Estudo do artigo:

‘Multivariate forecasting of road traffic flows in the presence of heteroscedasticity and measurement errors’ de Anacleto, O., Queen, C. e Albers, C. J.

André F. B. Menezes | Abril, 2020

O artigo

Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2013)
62, *Part* 2, *pp.* 251–270

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

Osvaldo Anacleto and Catriona Queen

The Open University, Milton Keynes, UK

and Casper J. Albers

University of Groningen, The Netherlands

[Received December 2011. Final revision June 2012]

Summary. Linear multiregression dynamic models, 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 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 due to data collection errors. This paper extends linear multiregression dynamic models to address these issues. Additionally, the paper investigates how close the approximate forecast limits that are usually used with the linear multiregression dynamic model are to the true, but not so readily available, forecast limits.

Keywords: Data collection error; Dynamic linear model; Linear multiregression dynamic model; Traffic modelling; Variance law

Organização

- Contextualização
- *Linear multiregression dynamic models*
- Acomodando Heteroscedasticidade
- Acomodando Erros de Medição
- Limites de Previsão
- Conclusões e Extensões

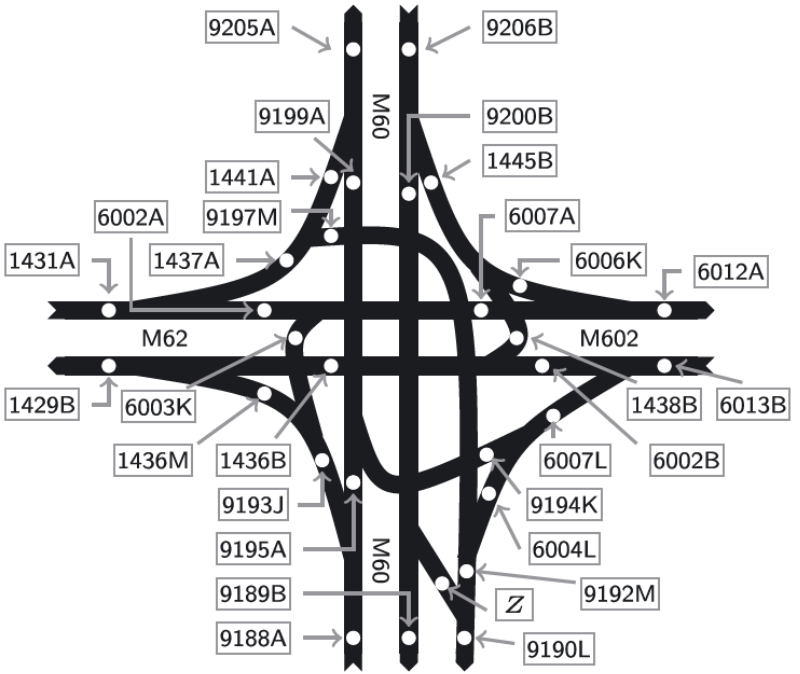
Contextualização

Tráfego de fluxo rodoviário

- **O que é?**
  - São uma "rede" (conjunto) de séries temporais de contagens de veículos que passam nos locais $S(1), \ldots, S(n)$.
  - Locais de fluxo de tráfego ao redor de $S(1)$ são informativos para prever o fluxo em $S(1)$.
- **Qual a importância?**
  - Auxiliar o sistema de gerenciamento de tráfego.
  - Controle de tráfego em tempo real (prever/gerenciar congestionamentos).
  - Sistemas de informação para viagens (tempo da viagem, acidentes).
- **Objetivo:** propor um modelo para prever fluxo de tráfego em Manchester, Reino Unido.

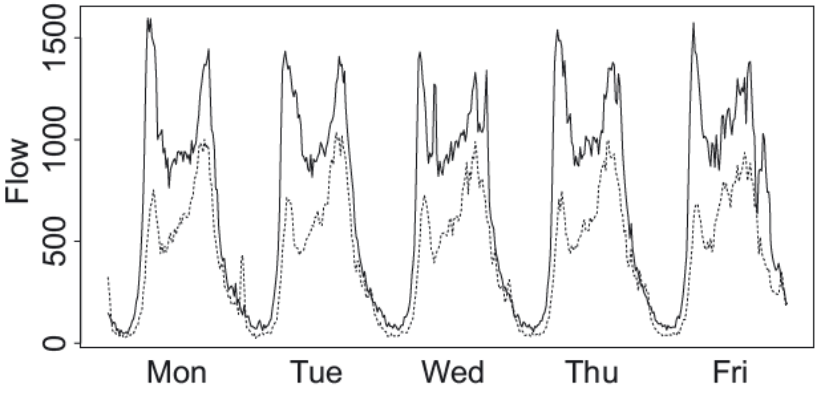
Os dados

- Contagem de veículos que passam na interseção entre as rodovias M60, M62 e M602 a cada intervalo de **15 minutos**.
- Coletados em fevereiro e novembro de 2010.
- Mesmo veículo pode passar por diferentes locais da rodovia $(\circ)$.

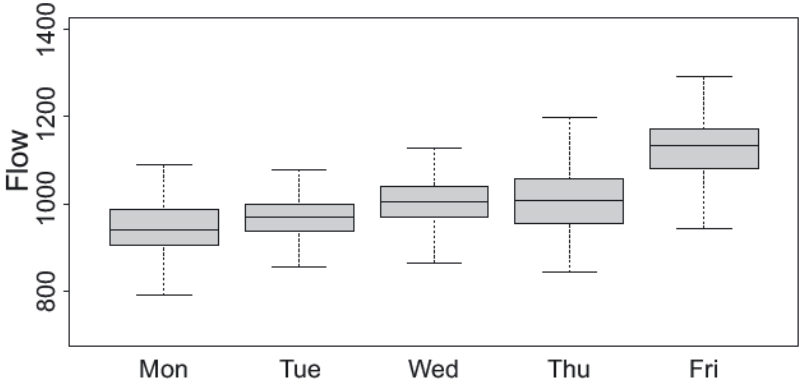


Os dados

- Série diária dos locais 1431A (solid) e 6013B (dotted) em Jun/2010.



- Local 1431A para período 2.00 p.m - 1.59pm de Mar a Nov/2010.



Linear multiregression dynamic models (LMDM)

O modelo

- Autores utilizam a teoria de grafos (*directed acycle graph* - DAG) para modelar a série temporal multivariada por meio de n modelos de regressão linear dinâmico (DLM).
- Seja $\mathbf{Y}_t = (Y_t(1), \ldots, Y_t(n))$ uma série temporal multivariada.
- $Y_t(i)$ é independente de $\{Y_t(1), \ldots, Y_t(i-1)\} \setminus \text{pa}\{Y_t(i)\}$ para cada $i = 2, \ldots, n$ e t.
- $\text{pa}\{Y_t(i)\}$ é o conjunto de variáveis *pais* (*parent*) de $Y_t(i)$.
- $Y_t(i)$ é *filho* (*child*) das variáveis $\setminus \text{pa}\{Y_t(i)\}$.
- Cada série utiliza seu **parente** como regressor.

Formalmente

- Equações de observação:

$$Y_t(i) = \mathbf{F}_t(i)^\top \theta_t(i) + v_t(i), \quad v_t(i) \sim N(0, V_t(i)),\ i = 1, \ldots, n.$$

- Equações de sistema:

$$\boldsymbol{\theta}_t = \mathbf{G}_t\, \theta_{t-1} + \mathbf{w}_t, \quad \mathbf{w}_t \sim N(0, \mathbf{W}_t)$$

- Informação inicial:

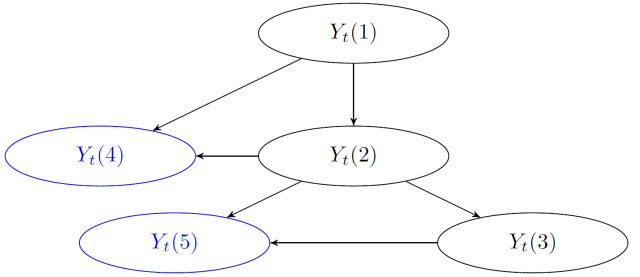
$$\boldsymbol{\theta}_{t-1} \mid D_{t-1} \sim N(\mathbf{m}_{t-1}, \mathbf{C}_{t-1})$$

O modelo

- Em que:
  - $\mathbf{F}_t(i)$ é um vetor s_i-dimensional com os parentes $\text{pa}\{Y_t(i)\}$ e/ou outras variáveis.
  - $\boldsymbol{\theta}_t(i)$ é o vetor de parãmetros s_i-dimensional para $Y_t(i)$.
  - $\boldsymbol{\theta}_t^\top = \left(\boldsymbol{\theta}_t(1)^\top, \ldots, \boldsymbol{\theta}_t(n)^\top\right)$.
  - $V_t(1), \ldots, V_t(n)$ são variâncias (escalares).
  - $\boldsymbol{m}_{t-1}$ e $\mathbf{C}_{t-1}$ são os momentos (a posteriori) para $\boldsymbol{\theta}_{t-1}$.
  - $v_t(1), \ldots, v_t(n)$ e $\mathbf{w}_t(1), \ldots, \mathbf{w}_t(n)$ são sequências independentes de erros independentes.
- Como $Y_t(i)$ e $\text{pa}\{Y_t(i)\}$ são observados simultaneamente a distribuição marginal preditiva de cada $Y_t(i)$ é necessária.
- No LMDM utiliza-se os momentos marginais da distribuição preditiva.

Exemplo

- Suponha que o seguinte grafo represente o fluxo de veículos por hora.



- $Y_t(4)$ e $Y_t(5)$ são conhecidos a partir de seus pais (variáveis lógicas).
- Equações de observação:

$$Y_t(i) = \mathbf{F}_t(i)^\top \boldsymbol{\theta}_t(i), \qquad v_t(i) \sim N(0, V_t(i)), \quad i = 1, 2, 3$$

em que

$$\begin{aligned}
\mathbf{F}_t(1) &= [1\, 0 \cdots 0], & \boldsymbol{\theta}_t(1)^\top &= [\theta_1(1), \ldots, \theta_{24}(1)] \\
\mathbf{F}_t(2) &= [y_t(1)\, 0 \cdots 0], & \boldsymbol{\theta}_t(2)^\top &= [\theta_1(2), \ldots, \theta_{24}(2)] \\
\mathbf{F}_t(3) &= [y_t(2)\, 0 \cdots 0], & \boldsymbol{\theta}_t(3)^\top &= [\theta_1(3), \ldots, \theta_{24}(3)]
\end{aligned}$$

Exemplo

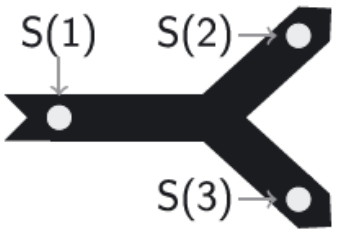
- Variáveis lógicas:

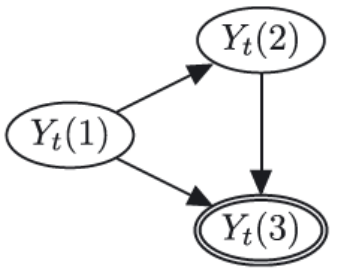
$$Y_t(4) = Y_t(1) - Y_t(2)$$
$$Y_t(5) = Y_t(2) - Y_t(3)$$

- Para acomodar ao padrão diário (sazonalidade) os vetores $\boldsymbol{\theta}_t(i)^\top$ apresentam um parâmetro para cada hora do dia.

Forks e *Joins*

- Redes de tráfego consistem em uma série de junções de dois tipos: *forks* e *joins*.
- *Fork*: veículos de um único local $S(1)$ movem-se para dois locais $S(2)$ ou $S(3)$.

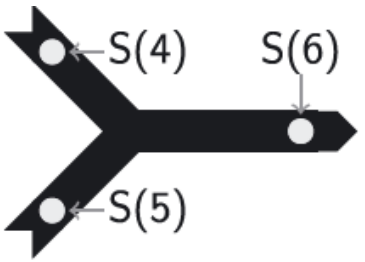


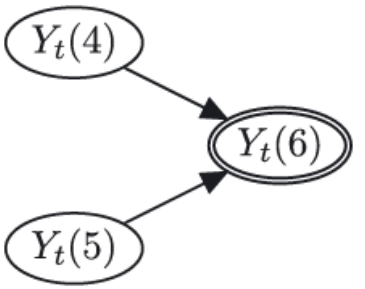


$$Y_t(1) = \mu_t(1) + v_t(1), \quad Y_t(2) = \alpha_t y_t(1) + v_t(2) \quad \text{e} \quad Y_t(3) = y_t(1) - y_t(2).$$

Forks e *Joins*

- *Join*: tráfego de dois locais, $S(4)$ e $S(5)$, se encontra em um único local $S(6)$.

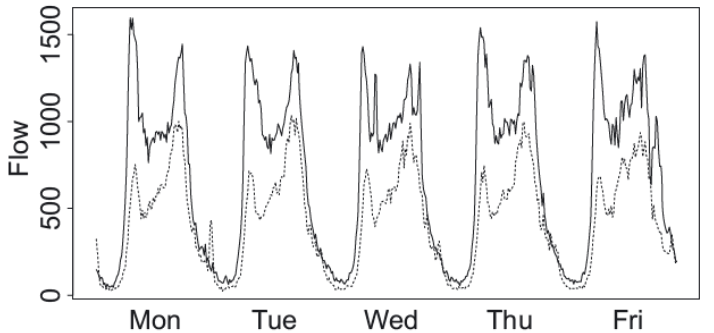


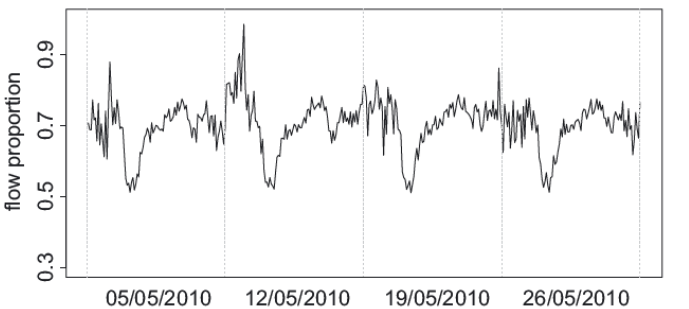


$$Y_t(4) = \mu_t(4) + v_t(4), \quad Y_t(5) = \mu_t(5) + v_t(5) \quad \text{e} \quad Y_t(6) = y_t(4) + y_t(5).$$

Parâmetros do modelo

- Os parâmetros $\mu_t(1)$, $\mu_t(4)$ e $\mu_t(5)$ descrevem o comportamento diurno cíclico do tráfego.
- Já α_t representa efeito da proporção de tráfego do local $S(1)$ para $S(2)$.





- Como modelar tal comportamento?

Propostas

- DLM com fator sazonal: parâmetro de nível médio para cada período de 15 minutos no dia.
    - 1 dia equivale a 96 períodos de 15 min, logo modelo tem 96 componentes.
- Representação de Fourier (modelo harmônico): ver livro de West e Harisson (1997) seção 8.6.
- Splines: representando suavemente o trafégo médio ao longo do dia.
- Obs: Splines e Fourier são preferíveis pela parcimônia. Fator sazonal possui maior interpretabilidade.

Relação linear entre *pais* e *filhos*

- É razoável supor que

$$Y_t(2) = \alpha_t y_t(1) + v_t(2)?$$

- Como avaliar? Empiricamente!

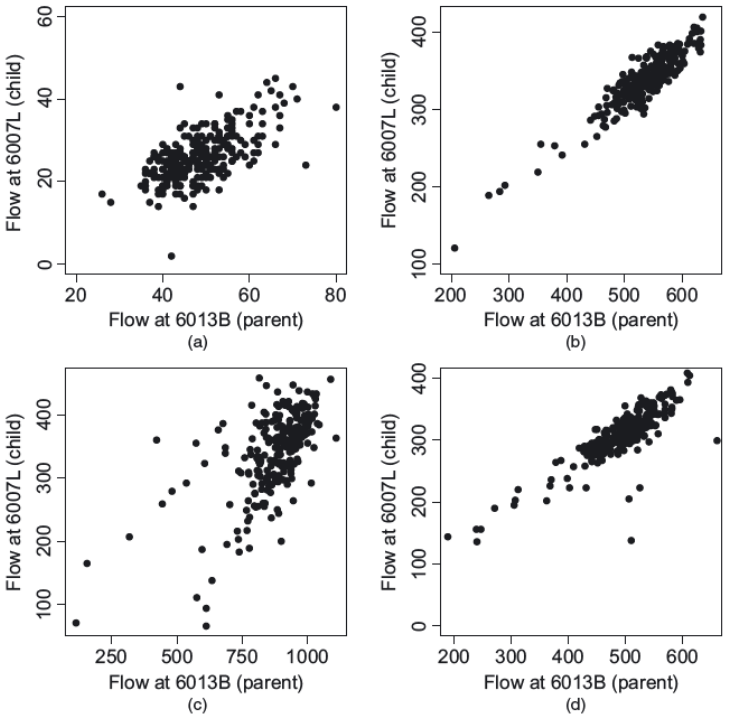


Fig. 7. Plot of the 15-min flows of parent 6013B *versus* the 15-min flows of its child 6007L for some periods of the day (the plots are on different scales): (a) 4.15 a.m.–4.29 a.m.; (b) 11.15 a.m.–11.29 a.m.; (c) 5.15 p.m.–5.29 p.m.; (d) 7.15 p.m.–7.29 p.m.

Representação DAG do modelo final

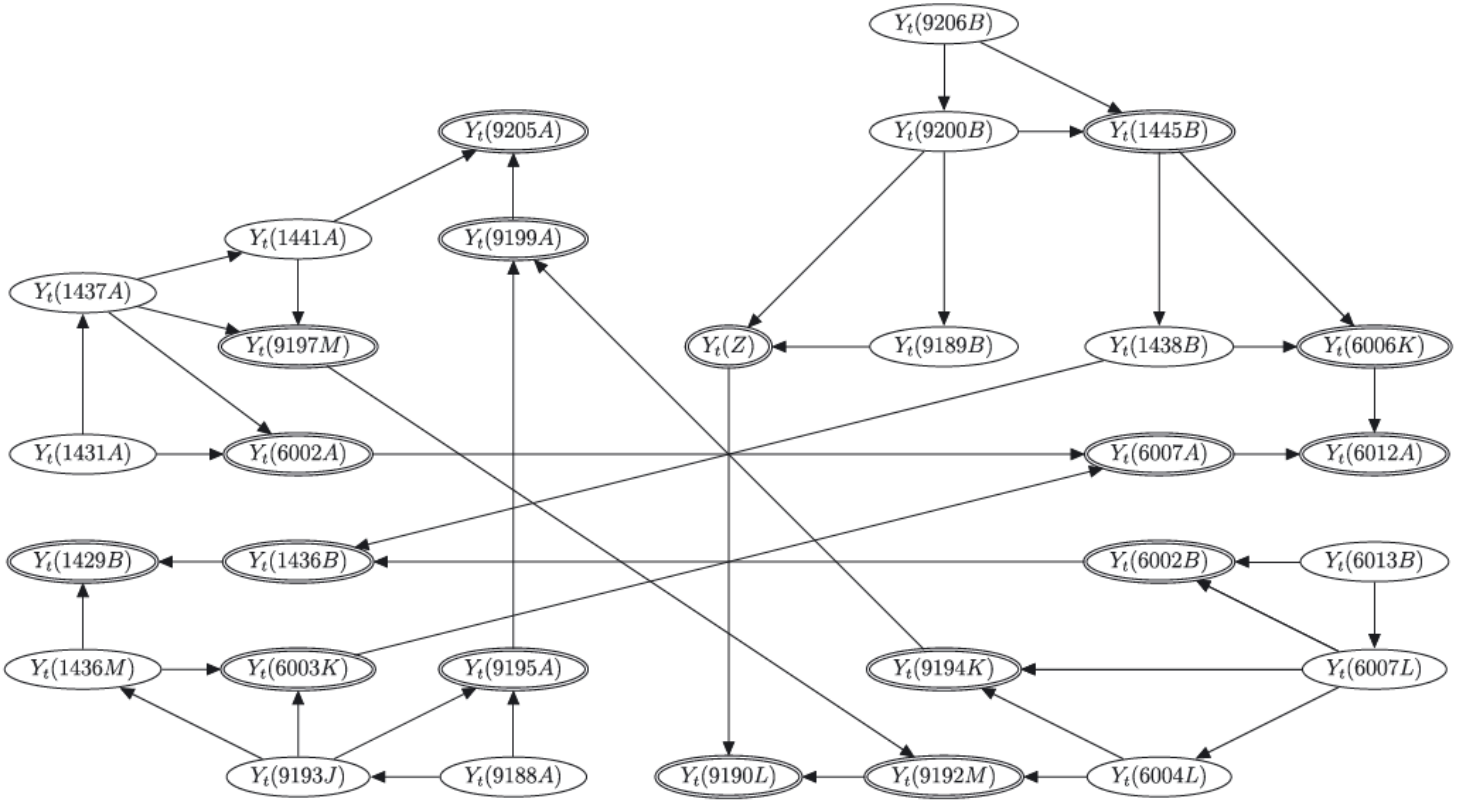
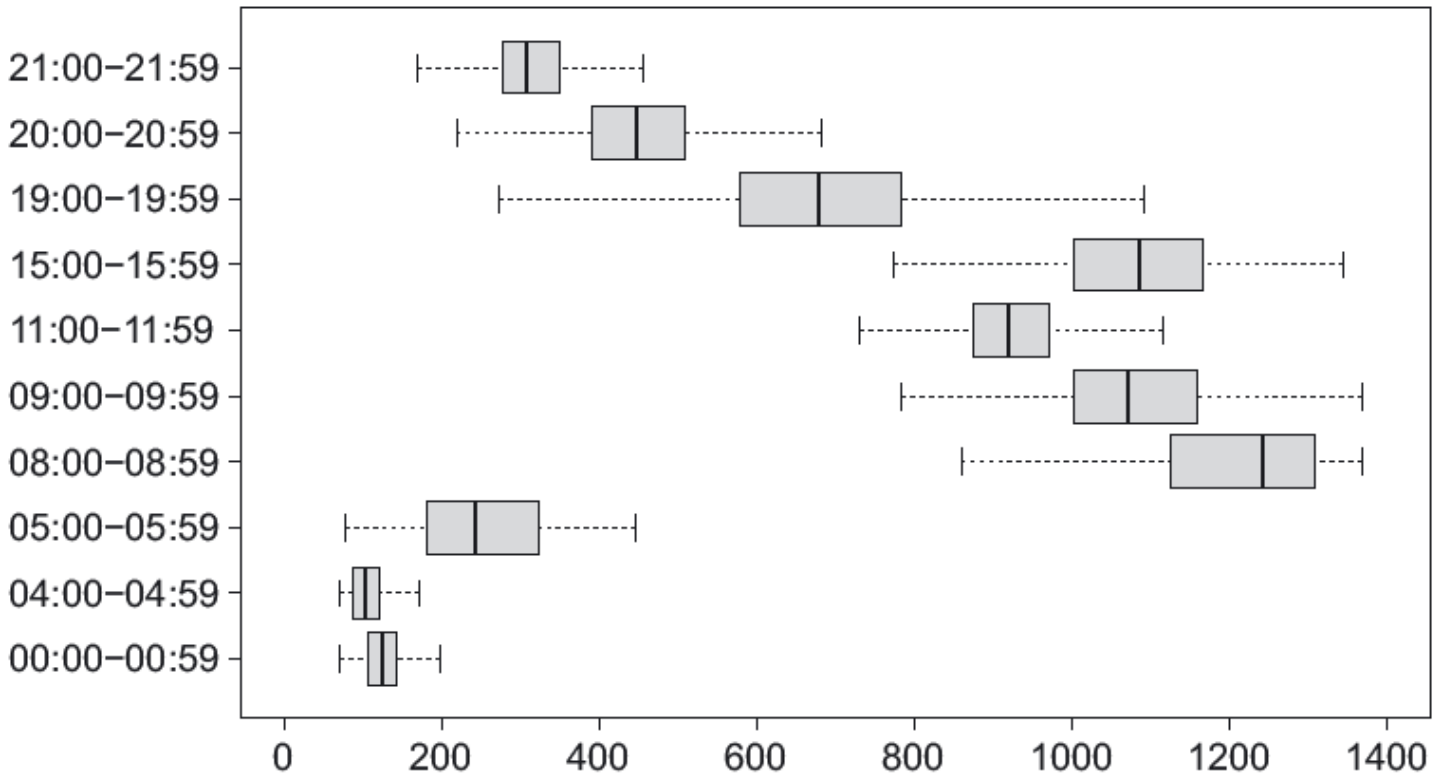


Fig. 5. DAG for traffic data collection sites in the Manchester network

Acomodando Heteroscedasticidade

O que diz os dados?

- É razoável supor que $V_t(i) = V(i), \forall\, t$?
- Comportamento do tráfego para diferentes horas do dia.



Modelando a variância (lei de variância)

- Seguindo West e Harrison (1997) o autor propõe o seguinte modelo

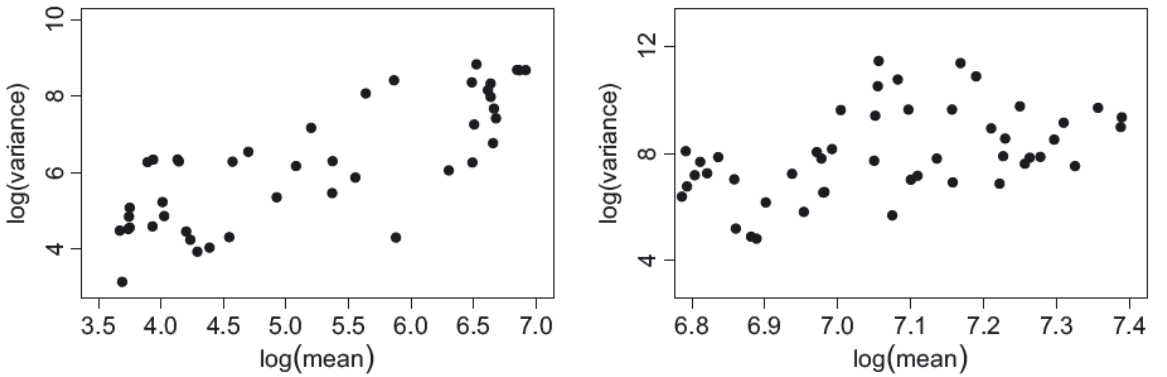
$$V_t(i) = k[\mu_t(i)] \cdot V(i)$$

em que

- $\mu_t(i)$ e $V(i)$ são, respectivamente, o nível médio e variância observacional da série $Y_t(i)$.
- $k[\mu_t(i)]$ representa a mudança na variância observacional associada ao nível $\mu_t(i)$.

Relação nivel média e variância

- Empiricamente foi observado:



- Motivando a seguinte relação:

$$V_t(i) = \exp[\beta \log f_t(i)]\, V(i)$$

em que

- β é um parâmetro que controla relação entre nivel e variância.
- $f_t(i)$ é a previsão para $\mu_t(i)$.

Evolução dinâmica da variância

- Além de supor um modelo para variância pode-se assumir que V evolui dinamicamente no tempo. Por conveniência trabalha-se com a precisão, isto é, $\phi_t(i) = V_t(i)^{-1}$.
- Dada a posteriori em $t-1$

$$\phi_{t-1}(i) \mid D_{t-1} \sim \text{Gamma}(a_{t-1}, b_{t-1})$$

a priori para $\phi_t(i)$ é dada por

$$\phi_t(i) \mid D_{t-1} \sim \text{Gamma}(\delta\, a_{t-1}, \delta\, b_{t-1}), \quad \delta \in (0, 1]$$

- A posteriori para $\phi_t(i)$ é obtida analiticamente (ver West e Harrison (1997) pág 359-361).
- **Importante:** modelar a variãncia por uma relação com a média é diferente de assumir que a variãncia evolui dinamicamente.

Modelos considerados

- Modelo A: assume variância constante $V(i)$.
- Modelo B: assume variância mudando no tempo $V_t(i)$ com lei de variância e evolução dinâmica de $V(i)$.
- Modelo C:
  - Período de 7.00 p.m.-6.59a.m. assume lei de variância e evolução dinâmica de $V(i)$.
  - Período 7.00 a.m.-6.59 p.m. assume somente que $V(i)$ evolui dinamicamente.
- Modelo D: assume variância mudando no tempo $V_t(i)$ via lei de variância.

Performance dos Modelos

- Para avaliar a performance preditiva do modelo as seguintes medidas foram utilizadas:
- Log-predictive likelihood (LPL), dada por

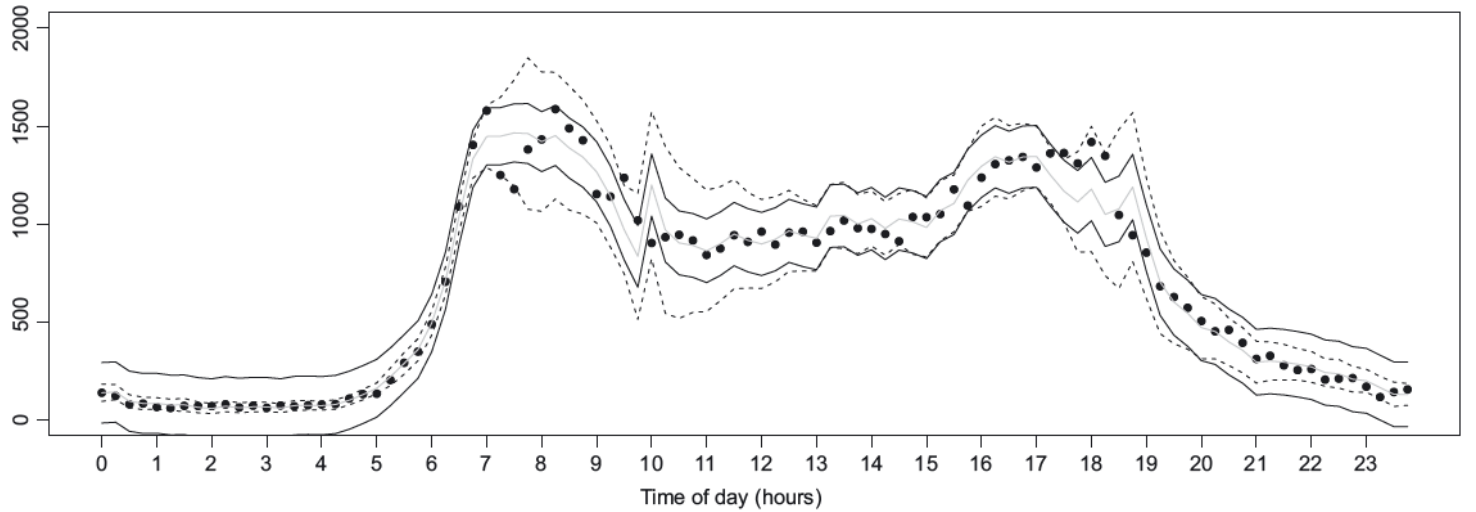
$$\text{LPL} = \sum_{t=1}^{T} \left\{ \sum_{i=1}^{n} \log f(y_t(i) \mid D_{t-1}) \right\}$$

- *Mean interval score* (MIS): função dos limites de previsão de cada observação (ver Gneiting and Raftery (2007)).

Series	*LPL for the following models:*				*MIS for the following models:*			
	A	*B*	*C*	*D*	*A*	*B*	*C*	*D*
(9206B, 9200B)	−10001	−10040	−10230	−10266	691	498	541	635
(9188A, 9193J)	−8010	−7710	−7852	−8394	407	294	336	396
(1431A, 1437A)	−9615	−9077	−9140	−9158	595	414	453	487
(6013B, 6007L)	−9137	−8466	−8724	−9157	441	272	347	385

Comparação Modelos A e B

- Limites de previsão baseados nos modelos A (solid) e B (dotted) para os local 1431A.

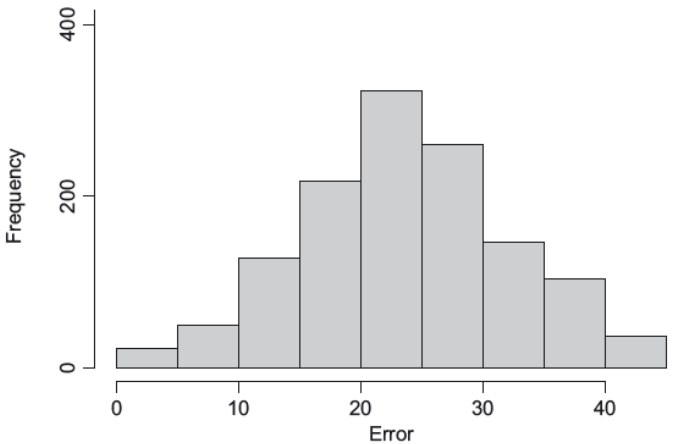


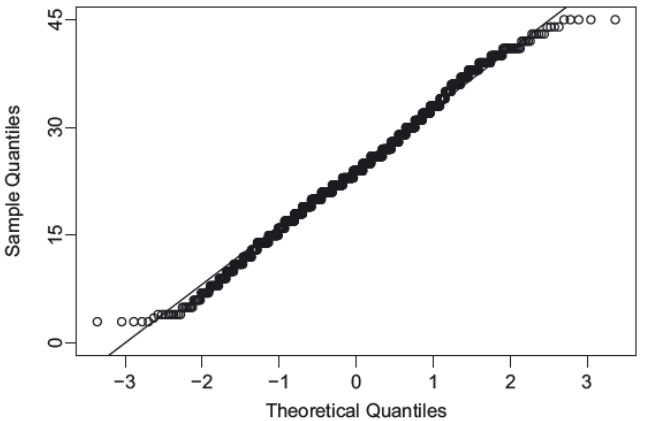
- Modelo B apresenta limites de previsão mais adequados, especialmente no período 00:00-6:59am.

Acomodando Erros de Medição

Erros de Medição

- O modelo especificado considera que as variáveis $Y_t(3)$ e $Y_t(6)$ que representam *forks* e *joins* podem ser modeladas por variáveis lógicas.
- No entanto, dispositivos de coleta de dados podem apresentar erros de medição.
- Assim, é razoável assumir que $Y_t(3) = y_t(1) - y_t(2)$ ou $Y_t(6) = y_t(4) + y_t(5)$?
- Empiricamente temos:





Modelo para Erros de Medição

- O modelo alternativo para acomodar erros de medição no caso de *forks*, é dado por:

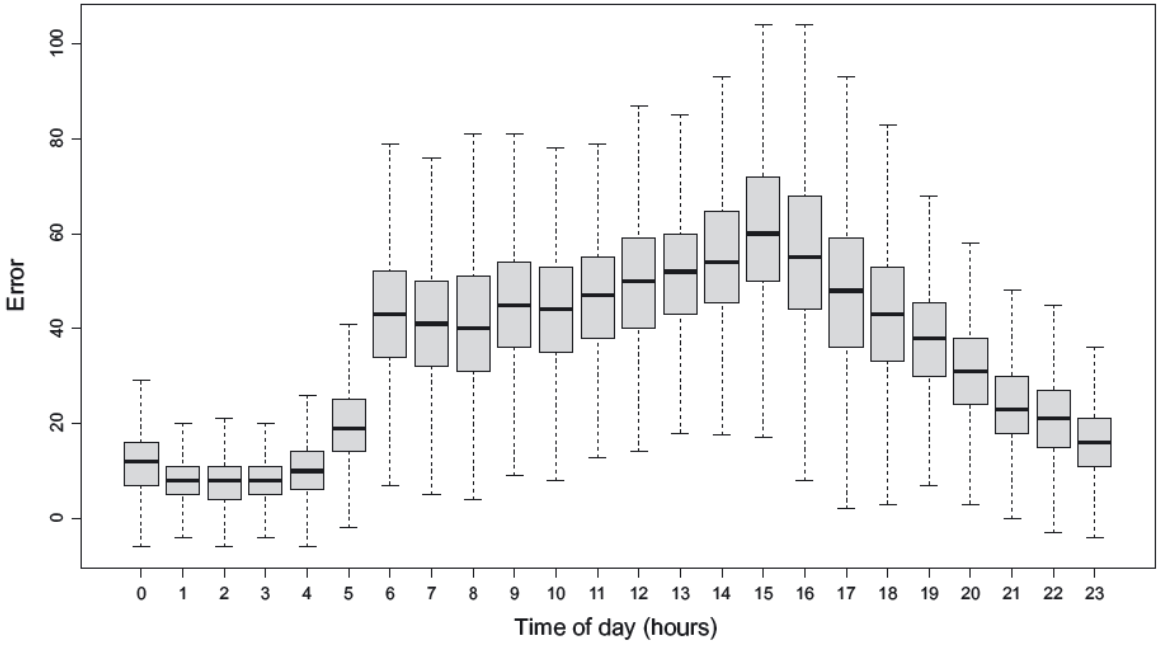
$$Y_t(3) = \theta_t(3)^{(1)} \ [y_t(1) - y_t(2)] + \theta_t(3)^{(2)} + v_t(3)$$

em que $\theta_t(3)^{(2)}$ é o nível do erro de medição e $v_t(3) \sim N(0, V_t(3))$.

- $V_t(3)$ pode acomodar lei de variância e evolução dinâmica.
- $\theta_t(3)^{(1)}$ representa o efeito do erro de medição observado.
- Uma priori adequada para $\theta_t(3)^{(1)}$ seria $N(1, \epsilon)$, com ϵ pequenho. Por que?
- Pois em um **fork** veiculos em $S(1)$ podem somente ir para $S(2)$ ou $S(3)$!

Modelo para Erros de Medição

- Qual estrutura assumir para o nível e variância?



- $\theta_t(3)^{(2)}$: fator sazonal para cada 15 min do dia.
- $v_t(3)$: lei de variância com evolução dinâmica.

Performance Preditiva

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

Series	*MedianSE for the following models:*		*Relative measurement errors*	
	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

Limites de Previsão

Limites de Previsão

- Usualmente, a incerteza relacionada a previsões é baseada em limites da forma

$$f_t(i) \pm 2 \times S_t(i)$$

em que $f_t(i)$ e $S_t(i)$ são média e desvio padrão marginais da distribuição preditiva, respectivamente.

- Se a distribuição preditiva é normal, então 95% das observações devem estar dentro deste limite.
- Contudo, no LMDM as distribuições marginais preditivas não são normais! Além disso, não possuem forma analítica!
- O quão razoável é essa aproximação em relação ao "verdadeiro" limite de previsão?

Comparação dos Limites

- Para obter a distribuição marginal preditiva autor utilizou métodos de Monte Carlo.

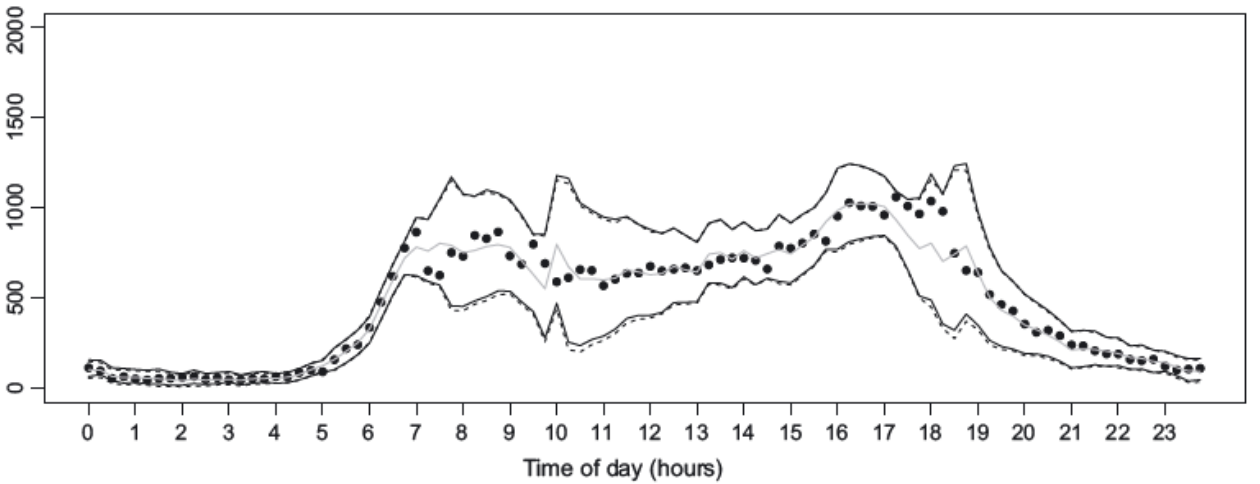


Fig. 13. Observed flows on May 19th, 2010 (•), along with forecast limits based on marginal moments (———) and simulated estimates of the true forecast limits (- - - - - - -) for site 1437A: ———, forecast mean

Considerações Finais

- Este artigo foi parte da tese desenvolvida por Osvaldo Anacleto
  - Bayesian dynamic graphical models for high-dimensional flow forecasting in road traffic networks
- Extensões do trabalho incorporando covariáveis (velocidade média, tempo de ocupação) e utilizando splines é apresentado em
  - Forecasting Multivariate Road Traffic Flows Using Bayesian Dynamic Graphical Models, Splines and Other Traffic Variables

Referências

- ANACLETO, O. **Bayesian dynamic graphical models for high-dimensional flow forecasting in road traffic networks.** Tese (Doutorado) - The Open University, 2012.
- ANACLETO, O.; QUEEN, C.; ALBERS, C. J. **Multivariate forecasting of road traffic flows in the presence of heteroscedasticity and measurement errors.** Journal of the Royal Statistical Society: Series C (Applied Statistics), v. 62, n. 2, p. 251–270, 2013.
- GNEITING, T.; RAFTERY, A. E. (2007) **Strictly proper scoring rules, prediction, and estimation**. J. Am. Statist. Ass., v. 102, N. 477, p. 359-378.
- WEST, M.; HARRISON, J. **Bayesian Forecasting and Dynamic Models.** Springer, 1997.

Obrigado!