Combining longitudinal data analysis with networks to examine spatio-temporal variation

Sarah C Gadd*[123], Peter W G Tennant[234], Mark S Gilthorpe[234] and Alison J Heppenstall[124]

[1]Centre for Spatial Analysis and Policy, School of Geography, University of Leeds, UK, LS2 9JT
[2]Leeds Institute for Data Analytics, University of Leeds, UK, LS2 9JT
[3]School of Medicine, University of Leeds, UK, LS2 9JT
[4]The Alan Turing Institute, London, UK, NW1 2DB
*S.C.Gadd@leeds.ac.uk

Abstract

Spatio-temporal networks are a useful tool for examining systems such as transport networks. However, it is relatively difficult to examine continuous temporal change in the properties of network connections. Spatially correlated time series analysis often uses only distance to estimate correlations between time series, which is not necessarily applicable to a network system.

Longitudinal data analysis methods that are common in epidemiology and psychology may be combined with spatio-temporal network data to capture complex temporal patterns at the level of individual observation-units, for example individual network objects. These can be used to distil complex temporal information into specific easily interpretable variables that represent a specific part, or feature, of a temporal pattern, such as the timing of maxima.

This paper illustrates how these methods could be combined with geographical methods to generate meaningful and interpretable results describing spatial variation in temporal patterns of temperature in a rail network. Longitudinal methods considered were: multilevel modelling and functional data analysis.

Results show differences across the longitudinal methods used that are likely down to necessary differences in model specification. The appropriate parameterisation of each method is one of several factors that will affect the utility of these methods to accurately capture temporal pattern features in a meaningful way. There is considerable scope for further investigation of the utility of these methods through simulation.

Keywords: Spatio-temporal data, modelling, transport networks.

1. Introduction

Spatio-temporal networks can be represented as series of networks at discrete time intervals (Williams and Musolesi, 2016). This allows temporal variation in the properties of objects in the network (nodes) and connections between them (edges). However, this does not represent continuous temporal variation, which may be of interest to researchers, and assumes all edge and node properties are measured simultaneously.

Spatially correlated time-series may be used to examine patterns in spatio-temporal data with more dense information in the time axis (Kyriakidis and Journel, 1999). Spatio-temporal models are useful in a wide range of situations, but often use the distance between objects to estimate spatial correlations (Kyriakidis and Journel, 1999, Min et al., 2009, Fotheringham et al., 2015). This restricts their use to temporal patterns associated with point locations and may not be suitable for properties of network connections as discussed above.

Several longitudinal data analysis methods used in disciplines such as epidemiology and psychology are capable of capturing complex temporal patterns at the level of individual observation-units, for example network connections or point locations (Bollen and Curran, 2005, Goldstein, 2011, Ramsay and Silverman, 1997). Different longitudinal methods have different advantages and disadvantages that affect their ability to accurately capture temporal patterns in different situations (table 1).

Method	Example application	Advantages	Disadvantages
Multilevel models	Modelling patterns of recovery in patients with bone fractures involving joints (Kwok et al., 2008)	Can capture complex random structures	Patterns must be parametric Error structures must be parametric Frequent problems with convergence
Latent growth curve models	Modelling non-linear patterns of height in childhood (Grimm et al., 2011)	Non-parametric patterns can be captured Can capture complex random variation	Error structures must be parametric Non-parametric forms do not allow interpolation, limiting the ways patterns can be represented
Functional data analysis	Investigating the effect of tele-interpersonal psychotherapy on depression (Woldu et al., 2019)	Easy estimation	Patterns must be parametric Does not capture complex random error variation
SITAR method	Investigating growth patterns in teenagers (Cole et al., 2010)	Interpretable summary of patterns	Inflexible representation of pattern – always as three set growth curve properties

Table 1: Summary of some advantages and disadvantages of three longitudinal data analysis methods examined in this paper. Example papers were selected to contain some discussion of the method involved as well as an example application. Other information from Blozis et al. (2007), Bollen and Curran (2005), Goldstein (2011), Ramsay and Silverman (1997), Sterba (2014), Cole et al. (2010).

Multilevel models (MLMs) and latent growth curve models (LGCMs) are capable of incorporating complex correlation structures into the model, such as those implied by the structure of a spatial network, whereas methods like functional data analysis (FDA) do not necessarily account for this (Bollen and Curran, 2005, Goldstein, 2011, Ramsay and Silverman, 1997). Information from these models could be combined with network analysis or other geographical methods to examine how temporal patterns vary spatially. Modelling temporal patterns with or without respect to the structure of a spatial network will likely affect the results.

Information representing the whole of each temporal pattern can be difficult to interpret, meaning spatial variation in the patterns is hard to understand. Therefore, it may be useful to identify specific parts of the pattern that are of interest (hereto referred to as *pattern features*). For example, the timing of the maximum point in the pattern (Aris et al., 2017). Information like this can often be recorded as a single numeric variable. Variables representing pattern features could more easily be combined information from geographical or network methods to visualise or model spatial variation in temporal patterns.

This paper illustrates the potential for combining longitudinal methods and network analysis to data describing daily patterns of temperature for journeys in a simulated rail network and discusses differences in results using two longitudinal methods: MLMs, which can account for complex random variation, and FDA, which does not.

2. Methods

2.1 Data simulation

A simulated scenario with a simple data structure was used for illustrating methods without the complications associated with real data. The known scenario gives some idea of accuracy, but this initial study *does not* involve simulations appropriate for fully assessing the accuracy and precision of methods. The outcome of interest was temperature – while this is perhaps not a common outcome, it is normally distributed and continuous, which allows for an uncomplicated illustration.

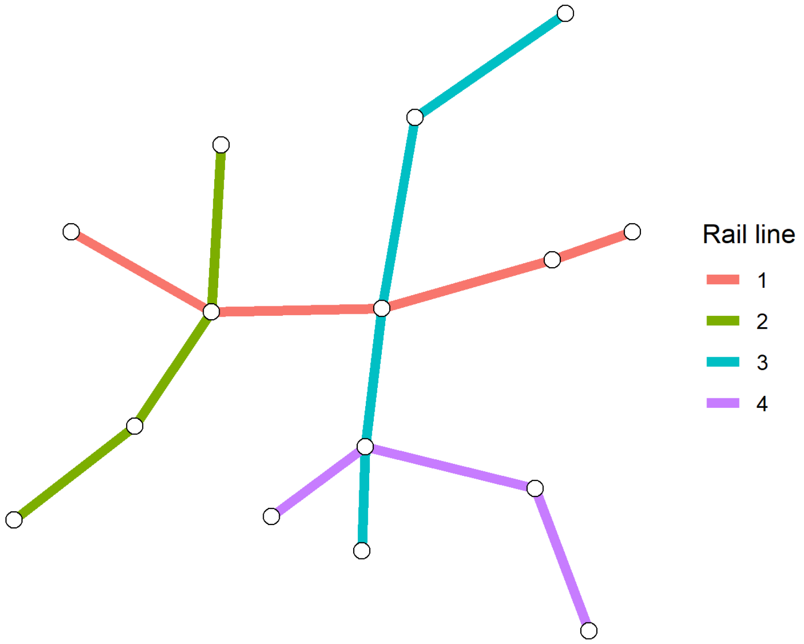

Figure 1: Map of the simulated rail network used to generate daily temperature data.

The scenario involves a small simulated underground rail network with four lines (figure 1). Like some real underground rail networks, high carriage temperatures often occur in summer, which can cause issues such as fainting for passengers. In this example, the network operators are interested in identifying where and when the highest temperatures occur to best target their resources to lower temperatures. For one summer day, they ask volunteers to carry digital thermometers with them and record the maximum carriage temperature they experience during their daily rail travel.

Longitudinal data were simulated using R to represent the maximum temperatures that volunteers recorded for each origin-destination pair (110 total) at various times of the day. The data had a hierarchical structure reflecting the rail lines travelled on (Gadd, 2019). Line 3 was simulated to reach a higher temperature than others, slightly earlier in the day.

2.2 Data analysis

In this example, analyses aimed to find the daily maximum temperature and when this occurred for each origin-destination pair. These data were combined with information from network analysis to identify if these varied according to which rail lines were used.

Non-parametric LGCMs and the SITAR method are not capable of identifying maxima in temporal patterns, so MLMs and FDA were used to model temporal patterns of temperature for each journey on each day. B-splines were used as a basis for both models, with three internal knots in the MLM and four in FDA (Pinheiro et al., 2019, Ramsay et al., 2018, Wang and Yan, 2018). The number was lower in the MLM to aid convergence. Model derivatives were used to record the maximum temperature reached for each journey-day combination and the time at which this occurred.

The R igraph package was used to identify which edges in the network were used to complete each journey in the data (Csardi and Nepusz, 2006). For each edge in the network, the mean maximum temperature and mean time of this maximum for all origin-destination paths that travelled through it was calculated. This information was visualised in network maps.

Using the path information, the rail lines that each journey used were identified. Maximum temperature and time of maximum were modelled with the inclusion of each rail line as explanatory variables.

3. Results

Figures 2 and 3 show average maximum temperatures and times of maximums for journeys using each edge in the network, respectively. For both MLMs and FDA, edges on line 3 reach higher maximum temperatures than other lines and the overall pattern of line temperatures is similar. However, temperature estimates from FDA tend to be higher. A difference in the time of maximum temperature on different lines is less apparent. The difference in estimates of time of maximum between the two models is quite large for some segments, generally towards the ends of the lines on segments used in fewer origin-destination paths.

Tables 2 and 3 show results from models investigating the relationship between rail lines used in each origin-destination path and their maximum temperature or time of maximum for each longitudinal method. For models of MLM- and FDA-estimated maximum temperature, the coefficients for Line 3 are very similar and larger than other lines, suggesting routes using line 3 have higher temperatures. The coefficients for other lines are more varied. The line 3 coefficient in both models for time of maximum temperature is the most negative, suggesting that the average time of maximum is earliest for journeys using line 3. However, the line 3 coefficients are more varied between models using MLM- and FDA- estimated time of maximum temperature than of maximum temperature.

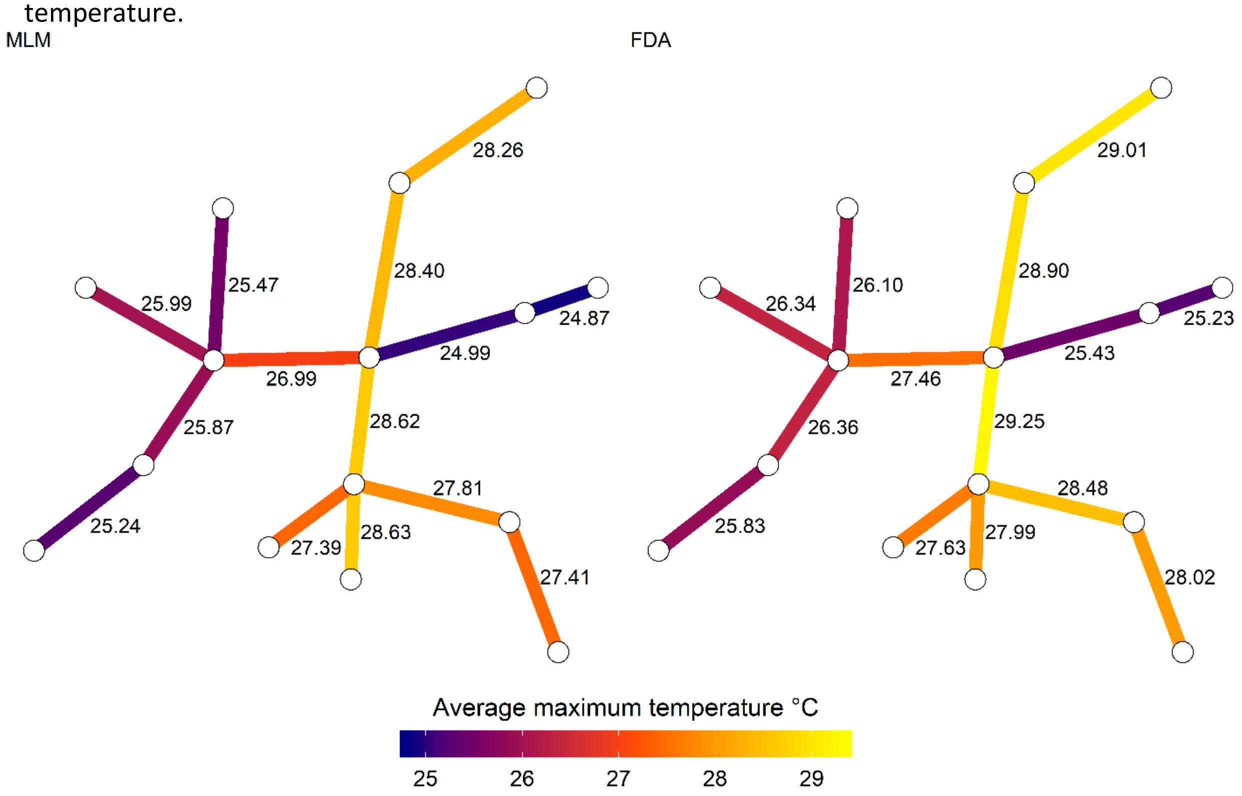

Figure 2: Maps with labels showing the average maximum temperature for journeys through each rail network edge, as estimated by multilevel modelling (MLM) and functional data analysis (FDA).

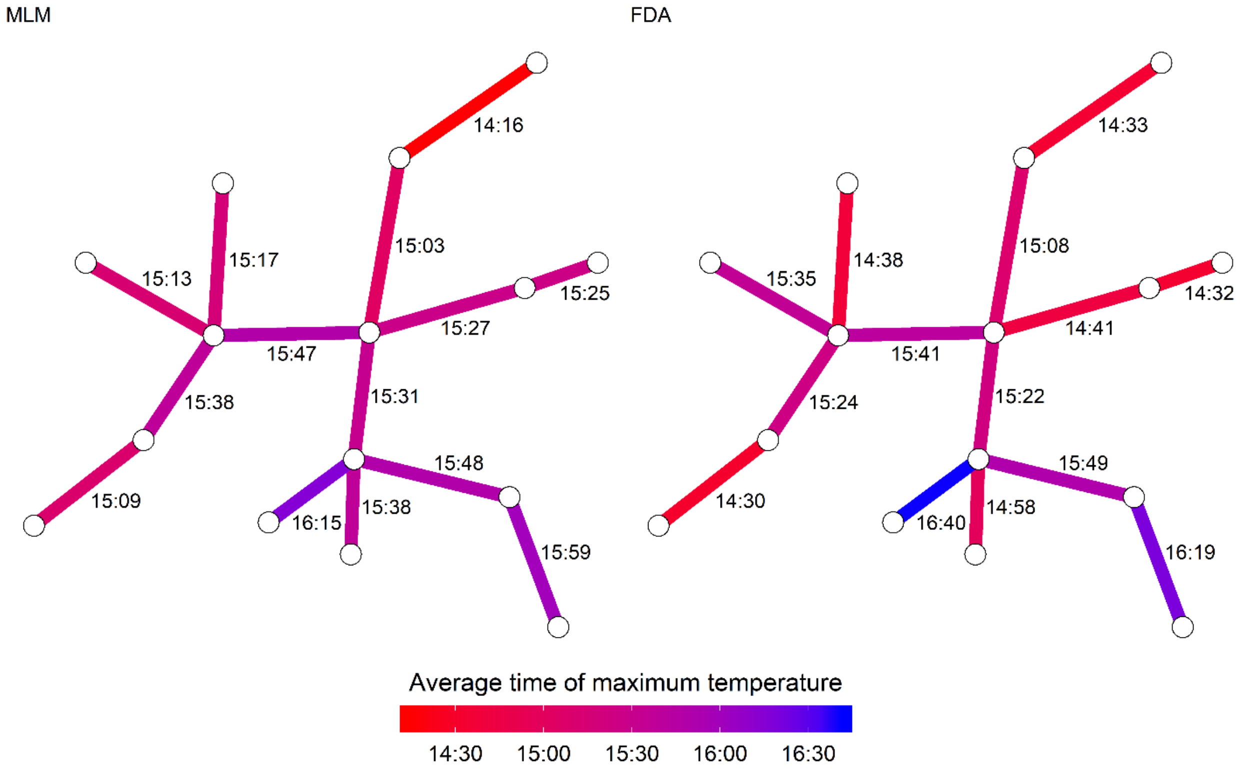

Figure 3: Maps with labels showing the average time of maximum temperature for journeys through each edge on the rail network, as estimated by multilevel modelling (MLM) and functional data analysis (FDA).

Covariate	Coefficient (95%CI)	
	Outcome: Maximum temperature	Outcome: Time of maximum
Intercept	22.93 (22.8,23.06)	14.71 (14.51,14.91)
Line1	-0.37 (-0.49,-0.25)	0.85 (0.66,1.03)
Line2	0 (-0.11,0.12)	0.25 (0.07,0.43)
Line3	5.83 (5.71,5.94)	-0.52 (-0.7,-0.34)
Line4	0.13 (0.01,0.24)	1.18 (1.01,1.36)

Table 2: Results from models investigating the relationship between lines used in journeys and the maximum temperature experienced or the time of this maximum, as estimated by MLMs.

Covariate	Coefficient (95%CI)	
	Outcome: Maximum temperature	Outcome: Maximum temperature
Intercept	22.57 (21.69,23.46)	15.6 (14.22,16.97)
Line1	0.21 (-0.61,1.02)	0.39 (-0.87,1.65)
Line2	0.28 (-0.51,1.07)	-0.34 (-1.57,0.89)
Line3	5.81 (4.99,6.62)	-1.19 (-2.45,0.07)
Line4	0.68 (-0.12,1.47)	1.41 (0.18,2.64)

Table 3: Results from models investigating the relationship between lines used in journeys and the maximum temperature experienced or the time of this maximum, as estimated by FDA.

4. Discussion

This paper provides an example application of longitudinal data analysis methods to examine spatio-temporal variation in network data. Information from the longitudinal methods was combined with information from network analysis to visualise and model variation in temporal patterns of temperature on train journeys according to network properties. The results were meaningful and easily interpretable suggesting the methods could provide a useful option for examining spatio-temporal variation. This example application focuses one method of combining longitudinal and geographical methods: combining information from network analysis (rail lines used for each journey) with information from longitudinal data analysis (maximum temperature) in models or visualisations. However, a wide range of other pattern features, types of network, network analysis methods and ways of combining them could be used.

Results from analyses using two different longitudinal methods (MLMs and FDA) found some similar patterns, but did not entirely agree. MLMs were specified with fewer splines than FDA to aid convergence. This could have altered their ability to accurately recover maximum temperatures and their timing, resulting in different estimates. The accuracy of the methods used here rely on the ability of longitudinal data analysis methods to recover pattern features such as maximum temperature. Their ability to do this accurately is likely to be affected by several factors including model specification, data structure and the complexity of the temporal pattern of interest. It is important to identify how different factors affect the accuracy of different longitudinal methods to make recommendations about when each method will provide optimal accuracy.

When examining temporal variation, MLMs accounted for the error structure in the data and FDA did not. This is unlikely to affect point estimates of pattern features, but will mean that MLMs estimate error variation more accurately than FDA (Goldstein, 2011). However, combining the longitudinal and network methods is a two-step process. In using point estimates of maximum temperature and time-of maximum in models or visualisation, we discard the uncertainty in these estimates, resulting in overly narrow confidence intervals for the results shown in tables 2 and 3 (Sayers et al., 2017). Further work is therefore needed to investigate the best way of accounting for this uncertainty and how important it is to capture the random structure in the longitudinal model.

5. Conclusion

This research presents new ideas for the combination of longitudinal data analysis with geographical methods to investigate spatio-temporal variation. The methods provided meaningful, easily interpretable results, but there were some differences between the two longitudinal methods considered. Future simulations should be considered to investigate which longitudinal methods extract pattern features most accurately for a range of different situations.

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.

7. References

Aris, I. M., Bernard, J. Y., Chen, L. W., Tint, M. T., Pang, W. W., Lim, W. Y., Soh, S. E., Saw, S. M., Godfrey, K. M., Gluckman, P. D., Chong, Y. S., Yap, F., Kramer, M. S. & Lee, Y. S. 2017. Infant body mass index peak and early childhood cardio-metabolic risk markers in a multi-ethnic Asian birth cohort. *Int J Epidemiol,* 46**,** 513-525.

Blozis, S. A., Conger, K. J. & Harring, J. R. 2007. Nonlinear latent curve models for multivariate longitudinal data. *International Journal of Behavioral Development,* 31**,** 340-346.

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

Cole, T. J., Donaldson, M. D. C. & Ben-Shlomo, Y. 2010. SITAR--a useful instrument for growth curve analysis. *International journal of epidemiology,* 39**,** 1558-1566.

Csardi, G. & Nepusz, T. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems,* 1695**,** 1-9.

Fotheringham, A. S., Crespo, R. & Yao, J. 2015. Geographical and temporal weighted regression (GTWR). *Geographical Analysis,* 47**,** 431-452.

Gadd, S. 2019. *Longitudinal Simulation Tool* [Online]. Available: github.com/sarahcgadd/longitudinal_simulation_tool.

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

Grimm, K. J., Ram, N. & Hamagami, F. 2011. Nonlinear Growth Curves in Developmental Research. *Child Development,* 82**,** 1357-1371.

Kwok, O.-M., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R. & Yoon, M. 2008. Analyzing Longitudinal Data With Multilevel Models: An Example With Individuals Living With Lower Extremity Intra-Articular Fractures. *Rehabilitation Psychology,* 53**,** 370-386.

Kyriakidis, P. C. & Journel, A. G. 1999. Geostatistical space–time models: a review. *Mathematical geology,* 31**,** 651-684.

Min, X., Hu, J., Chen, Q., Zhang, T. & Zhang, Y. Short-term traffic flow forecasting of urban network based on dynamic STARIMA model. 2009 12th International IEEE conference on intelligent transportation systems, 2009. IEEE, 1-6.

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D. & Team, R. C. 2019. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-139.

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

Ramsay, J. O., Wickham, H., Graves, S. & Hooker, G. 2018. fda: Functional Data Analysis. R package version 2.4.8.

Sayers, A., Heron, J., Smith, A., Macdonald-Wallis, C., Gilthorpe, M., Steele, F. & Tilling, K. 2017. Joint modelling compared with two stage methods for analysing longitudinal data and prospective outcomes: A simulation study of childhood growth and BP. *Statistical Methods in Medical Research,* 26**,** 437-452.

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

Wang, W. & Yan, J. 2018. splines2: Regression Spline Functions and Classes. R package version 0.2.8.

Williams, M. J. & Musolesi, M. 2016. Spatio-temporal networks: reachability, centrality and robustness. *Royal Society open science,* 3**,** 160196-160196.

Woldu, H., Heckman, T. G., Handel, A. & Shen, Y. 2019. Applying functional data analysis to assess tele-interpersonal psychotherapy's efficacy to reduce depression. *Journal of Applied Statistics,* 46**,** 203-216.