Combining network methods with longitudinal data analysis to examine spatio-temporal variation in bike sharing data

Sarah C Gadd
University of Leeds
Leeds, UK
S.C.Gadd@leeds.ac.uk

Peter Tennant
University of Leeds
Leeds, UK
P.W.G.Tennant@leeds.ac.uk

Mark S Gilthorpe
University of Leeds
Leeds, UK
M.S.Gilthorpe@leeds.ac.uk

Alison Heppenstall
University of Leeds
Leeds, UK
A.J.Heppenstall@leeds.ac.uk

Abstract

Much spatio-temporal data analysis focuses on variation over either time or space, and aggregates data the other one of these dimensions. Methods such as latent growth curve models, multilevel models and functional data analysis can be used to analyse fine-scale temporal variation. This study examines the performance of these methods when combined with information from network analysis to investigate the effect of London Underground strikes on the timing of morning peak commuter demand for bicycles from the London bicycle sharing scheme. Results indicate which methods recover pattern features most accurately and precisely and could be used to aid bicycle network management during London Underground strikes.
Keywords: Network analysis, spatio-temporal data, latent variable methods, functional data, origin-destination data

1 Background

Spatio-temporal data analysis is a complex problem. Many methods tend to aggregate data over space or time and focus on variation over either time or space, respectively (Corcoran et al., 2014, Fuller et al., 2012, Gebhart and Noland, 2014, Saberi et al., 2018).

Several statistical methods used in psychology and epidemiology can capture complex temporal patterns at an observation-unit, and whole-dataset, level. These include latent growth curve models (LGCM), multilevel models (MLMs) and functional data analysis (FDA). Each functions differently: MLMs fit an 'average' pattern and allow individuals to differ randomly from this (Goldstein, 2011); LGCMs function similarly to MLMs but can fit complex non-parametric curves by 'distorting' the time axis (Bollen and Curran, 2005); FDA uses a linear combination of spline functions to represent each curve individually, with no 'average' curve (Ramsay and Silverman, 1997).

These methods can be used to extract features of temporal patterns at the observation-unit level, for example, minima and maxima. Information about these features could be incorporated into network analysis as edge and node properties or modelled with information from network analysis to examine spatio-temporal variation in origin-destination data.

This study aims to: investigate the ability of LGCM, MLM and FDA to recover pattern features from longitudinal data; and apply these methods to investigate the effect of London Underground strikes in London on the timing of morning peak bicycle demand (MPBD), and how this relationship changes according to spatial and network properties of routes.

2 Application Area

This study focuses on origin-destination data from the London bicycle sharing scheme (LBSS) and its response to London Underground strikes. The LBSS uses a network of docking stations, from which bicycles are removed and replaced by users at the beginning and end of trips, with each trip recorded. Bicycles are frequently redistributed to limit the number of full or empty stations. External events, weather changes, and public transport strikes may affect demand for bicycles and redistribution (Corcoran et al., 2014, Fuller et al., 2012).

Previous studies examining the effect of such events have either focused on temporal variation, aggregating trip information over the whole bike network (Fuller et al., 2012, Gebhart and Noland, 2014); or spatial variation, aggregating data over whole days (Saberi et al., 2018) or several hours (Corcoran et al., 2014). Incorporating temporal variation on a finer timescale with spatial variation may better identify where and when to plan changes in bicycle redistribution in response to strike events.

3 Methods

3.1 Simulations

A preliminary simulation of 100 datasets containing longitudinal data (12 randomly timed measurements over 24 hours) for 100 observation-units was completed. Cubic spline MLMs, FDA and LGCMs (Sterba, 2014) were used to extract the number, timing and values of maxima and minima; hourly fitted values; and within observation-unit error variation. These were compared to features of recorded observation-unit level functions.

3.2 Empirical analysis

Bicycle trip data from Transport for London (cycling.data.tfl.gov.uk) for 14 London Underground strike days and 14 non-strike (7 days before each strike) were included in analysis. A random sample of 5% of routes was selected due to computational limitations. For each route on each day, cumulative bicycle counts were generated from midnight to 12pm.

FDA was used to model bicycle counts as the ideal multiple-failure, parametric survival models could not be implemented. This assumes that counts reflect an underlying continuous process of bicycle demand. The timing of the fastest increase in bike counts (MPBD) was identified using the derivative functions for each route. Average peak times were examined visually. Multilevel models were used to investigate the relationship between strikes the timing of MPBD, and if this varied according to the degree of origin or destination nodes or journey distance.

4 Results

4.1 Simulations

Table 1 shows distributions of differences between true and estimated error standard deviations and the number and location of maxima and minima. Due to ‘distortion’ of the time axis used to capture complex patterns, LGCMs are only defined at the measurement times. Therefore, they could not capture maxima and minima or hourly fitted values. LGCMs consistently overestimated error variation.

MLMs estimated the error variation most accurately and precisely, followed by FDA which slightly underestimated this. MLMs captured the correct number of maxima and minima in a greater proportion of observation-units, however FDA estimated the time and value of these more accurately.

4.2 Empirical data

Table 2 summarises MPBD times and journey distance for strike and non-strike dates. Peak time distributions are similar, journey distance is slightly longer on strike days. Figure 1 shows that average peak times for routes leaving each bicycle docking station on strike and non-strike dates appear to be similar, but journeys towards the edges of the map appear longer and more commonly directed towards the city center on strike days than non-strike days.

Table 3 shows results from three models estimating the relationship between London Underground strikes and MPBD. The biggest plausible change in peak times due to strike is approximately 5.5 minutes earlier. There is no evidence to suggest that this is altered by origin and destination node degree or journey distance. This information does not suggest altering timing of redistribution of bikes on strike days is necessary.

Table 2: Summary of MPBD and journey distance on strike and non-strike days

Strike	Median (IQR) peak time (hours)	Median (IQR) journey distance (km)
No strike	6.28 (5.53,7.82)	3.32 (1.71,4.36)
Strike	6.26 (5.35,7.82)	3.74 (1.87,4.94)

Source: Data from Transport for London (cycling.data.tfl.gov.uk)

5 Conclusions

In this preliminary simulation, MLMs and FDA recovered pattern features more accurately than LGCMs. However, further investigations with more simulations should be performed to help identify differences in the performance of MLMs and FDA under a range of scenarios with greater precision.

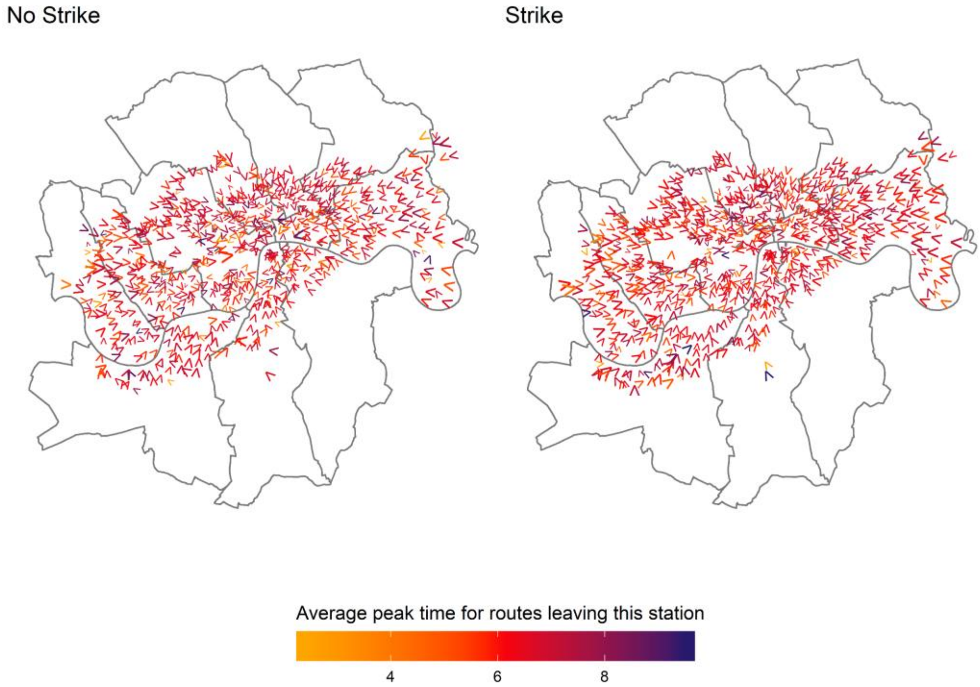
Any changes in MBPD time due to strikes were very small and on their own do not lead to any suggestions for improving bicycle redistribution on strike days. Changes did not appear to vary according to node degree of origin and destination nodes, but there was slight variation related to journey distance.

Table 1: Summary of differences between true simulated pattern features and those recovered by FDA, MLMs and LGCMs.

Method		FDA	MLM	LGCM
Percentage recovered correctly	Number of minima	24.196	48.225	-
	Number of maxima	16.956	27.643	-
Mean (95%CI) of difference between true simulated values and fitted values	Time of minima	0.64(-3.765,9.563)	0.563(-2.382,4.809)	-
	Time of maxima	3.53(-2.814,10.513)	4.102(-1.589,10.126)	-
	Value at minima	-0.32(-3.455,2.357)	-4.43(-7.97,-0.46)	-
	Value at maxima	0.846(-2.567,7.344)	-5.075(-10.377,1.042)	-
	Within person variation SD	-0.341(-0.774,0.161)	0.013(-0.485,0.714)	1.022(-0.24,3.335)
Modelling errors (out of 100 possible)		0	1	73

Source: Simulated data

Figure 1: Map showing summary of journeys from bicycle stations across London. Arrow size is proportional to average journey distance and direction depicts the circular mean journey direction. Average peak time for routes leaving each station is indicated by colour.



Source: Data from Transport for London (cycling.data.tfl.gov.uk)

Table 3: Results from four multilevel models examining the relationship between London Underground Strikes and peak morning travel times for routes in the London Bicycle Sharing Network. Model 2 investigates any modification of this relationship by the degree of the origin node, Model 3 by the degree of the destination node and Model 4 by the distance between origin and destination nodes.

Covariate	Fixed effects coefficient (95%CI)	Random effects standard deviation (95%CI)
Model 1		
Intercept	6.33 (6.279, 6.381)	0.914 (0.861, 0.966)
Strike	-0.031 (-0.088, 0.026)	
Residual		1.574 (1.549, 1.599)
Model 2		
Intercept	6.333 (6.282, 6.384)	0.915 (0.862, 0.967)
Strike	-0.032 (-0.089, 0.025)	
Strike*Origin Degree	-0.009 (-0.02, 0.002)	
Origin Degree	0.007 (-0.001, 0.015)	
Residual		1.573 (1.549, 1.598)
Model 3		
Intercept	6.331 (6.28, 6.382)	0.913 (0.86, 0.965)
Strike	-0.03 (-0.087, 0.027)	
Strike*Destination Degree	-0.013 (-0.025, -0.002)	
Destination Degree	0.004 (-0.004, 0.013)	
Residual		1.574 (1.549, 1.598)
Model 4		
Intercept	6.334 (6.283, 6.385)	0.914 (0.861, 0.966)
Strike	-0.029 (-0.086, 0.029)	
Strike*Journey Distance	-0.011 (-0.036, 0.015)	
Journey Distance	-0.006 (-0.028, 0.017)	
Residual		1.574 (1.549, 1.598)

Source: Data from Transport for London (cycling.data.tfl.gov.uk)

6 Acknowledgements

Sarah Gadd is funded by the Economic and Social Research Council (ESRC). Alison Heppenstall is funded by an ESRC-Alan Turing fellowship. Peter Tennant and Mark Gilthorpe are funded by the Alan Turing Institute. Data were provided by Transport for London.

References

Bollen, K.A. & Curran, P.J. (2005). Latent curve models: a structural equation perspective. John Wiley & Sons, Hoboken, NJ.

Corcoran, J., Li, T., Rohde, D., Charles-Edwards, E. & Mateo-Babiano, D. (2014). Spatio-temporal patterns of a Public Bicycle Sharing Program: The effect of weather and calendar events. Journal of Transport Geography, 41, 292-305.

Fuller, D., Sahlqvist, S., Cummins, S. & Ogilvie, D. (2012). The impact of public transportation strikes on use of a bicycle share program in London: interrupted time series design. Preventive medicine, 54, 74-76.

Gebhart, K. & Noland, R.B. (2014). The impact of weather conditions on bikeshare trips in Washington, DC. Transportation, 41, 1205-1225.

Goldstein, H. (2011). Multilevel statistical models. Wiley, Chichester, West Sussex.

Ramsay, J.O. & Silverman, B.W. (1997). Functional data analysis. Springer, London; New York.

Saberi, M., Ghamami, M., Gu, Y., Shojaei, M.H. & Fishman, E. (2018). Understanding the impacts of a public transit disruption on bicycle sharing mobility patterns: A case of Tube strike in London. Journal of Transport Geography, 66, 154-166.

Sterba, S.K. (2014). Fitting Nonlinear Latent Growth Curve Models With Individually Varying Time Points. Structural Equation Modeling-a Multidisciplinary Journal, 21, 630-647.