Multivariate forecasting of road traffic flows in the presence of heteroscedasticity and measurement errors

Osvaldo Anacleto[1]

The Open University, Milton Keynes, UK

Catriona Queen

The Open University, Milton Keynes, UK

Casper J Albers

University of Groningen, Groningen, The Netherlands

Summary

Linear multiregression dynamic models (LMDMs), which combine a graphical representation of a multivariate time series with a state space model, have been shown to be a promising class of models for forecasting of traffic flow data. Analysis of flows at a busy motorway intersection near Manchester, UK, highlights two important modelling issues: accommodating different levels of traffic variability depending on the time of day and accommodating measurement errors occurring due to data collection errors. This paper extends LMDMs to address these issues. Additionally, the paper investigates how close the approximate forecast limits usually used with the LMDM are to the true, but not so readily available, forecast limits.

Key words: data collection error; dynamic linear model; linear multiregression dynamic model; traffic modelling; variance law.

1 Introduction

Traffic flow data are now routinely collected for many roads. These data can be used as part of a traffic management system to assess highways facilities and performance over time, or for real-time traffic control to prevent and manage congestion. The data can also be used as part of a traveller information system. Good short-term traffic flow forecasting models are vital for the success of both traffic management and traveller information systems. This paper focuses on developing flow forecasting models particularly appropriate for assessing highways facilities and performance over time or for providing advanced traffic information for travellers.

[1] *Address for correspondence*: Department of Mathematics and Statistics, The Open University, Milton Keynes, MK7 6AA, UK. Email: o.anacleto-junior@open.ac.uk

Traffic flow data are time series of counts of vehicles passing data collection sites, $S(1), \ldots, S(n)$, across a network. Traffic flows at sites upstream and downstream to $S(i)$ are informative about the flows at $S(i)$. To make use of this, lagged flows at other sites have been used by some to help forecast flows at $S(i)$ (Tebaldi *et al.*, 2002; Kamarianakis and Prastacos, 2005; Stathopoulos and Karlaftis, 2003), while others use conditional independence so that lagged flows only at adjacent sites to $S(i)$ are required (Whittaker *et al.*, 1997; Sun *et al.*, 2006). However, when the time interval over which vehicles are counted is long enough for vehicles to register at more than one site in the network *in the same time period*, as is the case in this paper, then the flows at other sites at lag 0 are helpful for forecasting flows at $S(i)$. The proposed model, a dynamic graphical model called the linear multiregression dynamic model (LMDM) (Queen and Smith, 1993), takes advantage of this and uses information regarding upstream flows at time t for forecasting flow at $S(i)$ *at the same time* t (see Section 3 regarding how this is done).

While Carvalho and West (2007) use an undirected graph to represent conditional independence relationships in the covariance structure of a multivariate time series, the LMDM represents any conditional independence relationships *related to causality* across the time series by a directed acyclic graph (DAG). This DAG is used to break the multivariate model into simpler univariate components, each of which is (conditionally) a Bayesian regression dynamic linear model (DLM) (West and Harrison, 1997). In the context of traffic forecasting, as in Sun *et al.* (2006), the direction of traffic flow produces the causal drive in the system and the possible routes through the network are used to define a conditional independence structure across the time series.

Each univariate regression DLM in the LMDM uses contemporaneous upstream traffic flows as regressors. Tebaldi *et al.* (2002) also use regression DLMs when modelling traffic flows, with upstream traffic flows as linear regressors. However, their regressors are lagged flows, rather than contemporaneous flows, because they have 1-minute flows so that, unlike in this paper, vehicles are not counted at multiple sites during a single time period.

This paper specifies a DAG and associated LMDM for a busy motorway intersection near Manchester, UK. Although the advantages of the LMDM in the context of traffic

forecasting have been extensively explored (Whitlock and Queen, 2000; Queen *et al.*, 2007; Queen and Albers, 2009), modelling issues still remain, including accommodating different levels of traffic variability depending on the time of the day (Kamarianakis *et al.*, 2005) and accommodating measurement errors which can occur due to data collection errors (Bickel *et al.*, 2007). The methodology presented in this paper is developed to tackle these important practical issues. Additionally, the paper uses simulation to compare the approximate (easily calculated) forecast limits usually used in the LMDM with estimates of the true (not easily calculated) forecast limits.

Although this paper focuses on using the LMDM in the context of traffic flow forecasting, the model is potentially suitable for any application involving flows, such as electricity flows, signal flows in telecommunication networks, flows of packages over the internet, flows of goods in supply chains, and so on. It can also be applied to different types of multivariate time series problems such as sales forecasting (Queen, 1997). Farrow (2003) also focuses on sales forecasting using a model similar to the LMDM, while Fosen *et al.* (2006) and Guo and Brown (2001) use similar ideas to the LMDM to analyse hormone time series and cancer patients with liver cirrhosis, respectively.

The paper is structured as follows. Section 2 describes the data used throughout the paper. Section 3 gives a brief review of the LMDM while in Section 4 an LMDM is specified for the particular network of interest. Section 5 extends the LMDM so that it can accommodate the heteroscedasticity present in the usual pattern of traffic flows, while Section 6 adapts the proposed LMDM to accommodate measurement errors which frequently occur due to data collection errors. Section 7 investigates how close the approximate forecast limits usually used with the LMDM are to the true forecast limits. Finally, Section 8 offers some concluding remarks and discusses issues for future research.

2 The data

This paper focuses on developing a model for forecasting traffic flows at the intersection of three motorways — the M60, M62 and M602 — west of Manchester, UK. Figure 1(a) shows an aerial photograph of the network.

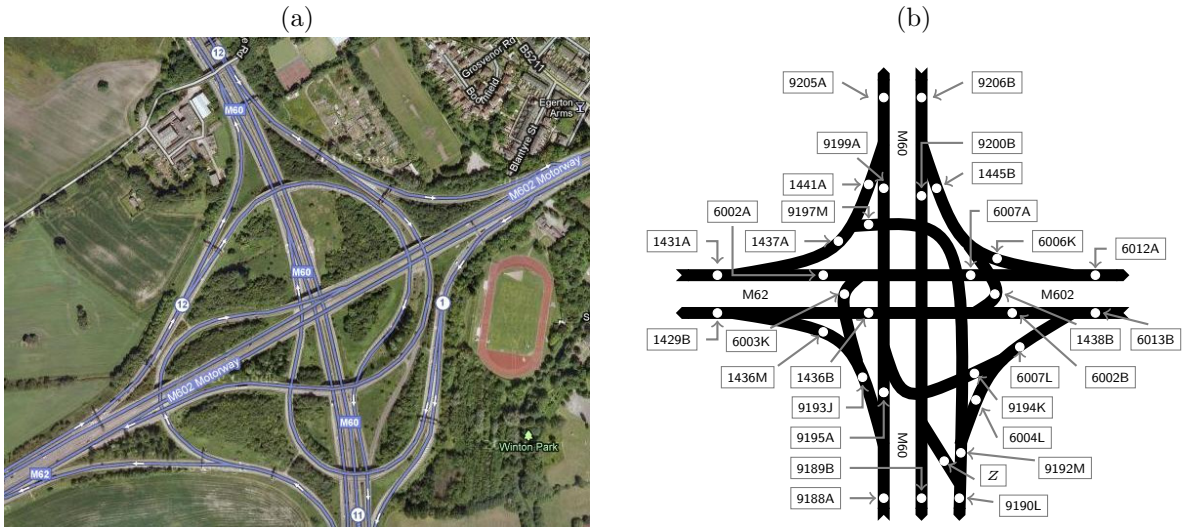

Figure 1: The Manchester network: (a) aerial photograph (©2012 DigitalGlobe, GeoEye, Infoterra Ltd & Bluesky, The GeoInformation Group, Map data ©2012 Google) and (b) schematic diagram.

The data are counts of vehicles passing over induction loops in the road surface at a number of data collection sites in the network. A schematic diagram of the Manchester network reflecting the layout of the data sites is given in Figure 1(b). Here, the arrows show the direction of travel and the data sites are labelled and indicated by circles. The data used in the paper were collected between March and November 2010 by the Highways Agency in England (http://www.highways.gov.uk/).

The data are in the form of minute counts. For traffic management systems for assessing

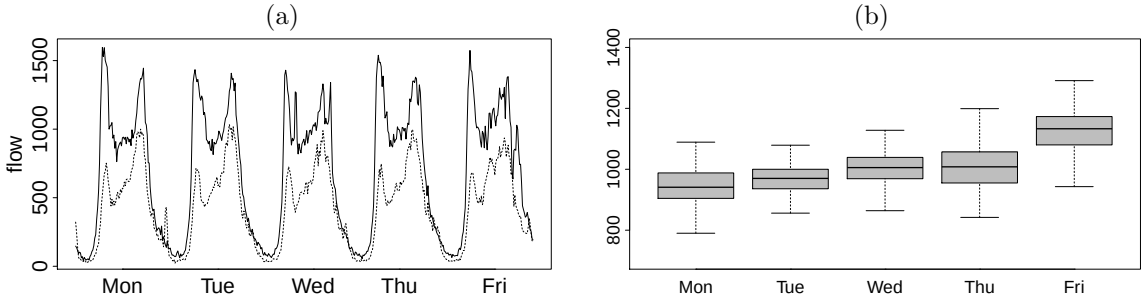

Figure 2: (a) 15-minute flows at site 1437A (solid line) and site 6013B (dashed line) for 07–11 June 2010. (b) Box-plots by weekdays using flows for the period 14:00-14:59 at site 1431A observed from March to November 2010.

highways facilities, the *Highways Capacity Manual* (2010) recommends aggregating data into 15-minute intervals. 15-minute intervals are also suitable for traveller information systems, traveller decisions being influenced by the expected conditions further along their route systems (Vlahogianni *et al.*, 2004). Thus in this paper the data have been aggregated into 15-minute intervals. Adapting the models for shorter time periods will be the focus of future research.

Figure 2(a) shows time series plots of 15-minute flows for a typical week for sites 1437A (solid line) and 6013B (dashed line). The daily patterns for both sites are similar with peaks in the morning and afternoon rush hours. Flows at all sites exhibit similar daily patterns.

Figure 2(b) shows box-plots of flows for each weekday from March to November 2010 at site 1431A for the period 14:00–14:59. These clearly show daily differences in level and variability of flows. These daily differences can be incorporated into the model, but for clarity of presentation, this paper will use flows for Wednesdays only (which doesn't cause problems with discontinuities because flows around midnight are very low and vary little).

It only takes a few minutes for a vehicle to traverse the network. So, for 15-minute data, vehicles are usually counted at several data sites within the same time period. The LMDM accommodates this, using information regarding the flows at sites upstream to a particular site $S(i)$ to help forecast the flow at $S(i)$ in the same time period.

3 Linear multiregression dynamic models

This section gives a brief overview of LMDMs (see Queen and Smith, 1993, for full details).

Consider a multivariate time series $\boldsymbol{Y}_t = (Y_t(1) \ \cdots \ Y_t(n))^\top$ with a conditional independence structure related to causality defined across it, so that for each $i = 2, \ldots, n$ and at each time t, conditional on variables $pa(Y_t(i)) \subseteq \{Y_t(1), \ldots, Y_t(i-1)\}$, $Y_t(i)$ is independent of $\{Y_t(1), \ldots, Y_t(i-1)\} \setminus pa(Y_t(i))$ (where "\" reads "excluding"). Each variable in the set $pa(Y_t(i))$ is a *parent* of $Y_t(i)$ and $Y_t(i)$ is a *child* of each variable in $pa(Y_t(i))$. Variable $Y_t(i)$ is a *root node* if $pa(Y_t(i)) = \varnothing$. The time series $\boldsymbol{Y}_t$ can then be represented

by a DAG at each time t with a directed arc to $Y_t(i)$ from each of its parents in $pa(Y_t(i))$.

The LMDM uses the DAG to model the multivariate time series by n separate regression DLMs: one each for $Y_t(1)$ and $Y_t(i)|pa(Y_t(i))$, $i = 2, ..., n$. Each time series has its parents as linear regressors, while root nodes are modelled by any suitable DLMs. As such, the LMDM is computationally simple and DLM techniques can be readily applied (see, for example, Queen and Albers, 2009).

Formally, denoting all available information at time $t-1$ by D_{t-1}, the LMDM is defined as follows.

$$\begin{aligned} Y_t(i) &= \boldsymbol{F}_t(i)^\top \boldsymbol{\theta}_t(i) + v_t(i), \;\; v_t(i) \sim N(0, V_t(i)), \;\; i = 1, \ldots, n, && (1) \\ \boldsymbol{\theta}_t &= \boldsymbol{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{w}_t, \qquad\qquad \boldsymbol{w}_t \sim N(0, \boldsymbol{W}_t), && (2) \\ \boldsymbol{\theta}_{t-1}|D_{t-1} &\sim N(\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}). && (3) \end{aligned}$$

The m_i-dimensional vector $\boldsymbol{F}_t(i)$ contains an arbitrary, but known, function of the parents $pa(Y_t(i))$ and possibly other known variables; $\boldsymbol{\theta}_t(i)$ is the m_i-dimensional parameter vector for $Y_t(i)$ and $\boldsymbol{\theta}_t^\top = (\boldsymbol{\theta}_t(1)^\top \; \cdots \; \boldsymbol{\theta}_t(n)^\top)$; $V_t(1), \ldots, V_t(n)$ are the scalar observation variances; $\boldsymbol{m}_{t-1}$ and $\boldsymbol{C}_{t-1}$ are the (posterior) moments for $\boldsymbol{\theta}_{t-1}$; matrices $\boldsymbol{G}_t$, $\boldsymbol{W}_t$, and $\boldsymbol{C}_{t-1}$ are block diagonal; $\boldsymbol{w}_t^\top = (\boldsymbol{w}_t(1)^\top \; \cdots \; \boldsymbol{w}_t(n)^\top)$, and $v_t(1), \ldots, v_t(n)$ and $\boldsymbol{w}_t(1), \ldots, \boldsymbol{w}_t(n)$, are independent sequences of independent errors.

Given the distribution (3), the prior distribution for $\boldsymbol{\theta}_t|D_{t-1}$ is obtained from (2). Forecast distributions for each $Y_t(i)$ conditional on $pa(Y_t(i))$ are then found separately via (1). However, as $Y_t(i)$ and $pa(Y_t(i))$ are both observed at the same time t, the *marginal* forecasts for each $Y_t(i)$ are required. Although the marginal forecast distributions cannot generally be calculated analytically, the marginal forecast moments are readily available using $E(X) = E\{E(X|Y)\}$ and $V(X) = E\{V(X|Y)\} + V\{E(X|Y)\}$. Essentially, in the LMDM, the marginal forecast moments of the parents of $Y_t(i)$ are used to obtain the marginal forecast moments for $Y_t(i)$, which in turn are used to find the marginal forecast moments of $Y_t(i)$'s children, and so on (see Queen and Smith, 1993 and Queen *et al.*, 2008). Finally, because of the structure of the LMDM, after observing $\boldsymbol{y}_t$, the distribution for each $\boldsymbol{\theta}_t(i)$ can be updated separately (in closed form) within the (conditional) DLM

for $Y_t(i)|pa(Y_t(i))$.

Some of the methodology developed in this paper directly affects the forecast variance. In order to evaluate the forecast performance of these methods, the joint log-predictive likelihood (LPL) is used rather than a measure based solely on forecast error. After observing $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, the LPL for the LMDM is calculated as:

$$\text{LPL} = \sum_{t=1}^{T} \left\{ \sum_{i=1}^{n} \log(f(y_t(i)|pa(y_t(i)), D_{t-1}) \right\}.$$

Because the forecast variance directly affects the forecast limits, an alternative, decision theoretically principled way of comparing forecast performance, is through the mean interval score (MIS), which is a function of the limits of the forecast interval for each observation, with a penalty when the observation lies outside the interval (for details, see Gneiting and Raftery, 2007). The MIS is then calculated over all observations in a time series. This idea can be extended to the multivariate LMDM setting by simply calculating the MIS over all observations for each time series.

4 Building an LMDM for the Manchester network

4.1 Forks and Joins

Traffic networks are basically a series of junctions of two types: forks and joins. A fork, in which vehicles from a single site $S(1)$ move to two sites $S(2)$ and $S(3)$, is illustrated in Figure 3(a). A join, in which traffic from two sites, $S(4)$ and $S(5)$, merge to a single site $S(6)$, is illustrated in Figure 3(b).

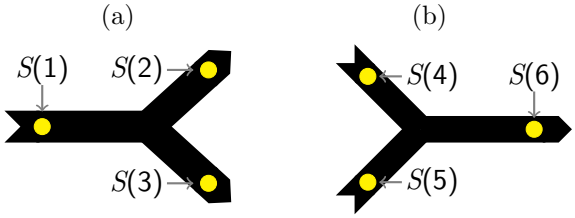

Figure 3: (a) a fork and (b) a join. In each diagram, the arrows denote the direction of travel and the circles are the sites.

Let $Y_t(i)$ be the number of vehicles passing site $S(i)$ during 15-minute period t. Following Queen *et al.* (2007), Equation (1) in LMDMs for $Y_t(1), Y_t(2), Y_t(3)$ and $Y_t(4), Y_t(5), Y_t(6)$ can be elicited to have the forms:

$$\begin{aligned} &Y_t(1) = \mu_t(1) + v_t(1), \quad Y_t(2) = \alpha_t y_t(1) + v_t(2), \quad Y_t(3) = y_t(1) - y_t(2), \\ &Y_t(4) = \mu_t(4) + v_t(4), \quad Y_t(5) = \mu_t(5) + v_t(5), \qquad Y_t(6) = y_t(4) + y_t(5). \end{aligned} \tag{4}$$

Thus, upstream flows are used in the models for downstream flows.

In (4), the $\mu_t(\cdot)$ parameters are level parameters, while parameter α_t represents the proportion of traffic flowing from $S(1)$ to $S(2)$, and $v_t(\cdot)$ are normal error terms. In Queen *et al.* (2007) the normality of the errors $v_t(\cdot)$ is justified by appealing to the Poisson approximation to normality for large means. While the data in this paper cannot be considered either Poisson or normal, as will be seen in Section 5, the variance does increase as a function of the mean. West and Harrison (1997) propose using a variance law within a normal DLM to model such non-normal data. Thus, in order to take advantage of the computational simplicity of the LMDM and the ease with which established DLM techniques can be incorporated into the model, normal errors will be used for $v_t(\cdot)$ and, in Section 5, the LMDM will be extended to incorporate a variance law to accommodate the non-normality of the data. Note that the data could have been modelled using non-normal errors via a generalisation of the LMDM known as the multiregression dynamic model (Queen and Smith, 1993), but that would be more computationally complex. The data could also have been transformed to normality, although that would lose model interpretability.

Following the terminology of the WinBUGS software (http://www.mrc-bsu.cam.ac.uk/bugs/), $Y_t(3)$ and $Y_t(6)$ are modelled as logical variables. This is because all traffic from $S(1)$ must flow to $S(2)$ and $S(3)$, while all traffic from $S(4)$ and $S(5)$ flows to $S(6)$. Of course, these logical relationships are not exactly true because some vehicles will be between sites at the start and end of the period. This error should, however, be small enough to make this model appropriate.

DAGs representing the fork and join are given in Figure 4. Because the model for $Y_t(2)$ depends on $Y_t(1)$, $Y_t(1)$ is a parent of $Y_t(2)$, and hence there is an arc from $Y_t(1)$ to $Y_t(2)$ in the DAG, and so on. Logical variables are denoted by double ovals in the DAG. Joining

together the DAGs of individual forks and joins provides a general method for eliciting a DAG and associated LMDM for an entire network. Figure 5 shows the full DAG for the Manchester network.

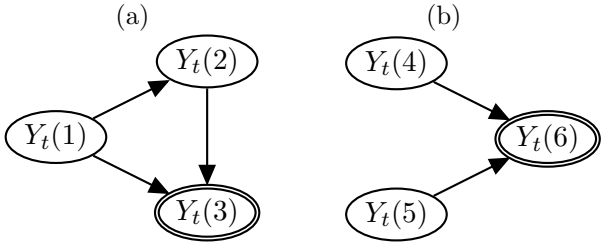

Figure 4: DAGs representing (a) a fork and (b) a join. The double ovals represent logical variables.

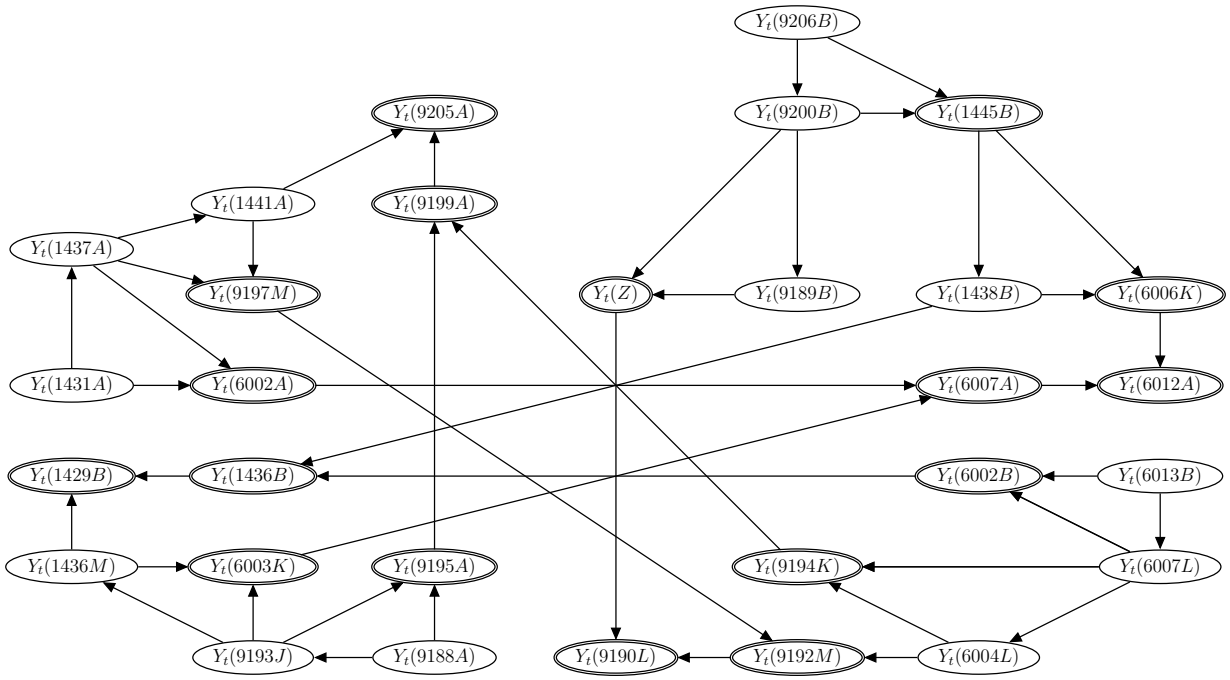

Figure 5: DAG for traffic data collection sites in the Manchester network.

4.2 Model parameters

Although each of the observation equations for $Y_t(1)$, $Y_t(4)$ and $Y_t(5)$ in (4) are algebraically the same for each time t, the actual parameters, $\mu_t(1)$, $\mu_t(4)$ and $\mu_t(5)$, will

exhibit a diurnal cycle, as clearly shown in Figure 2(a). This diurnal cycle can be modelled by a seasonal factor DLM (West and Harrison, 1997), in which there is a mean flow level parameter for each 15-minute period in the day (as described in Queen and Albers, 2009), or by a Fourier form DLM (West and Harrison, 1997, Section 8.6) or by considering splines to represent the smooth flow trend over the day (as in Tebaldi *et al.*, 2002). The advantage of a seasonal factor model is its interpretability, which, as demonstrated in Queen and Albers (2009), is especially helpful at times of modelling change via intervention (the technique of intervention allows information regarding a change in the time series to be fed into the model to maintain forecast performance — see West and Harrison, 1997, Section 11.2). When flow data are aggregated to small time intervals such as 5 minutes, a seasonal factor model can cause numerical instability problems with the Kalman filter computations because of the large number of parameters on possibly different scales. In this case, either a Fourier or a smooth trend model would be preferable for parsimony. However, for 15-minute data, a seasonal factor model does not have such problems and computation is fast and efficient.

The parameter α_t in (4) represents the proportion of traffic flowing from parent to child which, as illustrated in Figure 6, can vary systematically at different times of day. The diurnal pattern exhibited by the parameter α_t can also be modelled by a seasonal factor model as described in Queen and Albers (2009).

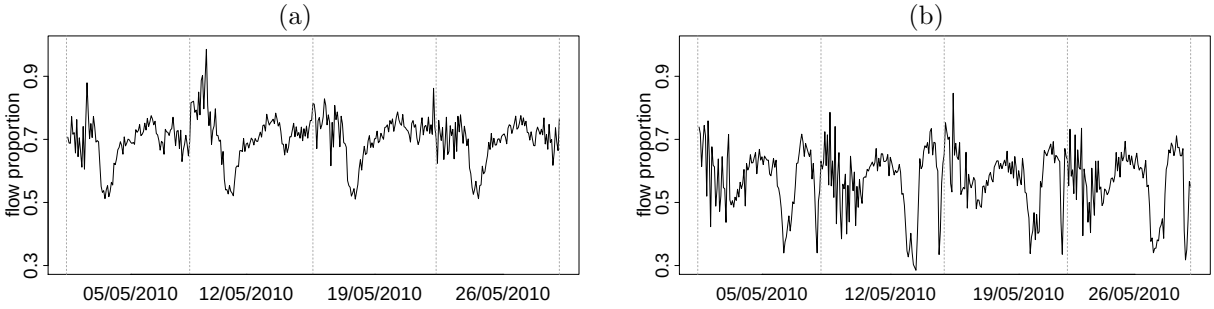

Figure 6: Proportion of traffic flowing from (a) parent 1431A to child 1437A and (b) parent 6013B to child 6007L during four Wednesdays in May 2010.

4.3 Linear relationship between parent and child

The LMDM equation for $Y_t(2)$ in (4) assumes a linear relationship between parent and child. Figure 7, showing typical plots of 15-minute flows for parent versus child at different times of the day, illustrates why this is a realistic assumption. A linear relationship would explain most of the variation between parent and child in each plot, although the relationship is not the same throughout the day. This is simply a consequence of the diurnal cycle of the proportion parameter α_t, as demonstrated in Figure 6. Notice that there seems to be two separate regimes in the plot for 17:15–17:29. This is the result of some unusual flows requiring intervention at the parent site.

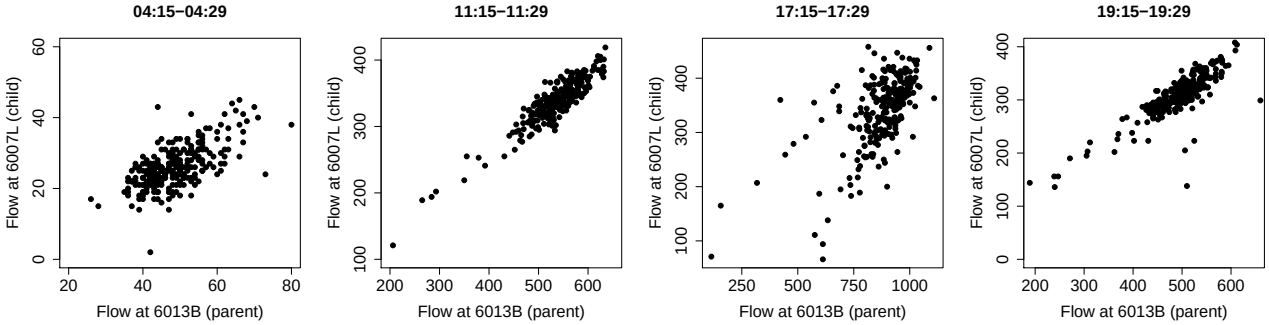

Figure 7: Plot of the 15-minute flows of parent 6013B versus the 15-minute flows of its child 6007L for some periods of the day (plots on different scales).

4.4 Contemporaneous flows as regressors

The LMDM uses univariate regression DLMs with contemporaneous flows as linear regressors. But could equally good forecasts be obtained for this application if univariate DLMs with lagged flows as regressors are used instead (as, for example, in Tebaldi *et al.*, 2002)?

To answer this question, both models were used to forecast 15-minute flows between 07:00–20:59 (ignoring the quiet night-time period) during May 2010 at sites 9206B and 9200B. The median squared error (MedianSE) and LPL were calculated for each model (the median squared error was used rather than the mean squared error because of the large number of possible outliers in traffic data — see Queen *et al.*, 2007.) The LMDM (with MedianSE = 1154 and LPL = −1288) did indeed perform better than univariate

DLMs with lagged flows as regressors (with MedianSE = 2876 and LPL = −7198). This result was also observed when forecasting 5-minute flows at these sites (LMDM MedianSE = 232, LPL = −3404 and lagged flows MedianSE = 914, LPL = −3834).

5 Modelling flow heteroscedasticity

In an LMDM, following standard variance learning methods for DLMs (see West and Harrison, 1997, Section 4.5), inference about the unknown, assumed constant, observation variances $V_t(i) = V(i)$, in (1) is based on a conjugate analysis for the associated precisions $\phi(i) = V(i)^{-1}$. However, as can be seen from the boxplots of flows shown in Figure 8, the assumption of constant observation variances is unrealistic here.

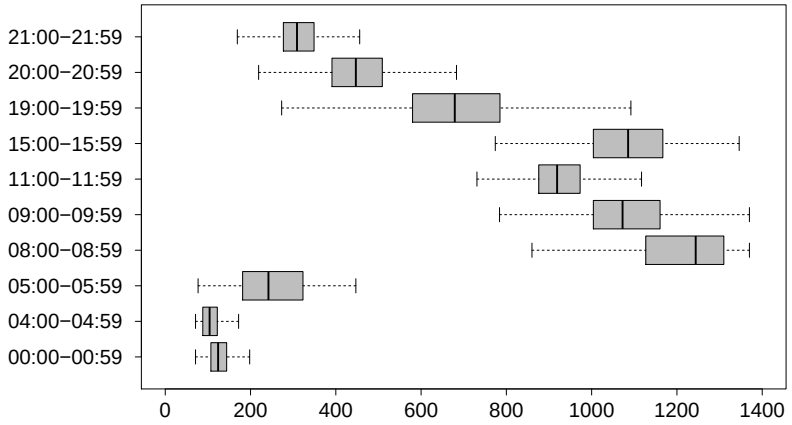

Figure 8: Boxplots of flows at site 1431A for various hours of the day using all Wednesdays from March to November 2010.

The assumption of constant variance may also not be reasonable when using the LMDM for other applications. As an example, the flow variability of goods to be distributed over a chain of supermarkets can be affected by seasonal effects due to holidays and seasons of the year. These seasonal effects can also, for example, be responsible for non-constant variability of electricity flow distribution to residential areas.

Section 4.1 proposed extending the LMDM to incorporate a variance law to enable non-normal data, in which the mean is related to the variance, to be modelled by an LMDM. Such a variance law would also accommodate a non constant $V_t(i)$. Since the LMDM

uses simple normal DLMs for each $Y_t(i)|pa(Y_t(i))$, the LMDM can easily be extended to incorporate a variance law into each conditional DLM, thus producing a novel approach for accommodating non-normal data and non constant $V_t(i)$ in multivariate state space models.

In a variance law model, write the observation variance at time t as $V_t(i) = k(\mu_t(i))V(i)$, where $\mu_t(i)$ and $V(i)$ are the underlying level and observation variance, respectively, of the series $Y_t(i)$, and $k(\mu_t(i))$ represents the change in observation variance associated with $\mu_t(i)$, which depends on the context and nature of the data (Migon *et al.*, 2005).

Figure 9 shows (different) roughly linear relationships between log mean and log variance of flows at site 9206B for two periods: 19:00-06:59 and 07:00–18:59. Similar relationships can also be observed at other sites. These empirical relationships suggest that, for each period,

$$\log(\mathrm{Var}(Y_t(i))) = \beta \log(\mu_t(i)), \tag{5}$$

where β takes different values for the two different periods.

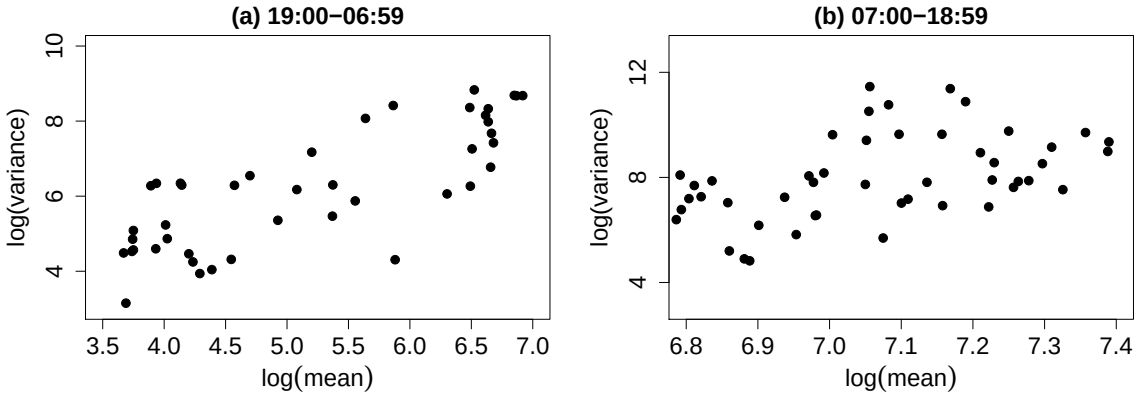

Figure 9: Flow mean versus flow variance (log scale, calculated using all Wednesdays in 2010) at site 9206B: (a) the 48 15-minute periods during 19:00–06:59 and (b) the 48 15-minute periods during 07:00–18:59 (plots on different scales).

As pointed out by West and Harrison (1997), what is important is that $k(\mu_t(i))$ "changes markedly as the flow level changes markedly", rather than determining precise values for $k(\mu_t(i))$. Thus, the empirical flow mean-variance relationship suggests modelling the

change in observation variance associated with level $\mu_t(i)$ by

$$k(\mu_t(i)) = \exp(\beta \log(\mu_t(i))),$$

with different β values for the two periods 19:00–06:59 and 07:00–18:59. (An alternative would be to have an intercept parameter in (5), but this wasn't found to improve model performance.)

The parameter $\mu_t(i)$ is the unknown mean of $Y_t(i)$. When considering a similar variance modelling issue in DLMs in the related application of road safety research, Bijleveld *et al.* (2010) use the observations themselves as proxies for the unknown mean. In this paper, where the emphasis is very much on forecasting, $\mu_t(i)$ is estimated by its forecast, denoted $f_t(i)$, obtained from the LMDM. This motivates a variance law in which $V_t(i)$ in (1) is replaced by

$$V_t(i) = \exp(\beta \log(f_t(i)))V(i). \tag{6}$$

The underlying observation variance $V(i)$ can be estimated on-line dynamically as data are observed using usual variance learning techniques (see West and Harrison, 1997, Section 4.5), whereas β can be estimated from flow means and variances using historical data, with different β values for the two periods 19:00–06:59 and 07:00–18:59.

In addition to the use of a variance law, the LMDM can be adapted further to allow the observation variances to evolve dynamically through time. Following methods developed for univariate DLMs, suppose that the precision for the model for $Y_t(i)|pa(Y_t(i))$ can change over time, so that given the posterior,

$$\phi_{t-1}(i)|D_{t-1} \sim \text{Gamma}\left(a_{t-1}, b_{t-1}\right), \tag{7}$$

the prior for $\phi_t(i)$ is given by,

$$\phi_t(i)|D_{t-1} \sim \text{Gamma}\left(\delta a_{t-1}, \delta b_{t-1}\right), \qquad \text{for } \delta \in (0, 1]. \tag{8}$$

While the prior mean for $\phi_t(i)$ is the same as the posterior mean for $\phi_{t-1}(i)$, the prior variance for $\phi_t(i)$ is larger than the posterior variance for $\phi_{t-1}(i)$, so that after observing $\boldsymbol{y}_{t-1}$, there is more uncertainty about $\phi_t(i)$ than $\phi_{t-1}(i)$. Smaller values of δ increase the

uncertainty more than larger values do. Thus smaller values of δ are suitable when the observation variance is unstable over time whereas larger values are suitable when the observation variance is more static. The updating equations to obtain the posterior for $\phi_t(i)$ are straightforward, as shown in West and Harrison (1997). This idea can also be used with the variance law so that, in (6), $V(i)$ can also evolve dynamically.

5.1 Some results

Four different LMDMs were used for forecasting in the Manchester network:

- Model A assumes a constant $V(i)$ and uses established variance learning techniques to estimate $V(i)$ on-line dynamically as data are observed;
- Model B assumes a time-varying $V_t(i)$ using the variance law (6) with a dynamically evolving underlying variance $V(i)$ as in (7) and (8);
- Model C assumes a time-varying $V_t(i)$ using the variance law (6) with a dynamically evolving underlying variance $V(i)$ as in (7) and (8) for period 19:00–06:59, while using a dynamically evolving underlying variance $V(i)$ as in (7) and (8) but no variance law for period 07:00–18:59 (because of the weaker mean-variance relationship in this period);
- Model D assumes a time-varying $V_t(i)$ using the variance law (6) only.

Historical data from February to April 2010 were used to estimate the two values of β in (6) for the two periods 19:00-06:59 and 07:00-18:59, and were also used, in the absence of expert information, to elicit priors. On-line one-step ahead forecasts were then obtained for Wednesday flows in May and June 2010.

As illustration of the parent and child forecast performance using Models A–D, Table 1 shows the LPL and MIS values when forecasting the 4 parent root nodes together with their associated (non logical variable) children. For models B and C, a value for δ is needed for each series. Following West and Harrison, 1997, the LPL could be used as an informal guide to choosing the δ which gave the best forecast performance for these data. However,

the LPL can be sensitive to outliers and so instead δ was chosen to minimise the MIS. (Models A-D all have the same forecast means and so only an assessment of the forecast limits is required.) The LPL and MIS quoted in Table 1 for each series for Models B and C, are those obtained when using the δ which minimised the MIS for that series and model. Although Model A performs the best in terms of LPL for the first pair of series in Table 1, Model B performs the best in terms of MIS for these series, and in all other cases, the best performing model is Model B, which uses the variance law and also allows the underlying variance $V(i)$ to evolve dynamically.

Table 1: LPL and MIS for forecasting using Models A–D.

	LPL				MIS			
Series	A	B	C	D	A	B	C	D
(9206B, 9200B)	$-10,001$	$-10,040$	$-10,230$	$-10,266$	691	498	541	635
(9188A, 9193J)	$-8,010$	$-7,710$	$-7,852$	$-8,394$	407	294	336	396
(1431A, 1437A)	$-9,615$	$-9,077$	$-9,140$	$-9,158$	595	414	453	487
(6013B, 6007L)	$-9,137$	$-8,466$	$-8,724$	$-9,157$	441	272	347	385

As another illustration of the forecast performance, Figure 10 shows the observed flows on a specific day for root node 1431A and its child 1437A, together with their one-step ahead forecast means $f_t(i)$ and one-step ahead forecast limits defined as $f_t(i) \pm 2\sqrt{\text{Var}(Y_t(i)|D_{t-1})}$. The forecasts were calculated considering Models A and B, since Model B performs the best amongst the time-varying models. The effect of the variance law and dynamically evolving underlying variance is clearly visible at both sites: for example, the range of the forecast limits given by Model B is much smaller than the range given by Model A during 00:00–06:59.

Note that there are some flows observed during the morning and afternoon peak periods that lie outside the forecast limits based on Model A but lie *inside* the forecast limits provided by Model B. As time t increases, in a variance law model, the observation variance estimate, $\widehat{V}_t(i)$, has the form of an exponentially weighted moving average of the forecast error (West and Harrison, 1997, p. 363), so that the most recent forecast error has a larger weight than the forecast errors observed in the past. The result of this is that, as the variance of the forecast distribution is scaled by $\widehat{V}_t(i)$, Model B will adapt more

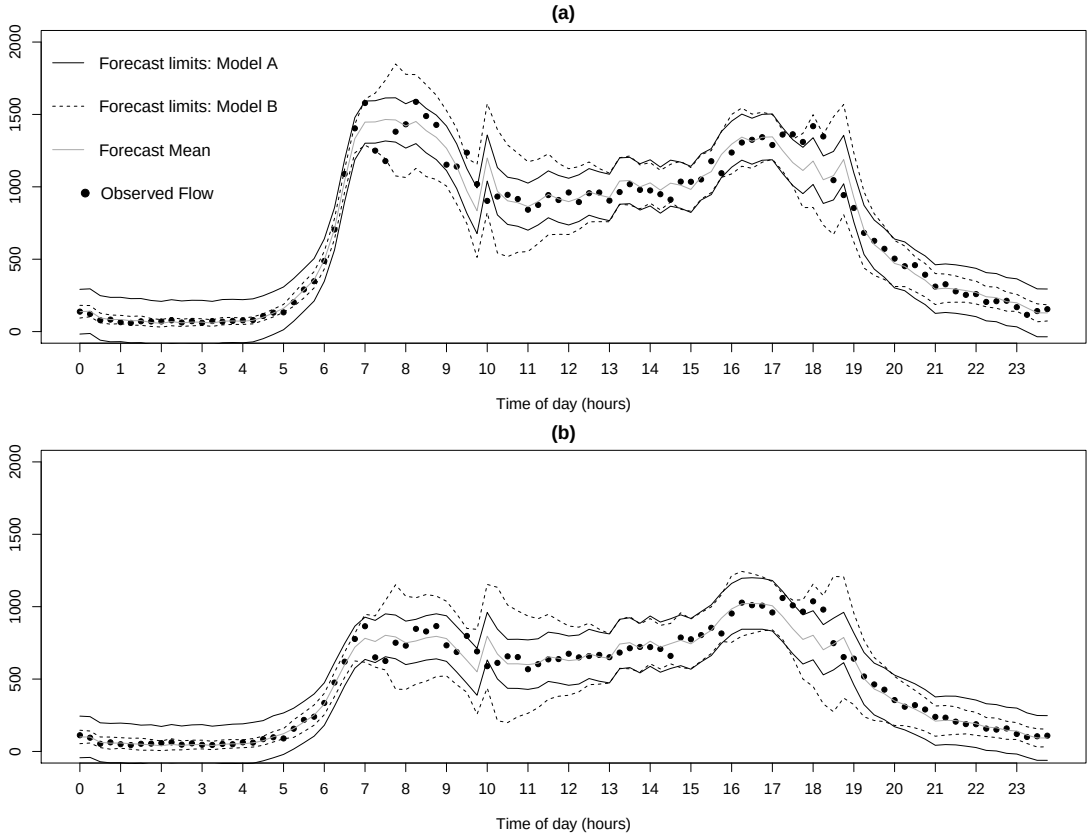

Figure 10: Observed flows on 19 May 2010, along with forecast means and forecast limits based on Models A and B — sites (a) 1431A and (b) 1437A.

quickly to correct for large forecast errors than will Model A. This means that a variance law model automatically increases uncertainty in the forecasts, which can be useful when intervention may be required but expert information is not available.

In Figure 10, the forecast limits are quite wide at times and most observations lie within them. However, for a well-calibrated model, approximately only 95% of observations should lie within the forecast limits. Over the whole forecast period, Model B actually is well-calibrated for the root nodes with roughly 95% of observations lying within the forecast limits for each series: the wide forecast limits in Figure 10(a) are a result of increased forecast uncertainty due to unexpected observations on that particular day. On the other hand, for each root node, Model A underestimates the forecast uncertainty with a coverage of roughly only 89%.

When forecasting child variables, however, Model B overestimates the forecast uncertainty with roughly 98% of observations falling within the forecast limits for each series, while this time Model A is well-calibrated with a coverage of roughly 95%. This suggests that, for child variables, there are factors affecting the variation that are not accounted for in Model B. One possible element missing from Model B, is the use of data for other traffic variables affecting flows. Anacleto *et al.* (2012) explore an adaptation of Model B also focusing on the Manchester network which uses these extra variables when forecasting flows and indeed that model is better calibrated with a coverage of roughly 95% for each of the four child variables and roughly 96% for each of the four grandchild variables considered in that paper.

6 Accommodating measurement error

6.1 Measurement errors

When building DAGs and MDMs for forks and joins in Subsection 4.1, $Y_t(3)$ and $Y_t(6)$ were both modelled as logical variables without errors. However, as is common for data in a variety of applications, loop detector data are prone to measurement errors due to device malfunctions (see Chen *et al.*, 2003 and Bickel *et al.*, 2007) so that modelling $Y_t(3)$ and $Y_t(6)$ as logical variables may not be a realistic assumption in practice.

To illustrate, consider the fork consisting of sites 1431A, 1437A and 6002A in Figure 1(b). As noted in Section 4, it would be unrealistic to expect $Y_t(6002\text{A})$ to be exactly equal to $Y_t(1431\text{A}) - Y_t(1437\text{A})$ because of time-lag effects. However, when examining the errors $Y_t(1431\text{A}) - (Y_t(1437\text{A}) + Y_t(6002\text{A}))$, it is apparent that modelling $Y_t(6002\text{A})$ as a logical variable really is too simplistic. Figure 11 shows a histogram and q-q plot of these errors observed in the period 21:00-22:59 during 2010 with 5% of the extreme errors excluded from the plot. (The most extreme errors were removed because these would be dealt with using intervention to maintain forecast performance and so the inclusion of such extreme errors in the plot gives an unrealistic picture of the errors.) From the histogram in Figure 11(a) it is clear that the errors are nearly all positive with some significant variability, while the

q-q plot in Figure 11(b), suggests that an assumption of normally distributed measurement error seems reasonable for 95% of the data and is worth considering as a simple model.

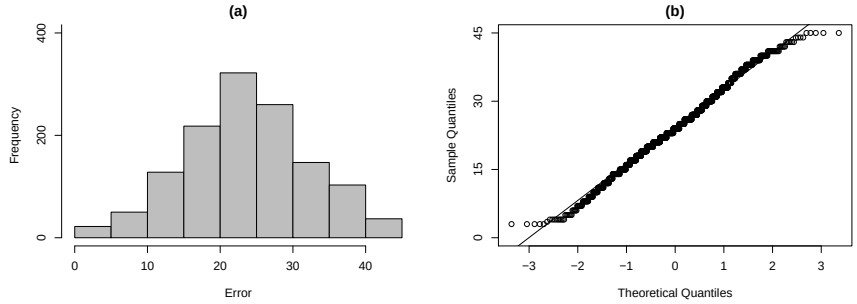

Figure 11: Histogram (a) and q-q plot (b) of errors $Y_t(1431\text{A}) - (Y_t(1437\text{A}) + Y_t(6002\text{A}))$ in the period 21:00-22:59 during 2010 (excluding 5% of the extreme errors).

6.2 Accommodating measurement error

Consider the fork of Figure 3(a). In Section 4.1, the layout of the sites and direction of traffic flow suggested that the model for $Y_t(3)|Y_t(1), Y_t(2)$ could be simply $Y_t(3) = y_t(1) - y_t(2)$. An alternative model which accommodates measurement error is of the form:

$$Y_t(3) = (y_t(1) - y_t(2))\theta_t(3)^{(1)} + \theta_t(3)^{(2)} + v_t(3), \tag{9}$$

where $\theta_t(3)^{(2)}$ is the level of the measurement error and $v_t(3) \sim N(0, V_t(3))$, for some $V_t(3)$. As vehicles from $S(1)$ can only go to $S(2)$ or $S(3)$, set the prior mean for $\theta_t(3)^{(1)}$ to be 1 with small prior variance. Note that the measurement errors at $S(1)$ and $S(2)$ are taken into account automatically through the model parameters and observation variances $V_t(1)$ and $V_t(2)$. The DAG representing this new model is the same as in Figure 4(a) except that the double oval (representing a logical variable) is now an ordinary single oval.

The distribution of the errors in the Manchester network actually differs with the time of day, as illustrated in Figure 12. The mean of the error follows the usual pattern of the flow observed during the day (see Figure 2(a)). To account for this, a seasonal factor model can be used for $\theta_t(3)^{(2)}$ in the same way as for modelling the diurnal cycle of $\mu_t(i)$ in Section 4.2. Figure 12 also shows the error variability changing through the day. In fact for the Manchester network, as with the flows themselves, there is a roughly linear

relationship between the logs of the means and variances of the errors during periods 19:00–06:59 and 07:00–18:59. Thus the variability of $V_t(3)$ can be accommodated by using a variance law LMDM as in (6), combined with a dynamically evolving underlying variance $V(3)$ as in (7) and (8).

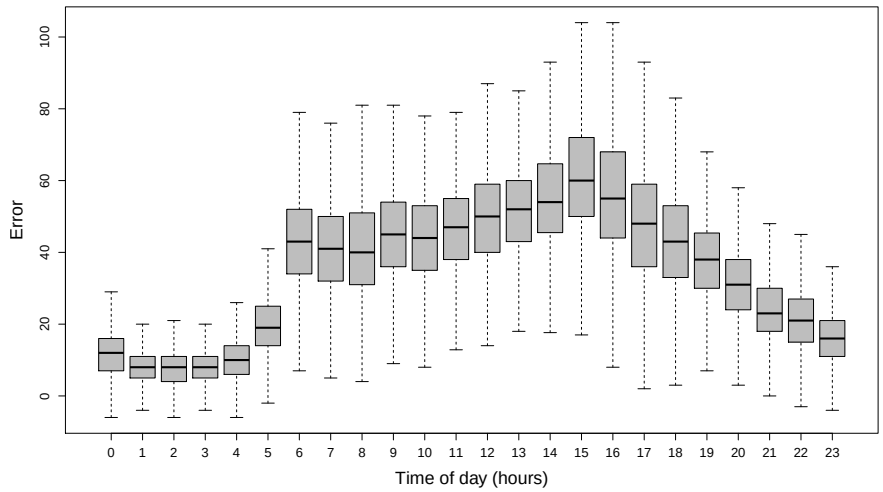

Figure 12: Box plot of the errors $Y_t(1431\text{A}) - (Y_t(1437\text{A}) + Y_t(6002\text{A}))$ in 2010.

An analogous model can be defined to allow for measurement error in a join.

6.3 Forecast performance

Model (9) and the logical model without an error term (as in (4)) were used to obtain one-step ahead forecasts for the four children of root nodes considered as logical variables in Figure 5: namely $Y_t(6002\text{A})$, $Y_t(1445\text{B})$, $Y_t(6002\text{B})$ and $Y_t(9195\text{A})$. A variance law and dynamically evolving observation variance as described in Section 5 were used with each model. As in the previous section, historical data from February to April 2010 were used to estimate the β parameters for the variance law model and for eliciting priors, while on-line one-step ahead forecasts were obtained for Wednesday flows in May and June 2010.

The MedianSE for each series when using these two models is shown in columns 2 and 3 of Table 2. (Neither the LPL nor the MIS are appropriate for model comparison here: the LPL cannot be calculated for the model without an error and the MIS is not appropriate because the error model naturally has wider forecast limits.) Table 2 also shows for each

series (in columns 4 and 5) the means and standard deviations of the relative measurement errors (that is, $100 \times$ (observed measurement error at time t)$/Y_t(i)$).

Table 2: MedianSE for the error model (9) and logical model without an error term, together with the means and standard deviations of the relative measurement errors.

	MedianSE		Relative measurement errors	
Series	Error model	Logical model	Mean	Standard devation
$Y_t(6002A)$	142	882	31.2	27.6
$Y_t(1445B)$	969	1211	9.0	59.8
$Y_t(6002B)$	180	159	-1.2	8.1
$Y_t(9195A)$	618	616	0.4	3.3

As can be seen in Table 2, the error model performs significantly better than the logical model in terms of MedianSE for two of these series and slightly worse for the other two series. Notice that the series which show the greatest improvement in using the error model in comparison to the logical model are those for which the relative measurement errors are high. However, although the error model gives greater improvement in forecast performance when the relative measurement errors are high, high relative measurement errors also mean an increase in the uncertainty of the resulting forecasts which, in turn, means that forecast limits are wider for series with high relative measurement error than for series with low relative measurement errors. Although the choice of which of two children at a fork should be considered to be the logical variable is arbitrary, the relative measurement errors for each of the children should be considered when making a decision.

As with the time-varying variance model of Section 5, the forecast limits for each (child) series in Table 2 overestimate the forecast uncertainty, with a coverage of roughly 97% for each series when using the logical model, and roughly 99% for each series when using the error model. Again, this is indicative that there are factors (such as extra traffic variables, possibly) affecting the variability which are not captured by the model.

Of course, the normal model used here for modelling the measurement error is only a simple model and other distributions may be more appropriate: for example, a mixture of distributions may work well. However, traveller information systems and some traffic management systems require real-time forecasts, and so the computational costs of considering

alternative approaches for error modelling must be carefully taken into account.

7 Forecast limits in the LMDM

When considering plots of forecasts together with the observed values, it is common to include an indication of the uncertainty associated with the forecasts. In this paper this has been done by considering the forecast limits as the marginal forecast mean $\pm(2 \times$ marginal forecast standard deviation). The uncertainty of the forecasts are often represented by forecast limits calculated in this way.

For normally distributed forecast distributions, roughly 95% of observations should lie within these forecast limits and the forecast limits are approximately 95% (equal-tailed) prediction intervals. However, the marginal forecast distributions in the LMDM are not normal and, what's more, they cannot usually be calculated analytically. Even though recent advances in MCMC and sequential Monte Carlo techniques can simulate estimates of the true forecast limits in real-time, the approximation based on marginal forecast moments is far simpler and faster. But, if the forecast limits are calculated using the marginal forecast moments in the LMDM, one question that remains is: how close is the approximation to the true 95% forecast limits?

To answer this question, consider once again the forecast limits of site 1437A obtained by this approximation (as shown in Figure 10). The 'true' 95% forecast limits of the marginal forecast distributions for site 1437A would be the 2.5% and 97.5% percentiles of the marginal forecast distributions. These can be estimated at each time t via simulation: simulate samples from the marginal forecast distributions by simulating the joint forecast distribution of parent $Y_t(1431\text{A})$ and child $Y_t(1437\text{A})$ via the normal forecast distribution for $Y_t(1431\text{A})$ and the conditional normal forecast distribution for $Y_t(1437\text{A})|Y_t(1431\text{A})$.

Figure 13 shows the approximate forecast limits for site 1437A based on marginal moments, together with the estimated 'true' forecast limits based on simulation. The plot shows the same day as was considered in Figure 10 in which there were some unusual traffic flows which created a high level of flow uncertainty. As can be seen, even when there is a lot

of forecast uncertainty, the forecast limits based on marginal moments are in fact close to the simulated true limits — certainly a good enough approximation given their ease and speed of calculation. When considering all flow series considered in Table 1, forecast intervals provided by both models also have similar MIS.

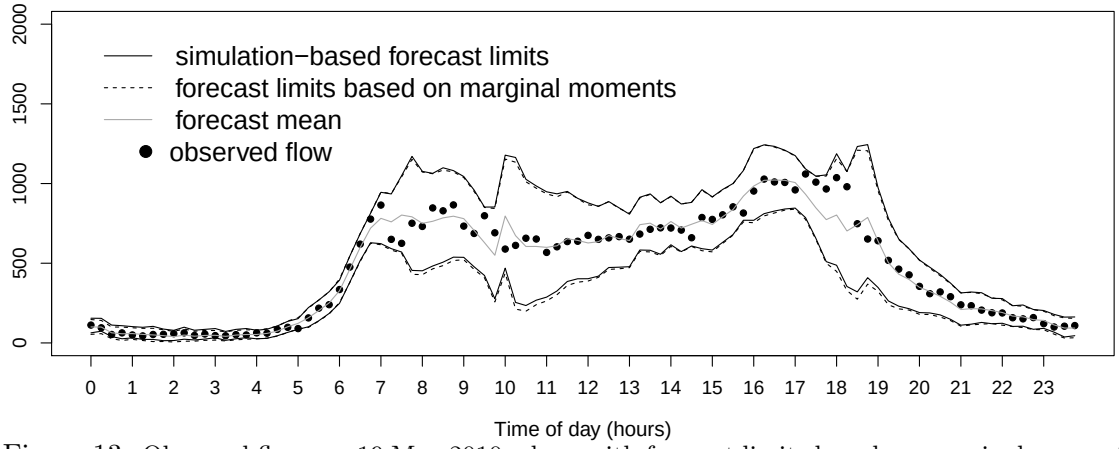

Figure 13: Observed flows on 19 May 2010, along with forecast limits based on marginal moments and simulated estimates of the true forecast limits for site 1437A.

8 Final Remarks

This paper developed models for forecasting multivariate traffic flow data, applying the proposed methodology to the problem of forecasting in a particular network.

A DAG and LMDM to represent the Manchester traffic network was elicited. New methodology has been developed allowing for time-varying observation variances in multivariate time series, extending the LMDM to incorporate variance laws and introducing methods for allowing the individual variances to evolve dynamically. Methods have also been developed for accommodating the non-negligible measurement errors which often occur in loop detector data. In addition, the paper used simulation to confirm that forecast limits approximated using the (readily available) marginal forecast moments are in fact close to estimates of the (not so readily available) true forecast limits calculated from the marginal forecast distributions.

An area of further research is the development of methods for using additional traffic

variables routinely collected along with flow data, such as average speed, headway and occupancy, to improve flow forecasts. Although current multivariate traffic forecasting models do not usually make use of the additional information from these variables, the fact that the LMDM breaks the multivariate model into a set of regression DLMs means that the incorporation of these variables into this particular multivariate model (using a combination of theoretical well-known relationships between traffic variables and data-driven approaches) is more straightforward. Research developing these ideas can be found in Anacleto *et al.* (2012).

Acknowledgements

The authors thank the Highways Agency for providing the data used in this paper and also Les Lyman from Mott MacDonald for valuable discussions on preliminary data analyses. The authors also would like to thank one of the Editors and a Referee for their constructive and helpful comments on an earlier version of the paper.

References

Anacleto, O., Queen, C.M. and Albers, C.J. (2012). Enhancing on-line multivariate flow forecasts for road traffic networks. Available at http://statistics.open.ac.uk/2012_technical_reports.

Bickel, P. Chen, C., Kwon, J., Rice, J., Van Zwet, E. and Varaiya, P. (2007). Measuring traffic. *Statistical Science*. **22(4)** 581-597.

Carvalho C. M. and West M.(2007). Dynamic matrix-variate graphical models. *Bayesian Analysis*. **2** 69-98.

Chen, C., Kwon, J., Rice, J, Skabardonis, A. and Varaiya, P. (2003). Detecting errors and imputing missing data for single loop surveillance systems. *Transportation Research Record*. **1855** 160-167.

Farrow, M. (2003). Practical building of subjective covariance structures for large complicated systems. *The Statistician*. **52** 553-573.

Fosen, J., Ferkingstad, E., Borgan, Ø., and Aalen, O. O. (2006). Dynamic path analysis — a new approach to analyzing time-dependent covariates. *Lifetime Data Analysis*. **12** 143-167.

Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*. **102** 359–378.

Guo, W., and Brown, M. B. (2001). Cross-related structural time series models. *Statistica Sinica*. **11** 961-979.

Highway Capacity Manual (2000). (Washington DC: Transportation Research Board, National

Research Council)

Kamarianakis, Y., Kanas, A. and Prastacos, P. (2005). Modeling traffic flow volatility dynamics in an urban network. *Transportation Research Record*, **1923**, 18-27.

Kamarianakis, Y., and Prastacos, P. (2005). Space-time modeling of traffic flow. *Computers & Geosciences*. **31** 119-133.

Migon, H.S., Gamerman, D., Lopes, H.F. and Ferreira, M. A. R. (2005). Dynamic models. In Dey, D. and Rao, C.R., editors, *Handbook of Statistics*. 553-88.

Queen, C.M. and Smith, J.Q. (1993). Multiregression dynamic models. *Journal of the Royal Statistical Society, B*. **55** No 4 849-870.

Queen, C.M. (1997). Model elicitation in competitive markets. In *The Practice of Bayesian Analysis* (eds S. French and J.Q. Smith) 229-243. Arnold, London.

Queen, C.M., Wright, B.J. and Albers, C.J. (2007). Eliciting a directed acyclic graph for a multivariate time series of vehicle counts in a traffic network. *Australian and New Zealand Journal of Statistics*. **49** (3) 221-239.

Queen, C.M., Wright, B.J. and Albers, C.J. (2008). Forecast covariances in the linear multiregression dynamic model. *Journal of Forecasting*. **27** 175-191.

Queen, C.M. and Albers, C.J. (2009). Intervention and causality: forecasting traffic flows using a dynamic Bayesian network. *Journal of the American Statistical Association*. **104** 669-681.

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

Stathopoulos, A. and Karlaftis, G.M. (2003). A multivariate state space approach for urban traffic flow modelling and prediction. *Transportation Research Part C*. **11** (2) 121-135.

Tebaldi, C., West, M. and Karr, A.K. (2002). Statistical analyses of freeway traffic flows. *Journal of Forecasting*. **21** 39-68.

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

Vlahogianni, E. I., Golias, J. C. and Karlaftis, M. G. (2004). Short-term traffic forecasting: overview of objectives and methods. *Transport Reviews*. **24(5)** 533-557.

Whitlock, M.E. and Queen, C.M. (2000). Modelling a traffic network with missing data. *Journal of Forecasting*. **19** 7 561-574.

Whittaker, J., Garside, S. and Lindveld, K. (1997). Tracking and predicting a network traffic process. *International Journal of Forecasting*. **13** 51-61.