The utility of multilevel models for continuous-time feature selection of spatio-temporal networks

Abstract

Many models for the analysis of spatio-temporal networks specify time as a series of discrete steps. This either requires evenly spaced measurement times or the aggregation of data into measurement windows. This can lead to the introduction of bias. An alternative is to use continuous-time models, for example, multilevel models. Models capturing complex spatio-temporal variation are often difficult to visualise and interpret. This can be addressed by simplifying the results, for example by extracting 'features' of interest (such as maxima or minima) of temporal patterns associated with different network connections.

This paper uses simulation to evaluate the accuracy and precision with which b-spline-based multilevel models (a flexible form of continuous-time model that can easily capture complex variation associated with a spatio-temporal network structure) capture the timing and extent of maximum delays to journeys made between pairs of stations in a small railway network.

On average models captured the timing and extent of maximum delay with small bias, but there was evidence of overestimation and underestimation of low and high values of these features, respectively. This systematic bias may have partially caused the undercoverage of credible intervals for the pattern features. Alternative model specifications – specifically to capture x-axis random variation, for example – should be considered in future work.

1 Introduction

Geographers are often interested in investigating the movement of people between or around cities. This typically involves public transport or other transport networks. Transport systems can be described by a series of connections, made by travel, between a set of locations. These connections between locations can be represented in the form of a network or graph (Newman 2018). For example, Yang et al. (2019) used spatio-temporal networks to represent a public bicycle sharing system; Cheng et al. (2011) represented a road system as a spatio-temporal network when predicting traffic counts; and Chen et al. (2014) represented a metropolitan rail system as a spatio-temporal network when studying accessibility.

Networks involve objects or locations represented as vertices connected by edges. For example, a rail network may include stations as vertices and edges where the stations are connected by railway lines. Each vertex and edge might be assigned properties, for example, a station might have a set capacity for people awaiting trains, or a railway line might have a speed limit (Newman 2003). Edges can be bidirectional or unidirectional. In the latter case, an edge between two vertices only specifies a connection in one direction; for example, trains may be able to travel from station A to station B, but not in reverse (Newman 2003).

Networks are used across a range of disciplines, for example: in studies of the social connections between people (X. Li and Griffin 2013); ecological networks, including animal interactions and river network settings (Anderson and Dragićević 2018; Neeson et al. 2012); transport networks (Abdelghany et al. 2001); or economic trade networks (Liu et al. 2018). Some networks (e.g. transport, trade and ecological) have a spatial aspect, where vertices are in specific locations. Many networks also have a temporal aspect, where the network connections or the properties of edges and vertices change over time (Blonder et al. 2012). When such properties change over time it may be of interest to model them, for prediction or causal inference purposes. The aim of this paper is to evaluate the accuracy and precision with which continuous-time multilevel models capture these temporal patterns within spatio-temporal network through simulation.

When analysing such spatio-temporal networks, it is important to account for the influence of temporal and spatial correlations on the processes of interest (Dubin 1998). Many models that aim to do this incorporate time as a series of discrete steps, rather than a continuous variable (Cheng et al. 2014). This limits the models to evaluating temporal properties of network objects at discrete time points, making it more challenging to use them for predictions or inferences at any time between these discrete points. As such, the length of these discrete time steps is of great importance and must be chosen carefully. There is a need for the time steps in the data to reflect existing knowledge of the underlying processes of interest in order to best capture these features (Comber and Wulder 2019). Steps of an inappropriately large length can mask intermediate temporal variations in the data (Freeman 1989; Weiss 1984). However, the choice of the length of time steps is often predetermined by the frequency of measurements in the data generating process; if measurements are taken every ten minutes, for example, time steps shorter than this cannot be used (Freeman 1989).

If data are not measured simultaneously for different network objects, or at regular intervals, data must be aggregated into common, regular time intervals for analysis (Freeman 1989; Hwang 2000). While this allows for some degree of choice in the length of the time steps (though still limited by the frequency of measurements), aggregation can also bias inference as it introduces error in the measurement times of observations, which many statistical models assume to be error-free (Freeman 1989; Hwang 2000). While discrete-time models have been used to great advantage in the context of spatio-temporal network analysis, it may be possible to avoid some of the issues outlined here by using continuous-time models (Niezink et al. 2019; Oud et al. 2012). Multilevel (or mixed effects) spline models are capable of modelling complex temporal patterns of properties for multiple observational units (Goldstein 2011). These models are able to account for spatial and temporal correlations among objects and could potentially be used to model temporal patterns of properties for a set of objects in a network (Besag et al. 1991).

Interpreting information regarding spatio-temporal networks is also difficult – the inclusion of multiple dimensions (i.e. two-dimensional space and time) makes visualisation challenging. T. Li and Liao (2016) developed an interesting solution to the visualisation of temporal patterns of edge properties – each edge was divided into a number of time segments which were coloured to indicate the value of some property during each time window. For multiple properties on one edge, each time segment was further subdivided into smaller coloured sections. This method has some shortcomings – first, time is still treated as a discrete variable made of a series of time steps; second, the length of each edge in the network can be expected to differ, so time is not represented on a consistent scale across the network. This may confuse interpretations from the visualisation when comparing temporal patterns of properties between edges of different lengths. Subdivision of time segments to represent multiple properties occurs in the same dimension as the division into time segments. This may make it appear as if each property is measured at a different time. Furthermore, if edges overlap, as they may for dense networks, some time periods of interest will be obscured for certain edges.

Colak et al. (2013) represented changes in temporal properties of network edges in a different manner, in the context of a road traffic network. This involved focusing on a feature of the temporal pattern that constituted an event of interest. In this case, this was the time at which each edge (road segment) in the network reached 80% of capacity, thought to be the threshold at which 'congestion' exists. The authors represented this feature of the temporal pattern of edge properties in a map of the network with each edge coloured to indicate the time at which it reached 80% capacity. Including only one feature of the patterns of interest helps simplify the visualisation and reduces issues associated with temporal scales and overlapping edges. While less information is included than in the proposal by T. Li and Liao (2016), the visualisation answers a specific research question – i.e. when congestion occurs for each edge in the network. Having a specific question of interest is important in developing an appropriate analysis based on prior knowledge of the processes concerned and for avoiding some problematic practices such as 'p-hacking' (Head et al. 2015).

The specification of time in Colak et al. (2013) was discrete, yet pattern features for properties of individual edges or vertices can still be recovered from continuous-time multilevel models. This process often involves the use of calculus to extract features – such as minima, maxima, and slopes of a model – once fitted to the data (Gadd et al. 2021). However, the accuracy and precision with which multilevel models can capture these features is unknown.

Assessing the performance of statistical methods can be carried out using algebraic results or repeated simulation and analysis of data (Morris et al. 2019). Unlike finding algebraic results regarding model performance, simulations only represent model performance in a specific situation, but they are often easier to conduct where algebraic solutions are difficult (or impossible) to produce and can be used to explore how changes in data structure, for example, can affect models (Morris et al. 2019). In addition, framing a simulation in a tangible example can aid understanding of the results describing model performance.

This paper aims to use simulations to evaluate the accuracy and precision with which continuous-time multilevel models recover features of temporal patterns of edge properties in a spatio-temporal network context. In particular, this paper focuses on the extraction of maxima (a relatively simple pattern feature that can be combined with other information to capture more complex features) from temporal patterns of journey durations (the time taken to travel between a pair of stations) in a rail network as this measure has potential practical uses in identifying times and locations with very large delays in a rail network.

2 Methods

This section first outlines the process of generating simulated data associated with a spatio-temporal rail network, based on a real rail network and common commuting patterns. This is followed by specification of the continuous-time multilevel model and process used to extract pattern features (maximum journey duration, time of maximum and maximum delay). Finally, the methods used to compare the estimated temporal pattern features to the simulated values in order to assess accuracy and precision are outlined.

The ability of models to recover maxima was tested on simulated data representing journey durations between pairs of rail stations over the course of a day (Gadd et al. 2021). A series of simulated datasets with known underlying values were generated and analysed. The accuracy with which models recover these underlying values can then be assessed by investigating the difference between simulated and modelled values (Bland and Altman 1999). The relationship between estimated uncertainty and simulated values was examined to assess precision. Simulations are a simple way of assessing model performance in specific circumstances, for example, with specific data structures and were chosen here as a way of assessing one such circumstance (Morris et al. 2019).

2.1 Data

Simulated data were based on a small section of the London Underground network comprising 12 stations, shown in Figure 1 (Transport for London 2009a; Transport for London 2009b). This area was chosen as it included stations with a range of numbers of connections – some being very well connected, like King's Cross and Euston, and some with few connections, like Mornington Crescent and Goodge Street.

For each of the 132 origin-destination pairs in the network, data representing the length of time taken to complete 50 journeys at random times in one 24-hour period were generated. Each origin-destination pair had a 'true' underlying function of journey durations over the 24-hour period, all based on the function shown in Figure 2. This function is intended to represent increased congestion, resulting in longer journey durations due to waiting for trains and making slower changes, during a morning and

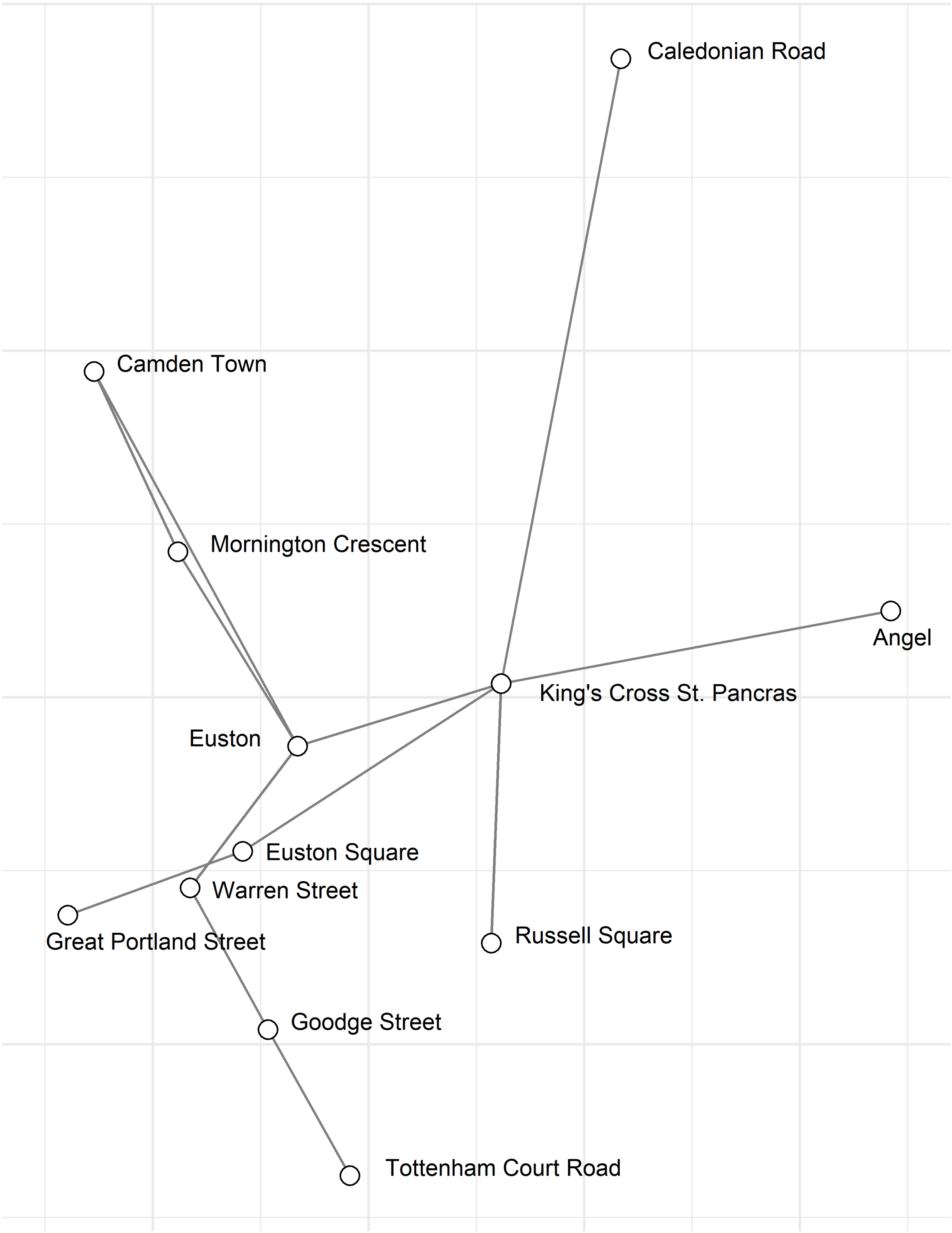

Figure 1: Map of the area of the London Underground network used to simulate data

afternoon rush hour. Note that the simulated durations include not only travel time on trains, but also travelling through the origin and destination stations between the barriers and platform and waiting for the next available space on a train – changes to this process during peak travel times can cause considerable delay. The simulated values, however, are not intended to reflect real travel times between this group of nearby stations, but just to reflect more general patterns of congestion, therefore some may be longer or shorter than expected for these particular stations.

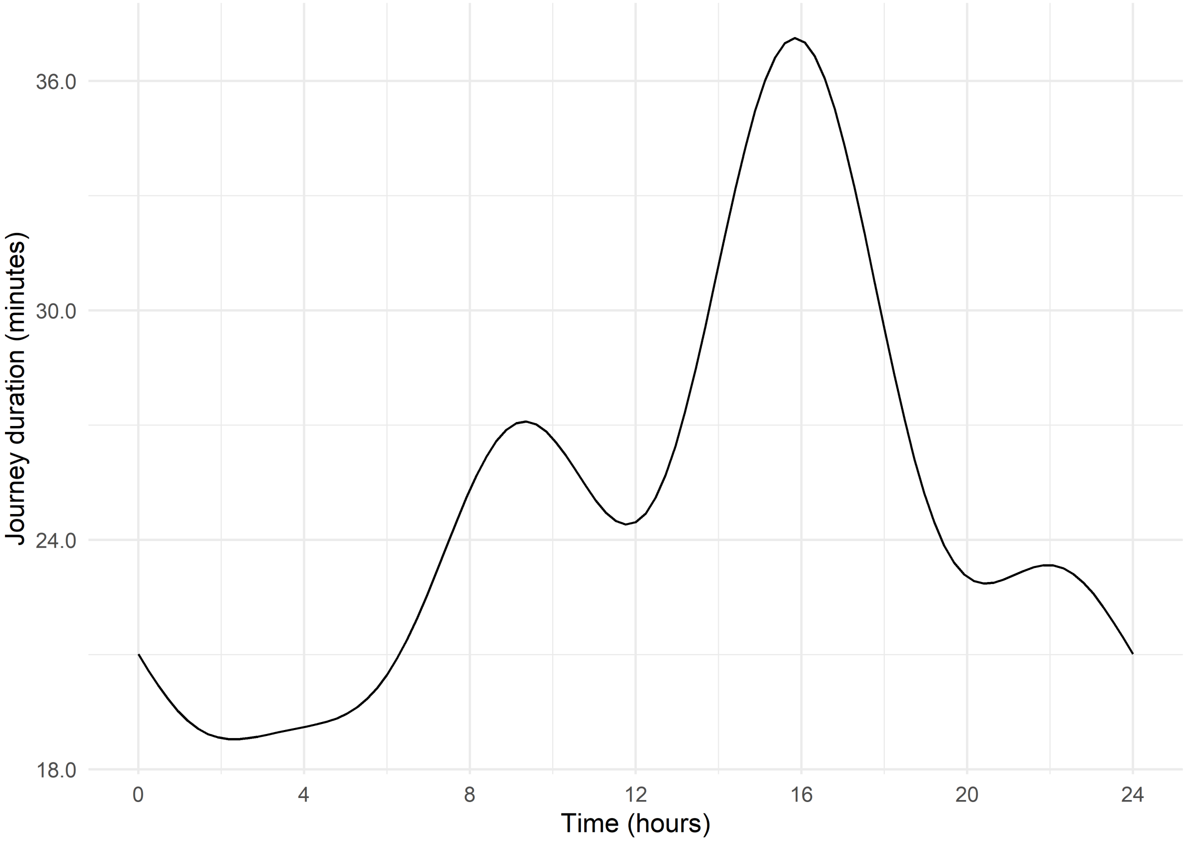

Figure 2: Basic function used to simulate temporal patterns of journey durations

There was random variation in the intercept and amplitude of this function between each origin-destination pair. Two pairs sharing an origin or destination were more similar than pairs with no shared stations.

The observed journey durations were allowed to vary randomly around the ‘true’ function for each origin-destination pair. This random variation was allowed to covary based on temporal and spatial proximity (of origin and destination stations) of the observations.

Some variation in the amplitude of temporal patterns of journey durations was introduced in relation to the properties of origin stations for each origin-destination pair. Journeys from stations with more connections (a higher degree) had greater changes between the shortest and longest journey durations (a larger amplitude), resulting in larger maximum delays for these stations. This is intended to reflect that these stations may become busier during rush hours than smaller stations, leading to more congestion and delay. This 'fixed' variation was included in the simulation process, but information was not retained for analysis – this is intended to represent latent or underlying subgroups in data that we do not have information about and to test the model's ability to cope with this situation.

An example set of simulated data is shown in Figure 3. Data were simulated using R (R Core Team 2020), code for simulation and analysis is available in the Supplementary Material.

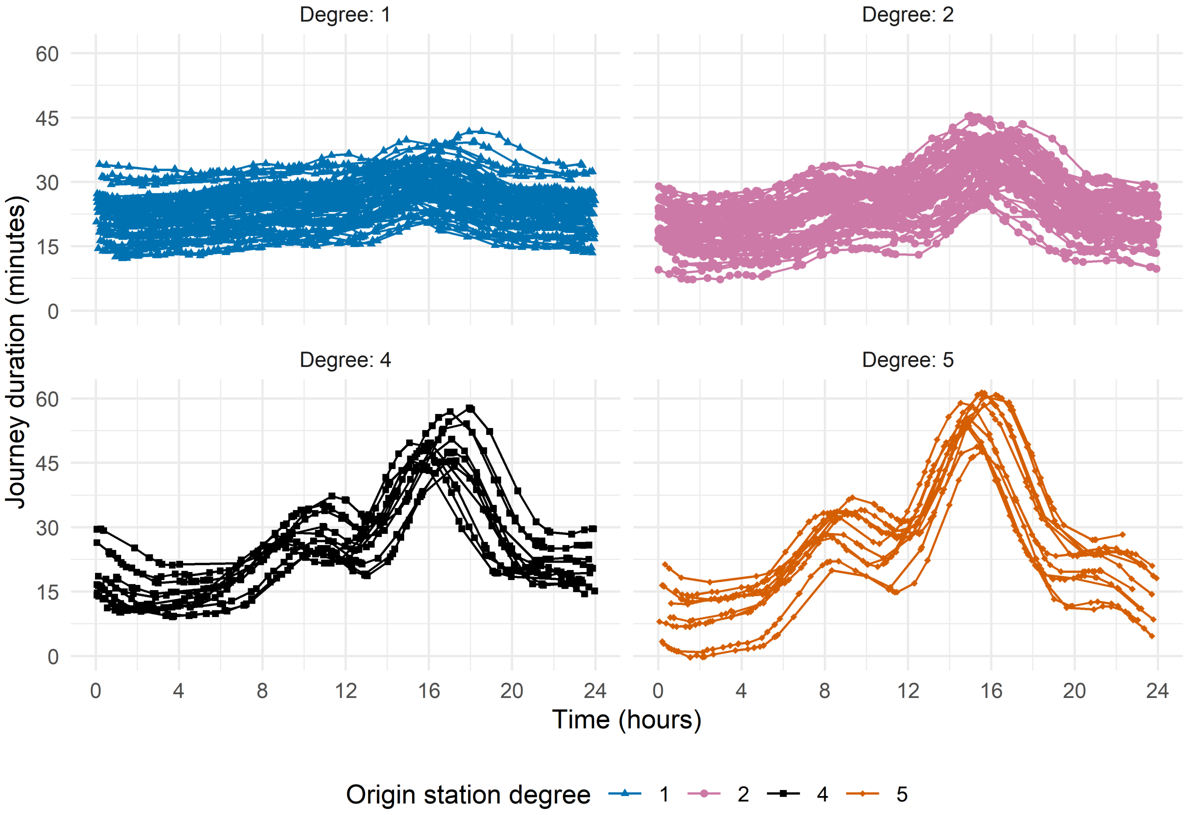

Figure 3: Example of simulated dataset representing temporal patterns of journey durations between different origin and destination stations

2.2 Model

The journey durations for each origin-destination pair were modelled over time using a multilevel model with a b-spline basis. Multilevel models were chosen because of their ability to model continuous temporal functions of properties for multiple observational units (origin-destination pairs), and their ability to incorporate complex random structures to account for spatial, temporal and network covariance of observations (Besag et al. 1991; Goldstein et al. 1994). A b-spline basis comprises a series of functions that take a positive value in part of the range of the data and are equal to zero elsewhere. When estimating a model with a b-spline basis, the linear combination of these functions that best fits the data is estimated (Perperoglou et al. 2019). B-splines were chosen over alternatives, such as penalised splines, because they are mathematically convenient; the value of each b-spline basis function can be pre-calculated for each observed measurement time and included in the model as a covariate, which is not the case for splines where a smoothing parameter must be estimated (Perperoglou et al. 2019).

The model (shown in Equation 1) for journey duration ($y_{i,from,to}$) included an intercept ($\beta_{0,from,to}$), coefficients ($\beta_{n,from,to}$) for each of k basis functions ($f_n(t_i)$) that make up the b-spline basis and an error term ($e_{i,from,to}$).

$$\begin{aligned}
y_{i,from,to} &= \beta_{0,from,to} + \sum_{n=1}^{k} \beta_{n,from,to} f_n(t_i) + e_{i,from,to} \\
\beta_{0,from,to} &= \beta_0 + u_{0,from} + u_{0,to} + v_{0,from} + v_{0,to} \\
\beta_{n,from,to} &= \beta_n + u_{n,from} + u_{n,to} + v_{n,from} + v_{n,to} \\
e_{i,from,to} &= \frac{ae_{i-1,from,to}}{(t_i - t_{i-1} + 1)}
\end{aligned} \tag{1}$$

The values of the intercept and all coefficients for each edge were allowed to vary randomly between origin-destination pairs. This random variation was split into parts (lines two and three of Equation 1): aspatial random variation associated with the origin vertex ($u_{0,from}$); aspatial random variation associated with the destination vertex ($u_{0,to}$); spatial random variation associated with the origin vertex ($v_{0,from}$); and spatial random variation associated with the destination vertex ($v_{0,to}$). The aspatial random variation was constrained to follow a normal distribution with mean zero. The spatial random variation was set to follow an intrinsic Gaussian conditional autoregressive (CAR) distribution such that random effects associated with spatially close vertices would be more highly correlated than spatially distant ones (Besag et al. 1991). This constrains the random effects to sum to zero. Having random effects associated with both the origin and destination stations meant that edges sharing an origin or destination vertex are likely to be more alike than those without a shared vertex.

In addition to the random effects, the error term in the model ($e_{i,from,to}$) was set to follow a first order continuous-time autoregressive structure to account for temporal autocorrelation of observations. This is detailed on the fourth line of Equation 1: the i^{th} error term for a given origin-destination pair ($e_{i,from,to}$) is a function of the previous error term for this origin-destination pair ($e_{i-1,from,to}$), multiplied by an autoregressive coefficient (a), divided by the time difference between the two observations plus one ($t_i - t_{i-1} + 1$ (the addition of one is to account for potential simultaneous observations)), plus a random error term, $w_{i,from,to}$, that is constrained to follow a normal distribution with mean zero. This means that the random variation for each pair of subsequent observations is correlated, with the extent of this correlation decreasing if the time difference between them increases (Goldstein et al. 1994).

The number of basis functions and the range in which each one is positive is specified using a series of knot points. The specification of these knot points was chosen separately for each known simulated pattern. Automatic knot point selection procedures are available for single level spline models (Yeh et al. 2020; Yuan et al. 2013). However, as no such procedure has been identified for a multilevel context, a selection of 40 possible knot placements were tested and the knot placement resulting in the lowest Deviance Information Criterion (a model fit criterion for Bayesian estimation) was chosen. The number of knots in the possible specifications ranged from one to 20, with knots either evenly spaced over time or placed at evenly spaced quantiles of the distribution of observations over time. Placing knots at equal spaces along the time-axis and at quantiles of the data distribution are common knot placement strategies (Ramsay and Silverman 1997). An alternative method is to place more knots at times where curvature in temporal patterns is high (Holmes and Mallick 2003; Ramsay and Silverman 1997). In this case, equally spaced knot placement strategies were chosen as the areas of high curvature will vary between origin-destination pairs. Using quantiles guarantees that all splines cover an equal number of observations, meaning estimates for all spline coefficients are supported by the same amount of data even if data are not evenly distributed (Howe et al. 2013; Ramsay and Silverman 1997). In this simulation, we expect data to be distributed evenly along the time-axis. The knot specification chosen was 19 knots spaced evenly along the time-axis.

Models were fitted using a Bayesian Markov chain Monte Carlo estimation procedure in R and OpenBUGS (Lunn et al. 2009; R Core Team 2020; Sturtz et al. 2005; A. Thomas et al. 2014). OpenBUGS is a modelling program that uses Markov chain Monte Carlo estimation (with a Gibbs sampling algorithm) to fit complex model structures (S. Geman and D. Geman 1984). This was chosen as Bayesian estimation methods are generally more effective at estimating complex model structures than maximum likelihood methods. OpenBUGS was chosen as alternative Bayesian engines did not have the capability to fit spatially correlated random effects (L. Muthén and B. Muthén 2010; Plummer 2003; Stan Development Team 2019). Bayesian estimation aims to estimate a probability distribution for model parameters, rather than a set of point values, and produces a series of posterior parameter samples. This is based on the model specification, data and prior distributions provided. Prior distributions represent current beliefs about which values are most likely for each parameter (Heck and S. Thomas 2015). In this case, loose priors that do strongly indicate any prior beliefs were specified as this is not an analysis intended to fully reflect Bayesian thinking, but rather one in which Bayesian estimation tools are the most suitable.

Initially, models with loose priors proved difficult to estimate, even with the aid of this tool. To aid estimation, models were initially run *without* the first order temporal autoregressive error structure included. This produced a set of posterior distributions that could be included as prior distributions for some parameters in the full model. As recommended when using intrinsic CAR random effects, the fixed effects parameters (β_0 and β_n in Equation 1) were set to follow a flat prior in this model (Besag et al. 1991; A. Thomas et al. 2014). The precision parameter for random effects distributions was set to follow a gamma prior distribution with shape and rate equal to two. The starting value for the precision parameter was set to one and starting values for fixed effects were sampled from a standard normal distribution. The initial model was run with four chains of 4000 iterations each, 3000 of which were discarded as model burn-in. This produced parameter distributions (to be used as priors in the final model) that allowed the final model (with temporally autocorrelated errors) to be consistently estimated with a low percentage of convergence failures (defined here as when the model estimation process failed to complete) in the simulations.

The posterior distributions for fixed effects from this initial model were used as priors and to choose starting values in the final model. The autocorrelation parameter was set to follow a gamma distribution with shape and rate equal to one, truncated at zero and one as positive autocorrelation was expected. The starting parameter was set to 0.5. The initial values and distributions for random effects precision parameters remained the same as the initial model. The final model was run with four chains of 20000 iterations each, 19000 of which were discarded as model burn-in. The remaining 4000 samples of parameter estimates provided empirical posterior distributions for each parameter and pattern feature from which means and 95% credible intervals were derived to compare with the known simulated values.

2.3 Comparison with simulated values

1000 simulations were generated, and the temporal patterns of edge properties analysed using the multilevel model specified above. The first derivative of the model for each individual edge was calculated to extract the maximum journey duration, the time of this maximum, and the range of the function (the difference between the maximum and minimum) during the 24-hour period (Gadd et al. 2021). The range of the function corresponds to the maximum delay that passengers experience travelling between a particular pair of stations. Extracting the pattern features was completed using mean model coefficients for each origin-destination pair to extract a point estimate. It was then repeated with coefficients from each of 4000 posterior samples. This generated posterior distributions of the maximum journey duration, delay and time (with one value for each posterior sample) from which quantiles could be taken to represent 95% credible intervals.

For each origin-destination pair, the maximum journey duration, maximum time and maximum delay were recorded, along with the width of the 95% credible interval. The simulated and modelled maximum journey duration, maximum time and maximum delay were compared. Whether the credible intervals for these pattern features included the pattern features from the simulated temporal patterns was recorded. The simulated maximum journey duration, time and delay was also converted into a z-score based on the mean modelled value and 95% credible interval for each feature. If the posterior distributions from which the 95% credible intervals are calculated are normally distributed, and the 95% credible intervals are of an appropriate width then we would expect the distribution of these z-scores to follow a standard normal distribution.

This work was undertaken on ARC3, part of the High-Performance Computing facilities at the University of Leeds, UK. Code for simulation and analysis of data is included as Supplementary Material.

3 Results

Across 1000 simulations, 98.5% of models ran successfully with only 15 failing to complete estimation. Running one simulation with each chain of the multilevel model estimated in parallel took approximately two days of computing time. The full simulation was run with other processes running in parallel and took approximately 14 days.

Table 1 compares simulated and modelled values for each of the three pattern features: maximum journey duration, maximum time and maximum delay. Maximum journey duration is the maximum time taken to complete a journey between a pair of stations during the day examined. Maximum time is the time of day at which this maximum journey duration occurred. Maximum delay is the difference between maximum journey duration and the minimum journey duration between a pair of stations during the day examined.

Feature	**Maximum time (hh:mm:ss, hours)**	**Maximum journey duration (minutes)**	**Maximum delay (minutes)**
Simulated mean	15:50:49	37.109	18.335
Modelled mean	15:50:02	35.788	17.617
Simulated SD	67.020	8.454	11.253
Modelled SD	0.844	6.362	9.530
Mean 95%CI width	2.095	3.978	5.321
SD 95% CI width	1.398	0.422	0.463
Median 95%CI width	1.900	3.950	5.288
IQR 95% CI width	(1.120,2.570)	(3.678,4.246)	(4.996,5.607)
Mean Bias	-0.013	-1.320	-0.719
95% Limits of Agreement	(-2.471,2.445)	(-9.58,6.939)	(-5.832,4.395)
Mean simulated value as z-score (no units)	0.023	1.333	0.532
SD simulated value as Z-score (no units)	3.598	4.137	1.868
Percentage of real values within CI (%)	59.533	36.174	68.613

Table 1: Table summarising and comparing simulated and estimated pattern feature values. SD = standard deviation, CI = credible interval, hh:mm:ss = hours, minutes, seconds time format. Mean bias is the mean difference between estimated and simulated values. 95% Limits of agreement represent limits of the bias we would expect to find in 95% of estimates of each feature.

Figure 4 shows a Bland-Altman plot (Bland and Altman 1999) for the agreement between simulated and modelled maximum time, including the mean bias and 95% limits of agreement. Figure 5 shows a density plot of the 95% credible interval widths for estimates of maximum time. Figure 6 shows a density plot, with mean and standard deviation indicated, of simulated maximum time features as a z-score of the normal distribution described by the estimated maximum time and 95% credible interval estimates. Note that this assumes that the posterior distribution of the pattern feature estimates is normally distributed.

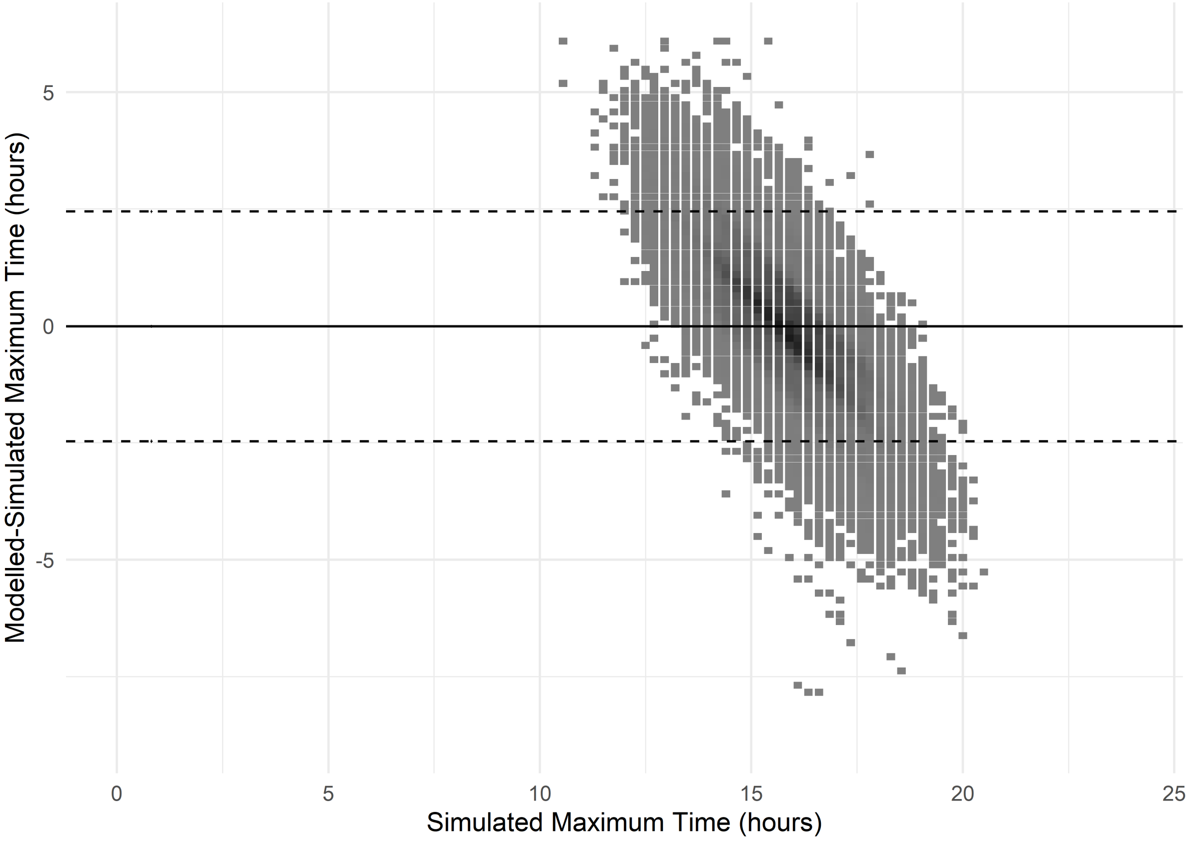

Figure 4: Bland-Altman plot showing the relationship between simulated maximum time and modelled minus simulated maximum time. The density of points is indicated by the colour of squares in the plot (darker grey equals greater density). The solid line indicates the mean bias and dashed lines indicated 95% limits of agreement.

In Figure 4, the simulated maximum times are mostly in the afternoon, with the majority of values falling between approximately 12:00 and 20:00. This reflects the simulated data we saw in Figure 3 and the average pattern in Figure 2, where the maxima take place around the afternoon commuting period. The mean bias is close to zero, suggesting that estimates of maximum time are, on average, unbiased. However, there is a negative correlation between simulated maximum time and modelled minus simulated maximum time. This suggests that maxima taking place early in the time period are being estimated as later than they are and vice versa. This may indicate that the model is failing to capture the full amount of variation in maximum times between edges in the network.

The 95% limits of agreement indicate the limits of the bias we would expect in 95% of estimates of maximum time from these models. These are wide, with a range of just less than 2.3 hours either side

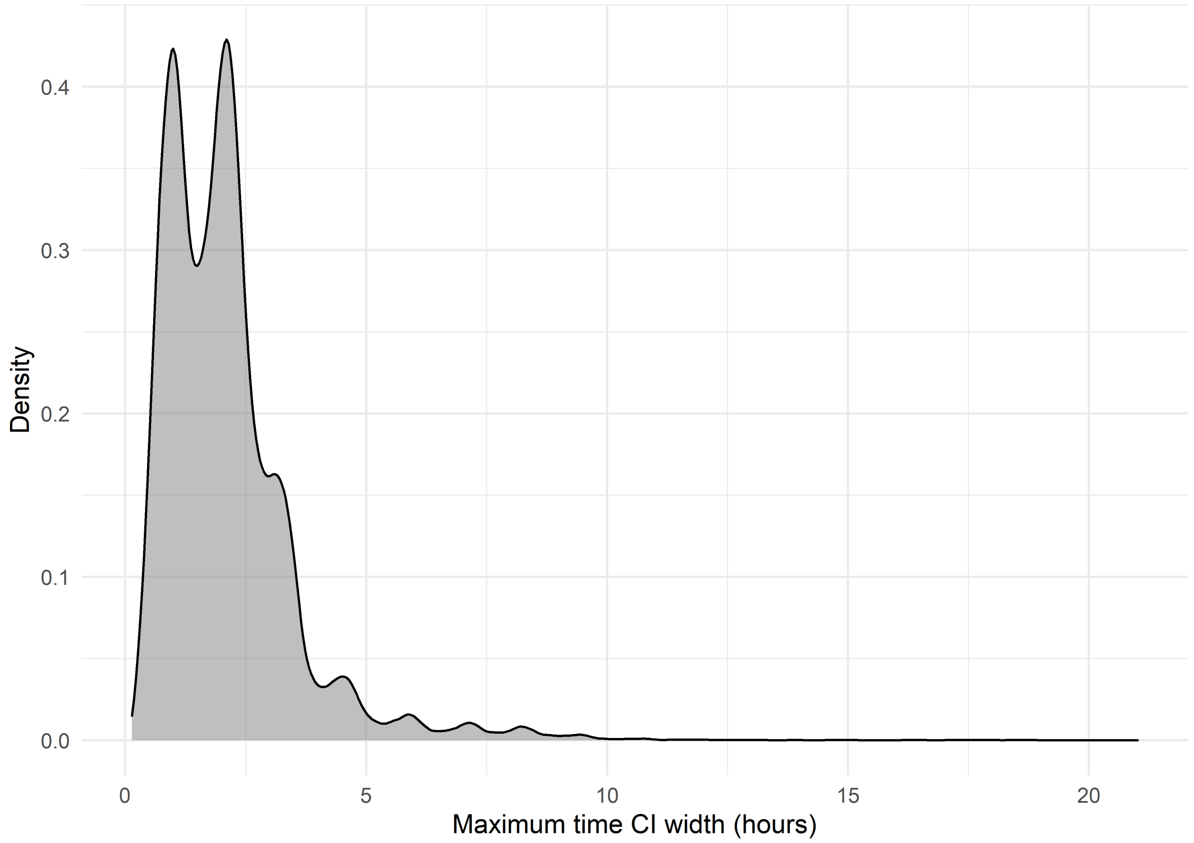

Figure 5: Density plot showing the distribution of 95% credible interval (CI) widths for maximum time estimates.

of zero. This indicates that, while the mean bias is small, there is the potential for individual estimates of maximum time to have considerable inaccuracy.

The 95% credible interval widths of maximum time are also wide: the most common width is slightly less than 2.5 hours with a large proportion of intervals wider than this. This indicates that the estimates of maximum time from the model are also imprecise. The density plot for maximum time credible interval widths does not have a smooth shape but rather a series of peaks of common credible interval widths. The distance between these peaks appears to approximately correspond to the distance between knots in the b-spline basis (1.2 hours, see Supplementary Figure 13).

If the 95% credible intervals had appropriate coverage, we would expect the distribution of simulated maximum times as z-scores of their estimated credible intervals to follow a standard normal distribution with mean zero and standard deviation one. The mean of the distribution is close to zero, as shown in Table 1, but the standard deviation, indicated by the dashed lines in Figure 6, is over 3.5. This suggests that the 95% credible intervals are too narrow and do not cover the appropriate amount of simulated maximum times. Table 1 shows that the percentage of simulated maximum times contained within the estimated credible intervals is just below 60%.

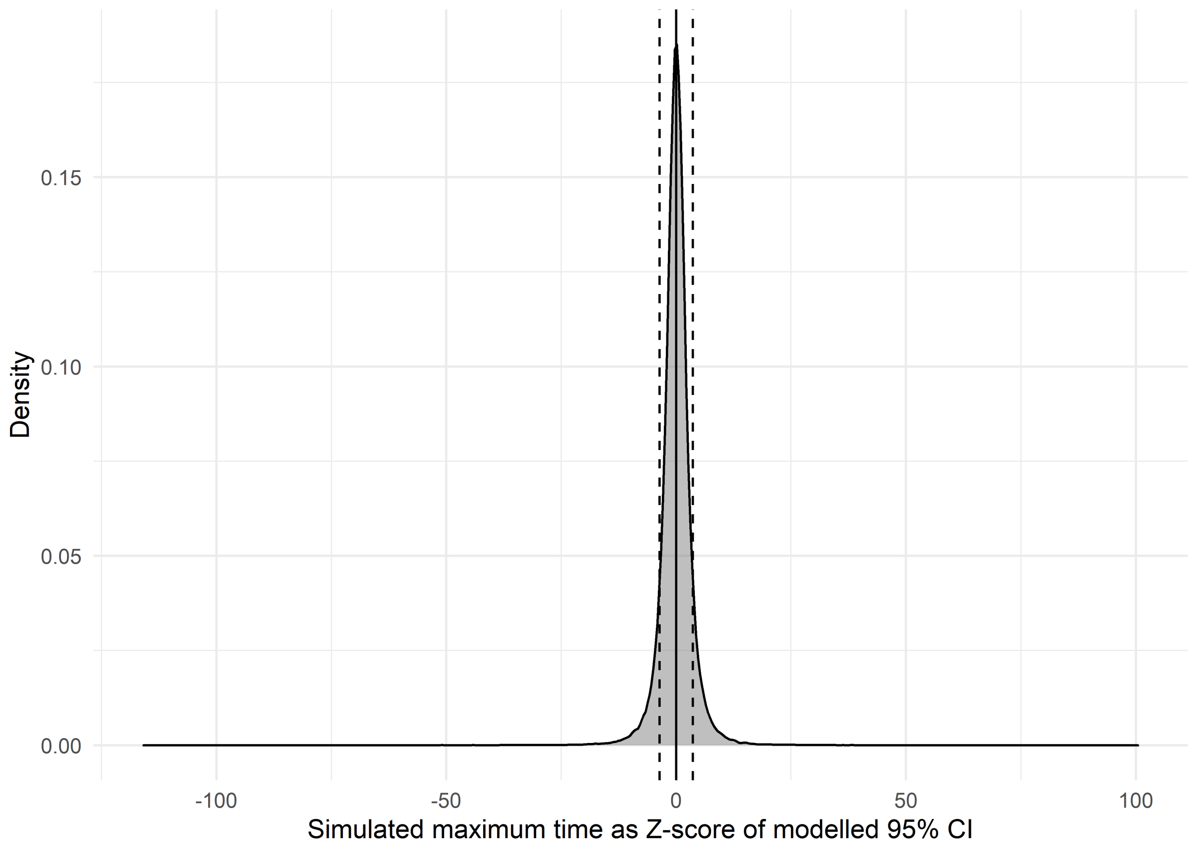

Figure 6: Density plot showing simulated maximum time values as a z-score of the relevant posterior distributions of estimates for maximum time. The solid line indicates the mean z-score and the dashed lines one standard deviation either side of the mean.

Figure 7 shows the Bland-Altman plot with limits of agreement for simulated and modelled maximum journey duration. Figure 8 shows the distribution of credible interval widths for this feature and Figure 9 shows the distribution of simulated maximum journey durations as a z-score of the credible interval distributions.

The distribution of simulated maximum journey duration is multimodal as it contains lower maximum journey durations for journeys originating at stations with few connections and higher values for journeys originating at stations with many connections. This matches the simulated data. In Figure 7 we see a negative correlation between the simulated maximum journey duration and the degree of over- or under-estimation by the models, like in Figure 4. In this case, however, there is a small negative bias in the estimation of maximum journey duration. On average, estimates from the models are 1.32 minutes below the simulated maximum journey duration. The 95% limits of agreement are relatively narrow indicating that 95% of estimated maximum journey durations are biased by less than 10 minutes in either direction from the true maximum journey duration.

95% credible interval widths, shown in Figure 8, are also narrower for the maximum journey duration compared to maximum time with almost all widths being under six minutes. However, the results

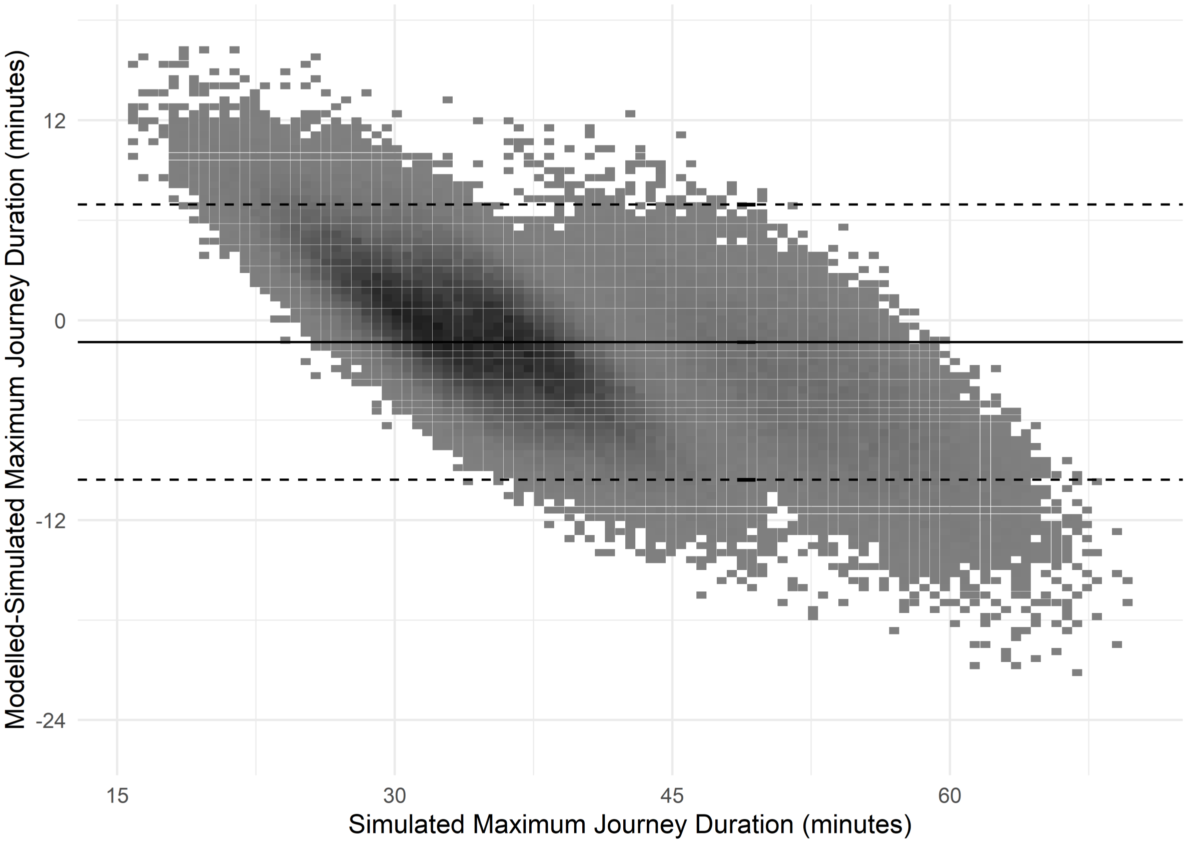

Figure 7: Bland-Altman plot showing the relationship between simulated maximum journey duration and modelled minus simulated maximum journey duration. The density of points is indicated by the colour of squares in the plot (darker grey equals greater density). The solid line indicates the mean bias and dashed lines indicated 95% limits of agreement.

in Figure 9 indicate that these may be somewhat too narrow. The standard deviation of simulated maximum journey durations as a z-score of the distributions defined by their estimated credible interval widths is over four and the percentage of simulated values covered by the credible interval is only 36%. This indicates severe undercoverage of these credible intervals.

Figure 10 shows the Bland-Altman plot for simulated and modelled maximum delay with the distribution of credible interval widths shown in Figure 11. Figure 12 shows a density plot of the simulated maximum delay values as z-scores of the distributions described by their corresponding modelled maximum delay and 95% credible interval estimates.

There are four distinct groups of simulated maximum delays corresponding to the four different numbers of connections for stations in Figure 1. Over all these groups, a negative correlation between simulated maximum delay and modelled minus simulated maximum delay is present, similar to the other pattern features observed. However, within some individual groups this correlation does not appear so strong. The mean bias is small and negative (-0.719 minutes, less than 45 seconds) and the 95% limits of agreement are within six minutes either side of zero.

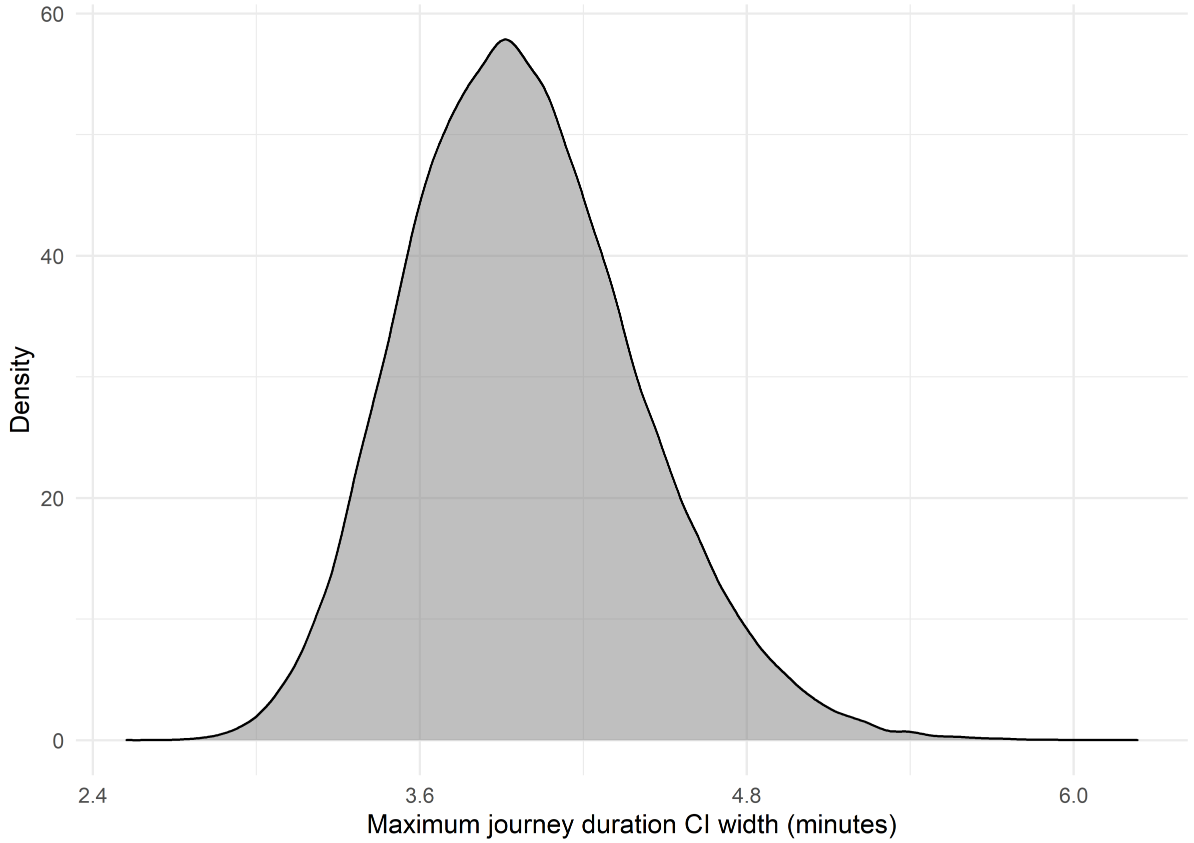

Figure 8: Density plot showing the distribution of 95% credible interval (CI) widths for maximum journey duration estimates.

The majority of estimated 95% credible interval width for maximum delay were less than 7.5 minutes wide, however there is still an undercoverage problem with these intervals: 68% of simulated maximum delays were covered by the estimated credible intervals and the standard deviation of simulated maximum delay as a z-score of the 95% credible interval widths is 1.868. The extent of this problem is less, however, than for the other two pattern features which had wider z-score distributions and lower coverage percentages.

4 Discussion

The mean bias for all pattern features was relatively low, with maximum time, maximum journey duration and maximum delay all being slightly underestimated by less than two minutes on average. However, for all features, a negative correlation of the simulated pattern feature and the modelled minus simulated pattern feature was observed. This suggests that lower values are being overestimated and higher values underestimated. This may be an indication that the models are not capturing the full extent of variation in pattern features between observational units. This may partially be due to a lack of flexibility in the model which prevents it from capturing more of the range of variability. Another cause may be shrinkage:

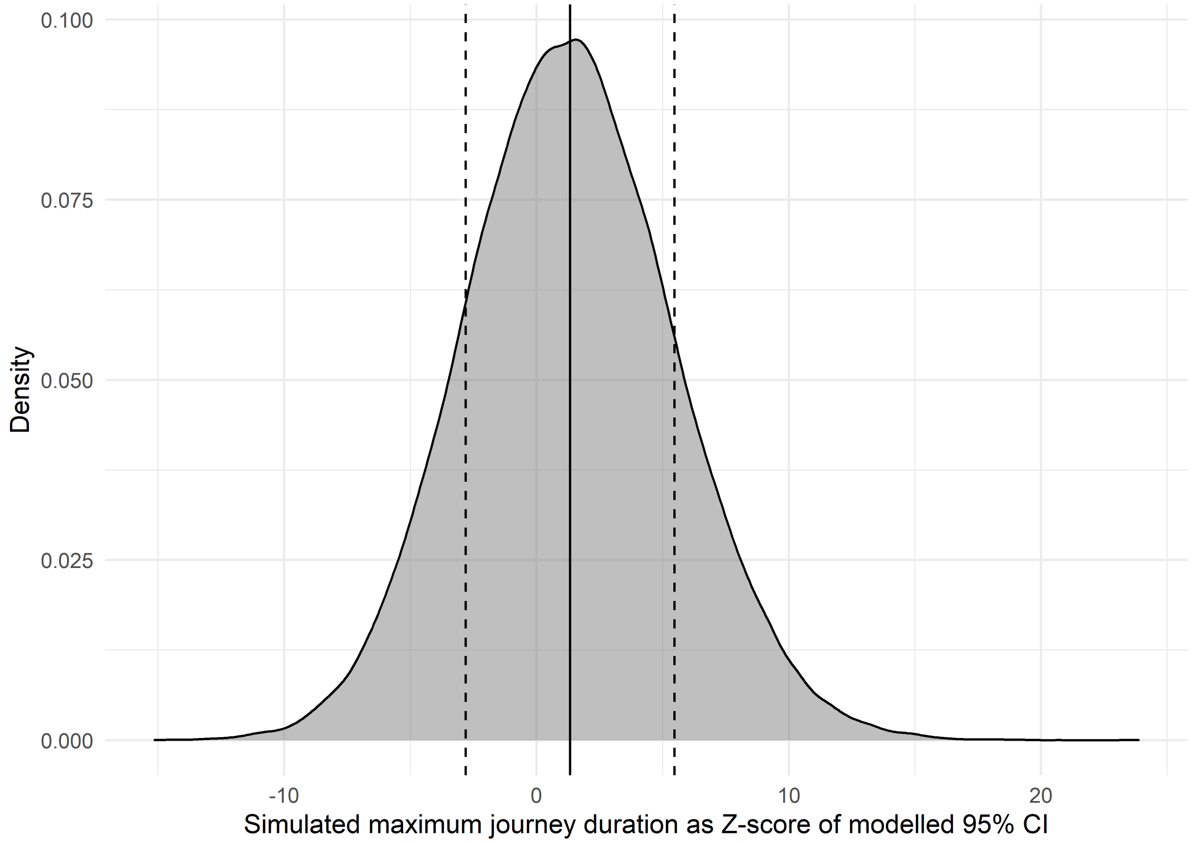

Figure 9: Density plot showing simulated maximum journey duration values as a z-score of the relevant posterior distributions of estimates for maximum journey duration. The solid line indicates the mean z-score and the dashed lines one standard deviation either side of the mean.

this is an effect in which any model will, on average, predict less variation than is present in the observed data (Copas 1983; Greenland 2000). This prevents the model from capturing the *full* range of variation in pattern features in the data, but it is possible that some changes to the model specification may help to capture *more* of the variation.

The model included no x-axis random variation and relied upon the flexibility of the underlying b-spline basis to capture variation between observational units in the x-axis. There are forms of multilevel models that include an x-axis random effect. For example, the Super Imposition by Translation And Rotation (SITAR) model is a form of multilevel model explicitly designed to investigate childhood growth curves (Cole et al. 2010). It includes three random effects: a y-axis random intercept changing the mean size of a child across the time measured, an x-axis random intercept representing changes in the timing of when a growth spurt occurs, and a random effect for the velocity of growth. Incorporating an x-axis effect to this model may help it to capture more of the variation in maximum time values by allowing horizontal translations of the modelled patterns between observations units – this reflects the simulation process underlying the data.

Including random effects in the x-axis may present some difficulties in fitting the model due to increased

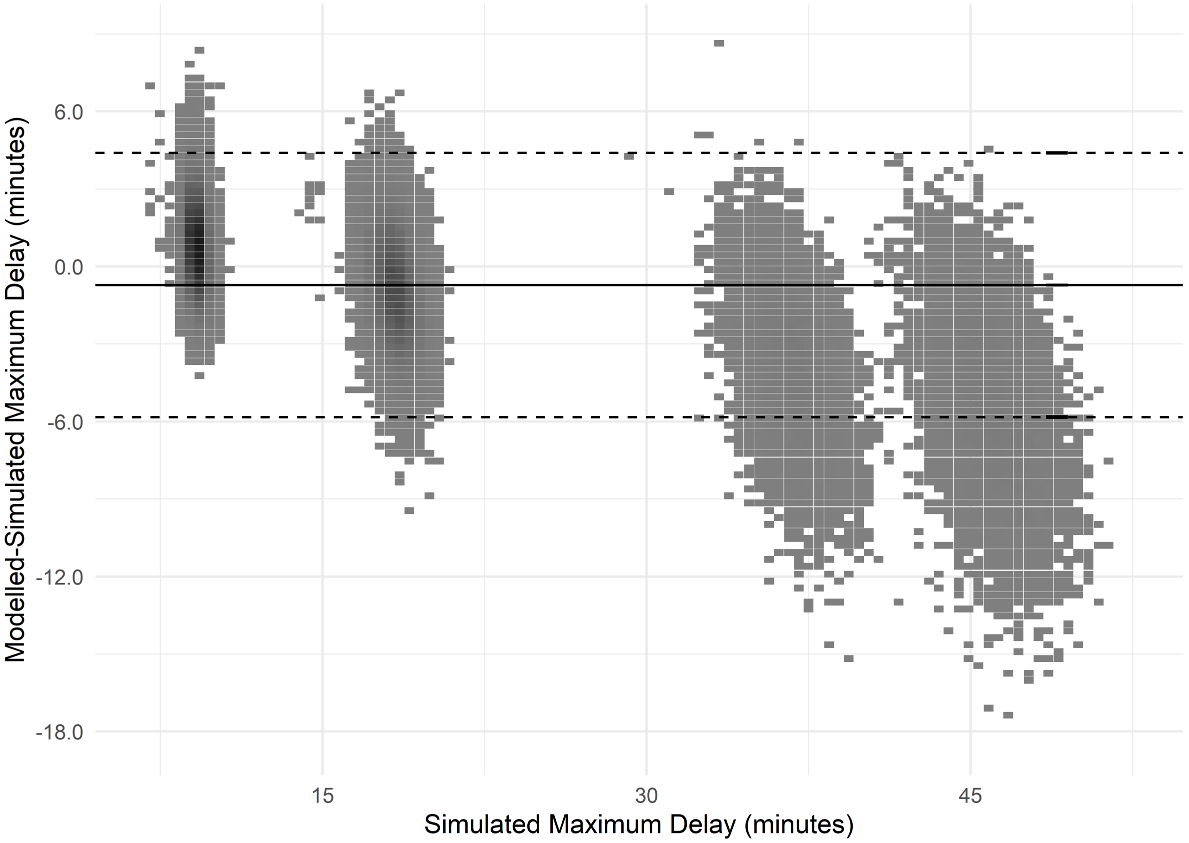

Figure 10: Bland-Altman plot showing the relationship between simulated maximum delay and modelled minus simulated maximum delay. The density of points is indicated by the colour of squares in the plot (darker grey equals greater density). The solid line indicates the mean bias and dashed lines indicated 95% limits of agreement.

complexity. Currently, the model code treats the b-spline basis as a series of fixed variables, the value of which at each observation time is calculated before estimating the model. This is convenient and represents an advantage of using b-splines over penalised alternatives. However, if an x-axis random intercept were included, the b-spline basis would be shifted horizontally for each edge in the network by an amount estimated by the model. This means that the value of each b-spline basis function at each observation time would depend on a quantity estimated in the model (the x-axis random effect) and could not be calculated before fitting. Incorporating such a random effect would likely be possible in OpenBUGS but would be complicated – this would probably increase computational intensity of the models and make them more difficult to construct.

However, it may be useful to consider alternative model specifications, particularly for the specification of random effects, that may allow an x-axis random effect to be included without increased complexity. In this simulation, spatial and aspatial random variation between observational units was incorporated into the coefficients for each b-spline function in the model. This aimed to allow variation in the form of the temporal patterns of journey durations for each observational unit while allowing observations closer

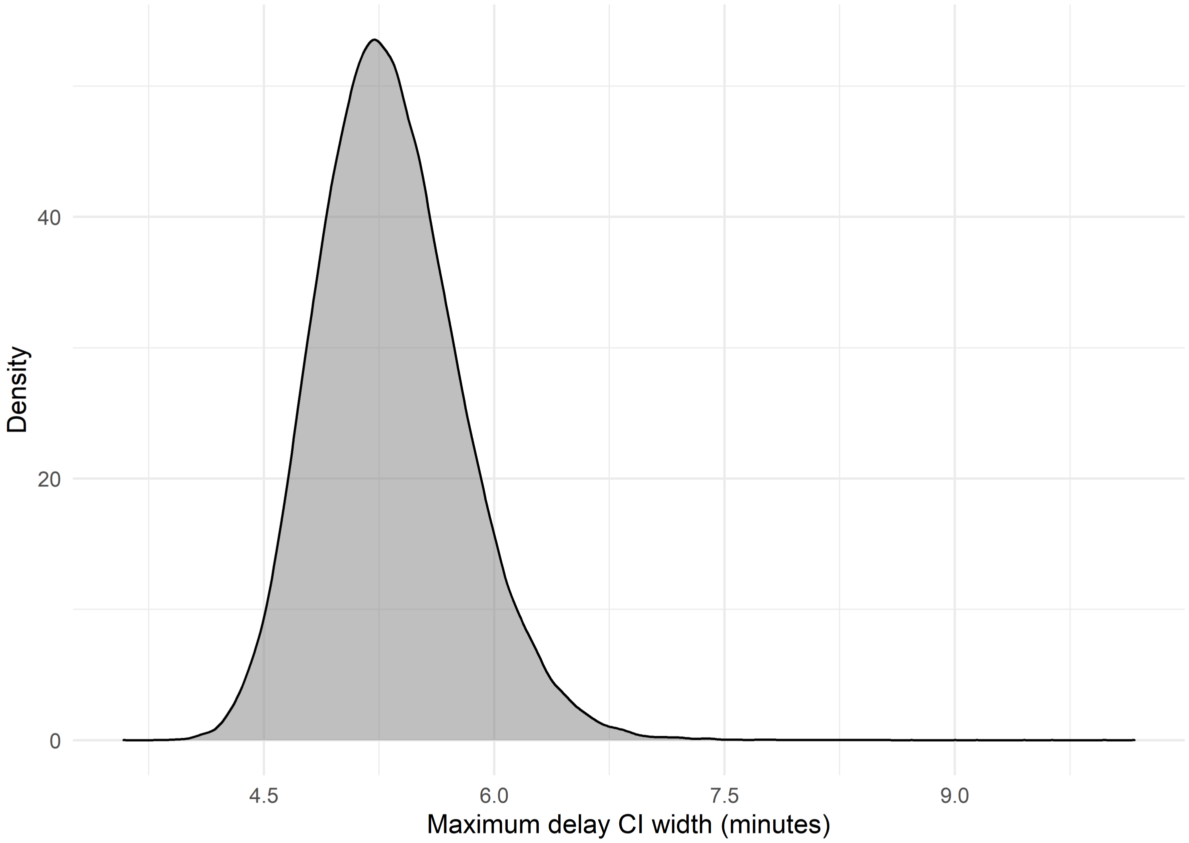

Figure 11: Density plot showing the distribution of 95% credible interval (CI) widths for maximum delay estimates.

together to be more similar. This requires estimation of a large number of random effects coefficients. The coefficients for the b-splines, however, do not have a clear real-world interpretation (Stimson et al. 1978) so perhaps the incorporation of random effects in these coefficients does not correspond completely to random variation in the temporal patterns, particularly spatial random variation, in the way we would expect. It may be more useful and interpretable to fix the b-spline coefficients and include a series of random effects to allow transformations of the average temporal pattern for each observational unit, for example including a random y-axis intercept and y-axis scaling factor, similar to the SITAR model (Cole et al. 2010). This would reduce the number of random effects that needed to be estimated by the model, allowing for additional x-axis random effects to be included without increasing model complexity, and also mean that the interpretation of these random effects was clearer – for example, similar coefficients for two observational units would clearly correspond to similarity in the temporal patterns of journey delays.

This approach, however, may have some shortcomings. The multilevel model in this paper assumes that the random coefficients for each b-spline are normally distributed (Besag et al. 1991; Goldstein 2011). As the simulation has incorporated some latent or underlying groups in the data with different journey delays, the distributions of simulated maximum delay and maximum journey duration are not normally

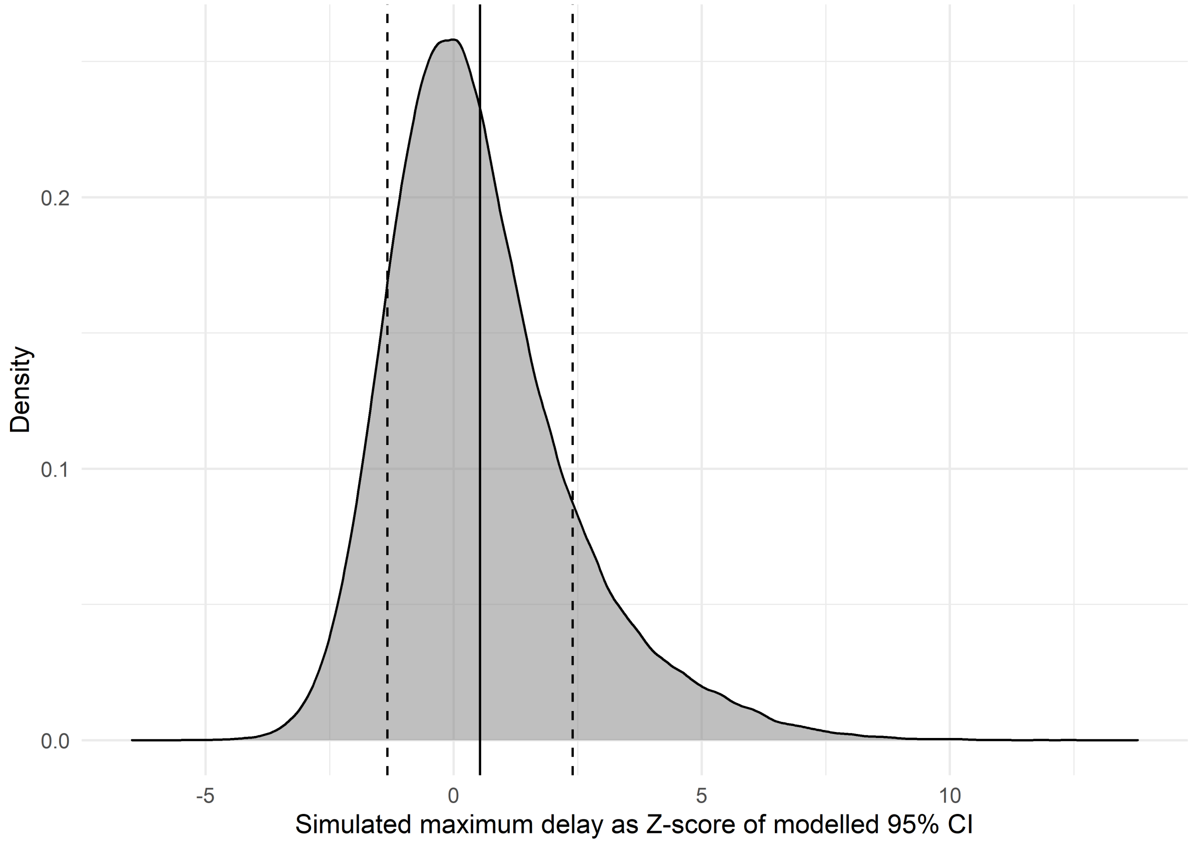

Figure 12: Density plot showing simulated maximum delay values as a z-score of the relevant posterior distributions of estimates for maximum delay. The solid line indicates the mean z-score and the dashed lines one standard deviation either side of the mean.

distributed. As the model used in this example does not directly estimate the random variation in these parameters, but rather random variation in the 'abstract' b-spline coefficients, this does not necessarily violate this assumption (though it may). However, if random effects more directly related to these pattern features were included this may more clearly violate these assumptions and cause the model to perform poorly (Besag et al. 1991; Goldstein 2011). It may be possible to account for this, for example by using models with latent (unmeasured) variables included (Kaplan et al. 2009). It is important for researchers to examine data carefully to see if there are any structures like this. This allows the researcher to attempt to account for this where necessary.

The mean width of the credible intervals estimated for maximum time were particularly large in relation to the 24-hour period covered by the data. The majority of credible intervals in Figure 5 are also fairly wide, which means that the estimates from these models often do not give a particularly precise indication of maximum time. This may not be adequate if the models were to be used in a real situation, for example, to choose a time to implement measures to reduce congestion in stations, or times to include more trains in a timetable. The distribution of credible interval widths in Figure 5, unlike those for maximum journey duration and maximum delay, has a series of local maxima at fairly regular intervals indicating that there are some credible interval widths that are more common than others. The distance between these local maxima appears to be similar to the distance between knot points in the b-spline basis specified for this model (1.2 hours, see Supplementary Figure 13). This may indicate that the structure of the temporal patterns estimated by the models strongly reflects the b-spline knot placements and that the shape of these patterns is influencing the width of posterior distributions that generate credible interval estimates. It would be useful for future work to investigate if this is the case and if alternative model specifications might change this pattern.

In this particular simulation, the simulated pattern has a clear maximum in the 24-hour period of interest that is far greater than the other local maxima included. However, it is possible that some temporal patterns may include multiple maxima with a smaller difference in their values, for example, patterns that include multiple days and have a periodic nature. It may be that, for patterns with multiple local maxima, the posterior distribution of maximum time estimates would contain more than one of these maxima, rather than the maximum with the largest value during the 24-hour period. This may produce a multimodal posterior distribution of maximum times which will result in a much wider credible interval than if the distribution is concentrated around a single maximum in the 24-hour period. While the usual interpretation of a credible interval based on a multimodal posterior distribution of maximum time would remain accurate, a 95% credible interval does not give such a precise description of the posterior distribution in this case and it may be assumed by readers that it is unimodal. It may be useful for future work to investigate if a multimodal posterior distribution for the maximum time pattern feature might be produced in certain situations and whether there is a more detailed way to represent uncertainty when this is the case.

For all the pattern features, even maximum time which had wide credible intervals, the coverage of credible intervals is far lower than the expected 95%. Coverage is lowest for maximum journey duration and highest for maximum delay. This indicates that the estimates of uncertainty for the pattern features are not reliable. This could suggest that the estimated credible intervals are too narrow, or the estimates of maximum journey duration are biased (or a combination of these). There is no clear indication here that the estimated credible intervals are too narrow – in fact, relative to the amount of variation in the real maximum journey duration, the credible intervals are relatively wide – but we cannot rule this out as a cause of the undercoverage of these estimates. Despite the mean bias estimates being low, there is clearly some bias in the estimation of low and high values of each pattern feature. This may have contributed to the size of the 95% limits of agreement for these features which are larger than half the mean credible interval width for all three pattern features. The size of the 95% limits of agreement in relation to the credible interval widths means that we would expect more than 5% of the credible intervals not to cover the simulated maximum journey duration, which is indeed the case. Where the credible interval widths are smaller compared to the 95% limits of agreement, we also see a decreased coverage. This may indicate that bias does contribute to the undercoverage of credible intervals. Further work could be carried out to investigate the contributions of different factors to the undercoverage seen in this simulation study. This would indicate how best to approach fixing the problem.

5 Conclusion

This simulation tested the ability of continuous-time multilevel models to capture maxima in temporal patterns of edge properties for a spatio-temporal network. Only one scenario with one data structure was investigated – this particular scenario was chosen to reflect a real data example to which this model has previously been applied (Gadd et al. 2021). Future work testing these models in the presence of different data structures (for example, the timing of measurements (evenly spaced, balanced/unbalanced, measured in waves), different degrees of temporal and spatial autocorrelation) should be carried out as the impact of different scenarios on the model performance should be assessed.

The models were able to capture the maximum time and maximum journey duration with a small *mean* bias; they tended to overestimate low simulated values and underestimate high simulated values for both features, leading to biased estimates for individual edges in the network with low or high maximum times/journey durations. The extent to which variation is not fully captured may have contributed to the relatively small proportion of credible intervals for these pattern features that contain the simulated pattern feature values. In practice, this means the credible intervals from this model in its current form cannot be interpreted as the range of most likely values for the pattern features. Future work concerning continuous-time multilevel models for temporal patterns of edge properties in spatio-temporal networks should further evaluate the application of continuous time models, especially considering incorporation of an x-axis random intercept and alternative prior and initial value specifications, which could potentially improve accuracy and precision.

References

Abdelghany, A., Mahmassani, H., & Chiu, Y. (2001). Spatial microassignment of travel demand with activity trip chains. *Transportation Research Record*, *1777*(1), 36–46.

Anderson, T., & Dragićević, S. (2018). A geographic network automata approach for modeling dynamic ecological systems. *Geographical Analysis*, *52*, 3–27.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*(1), 1–20.

Bland, J., & Altman, D. (1999). Measuring agreement in method comparison studies. *Stat Methods Med Res*, *8*(2), 135–60.

Blonder, B., Wey, T., Dornhaus, A., James, R., & Sih, A. (2012). Temporal dynamics and network analysis. *Methods in Ecology and Evolution*, *3*(6), 958–972.

Chen, S., Claramunt, C., & Ray, C. (2014). A spatio-temporal modelling approach for the study of the connectivity and accessibility of the Guangzhou metropolitan network. *Journal of Transport Geography*, *36*, 12–23.

Cheng, T., Wang, J., Haworth, J., Heydecker, B., & Chow, A. (2011). Modelling dynamic space-time autocorrelations of urban transport network. *Proceedings of the 11th international conference on Geocomputation 2011*, 215–210.

Cheng, T., Wang, J., Haworth, J., Heydecker, B., & Chow, A. (2014). A dynamic spatial weight matrix and localized space–time autoregressive integrated moving average for network modeling. *J Geographical Analysis*, *46*(1), 75–97.

Colak, S., Schneider, C., Wang, P., & Gonzalez, M. (2013). On the role of spatial dynamics and topology on network flows. *New Journal of Physics*, *15*, 113037.

Cole, T., Donaldson, M., & Ben-Shlomo, Y. (2010). SITAR - A useful instrument for growth curve analysis. *International Journal of Epidemiology*, *39*(6), 1558–1566.

Comber, A., & Wulder, M. (2019). Considering spatiotemporal processes in big data analysis: Insights from remote sensing of land cover and land use. *Transactions in GIS*, *23*(5), 879–891.

Copas, J. (1983). Regression, prediction and shrinkage. *45*(3), 311–335.

Dubin, R. (1998). Spatial autocorrelation: A primer. *Journal of Housing Economics*, *7*(4), 304–327.

Freeman, J. (1989). Systematic sampling, temporal aggregation, and the study of political relationships. *Political Analysis*, *1*, 61–98.

Gadd, S., Comber, A., Gilthorpe, M., Suchak, K., & Heppenstall, A. (2021). Simplifying the interpretation of continuous-time models for spatio-temporal networks. *Journal of Geographical Systems*.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741.

Goldstein, H. (2011). *Multilevel statistical models* (Fourth edition). Wiley.

Goldstein, H., Healy, M., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *J Statistics in medicine*, *13*(16), 1643–1655.

Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology*, *29*(1), 158–167.

Head, M., Holman, L., Lanfear, R., Kahn, A., & Jennions, M. (2015). The extent and consequences of P-hacking in science. *PLOS Biology*, *13*(3), e1002106.

Heck, R., & Thomas, S. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus* (Third edition). Taylor & Francis.

Holmes, C., & Mallick, B. (2003). Generalized nonlinear modeling with multivariate free-knot regression splines. *Journal of the American Statistical Association*, *98*(462), 352–368.

Howe, L., Tilling, K., Matijasevich, A., Petherick, E., Santos, A., Fairley, L., Wright, J., Santos, I., Barros, A., Martin, R., Kramer, M., Bogdanovich, N., Matush, L., Barros, H., & Lawlor, D. (2013). Linear spline multilevel models for summarising childhood growth trajectories: A guide to their application using examples from five birth cohorts. *Statistical Methods in Medical Research*, *25*(5), 1854–1874.

Hwang, S. (2000). The effects of systematic sampling and temporal aggregation on discrete time long memory processes and their finite sample properties. *Econometric Theory*, *16*(3), 347–372.

Kaplan, D., Kim, J.-S., & Kim, S.-Y. (2009). Multilevel latent variable modeling: Current research and recent developments. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology*. SAGE Publications Ltd.

Li, T., & Liao, Q. (2016). Dynamic networks analysis and visualization through spatiotemporal link segmentation. *IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 209–214.

Li, X., & Griffin, W. (2013). Using ESDA with social weights to analyze spatial and social patterns of preschool children's behavior. *Applied Geography*, *43*, 67–80.

Liu, S., Wang, L., Liu, C., & Destech Publicat, I. (2018). Visualized social network analysis on spatial dynamics of international trade between China and League of Arab States. *2018 3rd international conference on computational modeling, simulation and applied mathematics* (pp. 270–277).

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine*, *28*, 3049–3067.

Morris, T., White, I., & Crowther, M. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102.

Muthén, L., & Muthén, B. (2010). *Mplus user's guide* (Sixth edition).

Neeson, T., Wiley, M., Adlerstein, S., & Riolo, R. (2012). How river network structure and habitat availability shape the spatial dynamics of larval sea lampreys. *Ecological Modelling*, *226*, 62–70.

Newman, M. (2003). The structure and function of complex networks. *SIAM Review*, *45*(2), 167–256.

Newman, M. (2018). Fundamentals of network theory. *Networks* (Second edition). Oxford University Press.

Niezink, N., Snijders, T., & van Duijn, M. (2019). No longer discrete: Modeling the dynamics of social networks and continuous behavior. *Sociological Methodology*, *49*(1), 295–340.

Oud, J., Folmer, H., Patuelli, R., & Nijkamp, P. (2012). Continuous-time modeling with spatial dependence. *Geographical Analysis*, *44*(1), 29–46.

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in R. *BMC Medical Research Methodology*, *19*(1), 46.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *3rd International Workshop on Distributed Statistical Computing (DSC 2003); Vienna, Austria*, *124*.

R Core Team. (2020). R: A language and environment for statistical computing.

Ramsay, J., & Silverman, B. (1997). *Functional data analysis*. Springer.

Stan Development Team. (2019). RStan: The R interface to stan. R package version 2.19.2.

Stimson, J., Carmines, E., & Zeller, R. (1978). Interpreting polynomial regression. *Sociological Methods & Research*, *6*(4), 515–524.

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, *12*(3), 1–16.

Thomas, A., Best, N., Lunn, D., Arnold, R., & Spiegelhalter, D. (2014). GeoBUGS user manual.

Transport for London. (2009a). London Underground map December 2009.

Transport for London. (2009b). London Underground map September 2009.

Weiss, A. (1984). Systematic sampling and temporal aggregation in time series models. *Journal of Econometrics*, *26*(3), 271–281.

Yang, Y., Heppenstall, A., Turner, A., & Comber, A. (2019). A spatiotemporal and graph-based analysis of dockless bike sharing patterns to understand urban flows over the last mile. *Computers, Environment and Urban Systems*, *77*, 101361.

Yeh, R., Nashed, Y., Peterka, T., & Tricoche, X. (2020). Fast automatic knot placement method for accurate b-spline curve fitting. *Computer-Aided Design*, *128*, 102905.

Yuan, Y., Chen, N., & Zhou, S. (2013). Adaptive b-spline knot selection using multi-resolution basis set. *IIE Transactions*, *45*(12), 1263–1277.

6 Supplementary material

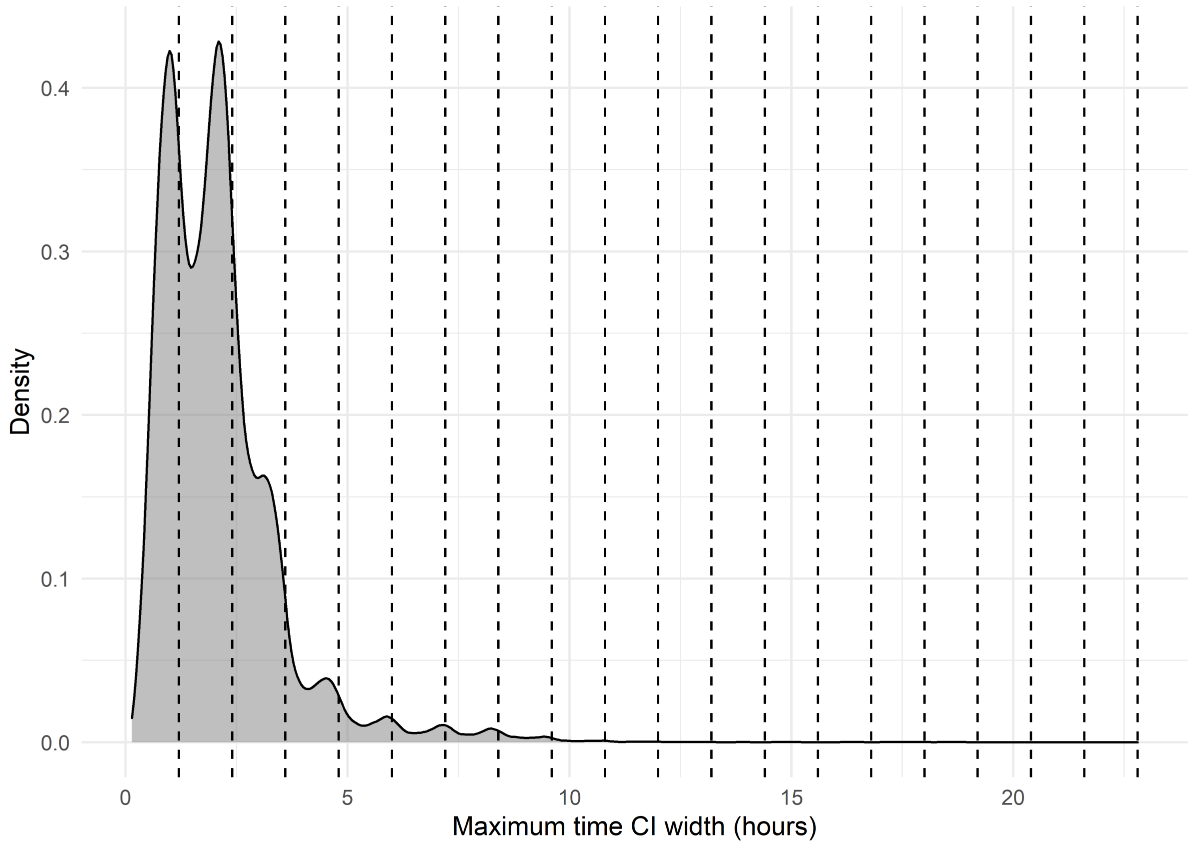

Figure 13: Density plot showing 95% credible interval widths for estimates of maximum time, with the distances between knot points marked as vertical dashed lines.