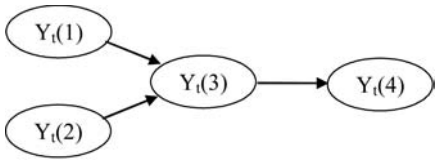

Figure 2: DAG representing four time series at time t, where $\text{pa}(Y_t(2)) = \varnothing$, $\text{pa}(Y_t(3)) = \{Y_t(1), Y_t(2)\}$ and $\text{pa}(Y_t(4)) = \{Y_t(3)\}$.

2 The Linear Multiregression Dynamic Model

In this section we shall give a brief (non-technical) overview of the LMDM. For a full account of the model, see Queen and Smith (1993).

Suppose that we have a multivariate time series $\boldsymbol{Y}_t = (Y_t(1), \ldots, Y_t(n))^T$. Suppose further that there is a conditional independence and causal structure defined across the series, so that, at each time $t = 1, 2, \ldots$, we have

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

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 $\text{pa}(Y_t(2)) = \varnothing$, $\text{pa}(Y_t(3)) = \{Y_t(1), Y_t(2)\}$ 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 joint one-step ahead forecast distribution of $\boldsymbol{Y}_t$ can be expressed as the product

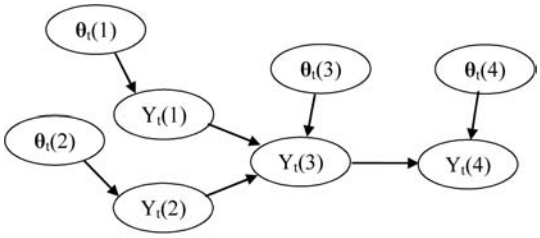

Figure 3: DAG from Figure 2 with independent model parameters included.

Figure 4: Two DAGs both representing $A_t \amalg C_t \,|\, B_t$, but yielding different LMDMs.

of $Y_t(1)$'s forecast distribution and the individual univariate conditional forecast distributions, $Y_t(i) \,|\, \text{pa}(Y_t(i))$, $i = 2, \ldots, n$. Even though regression is linear, 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.

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 $Y_t(i)$, $i = 1, \ldots, n$. We can consider $\boldsymbol{\theta}_t(i)$ as another parent of $Y_t(i)$ on the DAG. Further, the LMDM assumes independent 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.

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 $A_t \amalg C_t \,|\, B_t$. However, they would yield quite different LMDMs. For LMDMs the DAG needs to represent the conditional independence structure *related to causality*, so that (following Wermuth and Lauritzen (1990)) variables which are hypothesised 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 we shall first look at 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 vehicles may behave irrationally in this way, it is unlikely that such behaviour is common.

By considering the geographical layout of sites (Figure 1) all possible routes passing through the sites can be listed. These are given in Table 1 for each entry and exit point for the network. For example, we have already seen that vehicles entering the network travelling south on the A282 that leave the network via the southbound carriageway of the M25 pass the three sites 167, 170A and 173A. Note that there are routes through the network that do not pass through any sites. For example, vehicles travelling on the A2 who cross, but do not join, the M25 at Junction 2 are not counted. These vehicles are not an observable part of the traffic flow and do not enter the model.

From Table 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. We shall call this the flow diagram. Note that flows into and out of the network itself are also listed in Table 1 and shown in the flow diagram.

The flow diagram helps to give us 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, tells us that these time series will be highly correlated and that the sum of 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). It also tells us that the time series at 167 is hypothesised to

Entry point	Exit point	Route through sites
A282 southbound	M25 southbound	→ 167 → 170A → 173A →
	Junction 1b	→ 167 → 168 →
	A2 eastbound	→ 167 → 170B → 171 →
	A2 westbound	→ 167 → 170B → 161 →
Junction 1b	A282 northbound	→ 165 → 166 →
	M25 southbound	→ 169 → 173B →
	A2 eastbound	→ 169 → 171 →
	A2 westbound	→ 169 → 161 →
A2 westbound	A282 northbound	→ 172 → 164B → 166 →
	M25 southbound	→ 172 → 173B →
	Junction 1b	→ 172 → 163 →
A2 eastbound	A282 northbound	→ 162 → 164B → 166 →
	M25 southbound	→ 162 → 173B →
	Junction 1b	→ 162 → 163 →
M25 northbound	A282 northbound	→ 164A → 166 →
	Junction 1b	→ 160 → 163 →
	A2 eastbound	→ 160 → 171 →
	A2 westbound	→ 160 → 161 →

Table 1: Possible vehicle routes through the sites in the network for each entry and exit point.

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 and Queen (2000)). However, to allow future evaluation of our model we shall only consider modelling time series for which we have observed data. 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:

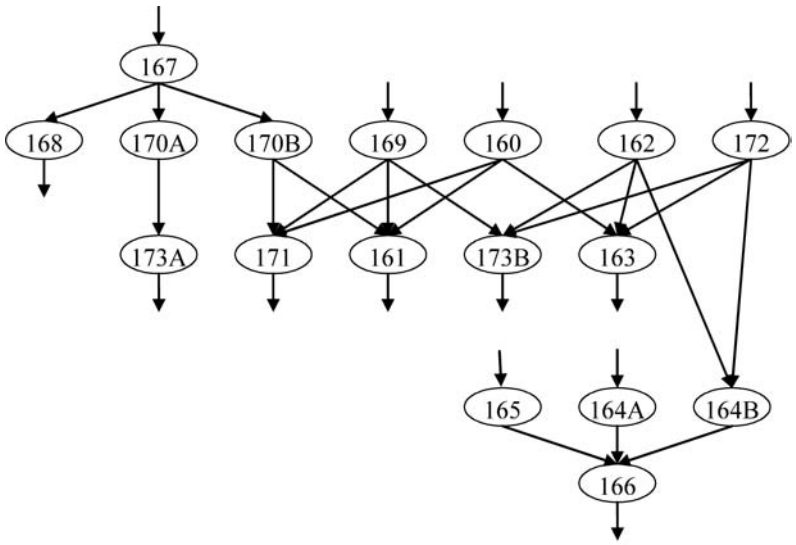

Figure 5: Flow diagram showing how vehicles pass through the data collection sites in the network. The data collection sites are represented by ovals and the flows between them by arrows.

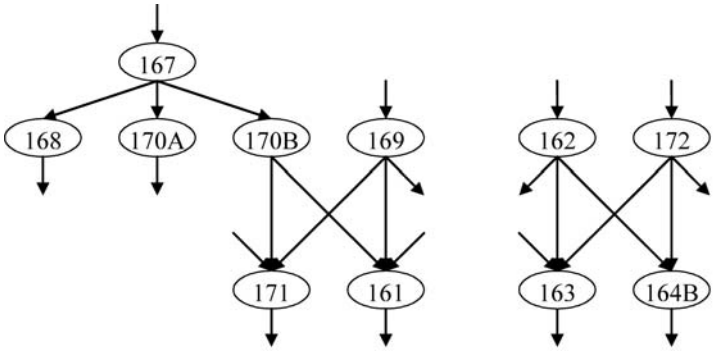

Figure 6: Flow diagram for the network with missing and trivial sites removed.

{167, 168, 170A, 170B}, {170B, 169, 171, 161}, {162, 172, 163, 164B}.

For each site s, denote the vehicle count at time t by $Y_t(s)$ and let $Y_t(s) \sim \text{Po}(\mu_t(s))$, where $\text{Po}(\mu_t(s))$ denotes a Poisson distribution with mean $\mu_t(s)$. In what follows we shall use the following results.

- 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

$$(X_1|X_1 + X_2) \sim \text{Bi}\left(X_1 + X_2, \frac{\mu_1}{\mu_1 + \mu_2}\right)$$

  where $\text{Bi}(n, p)$ denotes a binomial distribution from a sample of 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 $Y_t(170\text{B})$. We therefore have 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 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 for the (conditional) univariate model for $Y_t(168)$ in an LMDM would be of the form:

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

However, the parameters for the three conditional distributions for $Y_t(168)$, $Y_t(170\text{A})$ and $Y_t(170\text{B})$ are not independent. Consequently the parameters for each univariate model in

an LMDM could not be considered mutually independent. To ensure independent model parameters, we need to elicit the DAG 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 we have two alternative conditional distributions:

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

where

$$\alpha_t = \frac{\mu_t(170\text{A}) + \mu_t(170\text{B})}{\mu_t(167)} \qquad \text{and} \qquad \beta_t = \frac{\mu_t(170\text{B})}{\mu_t(170\text{A}) + \mu_t(170\text{B})}.$$

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 deterministic variables. Note that we could have chosen $Y_t(170\text{A}) + Y_t(170\text{B})$ and/or $Y_t(170\text{B})$ to be the deterministic 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.

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:

$$\begin{aligned} Y_t(167) &= \mu_t(167) + v_t(167), \qquad v_t(167) \sim N(0, V_t(167)) \\ Y_t(170\text{A}) + Y_t(170\text{B}) &= Y_t(167)\alpha_t + v_t(170\text{A} + 170\text{B}), \\ &\qquad v_t(170\text{A} + 170\text{B}) \sim N(0, V_t(170\text{A} + 170\text{B})) \\ Y_t(170\text{B}) &= \left(Y_t(170\text{A}) + Y_t(170\text{B})\right)\beta_t + v_t(170\text{B}), \\ &\qquad v_t(170\text{B}) \sim N(0, V_t(170\text{B})) \end{aligned}$$

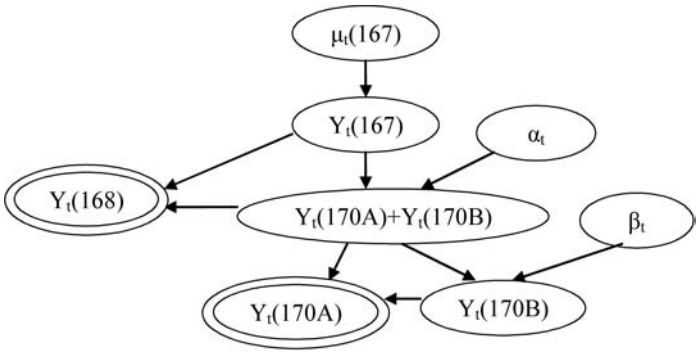

Figure 7: DAG representing $Y_t(167)$, $Y_t(168)$, $Y_t(170\text{A})$ and $Y_t(170\text{B})$, together with the model parameters: ◯ is a random variable, ◎ is a deterministic variable.

and

$$\begin{aligned} Y_t(168) &= Y_t(167) - (Y_t(170\text{A}) + Y_t(170\text{B})) \\ Y_t(170\text{A}) &= (Y_t(170\text{A}) + Y_t(170\text{B})) - Y_t(170\text{B}). \end{aligned}$$

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 $Y_t(170\text{B})$ and $Y_t(169)$. However, the parameters for the conditional distributions $Y_t(171)|Y_t(170\text{B}), Y_t(169)$ and $Y_t(161)|Y_t(170\text{B}), Y_t(169)$ are not independent. Thus we would not be able to model $Y_t(171)$ and $Y_t(161)$ as separate children of $Y_t(170\text{B})$ and $Y_t(169)$ using an LMDM. Again, an alternative DAG is required.

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,

$$Y_t(171) + Y_t(161) = X_t(u) + Y_t(170\text{B}) + X_t(169) \tag{4.1}$$

where

$$\begin{aligned} X_t(u) &= \text{vehicles at 171 or 161 at time } t \text{ inherited from the unobserved site} \\ X_t(169) &= \text{vehicles at 171 or 161 at time } t \text{ inherited from site 169} \end{aligned}$$

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

$$X_t(169)|Y_t(169) \sim \text{Bi}(Y_t(169), \gamma_t)$$

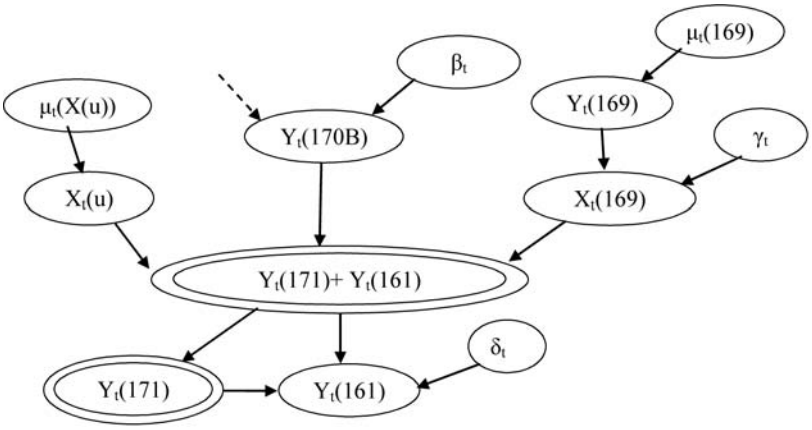

Figure 8: DAG representing $Y_t(170\text{B})$, $Y_t(169)$, $X_t(u)$, $X_t(169)$, $Y_t(171)$ and $Y_t(161)$, together with the model parameters: ◯ is a random variable, ◎ is a deterministic variable.

where

$$\gamma_t = \frac{\mu_t(X(169))}{\mu_t(169)}.$$

Also,

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

where

$$\delta_t = \frac{\mu_t(161)}{\mu_t(171) + \mu_t(161)}.$$

Both γ_t and δ_t are interpretable:

$$\begin{aligned} \gamma_t &= \text{proportion of vehicles at 169 flowing to 171 or 161} \\ &= \text{proportion of those vehicles travelling south from Junction 1b that join the A2} \\ \delta_t &= \text{proportion of those vehicles joining the A2 that travel west bound} \end{aligned}$$

This can all be represented by the DAG in Figure 8.

However, $X_t(u)$ and $X_t(169)$ cannot be observed. Approximate the Poisson distribution for $X_t(u)$ and the conditional binomial distribution for $X_t(169)|Y_t(169)$ to normality. Then using Equation 4.1, the conditional distribution for $Y_t(171)+Y_t(161)|Y_t(170\text{B}), Y_t(169)$ is approximately normal with mean:

$$E(Y_t(171) + Y_t(161)|Y_t(170\text{B}), Y_t(169)) = \mu_t(X(u)) + Y_t(170\text{B}) + Y_t(169)\gamma_t.$$

The DAG in Figure 8 can therefore be replaced with the DAG in Figure 9. An LMDM would then have the two regressors $Y_t(170\text{B})$ and $Y_t(169)$ in the DLM for $Y_t(171)+Y_t(161)$,

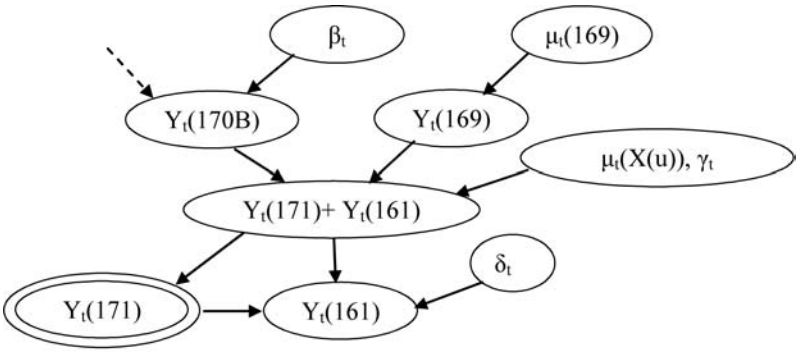

Figure 9: DAG representing $Y_t(170B)$, $Y_t(169)$, $Y_t(171)$ and $Y_t(161)$: ◯ is a random variable, ◎ is a deterministic variable.

with the regression parameter for $Y_t(170B)$ set to 1. However, doing this leads to problems with collinearity.

To tackle this problem, define

$$\begin{aligned} Z_t(1) &= \frac{Y_t(170B) + Y_t(169)}{2} \\ Z_t(2) &= \frac{Y_t(170B) - Y_t(169)}{2}. \end{aligned}$$

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 $Y_t(170B)$ and $Y_t(169)$. We can introduce $Z_t(1)$ and $Z_t(2)$ into the DAG as deterministic 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:

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

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 representing the first two groups of variables (i.e. the first part of the flow diagram) is given in Figure 10.

Approximate 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 10 then has the following observation equations for $Y_t(169)$, $Y_t(171) + Y_t(161)$ and

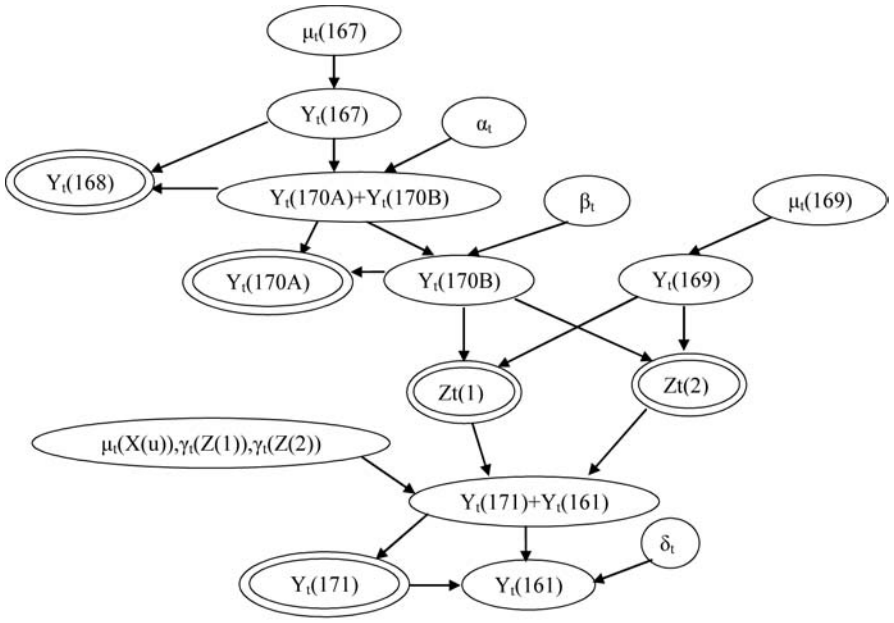

Figure 10: Final elicited DAG for the first part of the flow diagram: ◯ is a random variable, ◎ is a deterministic variable.

$Y_t(161)$:

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

and

$$Y_t(171) = (Y_t(171) + Y_t(161)) - Y_t(161).$$

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

The flow diagram for this group of four variables is almost identical in structure to that for the four variables in Section 4.2. As in that section, we cannot simply draw a DAG for an LMDM with $Y_t(163)$ and $Y_t(164\text{B})$ 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 $Y_t(164\text{B})|Y_t(162), Y_t(172)$ would not be independent. We therefore need to elicit an alternative DAG.

Consider the sum $Y_t(163) + Y_t(164\text{B})$. The vehicles making up this sum come from $Y_t(162)$, $Y_t(172)$ and the unobserved site 160. This time, for both sites 162 and 172,

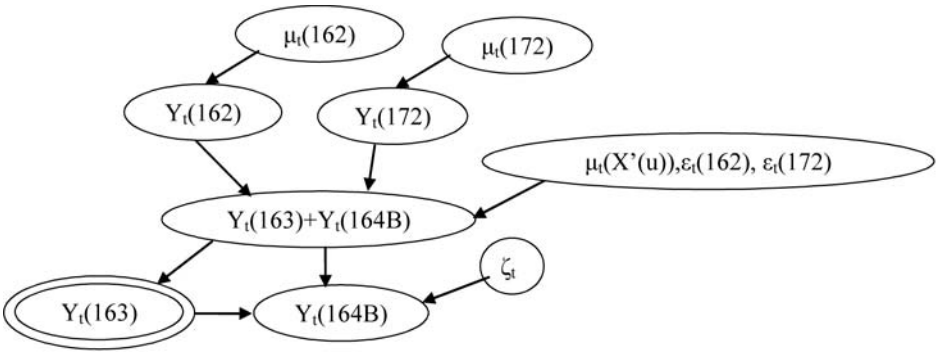

Figure 11: DAG representing $Y_t(162)$, $Y_t(172)$, $Y_t(163)$ and $Y_t(164\text{B})$: ◯ is a random variable, ◎ is a deterministic variable.

only a proportion of the vehicles flow to 163 or 164. Following exactly the same arguments as in Section 4.2, the conditional distribution $Y_t(163) + Y_t(164\text{B})|Y_t(162), Y_t(172)$ is approximately normal with mean

$$\mathrm{E}(Y_t(163) + Y_t(164\text{B})|Y_t(162), Y_t(172)) = \mu_t(X'(u)) + Y_t(162)\varepsilon_t(162) + Y_t(172)\varepsilon_t(172)$$

and

$$Y_t(164\text{B})|Y_t(163) + Y_t(164\text{B}) \sim \mathrm{Bi}(Y_t(163) + Y_t(164\text{B}), \zeta_t),$$

where the parameters are interpretable as follows:

$$\begin{aligned}
\mu_t(X'(u)) &= \text{mean number of vehicles in } Y_t(163) + Y_t(164\text{B}) \text{ inherited from the unobserved site} \\
\varepsilon_t(162) &= \text{proportion of vehicles at 162 flowing to 163 or 164B} \\
&= \text{proportion of those vehicles leaving the A2 eastbound that travel north} \\
\varepsilon_t(172) &= \text{proportion of vehicles at 172 flowing to 163 or 164B} \\
&= \text{proportion of those vehicles leaving the A2 westbound that travel north} \\
\zeta_t &= \text{proportion of those vehicles travelling north after leaving the A2} \\
&\quad \text{that join the M25}
\end{aligned}$$

The DAG to represent this is given in Figure 11. Unfortunately, there are again problems with collinearity, this time between $Y_t(162)$ and $Y_t(172)$ in the DLM for $Y_t(163)+Y_t(164\text{B})$.

Following the same idea as in Section 4.2, let

$$\begin{aligned}
Z_t(3) &= \frac{Y_t(162) + Y_t(172)}{2} \\
Z_t(4) &= \frac{Y_t(162) - Y_t(172)}{2}
\end{aligned}$$

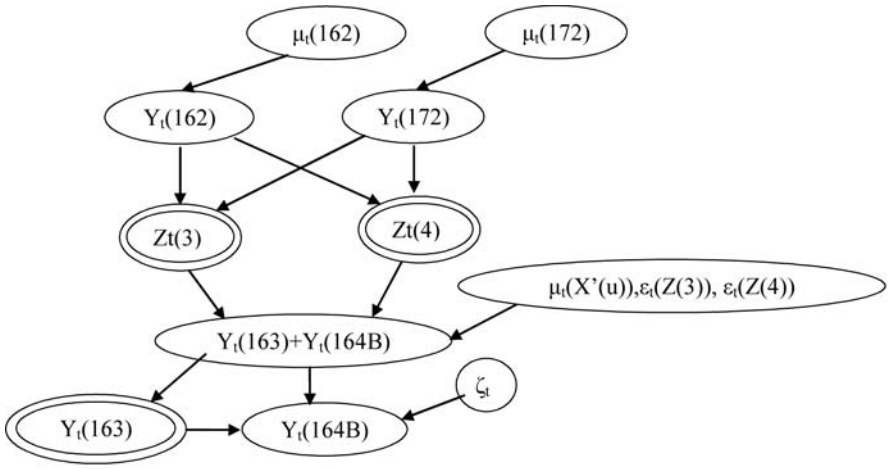

Figure 12: Final elicited DAG for the second part of the flow diagram: ◯ is a random variable, ◎ is a deterministic variable.

Introduce $Z_t(3)$ and $Z_t(4)$ into the DAG as deterministic children of $Y_t(162)$ and $Y_t(172)$ and parents of $Y_t(163) + Y_t(164\text{B})$. Then the conditional distribution for $Y_t(163) + Y_t(164\text{B})|Z_t(3), Z_t(4)$ is approximately normal with mean

$$\mathrm{E}(Y_t(163) + Y_t(164\text{B})|Z_t(3), Z_t(4)) = \mu_t(X'(u)) + Z_t(3)\varepsilon_t(Z(3)) + Z_t(4)\varepsilon_t(Z(4))$$

for parameters $\varepsilon_t(Z(3))$ and $\varepsilon_t(Z(4))$. Again we lose interpretability of the parameters, but retain the conditional independence structure and a computationally simple model. The resulting DAG representing the last part of the flow diagram is given in Figure 12.

Approximate the Poisson distributions for $Y_t(162)$ and $Y_t(172)$ and the conditional binomial distribution for $Y_t(164\text{B})|Y_t(163) + Y_t(164\text{B})$ to normality. Then the observation equations for an LMDM representing the DAG in Figure 12 are of the following form.

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

and

$$Y_t(163) = (Y_t(163) + Y_t(164\text{B})) - Y_t(164\text{B}).$$

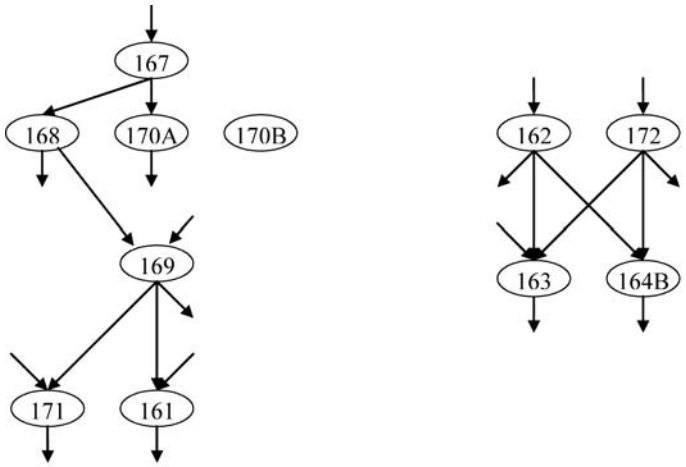

Figure 13: Flow diagram for the network when the road is blocked at site 170B.

5 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 of 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 informative priors for any new parameters. The following example will illustrate how this might be done.

5.1 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 cars wishing to leave the M25 southbound at Junction 2 are diverted via Junction 1B and sites 168 and 169.

Figure 13 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 14 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 12.

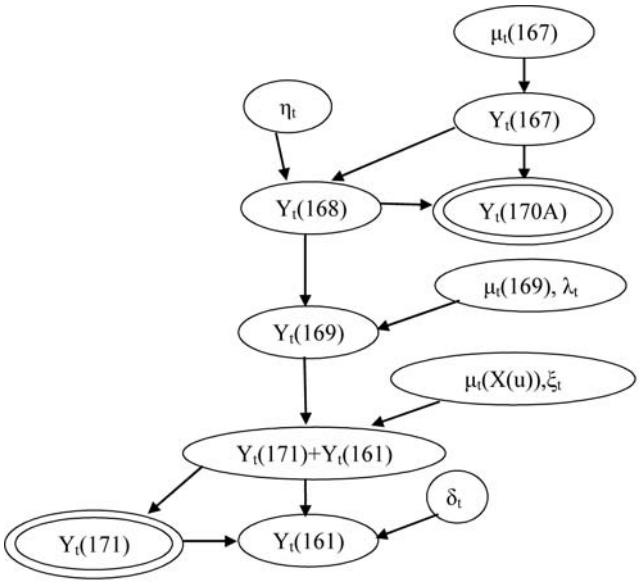

Figure 14: New DAG for the first part of the flow diagram when the road is blocked at site 170B: ◯ is a random variable, ◎ is a deterministic variable.

The new observation equations are as follows.

$$
\begin{aligned}
Y_t(167) &= \mu_t(167) + v_t(167), \qquad v_t(167) \sim N(0, V_t(167)) \\
Y_t(168) &= \eta_t Y_t(167) + v_t(168), \qquad v_t(168) \sim N(0, V_t(168)) \\
Y_t(169) &= \mu_t(169) + \lambda_t Y_t(168) + v_t(169), \qquad v_t(169) \sim N(0, V_t(169)) \\
Y_t(171) + Y_t(161) &= \mu_t(X(u)) + \xi_t Y_t(169) + v_t(171 + 161), \\
&\qquad\qquad v_t(171 + 161) \sim N(0, V_t(171 + 161)) \\
Y_t(161) &= \delta_t \left(Y_t(171) + Y_t(161)\right) + v_t(161), \\
&\qquad\qquad v_t(161) \sim N(0, V_t(161))
\end{aligned}
$$

and

$$
\begin{aligned}
Y_t(170\text{A}) &= Y_t(167) - Y_t(168) \\
Y_t(171) &= \left(Y_t(171) + Y_t(161)\right) - Y_t(161)
\end{aligned}
$$

There are three new parameters here and each is easily interpretable:

η_t = proportion of those vehicles at 167 leaving the M25 southbound at Junctions 1b or 2

λ_t = proportion of those vehicles leaving southbound at Junction 1b that are following the diversion to Junction 2

ξ_t = 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 informative priors for the new parameters η_t, λ_t and ξ_t as follows. Calculate the priors at time t for the original parameters α_t, β_t and γ_t and the one step ahead forecasts for $Y_t(170\text{B})$, $Y_t(168)$ and $Y_t(169)$ assuming no change in the DAG and model (i.e. assuming the DAG in Figure 10 still holds). Denote these by $\hat{\alpha}_t$, $\hat{\beta}_t$, $\hat{\gamma}_t$, $\hat{Y}_t(170\text{B})$, $\hat{Y}_t(168)$ and $\hat{Y}_t(169)$. Denoting the first $t-1$ observations by $\boldsymbol{y}^{t-1}$, prior mean estimates of the new parameters can be obtained using:

$$\begin{aligned}
\mathrm{E}(\eta_t|\boldsymbol{y}^{t-1}) &= 1-\hat{\alpha}_t+\hat{\alpha}_t\hat{\beta}_t \\
\mathrm{E}(\lambda_t|\boldsymbol{y}^{t-1}) &= \frac{\hat{Y}_t(170\text{B})}{\hat{Y}_t(170\text{B})+\hat{Y}_t(168)} \\
\mathrm{E}(\xi_t|\boldsymbol{y}^{t-1}) &= \frac{\hat{Y}_t(170\text{B})+\hat{Y}_t(169)\hat{\gamma}_t}{\hat{Y}_t(170\text{B})+\hat{Y}_t(169)}
\end{aligned}$$

Forecasts at time t for the changed network can then be calculated from the priors for the parameters in the new model.

6 Concluding remarks

In this paper we have demonstrated how we might elicit a DAG suitable for representing the time series of vehicle counts in a traffic network. The final DAG elicited for this particular traffic network comprises the two separate DAGs given in Figures 10 and 12.

It is important to elicit a DAG so that the parameters for each variable can be considered mutually independent a priori. This enables each variable to be modelled as a separate univariate DLM, making the multivariate problem far simpler.

With the exception of $\gamma_t(Z(1))$, $\gamma_t(Z(1))$, $\varepsilon_t(Z(3))$ and $\varepsilon_t(Z(4))$, all the model parameters in the elicited DAG are interpretable. This is helpful not only when setting

up the model and interpreting any output, but it is also extremely helpful when model intervention is required.

Unfortunately, in applications of this type, multiple parents may be highly correlated leading to problems with collinearity in the regression DLM for the child. Collinearity in conventional regression is often dealt with by dropping one of the correlated regressors. However, this is not an option here. Instead by transforming the parents to new uncorrelated regressors, the information from all the parents can be preserved.

Traffic networks do not always remain static through time and any traffic model needs to be adaptable. We have demonstrated here how this might be done with our model and how posterior information from the original DAG can be used to construct informative priors for the new DAG.

In this paper we have only focused on the crucial first stage in implementing the LMDM — namely, the elicitation of a simple model accommodating the multivariate structure of the series. The next stage in implementing the model is of course to use the elicited model for forecasting. In Wright (2005) the model is used with 21 weeks of data for this network and was found to perform well when compared with models of similar complexity.

References

Ahmed, M.S. and A.R.Cook (1979) Analysis of freeway traffic time-series data by using Box-Jenkins Techniques. *Transportation Research Record*, **722**, 1–9.

Dougherty, M. S. and Cobbett, M. R. (1997). Short-term inter-urban traffic forecasts using neural networks. *International Journal of Forecasting* **13**, 21–31.

Queen, C.M. and Smith, J.Q. (1993) Multiregression dynamic models. *Journal of the Royal Statistical Society, Series B*, **55**, 849–870.

Smith, J.Q. (1990) Statistical principles on graphs. In *Influence Diagrams, Belief Nets and Decision Analysis*, R.M. Oliver and J.Q.Smith eds. John Wiley and Sons Ltd.

Stephanedes, Y.J., Michalopoulos, P.G. and Plum, R.A (1981) Improved estimation of traffic flow for real-time control. *Transportation Research Record*, **795**, 28–39.

Wermuth, N. and Lauritzen, S.L. (1990) On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society, Series B*, **52**, 21–50.

West, M. and Harrison, P.J. (1997) *Bayesian Forecasting and Dynamic Models*. Springer-Verlag. New York. 2nd Edition.

Whitlock, M.E. and Queen, C.M. (2000) Modelling a traffic network with missing data. *Journal of Forecasting*, **19**, 561–574.

Wright, B.J. (2005) *A Bayesian dynamic approach to modelling flow through a traffic network*. PhD Thesis. The Open University.