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

Sarah C Gadd[*1,2], Peter Tennant[†1,3,4] , Mark S Gilthorpe[‡1,3,4], and Alison Heppenstall[§1,2,4]

[1]Leeds Institute for Data Analytics, Leeds, LS2 9JT
[2]School of Geography, University of Leeds, Leeds, LS2 9JT
[3]School of Medicine, University of Leeds, Leeds, LS2 9JT
[4]Alan Turing Institute for Data Science & AI, The British Library, London, NW1 2DB

January 10, 2019

Summary

Much spatio-temporal data analysis aggregates data over time or space, and focuses on variation in the other. 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 network analysis to investigate the effect of public transport strikes on the timing, length and volume of peak commuter demand for bicycles from the bicycle sharing scheme in London. Results indicate which methods are best for different data structures and may be used to aid bicycle network management during transportation strikes.

KEYWORDS: Network analysis, spatio-temporal data, latent variable methods, functional data, origin-destination data

1. Background

Analysis of spatio-temporal data is a complex problem. Many methods tend to focus on variation over space, while aggregating data over time, or on variation over time, while aggregating data over space (Corcoran et al., 2014, Fuller et al., 2012, Gebhart and Noland, 2014, Saberi et al., 2018). There are a number of statistical methods used in psychology and epidemiology that can analyse complex variation over a fine timescale while capturing variation between individual observation units, for example that due to spatial variation. These include latent growth curve models (LGCM), multilevel models (MLMs) and functional data analysis (FDA). These methods could be combined with information from network analysis to examine spatio-temporal variation in origin-destination data. This study aims to:

1. Investigate the performance of LGCM, MLM and FDA combined with network analysis when applied to spatio-temporal origin-destination data
    a. Compare the advantages and disadvantages of these three methods for this situation
    b. Examine the accuracy and precision of these methods

2. Apply these methods to investigate the effect of public transport strikes in London on the patterns of peak commuter demand for rental bikes, including
    a. The start time of peak demand
    b. The length of peak demand
    c. The volume of traffic at peak demand

[*] S.C.Gadd@leeds.ac.uk
[†] P.W.G.Tennant@leeds.ac.uk
[‡] M.S.Gilthorpe@leeds.ac.uk
[§] A.J.Heppenstall@leeds.ac.uk

2. Application Area

This study will focus on origin-destination data from the London bicycle sharing scheme and its response to London tube strikes. The London scheme uses a network of docking stations, from which bicycles are removed and replaced by users at the beginning and end of trips. Each trip is recorded, providing a particularly rich source of data. 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).

A range of methods have been applied to origin-destination data to examine the effect of such events. Two studies focused purely on temporal variation, aggregating trip information over the whole bike network (Fuller et al., 2012, Gebhart and Noland, 2014). Corcoran et al. (2014) examined spatial variation in trip data which was aggregated over a number of hours. Saberi et al. (2018) examined network properties of a bike sharing system, with data aggregated over whole days. Incorporating temporal variation on a finer timescale in combination with spatial variation may better identify where and when to make changes in resource management of a bicycle sharing system to in response to external events.

3. Methods

LGCMs, MLMs and FDA can be used to model patterns in longitudinal data over time at an observation unit level but approach this in different ways.

MLMs have a similar structure to a standard linear model, with time as an exposure and bicycle counts as an outcome. However, model coefficients are allowed to vary randomly between each bicycle route. Furthermore, clustering variables can be used to modify this variation, for example, by making the coefficients of routes with the same origin more similar to each other than those with a different origin. There are a number of ways that origin and destination data could be incorporated in a multilevel model structure which will be explored in this study (Goldstein, 2011).

LGCMs are similar to MLMs, but do not incorporate time as a variable in the same way. Instead, measurement times are incorporated as ‘factor loadings’ which specify the relationship between latent, or unmeasured, parameter variables and the bike count measurements. These parameter variables are equivalent to the coefficients in a multilevel model and are also allowed to vary between routes. Clustering variables, such as route origin, are allowed to affect the parameter variables (Bollen and Curran, 2005).

The use of factor loadings in LGCMs allows them to be extended to capture non-parametric patterns of bike counts over time in a latent-basis model (Bollen and Curran, 2005). In this form, some factor loadings are estimated by the model and may differ from the real measurement time. This effectively distorts the time axis of the model, and allows the pattern of data to differ substantially from linearity. Historically this model has only been possible if all measurements are made simultaneously, but it has recently been extended to account for variation in measurement times to some degree (Sterba, 2014). This study will investigate both parametric LGCMs and latent-basis models.

FDA uses additive combinations of several basis functions to define patterns over time for each observation unit as functional data (Ramsay and Silverman, 1997). Basis functions can take many forms; commonly they are positive for a small, user-defined period of time during the measurement period, and equal to zero for the remainder. The combination of a number of these can capture complex patterns of bicycle counts for individual stations.

LGCM, MLM and FDA can define temporal variation in bicycle counts as a function for each route. This can be used to identify pattern features, such as the start of peak commuter demand, using a definition that is consistent for all routes. Changes in these pattern features in response to strikes can be examined in a number of ways including multilevel modelling and network analysis.

Increasingly, network analysis is recommended and used for analysing origin destination data, such as that from bike sharing systems (Austwick et al., 2013, Saberi et al., 2017). Network analysis can represent a bicycle sharing network by using edges (which represent routes) between nodes (which represent stations). It is able to identify spatial features in a network other than those to do with geographical location, for example, the number of destinations travelled to from a given docking station or the average distance travelled from that station. This information is likely to drive some variation in bicycle demand on certain routes, and in the response to strike events. Information from network analysis could be incorporated into analysis of pattern features as explanatory variables. Alternatively, it may be possible to incorporate information about pattern features as weights on the edges on a network.

4. Data

4.1. Empirical

This study will use rental bike data from London, freely available from Transport for London (tfl.gov.uk). This data records the origin and destination docking stations of journeys made on rental bikes from September 2015. This data will be combined with data describing weather, pollution, sunrise/set, GMT/BST, socioeconomic variables, road quality and London Underground strike information.

4.2. Simulated

Similar origin destination datasets will also be simulated using a tool that simulation of spatio-temporal origin-destination data with a random structure that reflects this. The simulations record population and observation unit-level parameters to allow evaluation of methods that seek to capture observation-unit level patterns.

5. Results

This study will critically analyse the potential new methods arising from the combination of LGCMs, MLMs or FDA with network analysis. It will identify whether these methods perform well, and which methods are most appropriate for certain data structures. The new methods will also be compared to methods currently used to analyse spatio-temporal variation, particularly in origin destination data.

Applied results from this project will identify spatio-temporal variation in the response of the London bike sharing network to tube strikes. This could be used to inform management this bike sharing system when tube strikes occur.

6. Acknowledgements

Sarah Gadd is funded by the Economic and Social Research Council. Alison Heppenstall is funded by an Economic and Social Research Council-Alan Turing fellowship. Peter Tennant and Mark Gilthorpe are funded by the Alan Turing Institute. Bicycle sharing scheme trip data are provided by Transport for London.

References

Austwick M Z, O'brien O, Strano E & Viana M (2013). The Structure of Spatial Networks and Communities in Bicycle Sharing Systems. *Plos One*, 8.

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.

Saberi M, Mahmassani H S, Brockmann D & Hosseini A (2017). A complex network perspective for characterizing urban travel demand patterns: graph theoretical analysis of large-scale origin–destination demand networks. *Transportation*, 44, 1383-1402.

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

Biographies

Sarah C Gadd is a PhD student studying longitudinal data analysis methods for epidemiology and geography. She is funded by an ESRC advanced quantitative methods studentship.

Peter Tennant is a University Acadmic Fellow in Health Data Science. He focuses on adapting and translating new and emerging methods into applied health and social science research.

Mark S Gilthorpe is a Professor of Statistical Epidemiology and a Fellow of the Alan Turing Institute for Data Science and Artificial Intelligence. His interest centres on improving understanding of the observable world through modelling.

Alison Heppenstall is a Professor of Geocomputation specialising in the development of AI/ML approaches for solving complex geographical problems. She holds an ESRC-Alan Turing Fellowship.