Forecasting Network faults with Bayesian Spatio-temporal Statistical Models

Thomas Adam Statham

200936388

A Master's dissertation submitted to the faculty of Science and Engineering at the University of Liverpool, in partial fulfilment of the requirements for the degree of Geographic Data Science (MSc) in the department of Geography and Planning

Abstract

The focus of this study was to explore different methods of one step ahead forecasting of Virgin Media faults. Accurate forecasts of network faults is important for determining the number of engineers and truck call outs necessary to fix any service disruptions. In addition to the cost of broadband services, reliable services are also important for influencing consumer demand and customer churn. Moreover, understanding the number and spatial distribution of faults is necessary for determining the number of call centre and engineer employees for identifying and fixing network faults. This is important from an operational efficiency and network maintenance costs.

We extend the conventional time-series approaches in the telecommunication literature by incorporating space for one step ahead monthly forecasts of network faults. We applied four different model specifications: a stationary time-series process, a non-stationary time-series process, a spatio-temporal (ST) model and a spatial-temporal interaction (STI) model. To ensure customer confidentiality, we aggregated all network faults at the postcode level to the Middle Layer Super Output Area (MSOA) and space was incorporated using a Besag-York-Mollié (BYM) prior. We applied these models using a Bayesian Hierarchical model through the numerically efficient R-INLA package. We also looked at potential sociodemographic factors and their impacts on network faults. The study area was North West England, for areas with Virgin Media coverage and the models was applied using past network faults to forecast January to March 2018.

The results support the value of further incorporating space into conventional time-series approaches. Although the model fit of the most complex STI model was the best, the ST model had the highest average forecast accuracy or difference between observed and forecasted network faults. The ST model took significantly less time to estimate faults than the STI model, which took a similar amount of time as the non-stationary time series prior. We used the stationary temporal prior for ST and STI models because it had a significantly higher forecast accuracy. Moreover, incorporating space allows the forecaster to identify how the spatial distribution of faults changes over time at a much finer spatial scale than the common regional level of analysis. For example, there was a higher probability of forecasted faults exceeding 15 faults in North Liverpool, Saint Helens and Wigan.

This study demonstrates that it’s possible to obtain fast and accurate forecasts of network faults at this scale using spatio-temporal models. Moreover, this methodology could be extended to other applications, including forecasting broadband demand in areas which do not have coverage. There is scope for future research to examine different model specifications of space-time with and without a space-time interaction term and to apply this methodology over a larger study area. There is also a need to apply this methodology over a longer period, given the lack of literature in this field.

Acknowledgement

I would like to acknowledge and thank the following important people who have supported me during this project. In no particular order, I would like to express my gratitude to the Consumer Data Research Centre for providing the framework for linking leading UK retail companies with academia, which has for several years provided an excellent Master's Research Dissertation Programme. This leads me to thank Virgin Media for giving me access to secure data during a time where customer confidentiality is a sensitive topic. Specifically, I would like to take the time to recognise Andy Simpson and Robert Ogden from Virgin Media, for their regular support throughout this process. I am grateful to my Supervisor Gavin Dong for his technical guidance and his support during the course of this project and throughout my Master's degree.

Author's declaration

I declare that the ethics of this research was approved through the University of Liverpool Research Integrity and Ethics Committee, which complies with the new General Data Protection Regulation (GDPR) that came into force on 25th May 2018.

Reference: 3989

I declare that this dissertation was composed by myself, that the work contained herein is my own except where explicitly stated otherwise and that this work has not been submitted for any other degree or professional qualification.

T. Statham

I further authorize the University of Liverpool to reproduce this dissertation in total or in part, at the request of other institutions or individuals for scholarly research.

Contents

List of tables

List of figures

1 Introduction

Fast and reliable broadband services is increasingly important in the United Kingdom (UK), with the ongoing transition towards an information society. In 2017, the proportion of individuals using broadband was higher than any other key media service and it’s estimated that the average fixed-broadband data traffic for households has increased from 97 Gigabytes (GB) in 2015 to 132 Gb in 2016 (Ofcom, 2017). Superfast fibre and cable broadband offer significantly faster speeds than Asymmetric Digital Subscriber Line (ADSL) services, defined by the UK government as speeds greater than 24Mbits/s (Priestley & Baker, 2017). The Broadband Delivery UK (BDUK) is a government policy for rolling out superfast broadband to as much of the country. Another important aspect of the BDUK is to ensure that basic fixed-broadband services are affordable to all UK citizens and customers. Superfast broadband is important for enabling businesses to generate prosperity and to empower every citizen to take part in society by binding families and friends together and to keep us entertained (Cairncross, 2001). As broadband plays an increasingly critical role in our lives, the supporting network infrastructure must keep pace to ensure reliable services.

The market for Superfast Broadband Service Providers (BSP) has become increasingly competitive and there is a need to provide affordable and reliable services. To remain competitive, BSP rely on demand forecasts, using socio-economic information to target areas that contain high numbers of potential customers who are interested in and able to afford superfast broadband (Fildes & Kumar, 2002; Hopkins, et al., 1995). They also rely on past observations of service disruptions or failures to forecast where they are more likely to happen (Deljac, et al., 2011; Grubesic & Murray, 2002). Service disruptions happen because of conditions that results in a specified service going beyond those defined in the contract. Service disruptions are caused by a variety of faults and minimising faults is a top priority of BSP (Fildes & Kumar, 2002). Accurate network fault forecasts are important for network operational efficiency, including determining the number of engineers and truck call outs necessary to fix any service disruptions. In addition to the cost of broadband services, reliable services are also important for influencing consumer demand and customer churn.

In a highly competitive market, telecom problems have led to innovative time series forecast methods (Fildes & Kumar, 2002; Hopkins, et al., 1995). Typically, Autoregressive (AR) type models are used to capture the statistical characteristic of Temporal Dependency (TD), where the conditional expectation of future events is regressed on earlier values plus some noise. Despite the ubiquity of network faults, the forecasting literature on this topic is limited (Deljac, et al., 2011; Sandholm, 2007). Forecasting faults is a challenging task because of their stochastic nature and the rate they occur is much higher than in any other industry. Moreover, conventional forecasting methods rely on stable trends and seasonal patterns, which are not met by the stochastic nature of faults (Ozturkmen, 2000).

A small number of studies have analysed the spatial distribution of broadband access. it’s recognised that spatial disparities in broadband access exist (Grubesic & Murray, 2002). Spatial Autocorrelation (SA) is the statistical characteristic, where nearby observations in space tend to be more similar than those further apart (Tobler, 1970). Spatial information are defined by Geographical coordinates, which are typically applied in Geographic Information Systems (GIS). Since SA violates the independence and identically distributed (iid) assumption of many statistical methods, not accounting for this structure leads to biased parameter estimates (Anselin & Griffith, 1988). Therefore, a natural extension of the basic time-series models in the forecast literature is to incorporate space, within a hierarchical structure to account for Geographical variations in faults. One of the leading United States telecommunication companies AT&T, recognizes that spatio-temporal research will play an increasing role in the forecasting and strategic planning of future communication technologies (Volinksky, 2018). For example, the rollout of 5G connectivity will require more local planning of key infrastructure, to ensure complete coverage of network areas. Moreover, analysing the spatial distribution of faults allows the forecaster to identify how the spatial distribution of network faults changes over time at a much finer spatial scale than the common regional level of analysis. Moreover, also accounting for the uncertainty component of space should give a higher forecast accuracy of network faults.

Bayesian Markov Chain Monte Carlo (MCMC) algorithms are typically used for spatio-temporal modelling because they can explicitly introduce SA and TD simultaneously using a Bayesian Hierarchical Modelling Framework (BHMF), where each level can be modelled as a stochastic process. These models have been extensively applied in the disease mapping literature, to gain a better understanding of the processes driving the disease and to identify areas characterized by high or low relative risk (Knorr-Held, 2000; Knorr-Held & Besag, 1998; Lawson, 2013; Xia , et al., 1997). Typically, the spatial component is modelled as aggregated areal counts of disease risk instead of a continuous process, using Conditional Autoregressive (CAR) priors. However, the nature of modelling both space and time simultaneously using MCMC algorithms gives computational issues and so the application of spatio-temporal has been limited and restricted to small area studies.

We propose a Bayesian spatio-temporal method for modelling and forecasting faults for a BSP. It extends the limited literature on forecasting network faults by explicitly introducing space in a BHF. Statistical inference is achieved with the Integrated Nested Laplace Approximation (INLA) approach, as an efficient alternative to MCMC algorithms (Rue, et al., 2009). We include four different model specifications, with two conventional time-series priors, one also introducing space with a CAR prior and another with a space-time interaction term (Knorr-Held, 2000). This allows us to assess whether introducing space with time-series priors gives a higher forecast accuracy. Several covariates in the literature are included to explore their relationship with faults. Furthermore, we apply this methodology to North West England, a scale much larger than conventional small area studies in the literature of spatio-temporal modelling.

1.4 Aims and Objectives

The aim of this study is to extend the limited time-series forecasting literature on network faults by incorporating space. Specifically, we apply spatio-temporal areal unit modelling to forecast Virgin Media network faults in North West England. The objectives of this study involve;

1. Evaluate whether incorporating space with time-series priors gives a higher forecast accuracy of network faults, than conventional time-series forecasting.
2. Explore whether incorporating an additional spatio-temporal interaction term further improve the forecast accuracy of network faults compared to spatio-temporal models.
3. Assess what broadband inequality and socio-economic factors are most related to Virgin Media network faults.
4. To identify whether there is a spatial distribution of Virgin Media network faults in North-West England and how they change over time.

2 Literature Review

This chapter highlights the conventional time-series methods used in the telecommunication literature, as well as introducing spatial statistics from the wider literature. We begin by introducing some background information about network faults, including their distribution and association with socio-economic and inequality related factors (2.1). The second section (2.2) introduces conventional time-series in the literature and then considers how they have been applied in the telecommunication industry. The third section (2.3) introduces spatial statistics and specifically spatio-temporal models as a method for forecasting network faults. We finalise by stating our contributions to the forecasting network faults literature (2.4).

2.1 Network Faults

Network fault events do not happen continuously, and this stochastic nature is related to the many internal and external drivers of them. The most commonly cited network fault type is Customer Equipment (CE), related to problems with modems, ADSL splitters and other equipment (Deljac, et al., 2011). External factors that influence network faults include bad weather, which has been shown to increase Wiring and Radio Frequency related faults (Deljac, et al., 2011). For example, higher than average temperatures can overheat external cabinet boxes, resulting in bandwidths beyond their specifications.

Although network faults are a stochastic process, there are periods when network faults are more likely, which are associated with broadband usage. There exists daily, weekly and monthly periodicities of broadband usage and associated faults (Deljac, et al., 2011). For example, the average download speed for November 2015 across all United Kingdom connections was 27.0Mbits/s during the 8pm to 10pm weekday peak demand, which was 85% of the 31.6Mbit/s average maximum speed and 93% of the 28.9Mbit/s 24-hour average (Ofcom, 2016). They also found that the fastest download speeds are experienced between the hours of 2-3am. Another study found that monthly network faults do not vary significantly and exhibit a stationary process, where the probability distribution or mean and variance does not change when shifted over time (Deljac, et al., 2011).

Whilst (Cairncross, 2001) argued that space will be rendered virtually meaningless by near instantaneous communications, local Geographies still plays a pivotal role in access to broadband (Grubesic & Murray, 2002). Within a profit driven market, BSPs may target areas that contain a higher number of potential customers who are interested in and are able to afford superfast broadband (Grubesic & Murray, 2002). This suggests that a Geography of internet use exists, determined by their socioeconomic characteristics.

The 2018 Internet User Classification (IUC) presents one of the largest bespoke classifications for describing the Geography of internet use and engagement in England and Wales (Alexiou & Singleton, 2018). Understanding the Geography is important for both Government and corporate policy making, including mitigating digital inequalities and for identifying new customers. The IUC shows that demographic factors are the most influential on internet use and engagement, including age and ethnicity. For example, young populations are more engaged than those who are elderly, and they tend to reside in popular student areas, within proximity to universities in inner cities. Therefore, internet accessibility is also influenced by where we live.

Broadband access to urban areas with high-density populations has typically preceded that of rural locals with low-density populations (Grubesic & Murray, 2002). This has largely been driven by the profit-driven BSP market, where they target areas with higher broadband demand. Whilst this digital divide has narrowed to just 4% of UK households not receiving the Universal Service Obligation (USO), defined as a broadband download speeds of at least 10Mbit/s and an upload speed of at least 1Mbit/s (Ofcom, 2017), regional disparities still exist. For example, just 2% of Scottish urban properties don't receive the USO, whereas 27% of rural households do not receive the USO (Ofcom, 2017). Further inequalities in broadband access are characterized by the lowest income households, who are the least likely to take up fixed broadband services, due to increasing fixed monthly costs (Alexiou & Singleton, 2018; Ofcom, 2016).

2.2 Time-series forecasting

2.2.1 Time-series methods

A Time series exhibit a special case of the Gaussian process or Markov property, where x_t is serially dependent on other observations. Typically, time-series exhibit non-linear patterns, defined as a set of observations x_t, each one being recorded at a specific time t, which typically represents a stochastic process (Chatfield, 2016). Moreover, there is a decay of dependence, where x_{t+1} becomes increasingly near independence as $x_{t+1} \rightarrow \propto$. TD describes the similarity between observations as a function of time lag between them;

$$R(s,t) = \frac{E[(x_t - \mu_t)(x_s - \mu_s)]}{\sigma_t \sigma_s}, \qquad (1)$$

where E is the expected value operator. If the function R is well defined, its value lies within the range -1 to 1, with 1 indicating strong positive correlation, 0 indicating no association and 1 indicating strong negative association (Shumway & Stoffer, 2017). If x_t is a stationary process, then TD can be written as;

$$R(\tau) = \frac{E[(x_t - \mu)(x_{t+\tau} - \mu)]}{\sigma^2}. \qquad (2)$$

For a deterministic forecasting model, x_{t+1} is based only on the current state x_t of the time-series. In contrast, the Markovian property in probabilistic or stochastic models relaxes this assumption, where a noise component is added to account for all the unknown factors that are common in time-dependent observations (Brockwell & Davis, 1996). Therefore, the probability of future values is only partly explained and conditioned by past values.

Whilst time-series models can be used for a description of seasonal pattern and trends, we are interested in forecasting future values of that series. The objective of forecasting is to obtain a residual as close to zero or the difference between actual and forecasted values. Whilst network faults represent a continuous process that are realizations of a stochastic random process, the focus of this study is one step ahead monthly forecasts, so we focus on discrete time series analysis. A time series that is said to be discrete when observations are taken only at specific times and are usually equal spaced (Chatfield, 2016).

Autoregressive (AR) models are a broad class of time-series models, where the current value of the process is expressed as a finite, linear aggregate of previous values (Schabenberger & Gotway, 2005). Instead of estimating the whole conditional distribution of x_t, we just look at the conditional expectation where we regress x_t on earlier values. An AR model or order 1 (AR1) depends on the immediate, previous value, which can be expressed as;

$$y_t = \beta_0 + \beta_1 X_t + \varepsilon_t, \tag{3}$$

$$\varepsilon_t = p\varepsilon_{t-1} + \omega_t. \tag{4}$$

Here the iid assumption of the autoregressive error term ω_t follows the usual assumption about regression error terms;

$$\omega_t \sim iid\ N(0, \sigma^2). \tag{5}$$

The error at time t is a fraction of the error at time t plus some new perturbation ω_t. The relationship of y and X variables at time t being related to the y and X variable measurements at time t – 1, which are accounted into the error term at concurrent times. Moreover, the probability distribution or variance σ^2 and mean μ is the same for all values of t, independent of the time lag, expressed as a stationary process (Chatfield, 2016).

A Moving Average (MA) prior is a linear combination of past error terms. For a first order MA;

$$x_t = \mu + w_t + \theta_1 w_{t-1}, \tag{6}$$

$$w_t \sim iid\ N(0, \sigma_w^2). \tag{7}$$

Here w_t follows the iid assumption, which is normally distributed with a mean μ 0 and the same variance σ^2. AR and MA priors can be combined to form a general class of time-series models, Autoregressive Integrated Moving Average (ARMA) Models (Box & Jenkins, 1976);

$$x_t = \sum_{i=1}^{p} \varphi_i X_{t-1} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-1} + \varepsilon_t, \tag{8}$$

where x_t is the forecasted value φ and θ are the regression parameters for the calculated model, p and q determines the number of regression terms that are considered and ε_t characterizes error. In other contexts, non-stationary time-series can be expressed using an Autoregressive Integrated Moving Average (ARIMA) prior.

The Autoregressive Conditionally Heteroscedastic (ARCH) prior is used to describe the variance of the current error term as a function of the sizes of the previous time periods error terms. The most important extension of the ARCH process is the generalized ARCH (GARCH) process, which is said to be stationary if;

$$X_t = \sigma_t Z_t, \tag{9}$$

$$\{Z_t\} \sim IID\ N(0,1). \tag{10}$$

A Random Walk (RW) prior is a special case of AR models, which represents non-stationary time-series;

$$x_t = \sum_{j=1}^{t} w_j, \tag{11}$$

$$\gamma_x(s,t) = cov(x_s, x_t) = cov\left(\sum_{j=1}^{s} w_j, \sum_{k=1}^{t} w_k\right) = \min\{s,t\}\,\sigma_w^2, \tag{12}$$

The TD function of a RW prior depends on the time values s and t and not on the time lag. Moreover, the variance of a RW, increases without being bounded as time t increases.

2.2.2 Applications of time-series models in the telecommunication literature

In a highly competitive market, telecom problems have led to innovative time series methods for forecasting customer demand (Fildes & Kumar, 2002). One study applied these methods for assessing broadband demand within a local network system (Sandholm, 2007) and another forecasted customer demand at the national scale (Hopkins, et al., 1995). Despite the ubiquity of faults, there has been a limited number of studies who have applied such a methodology for forecasting network faults (Deljac, et al., 2011).

One study applied several autoregressive models, including the ARMA, ARIMA and GARCH models for short-term and long-term forecasting of network faults (Deljac, et al., 2011). For all forecast intervals, the forecast accuracy was highest for the ARIMA model, defined as having the smallest Cumulative Mean Square Error (MSE). Whereas the MSE was significantly higher for the GARCH and ARMA models for one day ahead forecasts, the ARIMA and ARMA models had similar forecasts for one step ahead monthly forecasts. They discussed that this is because faults represent the cumulative sum of faults over a longer time than daily and weekly forecasts, subject to a higher number of random factors that are unaccounted for. Moreover, another network study found that the RW1 prior outperformed the ARIMA models for one step ahead forecasting but the opposite for two and three steps ahead forecasts (Sandholm, 2007). This is an unsurprising result, given the smoothing effect of the ARIMA model. Overall, the selected forecast model depends on the purpose of the study and the underlying process.

2.3 Spatio-temporal models as a forecasting method

Although GIS and associated spatial models have been applied for evaluating the Geography of potential broadband customers (Grubesic & Murray, 2002) and national broadband access (Downes & Greenstein, 2005), they haven't been applied for forecasting network faults. The ability to represent, manipulate and analyse spatial information within a forecasting framework makes it possible to calculate areas with a higher probability of faults within the main study area. Furthermore, GIS enables one to examine various spatial relationships, including the use of the widely available UK Census information on socio-economic and demographic factors. This gap in the literature is unfortunate if we consider that local Geographies play a key role on broadband usage and network faults. Here we present spatio-temporal models from the wider literature.

2.3.1 Spatio-temporal models

Observations in spatial data are realizations of a stochastic processed indexed by space;

$$Y(s) = \{\, y(s), s \in D\} \tag{13}$$

For area and lattice data, $y(s)$ is a random aggregate value over each spatial unit within well-defined boundaries (Blangiardo & Cameletti, 2015). The former represents irregular units, such as administrative boundaries and the latter represents regularly space units (Cressie, 1993). SA is statistical characteristic, where nearby observations, defined in terms of Euclidean distance tend to be more similar than those further apart (Tobler, 1970). When present, SA violates the independence assumption in many statistical models and not accounting for this structure leads to biased parameter estimates (Anselin & Griffith, 1988).

The statistical modelling of spatio-temporal processes represents several decades of cross-field research in time-series analysis, spatial statistics and spatial econometrics (Elhorst, 2003; Huang, et al., 2010; Pfeifer & Deutsch, 1980). Space-time models consider correlated observations of a phenomena within fixed spatial and temporal areal units that change over time. A basic model assumes a Poisson distributed dependent variable in an infinite population with a small probability;

$$R(s,t) = \frac{E[(x_t - \mu_t)(x_s - \mu_s)]}{\sigma_t \sigma_s}, \tag{14}$$

$$y_{ij} \sim Pois(e_{ij}, \theta_{ij}), \tag{15}$$

$$Log(\theta_{ij}) = \alpha_0 + S_i + T_j, \tag{16}$$

where, S is the spatial term and T are the temporal term. The specification of the model can be extended to include an interaction between space and time or Space-Time Interaction (STI) δ_{it}, that accounts for any residuals not captured by the space-time model. Whilst (Bernardinelli, et al., 1995) extended the work of (Besag, et al., 1991) by incorporating a linear STI term, (Knorr-Held, 2000) presented four STI models that drop the assumption of linearity, capturing more commonly found non-linear time-series processes.

2.3.2 Frequentist vs Bayesian approaches

One study applied a spatio-temporal Poisson regression model using a Bayesian approach to investigate the link between ambient ozone and paediatric visits for asthma during the summers of 1993 to 1995 (Carlin, et al., 1999). The study compared their estimates against another study examining the same process but with a frequentist Poisson regression model (Tolbert, et al., 1997), which accounted for long-term temporal trends but not for TD and SA. The results indicated similar estimates of relative risk due to ozone exposure, but the Bayesian approach provides a natural framework for mapping the estimated values of risk.

Whilst both Bayesian and frequentist approaches incorporate the likelihood from a current study, what differentiates them is that prior information is combined with the likelihood to form the Posterior distribution. Under Bayes theorem;

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}, \tag{17}$$

the posterior distribution $p(\theta|y)$ represents the uncertainty about the parameter of interest θ after observing the data, thus conditioning on y. Therefore, uncertainty is described by the posterior density, where the prior and likelihood are modelled as stochastic processes. This conditional independence assumption forms the basis for Bayesian inference, where knowledge is gained about unknown parameters θ and the distribution of unknown process ϕ after observing the data y (Banjeree, et al., 2004; Blangiardo & Cameletti, 2015). Thus, we need to calculate the posterior marginal distribution $\pi(\phi|y)$ of each element of the latent model ϕ and the posterior marginal distribution $\pi(\theta|y)$ of the hyperparameter vector θ. The posterior probability of a parameter exceeding a specified threshold is also easily obtained from the posterior distribution, providing a more intuitive quantity than frequentist p-values. Bayesian inference is also appropriate when there are missing values, a common attribute of spatio-temporal data (Gelfand, et al., 2005; Haworth & Cheng, 2012). BHMFs account for the uncertainty of both space and time by modelling each parameter as a stochastic process at the next level (Banjeree, et al., 2004; Schabenberger & Gotway, 2005).

2.3.3 Markov Chain Monte Carlo methods

Bayesian statistics has become more available to researchers, in part due to advances in computational power and the popular software packages BUGS and winBUGS using Markov Chain Monte Carlo methods (MCMC). Here the marginal posteriors are computed using a simulation-based technique, based on the Markov chain, where the sequence of random variables $\theta_1, \ldots, \theta_n$ for which the distribution of the sampled draws depends only on the most recent value θ_{t-1}. Each simulation results in an improved approximation distribution, until it converges to a target distribution or stationary distribution, which must be defined (Gelman, et al., 2013). The main issue of such MCMC Bayesian inference is that such an approach can lead to several days of computing time. Another disadvantage is that the software packages require specialised programming that is non-trivial for applied researchers (Knorr-Held & Rue, 2002).

2.3.4 Integrated Nested Laplace Approximation

The Integrated Nested Laplace Approximation (INLA) is alternative, numerically efficient approximation method of computing the marginal posteriors of the latent variables and hyperparameters of the Gaussian latent model;

$$\pi(\xi_i|y) = \int \pi(\xi_i|\theta, y)\pi(\theta|y)d\theta, \tag{18}$$

$$\pi(\theta_j|y) = \int \pi(\theta|y)d\theta_{-j}. \tag{19}$$

This approximation is based on efficient combination of Laplace approximations to the full conditionals $\pi(\theta|y)$ and $\pi(\xi_i|\theta, y)$, $i = 1, .., n$ and numerical integration routines to integrate out the hyperparameters θ (Rue, et al., 2009; Blangiardo & Cameletti, 2015; Bivand, et al., 2011). INLA covers a wide range of models that use Latent Gaussian processes from generalized and dynamic linear models to spatial and spatio-temporal models (Blangiardo & Cameletti, 2015). The general form of an LGM is the likelihood;

$$y|x, \theta_2 \sim \prod_i p(y_i|\eta_i, \theta_2), \tag{20}$$

Latent field;

$$x|\theta_1 \sim p(x|\theta_1) = N(0, \textstyle\sum), \tag{21}$$

hyperparameters;

$$\theta = [\theta_1, \theta_2]^T \sim p(\theta). \tag{22}$$

Where y is an observed dataset, x is the joint distribution of all parameters in the linear predictor including itself and θ are the hyperparameters of the latent field that are not Gaussian. If we can assume conditional independence in x, then this latent field is a Gaussian Markov Random Field (GMRF). This property allows Bayesian inference without the need to use the computationally intensive MCMC algorithms, based on simulation. Whereas MCMC algorithms require hours to days to run, INLA provides precise estimates in seconds or minutes (Cressie & Wilke, 2011). Moreover, a user-friendly R environment (R Core Team, 2015) allows flexible INLA with other spatial packages.

2.3.5 Conditional Autoregressive (CAR) priors

When working with areal data, SA is modelled through a neighbourhood structure Q_{ij}, which are commonly represented by Conditional Autoregressive (CAR) priors in a lognormal Poisson model (Besag, 1974). CAR priors are also a special class of GMRFs, where the conditional distribution of observations depends only on the values of its neighbourhood, where $Q_{ij} = 0$ only if i and j are neighbours (Rue & Held, 2005). The most common CAR specifications include the intrinsic CAR (ICAR) and Besag-York-Mollié (BYM) priors (Besag, et al., 1991) but alternative specifications can be applied (Leroux, et al., 2000; Stern & Cressie, 1999). We focus on the commonly applied BYM or convolution model, which includes both a structured and unstructured spatial component, to capture spatial and non-spatial heterogeneity. The unstructured component ensures that if most of the variability is non-spatial, captured by θ, it does not lead to a significant overestimation of the conditional variance in the structured spatial component. Using the notation of (Banjeree, et al., 2004), the BYM is specified as;

$$Y_i|\psi_i \sim Poisson\left(E_i\, e^{\psi_i}\right), \tag{23}$$

For i $\in 1{:}N$, where;

$$\psi = x\beta + \theta + \phi. \tag{24}$$

The coefficients β are modelled as fixed effects, θ is an ordinary random-effects component for non-spatial heterogeneity and ϕ is an ICAR spatial component.

(Lee & Mitchell, 2012) proposed a different method to for capturing more localized spatial structures as an alternative to the single global level of spatial smoothing in space. They argue that this global level is too simplistic for real data, which exhibit sub-regions of stronger SA as well as locations at which the response exhibits a step-change. Moreover, the global level spatial smoothing effect also results in collinearity between any spatially and temporally smooth covariates, which can lead to poor estimates of the fixed effects. Nonetheless, since we have no prior knowledge about the spatial neighbourhood structure of faults, we focus on the most commonly used BYM specification.

2.4 Summary

Forecasting approaches in the telecommunication literature have focused on time-series models but these do not consider the statistical characteristic SA, which can lead to biased estimations. Our contributions to the forecasting network faults literature lie in several aspects. Primarily, we extend the conventional time-series models by incorporating space using the computationally efficient R-INLA. Specifically, we use BYM priors to incorporate space, as aggregated fault counts at the Middle Layer Super Output Area (MSOA) level. In total, four model specifications are defined, which include two models accounting for TD only, with the stationary AR1 prior and non-stationary RW1 prior and two also including SA, with one using just the BYM prior and another also including a spatio-temporal interaction term. For each model specification, we also include several covariates from the literature to quantify associations between broadband access inequality and socio-economic characteristics. To the best of our knowledge, no study has applied such a methodology and we evaluate whether incorporating space improves the model fit and forecast accuracy over conventional time-series models.

3 Methodology

This chapter presents the study area (3.1), data collection (3.2) and data analysis (3.3) procedures used for addressing the research aims and objectives. The main aim of this study was to extend the conventional time-series literature for forecasting faults by including space or spatial effects. The objectives of this study involve;

1. Evaluate whether incorporating space with time-series priors gives a higher forecast accuracy of network faults than conventional time-series forecasting.
2. Explore whether incorporating an additional spatio-temporal interaction term further improve the forecast accuracy of network faults compared to spatio-temporal models.
3. Assess what broadband inequality and socio-economic factors are most related to network faults.
4. To identify whether there is a spatial pattern of Virgin Media network faults in North-West England and how they change over time.

3.1 Study area

This study focused on North West England, one of nine regions in England. Specifically, we focused on areas with Virgin Media coverage, which covers approximately 4215km^2 and 600,000 Virgin Media customers. In total, 662 MSOAs were analysed, each with a mean population of 7,200, ranging from 5,000 to 15,000 (ONS, 2017). This Census areal unit was selected for its ability to be combined with a range of socio-economic related Census covariates. Whilst the smaller Local Layer Super Output (LSOA) areal unit could be used to give more detailed information, running the most complex STI model at this Geographical scale would require access to high performance computing. Therefore, we selected the MSOA level, which still gives a good level of detail, at a much finer spatial scale than the common regional level analysis.

3.2 Data and variables

All data used in this study are secondary data sources, aggregated to the MSOA level. Whilst we acknowledge the scaling and aggregation problems of areal data, we minimize these by selecting the smallest level of aggregation achievable with our resources (Openshaw, 1984).

3.2.1 Dependent variable

Virgin Media provided observed network faults from January 2017 to March 2018, including the time of fault, the location where the fault happened (LSOA) and the type of fault. The location of customer faults was provided by Virgin Media at the LSOA level to ensure customer confidentiality. Data pre-processing was applied to aggregate all faults to the MSOA level and by month, using the sum of observed faults.

3.2.2 Covariates

Based on the literature, we included several covariates in all model specifications (3.4), which was defined as fixed effects in the BHMF (Table 1). Level of income was included because this is typically modelled in conventional broadband demand forecasting models (Ozturkmen, 2000). Income is measured by income deprivation, one of several indices of relative deprivation for England (2015). The average fixed broadband usage (Ofcom, 2017) in Gigabytes (GB) was also included as another indicator of broadband demand. We also included the IUC because it best describes the Geography of socio-economic groups based on internet engagement (Alexiou & Singleton, 2018). Since this covariate represents 10 different socio-economic groups related to internet engagement, the first group is taken as a dummy variable and we aggregate the data from the LSOA to the MSOA level using the mode. We include the proportion of Elderly to evaluate the significance of this age group on faults. This is because this age group will increasingly represent a significant share of future broadband demand, as the UK age-structure is changing (Ofcom, 2016). Therefore, we assess the relationship between network faults and this age group over time. As Education is a network fault category for Virgin Media, we include Education deprivation to measure the association with network faults.

Table 1. Description of the Covariates

Covariate	Description
IUC Group	The Internet User Classification (IUC) is a bespoke classification that describes how different socio-economic groups living in England and Wales engage with the internet. This CDRC dataset was aggregated to the MSOA level using the mode.
Education Deprivation	This represents one of the Indices of Multiple Deprivation (IMD) in England. Education score was selected, where higher values indicate lower levels of education and generally the more deprived an area is. This CDRC dataset was aggregated from the LSOA to MSOA level using the average score.
Income Deprivation	This represents one of the IMD (2015). Income score was selected, where larger values indicate the more deprived an area is. The dataset was downloaded from the CDRC at the LSOA level and was aggregated to the MSOA level using the average score.
Elderly Population	The proportion of elderly population was calculated from the total population, defined as those aged 65 plus. This CDRC dataset was aggregated from the LSOA to MSOA level using the average score.
Broadband Use	This represents the average data usage in Gigabytes (GB) for fixed-line broadband connections. This formed part of the Ofcom Connection Nations Report 2017, an analysis of the major fixed telecommunication operators (BT, Virgin Media, Sky, Talk Talk, Vodaphone and KCOM). The dataset was aggregated from the postcode level to the MSOA level, taking the average data usage.

* CDRC represents the Consumer Data Research Centre, MSOA represents Middle Layer Super Output Areas (MSOA) & LSOA represents the Lower Layer Super Output Areas (LSOA), which are Census administrative boundaries.

3.3 Statistical modelling

We applied the programming language R for all data analysis and visualization in this study. Specifically, the “INLA” package (Rue, et al., 2009) was used for Bayesian inference and “ggplot2” (Wickham, 2016) for data visualization. Other packages required included the spatial “sp” (Bivand, et al., 2011), “spdep” (Bivand & Piras, 2015) and “rgdal” package (Bivand, et al., 2018).

3.3.1 Model specifications

In total, four model specifications were defined for forecasting Virgin Media faults (Table 2). Since we had no prior knowledge of Virgin Media network faults, we applied stationary and non-stationary priors to assess which gave the highest forecast accuracy. We include space with the “best” time-series prior to explore whether spatio-temporal methods yield a higher forecast accuracy than just time-series models. An additional space-time interaction term is included in the fourth model, to explore whether this increases the forecast accuracy further.

Table 2. Description of the Model Specifications

Model	Description
AR1	An Auto-Regressive correlation of order 1 (AR1) is incorporated for modelling for temporal dependency between consecutive months.
RW1	A Random Walk correlation of order 1 (RW1) is modelled for, to account for temporal dependency between consecutive months.
ST	This Space-Time (ST) model incorporates spatial autocorrelation between neighbouring areal units, using a Conditional Autoregressive Prior (CAR) and is correlated between consecutive months using the preferred AR1 model.
STI	An additional Space-Time Interaction (STI) interaction term is added to the specified ST model, to account for any residuals that are unaccounted for.

These model specifications were first applied to all network faults and then applied the top 5 broadband network fault types, to explore the heterogeneity of faults (Table 3). This second analysis was conducted to explore whether any model uncertainty for forecasting all faults is associated with a specific fault type.

Table 3. Description of the top 5 Broadband Network Fault Types

Fault Type	Description
Customer Equipment	This reflects any issues with the equipment of customers on their property. For example, a non-responsive modem, outdated software or faulty ADSL splitter.
Education	Education involves any reported broadband fault that do not reflect any software or hardware related issues but a lack of knowledge about how to operate their equipment.
Radio Frequency	This refers to any problem with the frequency ranges beyond the Virgin Media broadband specifications.
Wiring	This category reflects those faults relates to problems with copper and optical cable, network terminal points, main distribution frames and ADSL ports.
Other	This category represents all of the other fault types not listed above, including Unknown and No Access faults. Therefore, this category represents the largest proportion of faults.

All the model specifications include the covariates discussed in 3.2, which are modelled as fixed effects and the temporal and spatial main effects and space-time interaction term are modelled as random effects. Since the dependent variable is count data and skewed, we assume a Poisson distribution. To avoid the problem of identifiability in spatio-temporal models, each of the random effects include a structured and unstructured component (Knorr-Held & Besag, 1998; Knorr-Held, 2000). We specify the most complex STI model using the example in (Blangiardo & Cameletti, 2015);

$$log\left(\mu_{tj}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_4 x_{4i} + \beta_2 x_{2i} + \beta_5 x_{5i} + \gamma_t + \phi_t + \upsilon_i + \nu_i + \delta_{it}, \qquad (25)$$

where β_0 is the intercept, $\beta_1,..,\beta_5$ quantify the effects of the regression coefficients, $x_1,..,x_5$ represent the covariates modelled for on the dependent variable, γ_t and ϕ_t represents the temporally structured υ_i and ν_i represent the structured and unstructured space effects and δ_{it} represents the interaction between time and space. For the ST model, equation (25) can be rewritten by dropping the δ_{it}. For both the AR1 and RW1 model, equation (25) can also be rewritten by dropping υ_i and ν_i .

All models were defined within a BHMF, where each level is modelled as a stochastic process. As discussed in section 2.3.3, the main challenge of Bayesian inference using MCMC algorithms is that applying complex is computationally demanding (Irvine, et al., 2007). Instead, we use the R-INLA package to perform Bayesian inference, which is significantly more computationally efficient than MCMC algorithms (Rue & Held, 2005; Rue, et al., 2009). Therefore, this method allows us to fit complicated spatio-temporal models with the computational and time resources we have. We use the default simplified Laplace approximation when running INLA.

3.3.2 Temporal main priors

Since we are interested in forecasting based on past observations from the previous month, the first model includes the stationary AR process of order 1 (AR1) prior;

$$X_i = pX_{i_{-1}} + \epsilon_{ij}, \qquad (26)$$

where the current value X_i is a linear combination of p, times the most recent past values $X_{i_{-1}}$, plus a noise term ϵ_{ij}. With the correlation $|p| < 1$ and another restriction $\epsilon_{ij} \sim N(0, \tau^{-1})$ the AR1 prior is a stationary process (Rue & Held, 2005). As the slope parameter approaches 1, the AR1 exhibits higher persistence or a larger contribution from the previous term, relative to the noise.

The second model includes the none-stationary RW process of order 1 (RW1) prior;

$$X_i = X_{i_{-1}} + \Delta\, x_i, \qquad (27)$$

where the current value X_i is the previous values $X_{i_{-1}}$plus an increment term $\Delta\, x_i$. The restrictions for the AR1 prior include $\Delta\, x_i \sim N(0, \tau^{-1})$ and $\sum \Delta\, x_i = 0$. The RW1 model is a special case of the AR1 model, in which the slope parameter ϕ =1, so is a non-stationary process. The mean of a RW process is constant, but its variance is not. It also exhibits strong persistence, where past values have a big impact on current values.

3.3.3 Spatial prior

To incorporate space within the selected time-series prior, we apply a BYM prior (Besag, et al., 1991), using the notation from (Blangiardo, et al., 2013);

$$\upsilon_i \mid \upsilon_{j \neq i} \sim Normal(m_i, s_i^2), \tag{28}$$

$$m_i = \frac{\sum_{j \in \mathcal{N}(i)} \upsilon_j}{\#\mathcal{N}(i)} \text{ and } S_i^2 = \frac{\sigma_\upsilon^2}{\#\mathcal{N}(i)}, \tag{29}$$

where $\#\mathcal{N}(i)$ is the number of areas which share boundaries with its neighbours. This prior decomposes the spatial effect into the sum of the structured υ_i unstructured ν_i components. The unstructured one is modelled using an exchangeable prior $v_i \sim Normal(0, \sigma^2)$, to ensure that any random error within area i is not modelled as spatial correlation, preventing any misleading estimates (Breslow, et al., 1998). Since little information is known about the prior distribution of faults, we use the default INLA minimally informative priors for the log of the unstructured effect precision;

$$\log(\tau_v) \sim logGamma(1{,}0.0005), \tag{30}$$

and log of the structured effect precision;

$$\log(\tau_v) \sim logGamma(1{,}0.0005). \tag{31}$$

Whilst it is reasonable to regard areas i and j as neighbours if they share a common border (Best, et al., 2001), denoted as $i \sim j$, this is not appropriate when areal units are not distributed evenly in the study area (Earnest, et al., 2007; Wall, 2004; Wakefield, 2007). Since the MSOAs in this study area are not regularly arranged, we selected a distance-based neighbourhood structure. We assume i and j are neighbours whenever j falls within a critical distance band from i. More specifically, $w_{ij} = 1\ when\ d_{ij} \leq \delta$, and $w_{ij} = 0$ otherwise where δ is a critical distance cut-off. We calculated the K-Nearest Neighbour (KNN) or minimum Euclidean distance between the MSOA centroids, to avoid isolates or observations with no neighbours and this information defined δ. Finally, we specified a sparse precision weight, which is computationally efficient because they are specified entirely through neighbourhood structures and not the full covariance.

3.3.4 Space-time interaction term

We specify a non-parametric type II space-time interaction δ_{it} term, which combines the structured temporal main effect γ_t and the unstructured spatial effect υ_i (Knorr-Held, 2000). We favour a non-parametric term because a linear time trend (Bernardinelli, et al., 1995) does not accurately model the linear time-series process of faults (Schrödle & Held, 2010). The parameter vector δ follows a Gaussian distribution with a precision matrix given by $\tau_\delta \boldsymbol{R}_\delta$, where τ_δ is an unknown scalar and $\boldsymbol{R}_\delta$ is the structure matrix, identifying the type of TD and/or SA between the elements of δ (Blangiardo & Cameletti, 2015). We write the structure matrix of a type II δ_{it} as;

$$\boldsymbol{R}_\delta = \boldsymbol{R}_\upsilon \otimes \boldsymbol{R}_\gamma \tag{32}$$

Where $\boldsymbol{R}_\upsilon$ and $\boldsymbol{R}_\gamma$ is the neighbourhood structure specified by the selected time-series prior. This leads to the assumption for the ith area the parameter vector $\{\{\delta_{i1,\dots,\delta_{iT}}\}$ has an autoregressive structure on the time component, independent from the ones of other MSOAs. The matrix $\boldsymbol{R}_\delta$ has a rank of $n(T-1)$ for the selected prior.

3.5 Forecasting procedure

For one step ahead monthly forecasts in INLA, we apply posterior predictive simulations to compute the linear predictor (Wang, et al., 2018). Firstly, the datasets used by INLA requires some pre-processing. For each forecasted month, we apply past observations from the previous months and add missing values or NA values for those observations in the month we want to forecast. For example, for January 2018, we use past observations from January to December 2017 and then add in NA values for the observations we want to forecast. To account for uncertainty in the estimated values of faults for each MSOA, we simulated from the posterior joint distribution of the model to obtain 1000 samples. In total, three forecasts was made from January to March 2018.

3.5 Model selection

Since we are primarily interested in comparing the forecast accuracy of the model specifications, we use the residual or difference between observed and forecasted faults. Therefore, we select the model based on the residual closest to zero. We are also interested in the model fit or how well a statistical model describes how well it fits a set of observations. We use the Deviance Information Criterion (DIC) to measure the model, where the smallest value has the best fit (Spiegelhalter, et al., 2002). We define deviance as;

$$D(\theta) = -2\log(p(y|\theta)) \tag{33}$$

In a Bayesian model, this is a random variable, so we use the expected deviance $(E(D(\theta))$under the posterior distribution as a measure of fit. For counting parameters, we introduce effective number of parameters;

$$p_D = E\big(D(\theta)\big) - D\big(E(\theta)\big) = \overline{D} - D(\bar{\theta}) \tag{34}$$

And then DIC is:

$$DIC = \overline{D} + p_D \tag{35}$$

Nevertheless, no model fit criteria is “good” and DIC has problems if the fitted posterior distribution is not well represented by its posterior mean (Zurr, et al., 2017).

4 Results

This section begins by comparing the forecast accuracy and model fit of the four specified forecasting models to select a “best model” at the North West scale (4.1.1). We also compare the fixed effect estimates from the different models (4.1.2). The next subsection involves exploring the random (4.2.1) and fixed effects (4.2.2) of the preferred model. We then explore the spatial distribution of all faults at the MSOA scale (4.2.3). The preferred model was then applied to the top 5 most common network fault types, to assess whether any forecast uncertainty for all network faults is associated with a specific fault type (4.3).

4.1 Model estimation results

4.1.1 Model fit

We begin by assessing the two time-series models. For all forecasted months, the model fit of the AR1 model was higher than the RW1 specification, as indicated by having the smallest DIC value (Table 4). The AR1 model generally had a higher forecast accuracy and the model took substantially less time to run. The only anomaly to the pattern was the forecasted month January, where the RW1 had the smallest residual. This was because the observed count of faults deviated away from the trend of decreasing faults from November to December 2017. Although, both models under-estimated the number of faults, the residual for the RW1 was slightly lower. We therefore conclude this pattern an anomaly and based on these results, we select the AR1 as the temporal prior for both spatial models.

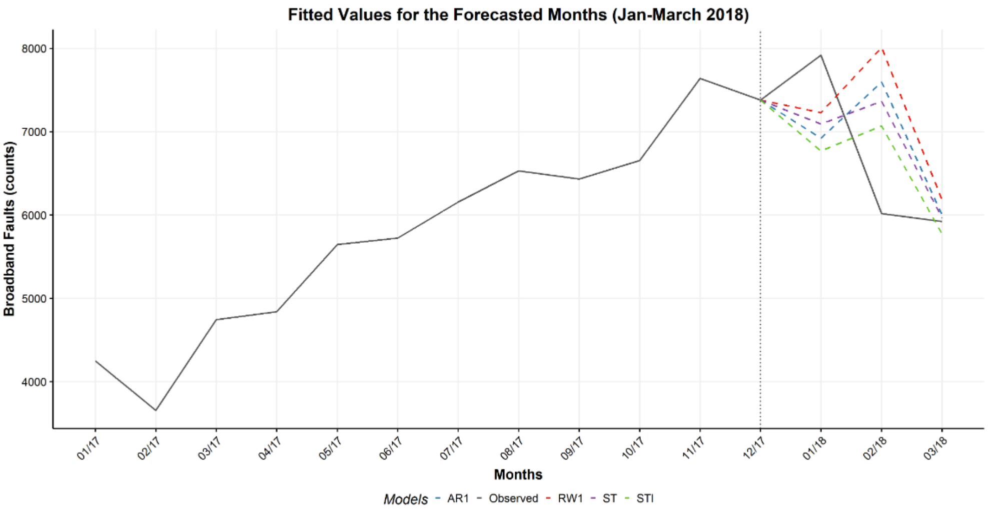


Figure 1. *Comparison of the fitted faults for the North West study area, by each model forecast specifications*

Including the spatial effect significantly increased the model fit for all forecasted months (Figure 1) and had the highest forecast accuracy for March in addition to having the highest average forecast accuracy (Table 4). Whilst the model run-time of the ST model was higher than the AR1 specification, it was comparable to the RW1 one. Furthermore, adding the spatial effect allows the spatial distribution of forecasted faults at the MSOA level to be analysed, allowing further spatial analysis. For the fourth model specification, including the STI term significantly increased the model fit further and had the best model fit for all models (Table 4). However, this model had the lowest average forecast accuracy because the average residual was the furthest from zero. Additionally, the model run time for this specification was significantly the highest because it had the highest number of effective parameters.

Table 4. Forecasting accuracy for all Model Specifications

	2017		2018											
	Nov	*Dec*	*Jan*				*Feb*				*Mar*			
			AR1	**RW1**	**ST**	**STI**	**AR1**	**RW1**	**ST**	**STI**	**AR1**	**RW1**	**ST**	**STI**
Observed	7,641	7,380	7,920				6,016				5,923			
Predicted			6,923	7,228	7,091	6,770	7,597	8,011	7,366	7,070	6,000	6,181	5,966	5,771
Residual (%)			-12.58	-8.74	-10.46	-14.52	26.28	33.15	22.45	17.52	1.30	4.36	0.73	-2.56
DIC			58,270	58,270	42,278	40,407	63,876	63,876	46,228	44,041	68,872	68,872	49,998	47,473
PD			25	25	652	3,069	26	26	655	3,359	27	27	658	3,654
Model run-time (secs)			33	198	209	1,116	44	188	216	1,243	45	236	232	1,460

* Observations for November and December are included for explanatory purposes only

4.1.2 Fixed Effects

We are also interested in how the parameters for each of the fixed effects between the different models. The widths of the 95% Credible Intervals (CI) for spatial models are significantly wider than the temporal models (Figure 2). This is because the spatial models for space also account for the extra residual component. In other words, there is higher uncertainty in the association between faults and covariates when space is considered. For example, education deprivation was significant for the AR1 and RW1 models but not significant for spatial ones because their 95% CI do not include zero.

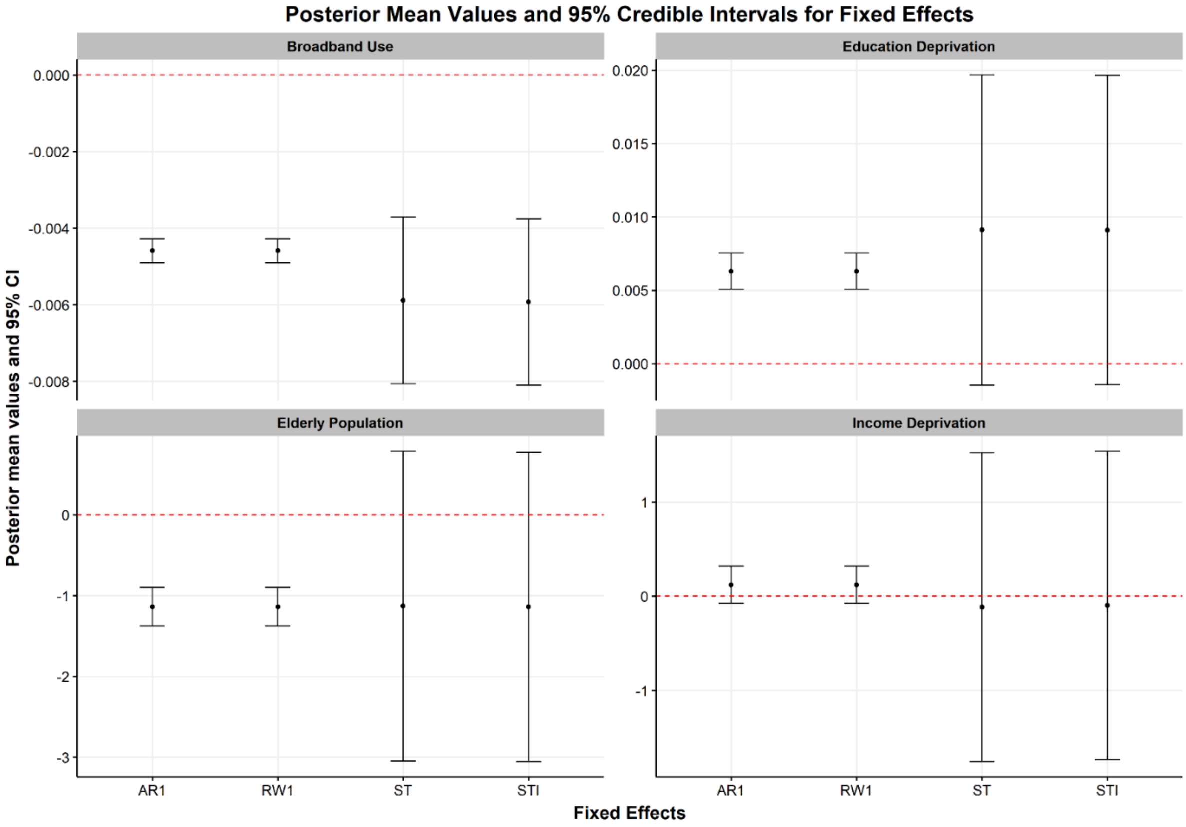


***Figure 2**. Comparison of the fixed effects (continuous covariates) for the different forecast model specifications for March 2018*

Overall, the results supports the value of incorporating the spatial effect using the BYM prior in addition to the main temporal component. Whilst the STI model had the highest model fit, the forecast accuracy was highest for the ST model. Therefore the additional computational demand of the STI model is not valuable and we prefer the less complex ST Model. The following sections explore the ST model further with the random and fixed effects and spatial distribution of the 662 MSOAs in our study area.

4.2 Model estimations for the Space-time model

4.2.1 Random Effects

In this section, we further explore the random effects for the ST model. By default, the posterior summaries of the precision τ are given for the random effects but we calculated the posterior standard deviation σ because they are easier to interpret, using the “Brinla” package (Wang, et al., 2018). We also calculated the proportion of variance explained by each of the random effects using the package “INLAOutputs”, so that both the structured and unstructured spatial components can be directly compared (Blangiardo & Cameletti, 2015).

The results of this study illustrate that there is strong TD in network faults because the highest proportion of variance is explained by this component (Table 5). As expected, the temporal unstructured effect explains none of the variance in faults for North West England. Unexpectedly, the unstructured spatial component explains a higher proportion of variance than the structured one. Furthermore, there is a higher count of MSOAs that are not significant because the 95% CI include zero (Figure 3). Overall, most of the variance in faults is captured by the time-series prior but the spatial component is still significant.

Table 5. *Proportion of Variance Explained by the Random Effects Components in the ST model*

	Forecasted Months		
Random Effect	**Jan**	**Feb**	**Mar**
Spatially Unstructured	0.284	0.304	0.296
Spatially Structured	0.001	0.001	0.001
Temporal Structured	0.039	0.043	0.038
Temporal Unstructured	0.000	0.000	0.000
Rho	0.676	0.651	0.666

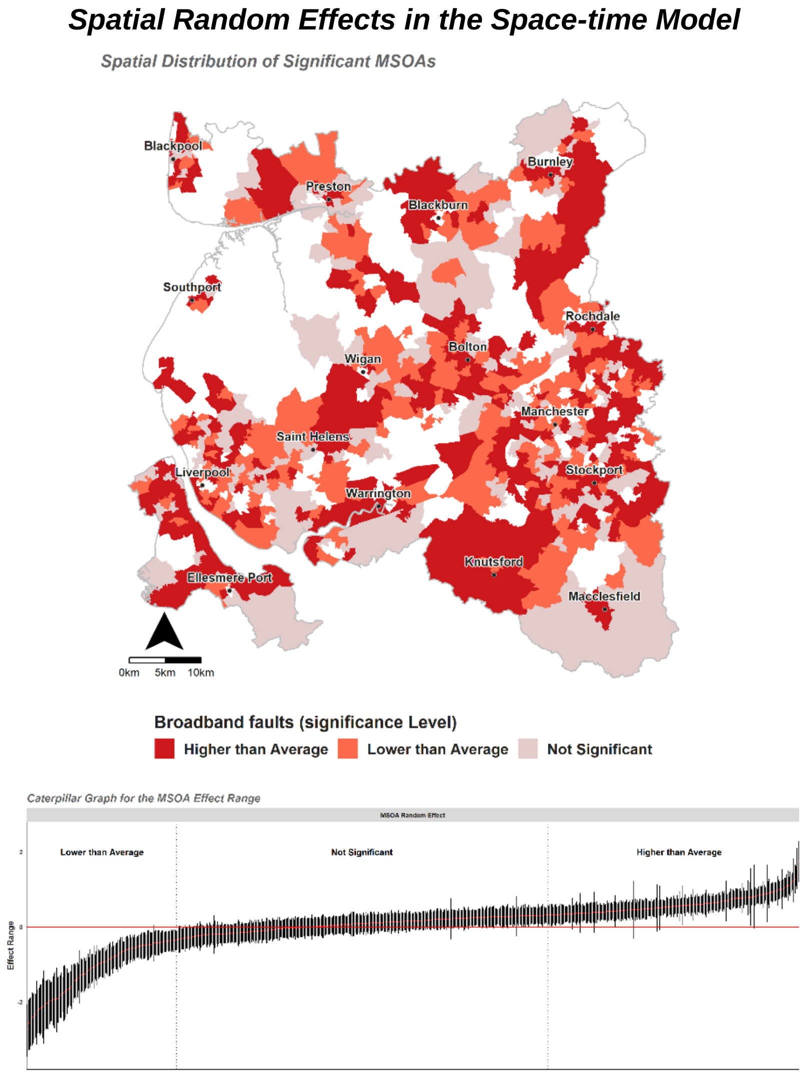


Figure 3. Map (top) and Caterpillar graph (bottom) to visualise the significance of each of the areal units (MSOAs) in the Space-time model

4.2.2 Fixed Effects

We included several covariates to measure the association of network faults with inequality and socio-economic factors. The results show that several covariates are significant predictors of faults, indicated as those where the 95% CI does not include zero (Table 6). Although the association between broadband usage and faults was not as strong as other covariates, the effect on faults is still significant. Moreover, the variability or standard deviation for this effect is very low and is significant for all model specifications in this study (Figure 2). Overall, MSOAs that have higher broadband usage exhibit lower faults, holding everything else constant.

The proportion of elderly had a negative association with faults (Table 6). In other words, MSOAs that have a higher proportion of elderly population have fewer faults. However, this effect is not significant because the 95% CI include zero. Moreover, this covariate had the highest variation in the association with faults, defined by a high standard deviation. Interestingly, the proportion of elderly population was significant for the two models accounting only for TD and not for the spatial models. This is expected after accounting for the residual SA component (Table 7). There is a mixed association between income deprivation and reported faults for the forecasted months using the ST model (Table 6). For the forecasted month January, this relationship is positive but for February March there is a negative relationship. Moreover, this effect was not significant for all forecasted months because the 95% CI includes zero. Furthermore, the effects of income deprivation were only significant for the forecasted month January in the model specifications accounting for TD only (Table 7). The effect of education deprivation on faults was a positive one (Table 6). In other words, as faults increase levels of education decrease. However, this covariate is not significant after controlling for the other fixed effects because the 95% CI do not include zero.

The Passive and Uncommitted IUC group had the highest positive effect on network faults out of the IUC groups (Table 6). This relationship was also significant because the 95% CI do not include zero. This group represent blue collar workers, who use the internet less than other IUC groups, but they use the internet during busy times and their usage is high bandwidth, consisting of social networking, gaming and online shopping. Therefore, they are more likely to feel the effects of bandwidth restrictions and more therefore more likely to report a fault or service beyond specified in their contract. The E-Withdrawn IUC group also had a high positive effect on faults (Table 6). This effect was also significant because the 95% CI does not include zero. Interestingly, the socio-economic profile of this group is characterized by less affluent mixed ethnic groups and is associated by more deprived areas in the outer city (Alexiou & Singleton, 2018). Both IUC groups are characterized by individuals who are less engaged with the internet.

The Settled Offline Communities group had the smallest effect on faults out of the IUC groups which was negative (Table 6). This is expected because this represents elderly White British, individuals who rarely access or do not access the internet. Interestingly, the second smallest effect on faults by IUC group is represented by the E-Professional group. This group is characterized by young urban professionals, who represent the highest levels of internet engagement. The E-Professional group is also highly educated and experienced users of the internet. However, both groups are not significant because the 95% CI contain zero.

Table 6. Space-time model Estimation results

	Jan				*Feb*				*Mar*			
Variables	**Mean**	**SD**	**2.5%**	**97.5%**	**Mean**	**SD**	**2.5%**	**97.50%**	**Mean**	**SD**	**2.5%**	**97.5%**
Intercept	2.286	0.411	1.476	3.093	2.276	0.424	1.440	3.109	2.265	0.415	1.444	3.075
E-Professionals	0.135	0.326	- 0.505	0.776	0.154	0.327	- 0.488	0.796	0.117	0.328	- 0.527	0.762
E-Veterans	**0.636**	0.302	0.045	1.230	**0.669**	0.303	0.075	1.264	**0.650**	0.304	0.054	1.248
Youthful Urban Fringe	0.361	0.373	- 0.371	1.093	0.403	0.374	- 0.331	1.138	0.371	0.376	- 0.366	1.109
E-Rational Utilitarian’s	0.302	0.312	- 0.310	0.916	0.330	0.313	- 0.284	0.945	0.293	0.314	- 0.323	0.911
E-Mainstream	**0.863**	0.295	0.286	1.442	**0.886**	0.296	0.307	1.467	**0.861**	0.297	0.280	1.444
Passive and Uncommitted	**0.971**	0.298	0.387	1.557	**0.992**	0.299	0.406	1.580	**0.968**	0.300	0.380	1.558
Digital Seniors	**0.754**	0.324	0.119	1.390	**0.777**	0.325	0.140	1.416	**0.750**	0.326	0.110	1.391
Settled Offline Communities	- 0.065	0.491	- 1.029	0.898	- 0.044	0.493	- 1.012	0.923	- 0.099	0.495	- 1.072	0.873
E-Withdrawn	**0.894**	0.315	0.277	1.512	**0.907**	0.316	0.288	1.528	**0.882**	0.317	0.260	1.505
Education Deprivation	0.009	0.005	- 0.002	0.019	0.009	0.005	- 0.002	0.019	0.009	0.005	- 0.002	0.020
Income Deprivation	0.047	0.825	- 1.574	1.666	- 0.001	0.830	- 1.631	1.626	- 0.115	0.835	- 1.755	1.523
Elderly Population	- 1.104	0.966	- 3.001	0.790	- 1.134	0.970	- 3.040	0.770	- 1.127	0.976	- 3.045	0.788
Broadband Use	**- 0.006**	0.001	- 0.008	- 0.004	**- 0.006**	0.001	- 0.008	- 0.004	**- 0.006**	0.001	- 0.008	- 0.004
Spatial (σ) - unstructured	0.706	0.018	0.674	0.745	0.732	0.024	0.690	0.782	0.737	0.023	0.693	0.783
Spatial (σ) - structured	0.034	0.021	0.012	0.090	0.035	0.021	0.012	0.090	0.026	0.012	0.010	0.058
Temporal (σ) - structured	0.232	0.081	0.132	0.443	0.246	0.087	0.138	0.471	0.238	0.075	0.139	0.428
Rho for month	0.848	0.089	0.641	0.973	0.881	0.070	0.714	0.977	0.822	0.095	0.601	0.958
Temporal (σ) - unstructured	0.010	0.007	0.004	0.029	0.010	0.007	0.004	0.029	0.011	0.007	0.004	0.028
DIC	42,278				46,228				49,998			
PD	652				655				658			

Note: SD and σ represents standard deviation, bold represents covariates that are important at the 95% credible interval; DIC represents Deviance Information Criterion, PD represents Effective Number of Parameters

Table 7 Income Deprivation and Proportion of Elderly Population Covariates, by Model Type

	Jan				*Feb*				*Mar*			
	Mean	**SD**	**2.5%**	**97.5%**	**Mean**	**SD**	**2.5%**	**97.5%**	**Mean**	**SD**	**2.5%**	**97.5%**
						RW1						
Income Deprivation	**0.279**	0.110	0.064	0.494	0.199	0.104	- 0.005	0.403	0.121	0.100	- 0.076	0.317
Elderly Population	**- 1.055**	0.133	- 1.317	- 0.793	**- 1.140**	0.126	- 1.388	- 0.893	**- 1.136**	0.122	- 1.375	- 0.898
						AR1						
Income Deprivation	**0.279**	0.110	0.064	0.494	0.199	0.104	- 0.005	0.403	0.121	0.100	- 0.076	0.317
Elderly Population	**- 1.055**	0.133	- 1.317	- 0.793	**- 1.140**	0.126	- 1.388	- 0.893	**- 1.136**	0.122	- 1.375	- 0.898
						ST						
Income Deprivation	0.047	0.825	- 1.574	1.666	- 0.001	0.830	- 1.631	1.626	- 0.115	0.835	- 1.755	1.523
Elderly Population	- 1.104	0.966	- 3.001	0.790	- 1.134	0.970	- 3.040	0.770	- 1.127	0.976	- 3.045	0.788
						STI						
Income Deprivation	0.066	0.830	- 1.566	1.695	0.022	0.831	- 1.611	1.652	- 0.096	0.833	- 1.733	1.539
Elderly Population	- 1.131	0.972	- 3.040	0.775	- 1.151	0.972	- 3.061	0.756	- 1.137	0.975	- 3.053	0.775

**SD represents standard deviation and bold estimates represent those that are significant*

4.2.3 Spatial distribution of faults

This section describes the spatial distribution of faults at the MSOA scale in the study area. Firstly, faults tend to be higher in MSOAs distributed in urban areas than rural ones. However, this pattern is not straightforward and there is a clear “island effect” in city, where central business district and inner cities of Liverpool and Manchester have low faults, surrounded by MSOAs in the inner suburbs with higher faults (Figures 4:5). However, this structure is not present in MSOAs within towns in this study area.

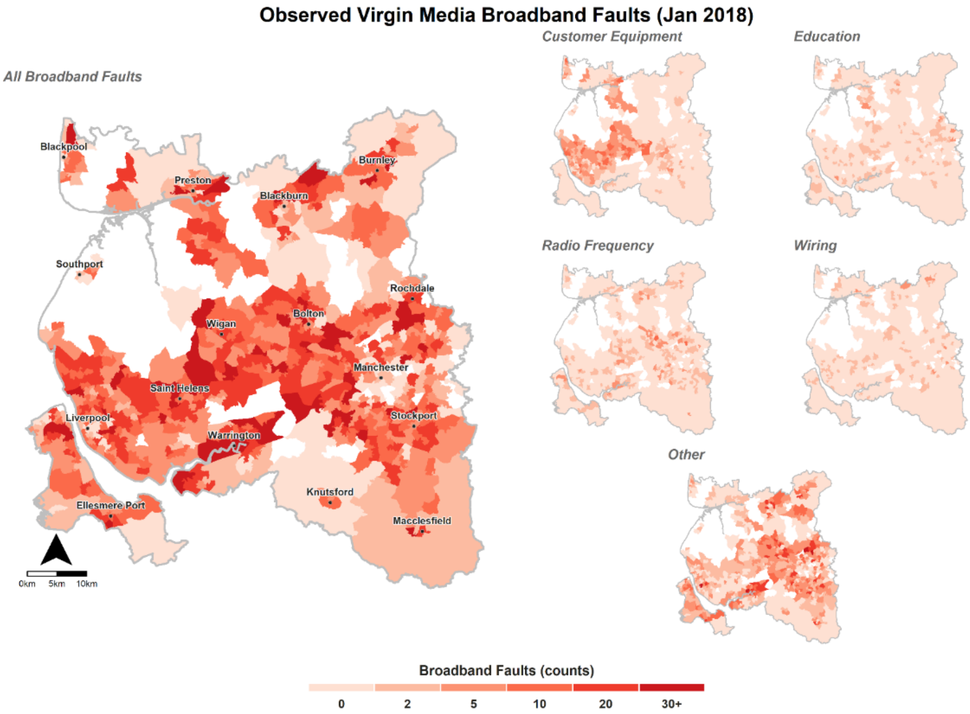


Figure 4. All observed (left) and by type network faults North West England (Jan 2018)

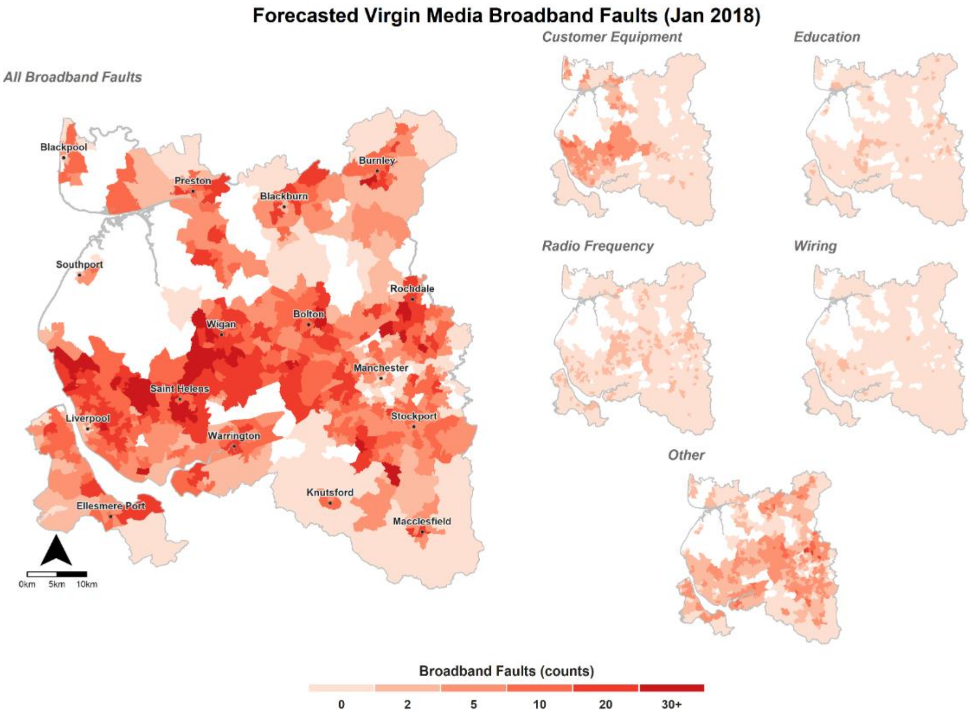


Figure 5. All forecasted (left) and by type network faults North West England (Jan 2018)

As discussed, the forecast accuracy for February 2018 using all model specifications was the lowest. This is because the forecast was based on the previous months anomalous network fault count. Generally, there is a clear pattern of higher forecasted than observed faults across the whole study area (Figure 6:7). Unsurprisingly, the spatial distribution of observed and forecasted MSOA level faults in March 2018 matched the closest (Figure 8:9). For all months, observed faults tend to be highest for the Merseyside County.

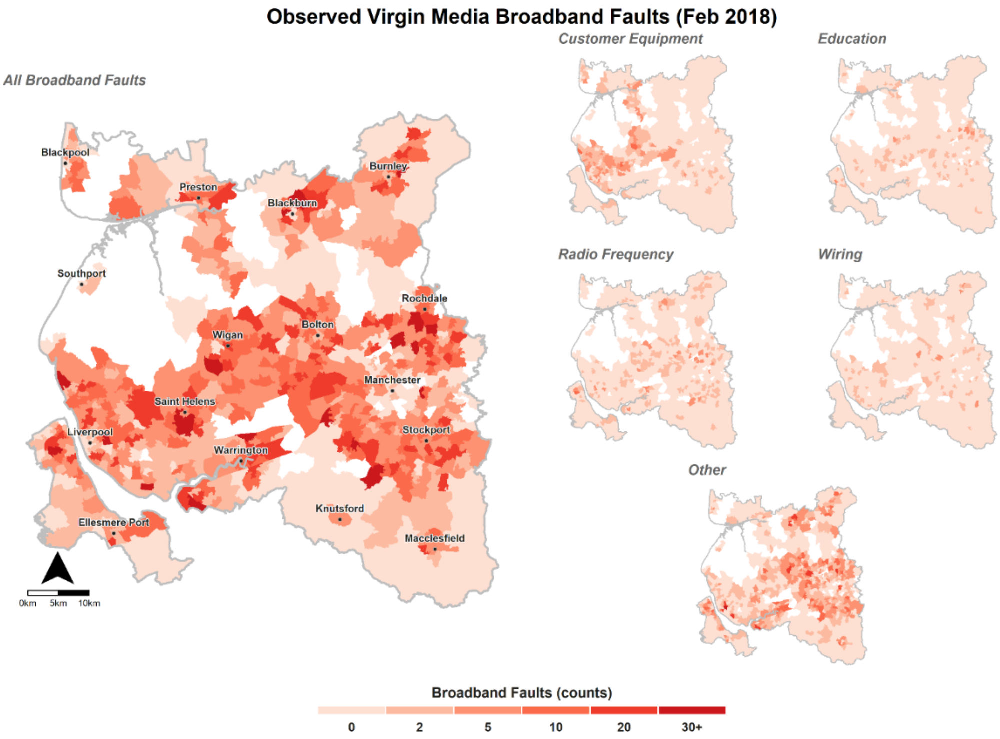


Figure 6. All observed (left) and by type network faults North West England (Feb 2018)

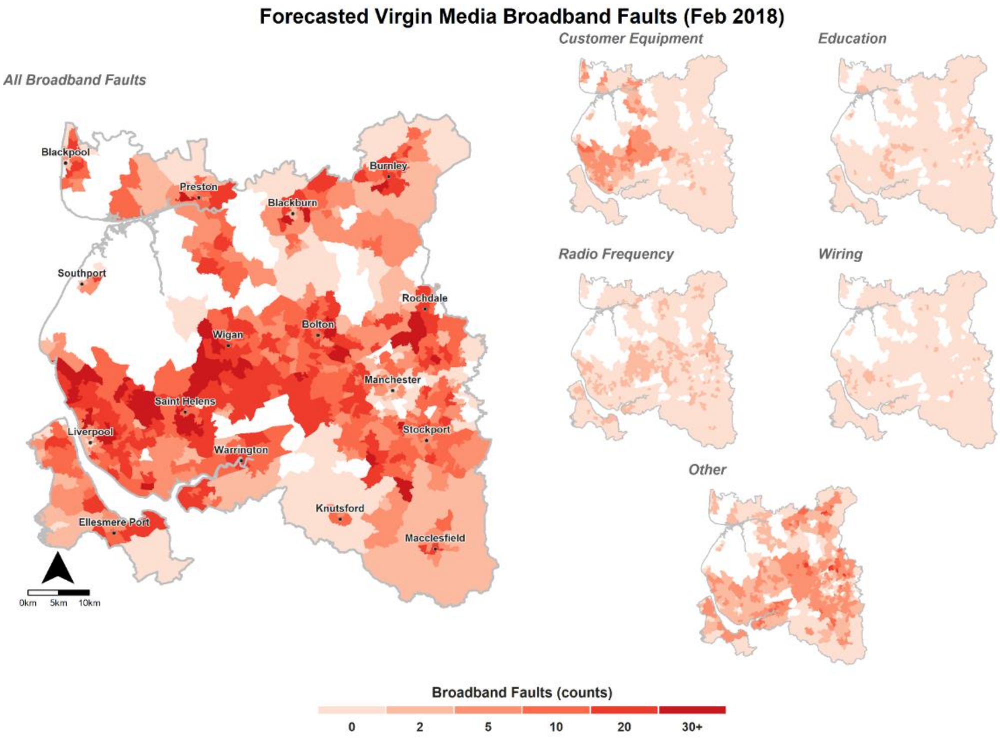


Figure 7. All forecasted (left) and by type network faults North West England (Feb 2018)

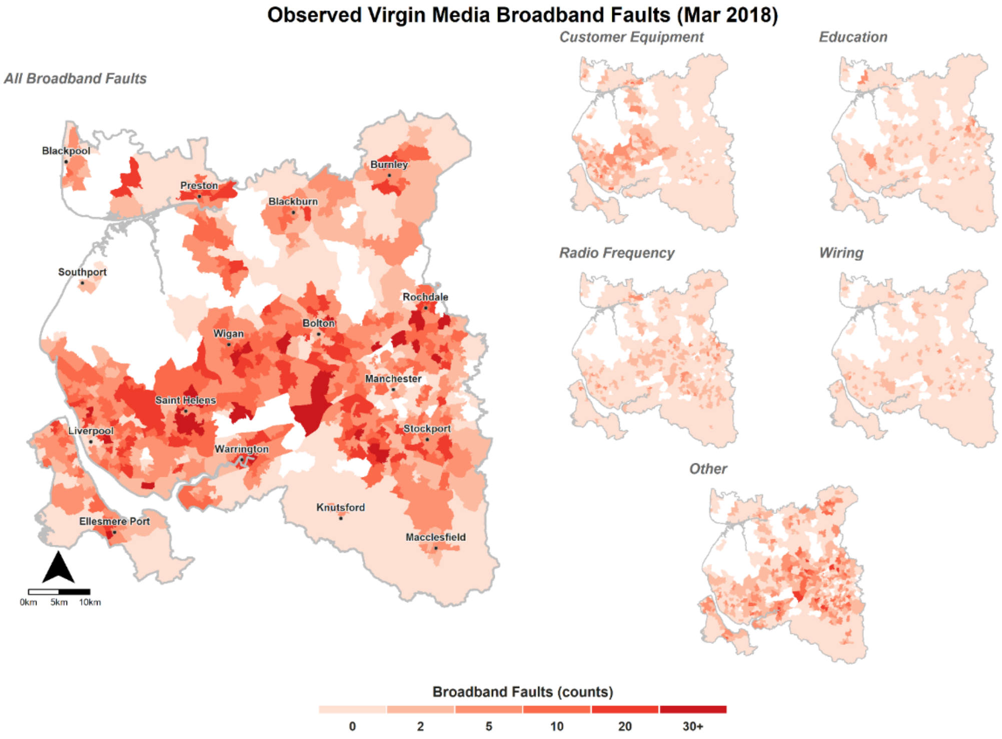


Figure 8. All observed (left) and by type network faults North West England (Mar 2018)

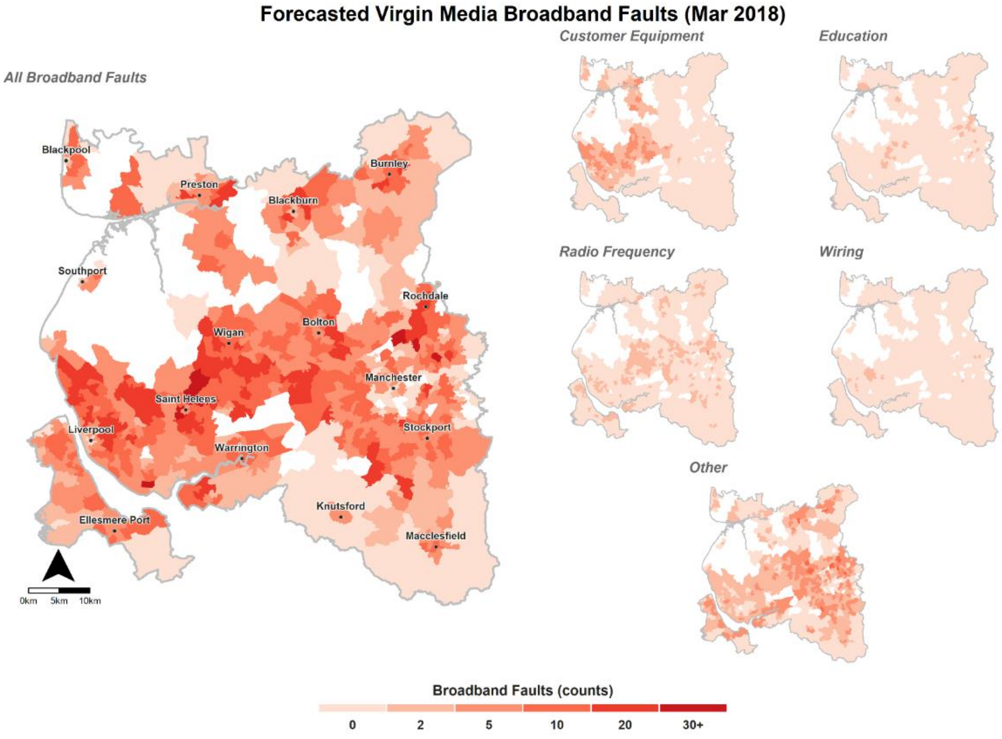
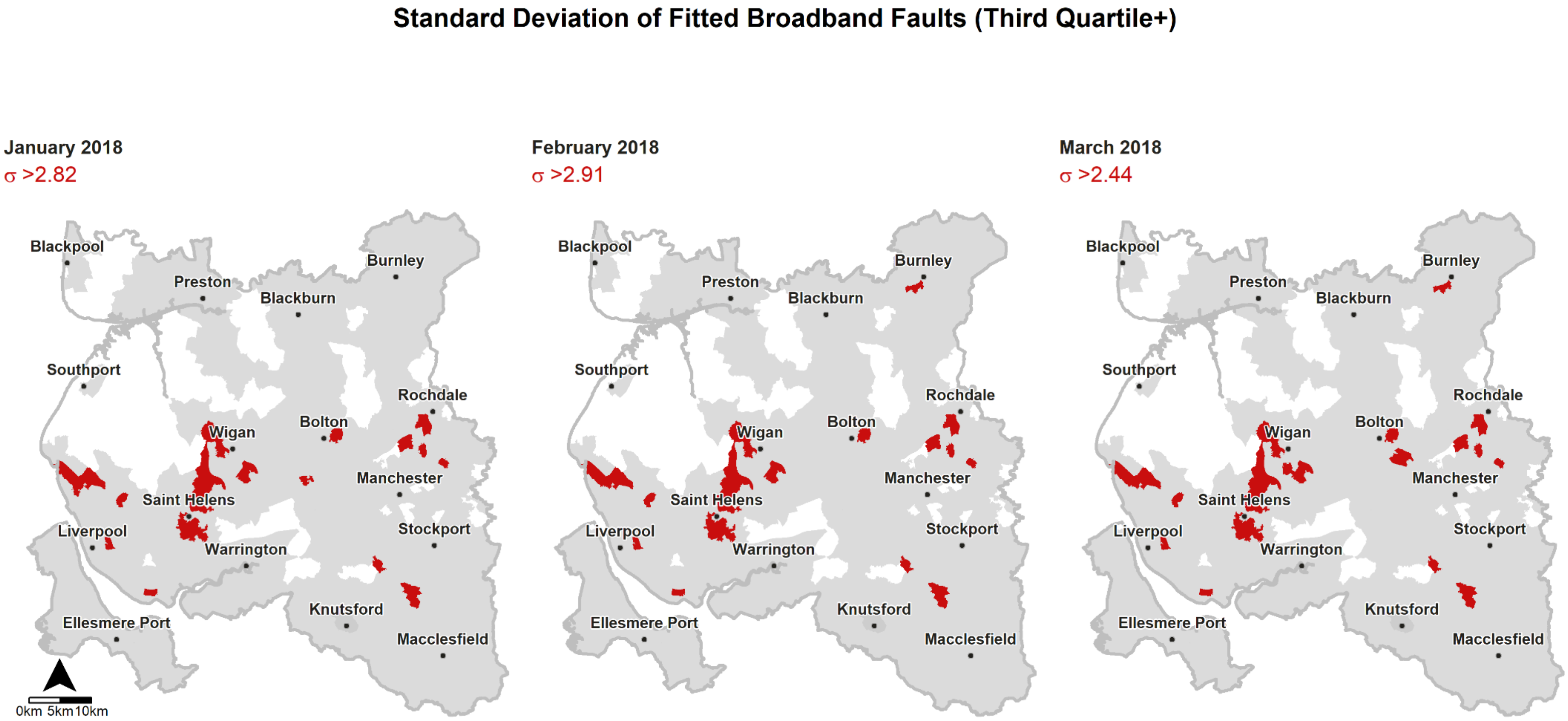


Figure 9. All forecasted (left) and by type network faults North West England (Mar 2018)

Forecast uncertainty was also calculated for the fitted values of each of the forecasted months, defined as MSOAs with a standard deviation in the third quantile or higher (Figure 10). This was essential to identify which MSOAs have the highest variance away from the mean. As expected, model uncertainty was lowest for the forecasted month March and highest for February. Some MSOAs had high model uncertainty for all forecasted months, included those near Saint Helens, Wigan, South Liverpool and in-between Rochdale and Manchester. This is further supported by Figure 11, which shows that the highest probability of MSOAS exceeding the threshold of 15 faults matched a similar spatial distribution as Figure 10.



Standard Deviation of Fitted Broadband Faults (Third Quartile+)

January 2018
σ >2.82

February 2018
σ >2.91

March 2018
σ >2.44

Figure 10. Areal Units (MSOAs) with the highest forecast uncertainty for the Space-time model

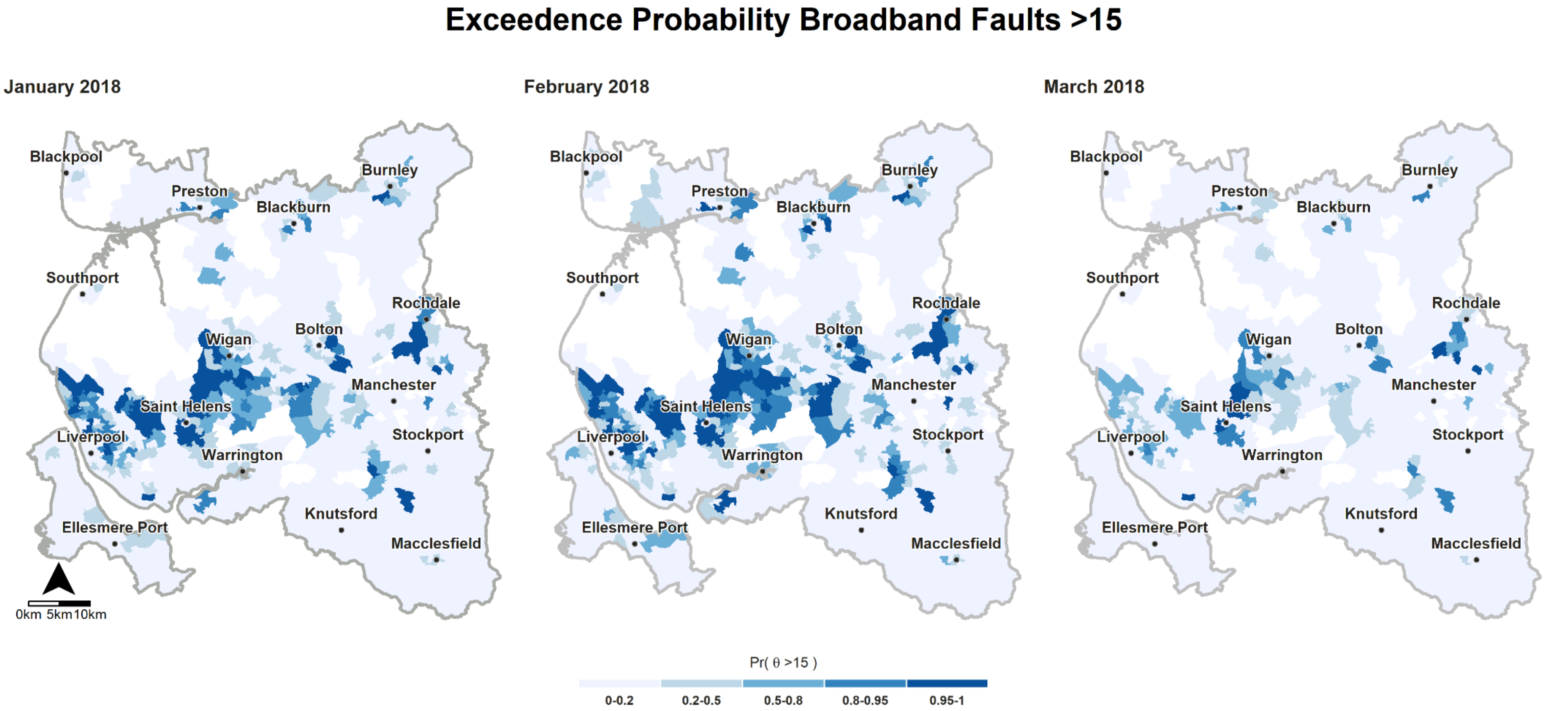


Figure 11. Probability of areal units (MSOAs) with forecasted network faults exceeding 15 faults or the third quantile for the Space-time model

4.3 Model estimation results for the Space-time model, by broadband fault types

In this section, we present the results from the analysis of network fault types. Due to the anomalous observed faults in January 2018, the model significantly overestimated the forecast for February 2018 (Table 8). By applying the selected ST model to the top 5 network fault types, we found that all network fault types were overestimated but most significant for Customer Equipment faults, with the highest residual. This suggests that the low forecast accuracy for all network faults in February 2018 is associated a significantly higher count of Customer Equipment faults in January. Furthermore, this fault type was associated with a higher count of faults in Merseyside (Figure 6:7).

Table 8. Space-time model estimation results: Top 5 Broadband Network Fault Types

	Customer Equipment	Education	Radio Frequency	Wiring	Other
Jan					
Observed	1,645.00	703.00	1,175.00	689.00	3,653.00
Predicted	1,505.75	677.70	938.11	594.00	3,162.61
Residual (%)	- 8.47	- 3.60	- 20.16	- 13.79	- 13.42
DIC	21,392.65	19,000.35	22,903.48	18,331.82	33,860.09
PD	574.31	468.47	515.59	433.66	622.57
Feb					
Observed	1,067.00	668.00	1,035.00	571.00	2,699.00
Predicted	1,573.00	708.26	1,055.33	633.61	3,373.94
Residual (%)	47.42	6.03	1.96	10.97	25.01
DIC	23,411.19	20,966.11	25,187.13	20,117.45	37,082.57
PD	578.80	473.44	529.57	444.26	627.61
Mar					
Observed	1,042.00	951.00	1,080.00	608.00	2,576.00
Predicted	1,082.44	632.90	974.63	560.74	2,651.99
Residual (%)	3.88	- 33.45	- 9.76	- 7.77	2.95
DIC	25,186.67	22,679.08	27,300.12	21,783.50	40,065.42
PD	582.84	484.51	539.06	453.65	631.85

4.4 Summary

The main findings of this study is that the ST model had the highest average accuracy for forecasting Virgin Media network faults (Table 4). Including the STI term improved the model fit further but did not increase the forecast accuracy. Including the spatial effect allows an analysis of the spatial distribution of network faults at a much finer scale than conventional time-series models. Although including space with the stationary time-series prior improved the forecast accuracy, the highest proportion of variance was still explained by the temporal prior. Whilst the model fit of the STI model was the highest, the forecast accuracy was the lowest for this model. These results supports the value of incorporating the spatial effect in addition to the main temporal component.

For the forecasted month February 2018, all models overestimated network faults due to the anamalous faults in January (Table 4). This highlights the stochastic nature of faults and the autoregressive characteristic of t_{-1}, where forecasts are based on the previous months broadband fault counts. The analysis of faults types showed that this was related to Customer Equipment faults because the faults decreased significantly from January to February (Table 8). Furthrmore, there was a clear spatial distirbution of Customer Equipment faults being higher in the Merseyside area (Figure 4:9).

The significance of the relationship between some fixed covariates changed by adding the spatial effect. This included the proportion of elderly and income deprivation, however broadband usage was significant for all model specifications This was because the 95% CI was wider for the spatial models, for this residual component. Out of the IUC groups, network faults and both the Passive and Uncommitted and E-Withdrawn groups had the highest positive association, which was significant (Table 6). These groups represent blue-collar workers and those in more deprived neighbourhoods. In contrast, the Settled Offline Communities and E-Professional group had the smallest effect on faults. Interestingly, these represent geodemographic groups who have the smallest and highest engagement with the internet but these effects were not significant.

5 Discussion

We begin this section by discussing the main findings of this study and how they relate to the research aims and objectives (5.1). The second subsection (5.2) draws on these results and engages these findings with the wider literature. The next section explores the significance of results, as well as the limitations of this study (5.3). We finalise by providing a summary of this discussion and provide some recommendations for future research in this field (5.4).

5.1 Main study findings

We begin by restating the study aims and objectives. The aim of this study was to extend the time-series forecasting literature for network faults, by incorporating space. The objectives of this study were to;

1. To evaluate whether incorporating space with conventional time-series priors gives a higher forecast accuracy of network faults, than just the time-series priors.
2. Explore whether incorporating an additional spatio-temporal interaction temr in spatio-temporal models further improve the forecast accuracy of network faults
3. Assess which broadband inequality and socio-economic factors are most related to network faults.
4. To identify whether there is a spatial pattern of network faults in North-West England and how they change over time.

Both spatial models had a significantly higher model fit than the temporal priors only (Table 4) In other words the spatial models fitted the faults more closely than the AR1 and RW1 prior. Whilst the STI model had the best model fit for all forecasted months, the ST model had the highest forecast accuracy. Nevertheless, the ST model didn't have the highest forecast accuracy for all forecasted months, due to the anomalous January 2018 faults but had the highest average forecast accuracy. Interestingly, most of the variance in faults was captured through the main temporal prior but the variance captured by the BYM prior was still significant (Table 5). We must also consider the principle of parsimony or the model with the smallest number of parameters that adequately represent the underlying time series (Chatfield, 1996). This is because a simpler model allows the possibility to scale up the analysis to a larger study area. Overall, the results of this study suggest the value of incorporating space into the traditional time-series framework.

We applied the ST model to network fault types to identify whether the anomalous January and February 2018 observations are associated with this pattern. In other words, to identify whether any difference between the observed and forecasted faults is associated with a network fault type. The most significant finding of this analysis was that Customer Equipment faults was significantly higher for January 2018 compared to the other months. Therefore, we conclude the anomalous results in January and February 2018 was associated with this network fault type.

The proportion of elderly represented a negative association with network faults but was not significant. Likewise, the Settled Offline Community IUC group, which also represents retired White British, who reside in semi-rural areas have a negative but smaller association with faults. However, these associations was not significant because the 95% CI included zero. Conversely, the Passive and Uncommitted group, which represents semi-skilled and blue-collar occupations who live in the suburbs had the positive association with faults, which was significant. Interestingly, broadband usage was significant for all model specifications, despite having a small, negative association with network faults (Table 6). In contrast, income deprivation had a small positive association, which was not significant for all model specifications.

Another advantage of incorporating space with time-series models is analysing the spatial distribution of network faults at a finer scale than the regional scale. For example, there is a clear pattern of Customer Equipment related faults being significantly higher in the Merseyside area relative to the rest of the study area (Figures 4:9). This could be due to regional differences in broadband packages offered to customers and associated equipment or a way in which the faults are reported by the engineers. In contrast, the other fault types have less of a clear spatial pattern. We also calculated and visualized which MSOAs had the highest probability of MSOAs exceeding 15 broadband or the third quantile. For all forecasted months, MSOAs in urban areas tend to have a higher number of faults than rural ones. Specifically, MSOAs surrounding MSOAS in the outer city tend to have the highest number of observed faults. The inner cities of Liverpool and Manchester had significantly lower faults than surrounding MSOAs, representing an island effect.

The e-Professionals and Youthful Urban Fringe IUC groups make up most socio-economic groups in the inner city (Figure 12). These represent a young socio-economic group, who are typically students and young professionals who are actively engaged with the internet. As expected, there is a low association between these groups and faults and this association was not significant. Another reason why faults in the inner city are lower is because network infrastructure is likely to be the most modern and subject to more routine maintenance to support the central business district. Moreover, Universities which are located in inner cities tend to have excellent network infrastructure and so MSOAs surrounding Universities piggy-back onto the supporting University network infrastructure. However, it’s unclear whether it’s the characteristics of these socio-economic groups that influence faults or whether it’s the Geography that indirectly influences network faults. By this we mean that, certain areas attract certain socio-economic groups, whether this is related to job opportunities, financial restrictions, and closeness to certain ethnic groups or age structures.

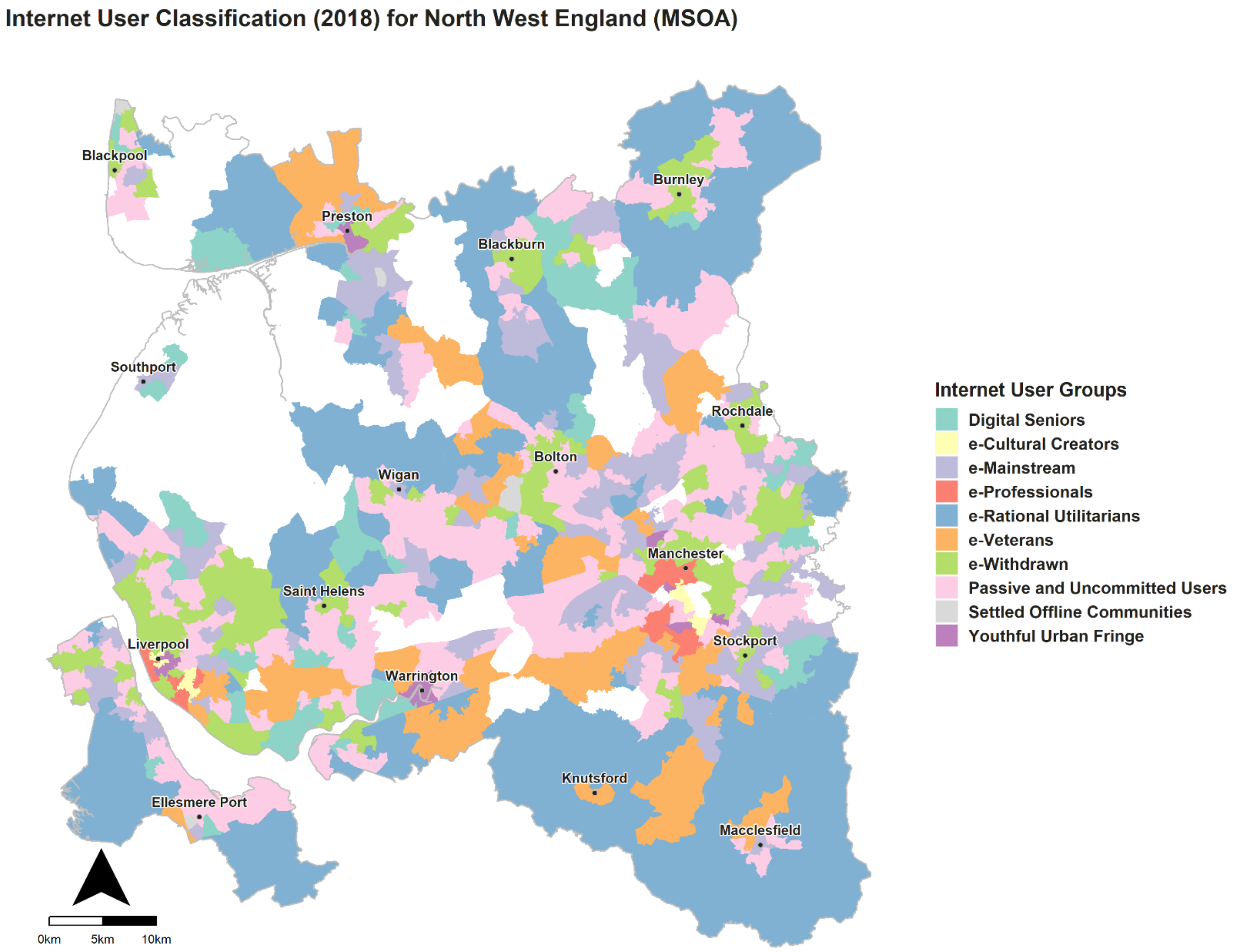


Figure 12. Internet User Classification (2018) for North West England, aggregated to the Middle Layer Super Output Area (MSOA)

5.2 Comparison of the main study findings to the wider literature

The results of this study support the literature in incorporating space with time because the ST model had the highest forecast accuracy. However, the results also contradict the wider literature on STI models, which includes an additional spatial interaction term (Knorr-Held 2000). Whilst the STI models had the best model fit, they had the lowest forecast accuracy. Model selection in these studies are generally based on just the model fit of known observations. Instead, we are interested in achieving the highest forecast accuracy of network faults with unknown quantities or one step ahead forecasting. The results of this study also support the results of (Deljac, et al., 2011), where monthly faults was best forecasted using a stationary time process.

By including several covariates from the literature, we explored whether there was an association with network faults. Firstly, broadband usage was significant for all model specifications, with a negative association (Figure 3). In other words, as the number of faults increase, broadband usage decreases. Furthermore, the association between levels of income and faults was only significant for the forecasted month January 2018in the model specifications accounting for TD only (Table 7). Interestingly, the association between the proportions of elderly was significant for just the temporal models but these associations was not significant for the models accounting for space.

The previous subsection (5.1) highlighted there is a spatial inequality of network faults, which supports the literature. For example, network faults tends to be higher in urban areas than rural ones, supporting the digital divide in the literature. As already discussed, the inner city is characterized by a higher proportion of young students and working professionals, where there is a lower count of faults, supported by an excellent network infrastructure. In comparison, the outer city is characterized by both E-withdrawn and E-mainstream IUC groups, who represent lower income and an ethnically diverse group, where the network infrastructure is not as well developed. Although spatial inequalities exist, it’s unclear whether it’s the characteristics of socio-economic groups that influence network faults or whether it’s the local Geography or network infrastructure.

5.3 Significance and limitations of this study

The main finding of this study was that incorporating space with conventional time-series priors yields a higher forecast accuracy and model fit than just accounting for TD. The significance of this approach is that it allows one to simultaneously apply one step ahead monthly forecasts to the whole study area, as well as analyse the spatial distribution of faults at a MSOA scale, a significantly finer scale and how they change over time. This spatio-temporal areal modelling approach could be extended to other applications of forecasting.

One of the limitations of this study is that this methodology was only applied over a short time period of three forecasted months, so there is a need to conduct the same methodology over a longer period. More specifically, these results represent the seasonality of winter, which are likely to be significantly different from summer. Additionally, all the data sources in this study was aggregated to the MSOA scale, which was selected as a compromise between level of detail and time taken to run the most complex STI model. Although aggregation was an important part of the research design, to ensure customer confidentiality, information is lost. Therefore, future studies should apply smaller areal unit scale such as the LSOA level in ST modelling. On the other hand, it would also be useful for Virgin Media to apply the ST model with MSOAs units over a larger study area such as North England, to analyse the distribution of faults over a larger scale.

5.4 Summary

This study has shown the value of incorporating space with conventional time-series methods. However, it’s worth noting that this requires the forecaster to understand GIS and spatial analysis. For example, a misspecification of the neighbourhood structure for the BYM prior could yield very different estimates. It would be interesting for future studies of forecasting faults to experiment with different priors and specifically the spatial one because the spatial structured component did not explain much of the variance in network faults. Furthermore, (Lee & Mitchell, 2012) proposed a method for capturing more localized spatial structures as an alternative to the single global level of spatial smoothing but at the time of this study, only a limited number of spatial priors can be implemented in INLA. We also recommend Virgin Media to conduct a separate analysis, to examine why Customer Equipment network faults were significantly higher for January 2018 than the other forecasted months.

6 Conclusion

This study developed a new method for one step ahead, monthly forecasting of network faults, by extending the conventional time-series approach by combining spatial statistics. Accurate forecasting of broadband faults with a high level of accuracy is important for budget purposes, operational efficiency and minimizing customer churn. We compared the forecast accuracy of four model specifications, two which accounted for just a stationary and non-stationary time process, another included space and another included a STI term (Table 1). We adopted a BHMF to account for the different levels of uncertainty, which was achieved using the computationally efficient R-INLA package.

The results of this study support the value in incorporating space with time together in a hierarchical structure because the ST model had the highest forecast accuracy. However, the results also contradict the wider literature on STI models, which also include a spatio-temporal interaction term (Knorr-Held 2000). Whilst the ST model didn't have the highest forecast accuracy for other months, due to anomalous January 2018 network faults, the model had the highest average forecast accuracy. The less complex ST model also allows for easier interpretation of results and took significantly less time to run, than the STI model.

We then applied the ST model to the top 5 network fault types over the same study area and period. This was to examine whether any of the low forecast accuracy for February is associated with a specific network fault type. This analysis showed that Customer Equipment faults was significantly higher in January than the other forecasted months, February and March 2018. We concluded that the anomalous results in January 2018 for all broadband faults were associated with this fault type. This could be related to different equipment's used by customers, which are offered as part of different packages or a way in which the faults are reported by the engineers. We conclude that a separate study is required to understand what factors contributed to the differences between observed Customer Equipment faults in January and February 2018.

The incorporation of space also allows the forecaster to identify how the spatial distribution of network faults changes over time at a much finer spatial scale than the common regional level analysis. We calculated that MSOAS with the highest probability of exceeding 15 faults was in Merseyside and Wigan. Similarly, a higher count of Customer Equipment faults are associated with MSOAs in North Liverpool, Saint Helens and Wigan. Moreover, the results support the literature on the digital divide between urban and rural areas, where generally, urban areas have higher faults. However, it was also found that network faults was lower in the inner city, surrounded by MSOAs in outer city and inner suburban areas with higher fault counts.

Gaining insight into the relationship between network faults and different socio-economic groups is significant to identify market demands and the strategic placement of corporate resources to both prevent and mitigate future network faults. This is particularly important in the competitive market of superfast BSPs. As expected, the proportion of elderly had a negative association with faults. In other words, MSOAs with a higher proportion of elderly population have fewer faults. Furthermore, there is a low association between the e-Professionals and Youthful Urban Fringe IUC groups and network faults. These represent areas in the inner city where faults are less likely. Conversely, there was a positive association for the Passive and Uncommitted IUC group and E-Withdrawn groups with network faults. The spatial distribution of these groups are associated with areas in the outer city and inner suburb, which represent low income, ethnically diverse socio-economic groups.

We have presented this forward-thinking methodology to explore whether introducing space can achieve higher forecast accuracies than conventional time-series forecasting of faults. The results support the wider literature on spatio-temporal modelling and highlights the limitations of conventional time-series forecasting approaches. Moreover, such a methodology could be extended to other forecasting. Therefore, based on the findings of this study, we suggest Virgin Media to explore this methodology further and other applications of spatial analysis. Whilst leveraging Geographic information within modelling approaches is still in their infancy, applying spatial analysis could give a competitive advantage over their competitors.

References

Alexiou, A. & Singleton, A., 2018. *Indicators of Internet Use and Engagement: 2018 Internet User Classification User Guide,* Liverpool : Dept of Geography and Planning, University of Liverpool.

Anselin, L. & Griffith, D. A., 1988. Do spatial effects really matter in regression analysis. *Papers in Regional Science,* 65(1), pp. 11-34.

Banjeree, S., Carlin, B. P. & Gelfand, A. E., 2004. *Hierarchical Modelling and Analysis of Spatial Data.* New York: CRC press.

Bernardinelli, L., Clayton, D. & Pascutto , C., 1995. Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine,* 14(1), pp. 2433-2443.

Besag, J., 1974. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society, Series B,* 36(2), pp. 192-236.

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), pp. 1-20.

Best, N. et al., 2001. Ecological regression analysis of environmental benzene exposure and childhood leukemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society,* 164(1), pp. 155-174.

Bivand, R. et al., 2018. *rgdal: Bindings for the Geospatial Data Abstraction Library.* Version 1.3-4: https://www.gdal.org/.

Bivand, R. & Piras, G., 2015. Comparing Implementations of Estimation Methods for Spatial Econometric. *Journal of Statistical Software,* 63(18), pp. 1-36.

Bivand, R. S., Pedesma, E. J. & Gómez-Rubio, V., 2011. *Applied Spatial Data Analysis.* New York: Springer.

Blangiardo, M. & Cameletti, M., 2015. *Spatial and Spatio-temporal Bayesian Models with R-INLA.* Chichester: John Wiley & Sons.

Blangiardo, M., Cameletti, M., Baio, G. & Rue, H., 2013. *A tutorial in spatial and spatio-temporal models with R-INLA.* [Online]
Available at: http://discovery.ucl.ac.uk/1415919/1/Baio_BlaCamBaiRue.pdf
[Accessed 02 07 2018].

Box, G. E. P. & Jenkins, G. W., 1976. *Time Series Analysis: Forecasting and Control.* San Fransisco: Holden-Day.

Breslow, N., Leroux, B. & Platt, R., 1998. Approximate hierarchical modelling of discrete data in epidemiology. *Statistical Methods in Medical Research,* 7(1), pp. 49-62.

Brockwell, P. J. & Davis, R. A., 1996. *Introduction to Time Series and Forecasting.* New York: Springer.

Cairncross, F., 2001. *The Death of Distance 2.0: How the Communications Revolution will Change our Lives.* London: Texere Publishing Limited.

Carlin, B. P. et al., 1999. Spatio-Temporal Hierarchical Models for Analyzing Atlanta Pediatric Asthma ER Visit Rates. In: *Case Studies in Bayesian Statistics.* New York: Springer, pp. 303-320.

Chatfield, C., 1996. Model uncertainty and forecast accuracy. *Journal of Forecasting,* 15(1), pp. 495-508.

Chatfield, C., 2016. *The Analysis of Time Series: An Introduction.* London: CRC Press.

Cressie, N., 1993. *Statistics for Spatial Data.* New York: John Wiley & Sons.

Cressie, N. & Wilke, C. K., 2011. *Statistics for Spatio-temporal Data.* New York: Wiley.

Deljac, Z., Kunstic, M. & Spahija, B., 2011. *Using temporal neural networks to forecasting of broadband network faults.* Split, Croatia, IEEE.

Downes, T. & Greenstein, S., 2005. *Understanding Why Universal Service Obligations May be Unnecessary: The Private Development of Local Internet Access Markets,* Massachussetts: Tuffs University.

Earnest, A. et al., 2007. Evaluating the effect of neighbourhood weight matrices on smoothing properties of conditional autoregressive models. *International Journal of Health Geographies,* 6(54), pp. 1-13.

Elhorst, J. P., 2003. Specification and estimation of spatial panel data models. *International Regional Science Review,* 26(3), pp. 224-268.

Fildes, R. & Kumar, V., 2002. Telecommunications demand forecasting - A review. *International Journal of Forecasting,* 18(4), pp. 489-522.

Gelfand, A. E., Banerjee, S. & Gamerman, D., 2005. Univariate and multivariate dynamic spatial modelling. *Environmetrics,* 16(5), pp. 465-479.

Gelman, A. et al., 2013. *Bayesian Data Analysis.* London: CRC Press.

Grubesic, T. H. & Murray, A. T., 2002. Constructing the divide: Spatial disparities in broadband access. *Papers in Regional Science,,* 81(2), pp. 197-221.

Haworth, J. & Cheng, T., 2012. Non-parametric regression for space-time forecasting under missing data. *Computers, Environment and Urban Systems,* 36(6), pp. 538-550.

Hopkins, M. et al., 1995. A multi-faceted approach to forecasting broadband demand and traffic. *IEEE Communications Magazine,* 33(2), pp. 36-42.

Huang, B., Wu, B. & Barry, M., 2010. Geographically and temporally weighted regression for modelling spatio-temporal variation in house prices. *International Journal of Geographical Information Science,* 24(3), pp. 383-401.

Irvine, K., Gitelman, A. I. & Hoeting, J. A., 2007. Spatial designs and properties of spatial correlations: effects on covariance estimation. *Journal of Agricultural, Biological and Environmental Statistics,* 12(1), pp. 450-469.

Knorr-Held, L., 2000. Bayesian modelling of inseperable space-time variation in disease risk. *Statistics in Medicine,* 19(17), pp. 2555-2567.

Knorr-Held, L. & Besag, J., 1998. Modelling risk from a disease in space and time. *Statistics in Medicine,* 17(18), pp. 2045-2060.

Knorr-Held, L. & Rue, H., 2002. On Block Updating in Markov Random Field Models for Disease Mapping. *Scandanavian Journal of Statistics,* 24(4), pp. 597-614.

Lawson, A., 2013. *Bayesian Disease Mapping: Hierarchical Modelling in Spatial Epidemiology.* Boca Raton: CRC Press.

Lee, D. & Mitchell, R., 2012. Boundary Detection in Disease Mapping Studies. *Biostatistics,* 13(3), pp. 415-426.

Leroux, B., Lei, X. & Breslow, N., 2000. Estimation of disease rates in small areas: A New Mixed Model for Spatial Dependence. In: *Statistical Models in Epidemiology, the Environment and Clinical Trials.* New York: Springer, pp. 179-191.

Ofcom, 2016. *UK Home broadband performance: The performance of fixed-line broadband delivered to UK residential consumers,* London: Office of Communications.

Ofcom, 2017. *Connected Nations,* London: Office of Communications.

ONS, 2017. *Overview of the UK Population: July 2017, Office of National Statistics.* [Online]
Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/overviewoftheukpopulation/july2017
[Accessed 08 08 2018].

Openshaw, S., 1984. The Modifiable Areal Unit Problem. *Concepts and Techniques in Modern Geography,* 38(1), pp. 1-22.

Ozturkmen, Z. A., 2000. Forecasting in the Rapidly Changing Telecommunications Industry: AT&T Experience. *The Journal of Business Forecasting Methods and Systems,* 19(3), pp. 1-11.

Pfeifer, P. E. & Deutsch, S. J., 1980. A three-stage iterative procedure for space–time modelling. *Technometrics,* 22(1), pp. 35-47.

Priestley, S. & Baker, C., 2017. *Superfast broadband Coverage in the UK,* London: House of Commons Library.

R Core Team, 2015. *A langugage and Environment for Statistical Computing, R Foundation for Statistical Computing.* [Online]
Available at: https://www.r-project.org/
[Accessed 01 06 2018].

Rue, H. & Held, L., 2005. *Gaussian Markov Random Fields: Theory and Applications.* Boca Raton: Chapman & Hall.

Rue, H., Martino, S. & Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B,* 71(2), pp. 319-392.

Sandholm, T., 2007. *Autoregressive Time Series Forecasting of Computational Demand.*
[Online]
Available at:

https://www.researchgate.net/profile/Thomas_Sandholm/publication/1756232_Autoregressive_Time_Series_Forecasting_of_Computational_Demand/links/0deec516de739240cc000000/Autoregressive-Time-Series-Forecasting-of-Computational-Demand.pdf?origin=publication_de
[Accessed 19 07 2018].

Schabenberger, O. & Gotway, C. A., 2005. *Statistical Methods for Spatial Analysis.* Boca Raton: Chapman & Hall.

Schrödle, B. & Held, L., 2010. Spatio-temporal disease mapping using INLA. *Environmetrics,* 22(6), pp. 725-734.

Shumway, R. H. & Stoffer, D. S., 2017. *Time Series Analysis and its Applications: With R Examples.* New York: Springer.

Spiegelhalter, D. J., Best, G., Carlin, B. P. & Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B,* 64(4), pp. 583-639.

Stern, H. & Cressie, N. A., 1999. Inference for extremes in disease mapping. In: *Disease Mapping and Risk Assessment for Public Health.* Chichester: Wiley, pp. 63-84.

Tobler, W., 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography,* 46(1), pp. 234-240.

Tolbert, P. et al., 1997. Spatio-temporal analysis of air quality and pediatric asthma emergency room visits. *American Journal of Epidemiology,* 151(8), pp. 798-810.

Volinksky, C., 2018. *Data Science's Impact on Telecom (Transcript)* [Interview] (12 02 2018).

Wakefield, J., 2007. Disease Mapping and Spatial Regression with Count Data. *Biostatistics,* 8(1), pp. 158-183.

Wall, M. W., 2004. A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference,* 121(1), pp. 311-324.

Wang, X., Yu, Y. R. & Faraway, J. J., 2018. *Bayesian Regression Modelling in INLA.* Boca Raton: CRC Press.

Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag.

Xia , H., Carlin, B. P. & Waller, L. A., 1997. Hierarchical models for mapping Ohio lung cancer rates. *Environmetrics,* 8(1), pp. 107-120.

Zurr, A. F., Leno, E. N. & Saveliev, A. A., 2017. *Beginner's Guide to Spatial, Temporal and Spatial-temporal Ecological Data Analysis with R-INLA.* Volume I: Using GLM and GLMM ed. Newburgh: Highland Statistics Ltd.

Appendix

We include a simplified R code we used to specifying the four different model specifications. This shows how we applied it to all broadband faults. The same principle is applied for forecasting fault types but not shown for parsimony in addition to all data visualizations, where the the R-package ggplot2 was used.

```
# 1.0 Broadband Service Provider (BSP) data pre-processing
# read BSP data
BSP <- read.csv('./Data/NW.csv', stringsAsFactors = FALSE) # load data as
characters, not factors
# create new column with months only
BSP$month <- format(as.Date(BSP$FAULT_DATE, format="%d/%m/%y"), "%m")
# create new column with year only
BSP$year <- format(as.Date(BSP$FAULT_DATE, format="%d/%m/%Y"), "%Y")
# remove all LSOAs with NA values - 49 in total
BSP1 <- BSP[!BSP$LSOA == "#N/A", ]
# convert to character
BSP1$month <- as.character(BSP1$month)
BSP1$year <- as.character(BSP1$year)
# change values for jan - mar 18 for identifiability later
BSP1$month[BSP1$month %in% "01" & BSP1$year %in% "2018"] <- "13"
BSP1$month[BSP1$month %in% "02" & BSP1$year %in% "2018"] <- "14"
BSP1$month[BSP1$month %in% "03" & BSP1$year %in% "2018"] <- "15"
# create count column
BSP1$count <- 1
# rename category
names(BSP1)[8] <- 'Fault_cat'
# subset columns for modelling
BSP1 <- BSP1[c(8:12)]
# aggregate by LSOA
agg <- aggregate(count~ LSOA+month, data=BSP1,FUN=sum)
# Aggregate data by MSOA - Fault count
# find lookup LSOA-MSOA online  https://data.gov.uk/dataset/9b090605-9861-4bb4-
9fa4-6845daa2de9b/postcode-to-output-area-to-lower-layer-super-output-area-to-
middle-layer-super-output-area-to-local-authority-district-february-2018-lookup-in-
the-uk
lookup <-
read.csv('./data/Postcode_to_Output_Area_to_Lower_Layer_Super_Output_Area_to
_Middle_Layer_Super_Output_Area_to_Local_Authority_District_February_2018_L
ookup_in_the_UK.csv',stringsAsFactors = F)
# subset only LSOA and MSOA columns
lookup <- lookup[,c(3,8,9)]
# rename columns
names(lookup) <- c('Postcode','LSOA','MSOA')
# subset columns for only LSOA and MSOA
```

```
lookup2 <- lookup[c(2,3)]
# remove duplicated LSOAs
lookup2 <- lookup2[!duplicated(lookup2$LSOA),]
# merge lookup with 2017 data
mer <- merge(agg,lookup2,by='LSOA',all.x=T)
# check for non-matches
sum(is.na(mer$MSOA))
# check original count of faults against new df
sum(mer$count)
# aggregate by MSOA
agg2 <- aggregate(count~ MSOA+month, data=mer,FUN=sum)
# check original count of faults against new df
sum(agg2$count)
# load library to reshape a data frame by aggregated form
library(reshape)
# cast with aggregation - this ensures that all MSOAs in each month have a value - those with no observations = 0
md <- as.data.frame(cast(agg2,MSOA~month,sum))
##################
# 2.0 Covariate data pre-processing & merge with BSP data
## 2.1 INTERNET USER CLASSIFICATION GROUP - 2018
# read csv file in
iuc <- read.csv('./Data/iuc2018.csv')
# select LSOA & GRP Group&Label
iuc <- iuc[,c(2,4)]
# change column names
names(iuc) <- c('LSOA','IUC_group')
# merge lookup with iuc data
iuc <- merge(iuc,lookup2,by='LSOA',all=TRUE)
# write function mode
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
  }
# aggregate
iuc <- aggregate(IUC_group ~ MSOA,data=iuc,FUN=Mode)
# merge fault type with lookup
md <- merge(md,iuc,by='MSOA',all.x=T)
## 2.2 INCOME & EDUCATION DEPRIVATION
# read IMD data for England - CDRC
imd <- read.csv('./data/imd2015eng.csv',stringsAsFactors = TRUE)
# select LSOA and income deprivation rate only
imd <- imd[,c(1,5,8,14)]
# rename columns - LSOA and income dep rate
names(imd) <- c('LSOA','IMD_score','Income_score','Education_dep')
# merge lookup with 2017 data
```

```
imd <- merge(imd,lookup2,by='LSOA')
# ed deprivation
ed_dep <- aggregate(Education_dep ~ MSOA,data=imd,FUN=mean)
# income rank
income_dep <- aggregate(Income_score ~ MSOA,data=imd,FUN=mean)
# merge fault type with lookup
md <- merge(md,ed_dep,by='MSOA',all.x=T)
# merge fault type with lookup
md <- merge(md,income_dep,by='MSOA',all.x=T)
## 2.3 Aged65+
# Infuse 2011 population stats - MSOA - North West - Age
age <- read.csv('./data/Data_AGE_UNIT.csv',header=TRUE,stringsAsFactors =
FALSE)
# remove first row
age <- age[-1,]
# convert columns to numeric
age[,c("F105","F167","F180","F181","F182","F183")] <-
as.numeric(as.character(unlist(age[,c("F105","F167","F180","F181","F182","F183")]))
)
# add columns
age$plus65 <- age$F105+age$F180+age$F181+age$F182+age$F183
# calculate proportion of elderly
age$eld_prop <- age$plus65/age$F167
# subset columns
age <- age[c(2,14)]
# change name
names(age)[1] <- 'MSOA'
# merge fault type with lookup
md <- merge(md,age,by='MSOA',all.x=T)
## 2.4 Ofcom - data usage
### Ofcom data - data usage by LSOA
# read in data
data_use <- read.csv('./Data/combined_data2.csv', stringsAsFactors = F)
# merge lookup with data_usage
mm <- merge(lookup,data_use,by='Postcode')
# convert data usage to numeric
mm$av_data_usage <- as.numeric(mm$av_data_usage)
# get list of MSOAs in data
rm <- md$MSOA
# apply list to data use file to only include MSOAs in study site
mm<- mm[ mm$MSOA %in% rm, ]
mm2 <- aggregate(av_data_usage ~ MSOA, data=mm, FUN=mean)
# aggregate to main data
md <- merge(md, mm2,by='MSOA',all.x=T)
# check for non-matches - NONE
sum(is.na(md$av_data_usage))
```

```r
# 3.0 Spatial data wrangling
## 3.1 Load Spatial shapefile & subset only MSOAs with VM faults
# package for loading spatial information
library(rgdal)
# read MSOA shapefile
NW <- readOGR('./infuse_msoa_lyr_2011_clipped.shp', stringsAsFactors = F)
# check CRS - OSGB36
proj4string(NW)
# subset MSOA column from shapefile
NW <- NW[,c(1)]
# rename column
names(NW)[1] <- c('MSOA')
# create list of MSOAs in VM dataset
rm <- md$MSOA
# apply list to shapefile to only include MSOAs in study site
NW_VM <- NW[ NW$MSOA %in% rm, ]
# check study site
plot(NW_VM)
## 3.2 Create spatial weights for INLA
# load package for creating spatial weights matrix
library(spdep)
# get centroid coordinates of df
coords <- coordinates(NW_VM)
# get nearest neighbour - minimum distance to have at least 1 neighbour
knb <- knn2nb(knearneigh(coords, k = 1))
# calculate maximum distance for each MSOA to have at least 1 neighbour
dist <- unlist(nbdists(knb, coords))
# show summary of distances
summary(dist)
# get maximum distance
max_d <- max(dist)
# calculate distance-based neighbours - specified distance
dnb1 <- dnearneigh(coords, d1 = 0, d2 = max_d)
## check plot
plot(dnb1,coords)
# create binary  matrix - by specifying style="B" - Binary is necessary for INLA to work ok
adj <- nb2mat(dnb1, style="B", zero.policy=TRUE) # ensures matrix is computed, even if there are islands or no-neighbour areas
# create spare matrix
adj <- as(adj, "dgTMatrix")
# 4.0 Modelling via INLA
## 4.0.1 Create dataframe elements for INLA
data.MSOA <-  attr(NW_VM, "data")
# Order first based on the map
order <- match(data.MSOA$MSOA,md$MSOA)
```

```
# ordered data
data.or<- md[order,]
# transform data in the right format for INLA
# VM fault count
y <- as.vector(as.matrix(data.or[,2:16]))
# get spatial component
MSOA<- as.factor(rep(data.or[,1]))
MSOA1<- as.factor(rep(data.or[,1]))
# month
month <- numeric(0)
for(i in 1:15){
  month<- append(month,rep(i,dim(data.or)[1]))
}
# covariates
iu <- as.factor(rep(data.or[,17]))
ed <- as.numeric(rep(data.or[,18]))
inc <- as.numeric(rep(data.or[,19]))
eld_prop <- as.numeric(rep(data.or[,20]))
av_d <- as.numeric(rep(data.or[,21]))
# JAN
# get length of all rows minus 3 months -662 MSOAs
length(y) - (662*3)
# apply one month of NA values
yyy = rep(NA,662)
# use length of rows minus 3 months + the NA values
y_train1 <- c(y[1:(7944)],yyy)
# same n months as observations
month1 <- c(month[1:8606])
# create INLA df for jan
data1 <-
data.frame(y=y_train1,MSOA=MSOA,MSOA1=as.numeric(MSOA),month=month1,m
onth1=month1, area.year = seq(1,length(MSOA)),
MSOA.int=as.numeric(MSOA),month.int=month1,
x1=iu,x2=ed,x3=inc,x4=eld_prop,x5=av_d)
#FEB
# get length of all rows minus 3 months -662 MSOAs
length(y) - (662*2)
# use length of rows minus 3 months + the NA values
y_train2 <- c(y[1:(8606)],yyy)
# same n months as observations
month2 <- c(month[1:9268])
# create INLA df feb
data2 <-
data.frame(y=y_train2,MSOA=MSOA,MSOA1=as.numeric(MSOA),month=month2,m
onth1=month2, area.year = seq(1,length(MSOA)),
MSOA.int=as.numeric(MSOA),month.int=month2,
x1=iu,x2=ed,x3=inc,x4=eld_prop,x5=av_d)
# see model output
#MAR
# get length of all rows minus 3 months -662 MSOAs
length(y) - (662)
# get number of MSOAs - to one month
yyy = rep(NA,662)
# use length of rows minus 3 months + the NA values
y_train3 <- c(y[1:(9268)],yyy)
# same n months as observations
month3 <- c(month[1:9930])
# create INLA df for mar
data3 <-
data.frame(y=y_train3,MSOA=MSOA,MSOA1=as.numeric(MSOA),month=month3,m
onth1=month3, area.year = seq(1,length(MSOA)),
MSOA.int=as.numeric(MSOA),month.int=month3,
x1=iu,x2=ed,x3=inc,x4=eld_prop,x5=av_d)
## 4.2 RW1 model
# run INLA
library(INLA)
# RW1 model 1 - jan
r1 <- y ~ 1 + x1+x2+x3+x4+x5+
   f(month,model='rw1') + # temporal structured component
   f(month1,model='iid')
system.time(rw1 <- inla(r1,family="poisson",data=data1,
                  control.compute = list(dic = TRUE,config=TRUE),
                  control.predictor = list(link = 1)))
# get model output
summary(rw1)
# RW1 model 2 - feb
r2 <- y ~ 1 + x1+x2+x3+x4+x5+
   f(month,model='rw1') + # temporal structured component
   f(month1,model='iid')
system.time(rw2 <- inla(r2,family="poisson",data=data2,
               control.compute = list(dic = TRUE,config=TRUE),
                  control.predictor = list(link = 1)))
# get model output
summary(rw2)
# RW1 model 3 - mar
r3 <- y ~ 1 + x1+x2+x3+x4+x5+
   f(month,model='rw1') + # temporal structured component
   f(month1,model='iid')
system.time(rw3 <- inla(r3,family="poisson",data=data3,
               control.compute = list(dic = TRUE,config=TRUE),
                  control.predictor = list(link = 1)))
```

```
# get model output
summary(rw3)
## 4.2  AR1 model
# AR1 model 1 - jan
a1 <- y ~ 1 + x1+x2+x3+x4+x5+
   f(month,model='ar1') + # temporal structured component
   f(month1,model='iid')
system.time(ar1 <- inla(a1,family="poisson",data=data1,
                   control.compute = list(dic = TRUE,config=TRUE),
                       control.predictor = list(link = 1)))
# get model output
summary(ar1)
# AR1 model 2 - feb
a2 <- y ~ 1 + x1+x2+x3+x4+x5+
   f(month,model='ar1') + # temporal structured component
   f(month1,model='iid')
system.time(ar2 <- inla(a2,family="poisson",data=data2,
                   control.compute = list(dic = TRUE,config=TRUE),
                       control.predictor = list(link = 1)))
# get model output
summary(ar2)
# AR1 model 3 - mar
a3 <- y ~ 1 + x1+x2+x3+x4+x5+
   f(month,model='ar1') + # temporal structured component
   f(month1,model='iid')
system.time(ar3 <- inla(a3,family="poisson",data=data3,
                   control.compute = list(dic = TRUE,config=TRUE),
                       control.predictor = list(link = 1)))
# get model output
summary(ar3)
# 4.4 space-time model
# run model - jan
s1 <- y ~ 1 + x1+x2+x3+x4+x5+
  f(MSOA1, model="bym", graph =adj) + # modellng structured + unstructured spatial
components
  f(month,model='ar1') + # temporal structured component
  f(month1,model='iid') # temporal unstructured component

system.time(st1 <- inla(s1,family="poisson",data=data1,
                       control.compute = list(dic = TRUE,config=TRUE),
                       control.predictor = list(link = 1)))
# see model output
summary(st1)
# run model - feb
s2 <- y ~ 1 + x1+x2+x3+x4+x5+
  f(MSOA1, model="bym", graph =adj) + # modellng structured + unstructured spatial
components
  f(month,model='ar1') + # temporal structured component
  f(month1,model='iid') # temporal unstructured component

system.time(st2 <- inla(s2,family="poisson",data=data2,
                             control.compute = list(dic = TRUE,config=TRUE),
                             control.predictor = list(link = 1)))
# see model output
summary(st2)
# run model - mar
s3 <- y ~ 1 + x1+x2+x3+x4+x5+
  f(MSOA1, model="bym", graph =adj) + # modellng structured + unstructured spatial
components
  f(month,model='ar1') + # temporal structured component
  f(month1,model='iid')  # temporal unstructured component

system.time(st3 <- inla(s3,family="poisson",data=data3,
                             control.compute = list(dic = TRUE,config=TRUE),
                             control.predictor = list(link = 1)))
# see model output
summary(st3)
## 4.5 space-time interaction predictive model
# run model - jan
formula1 <- y ~ 1 + x1+x2+x3+x4+x5+
  f(MSOA1, model="bym", graph =adj) + # modellng structured + unstructured spatial
components
  f(month,model='ar1') + # temporal structured component
  f(month1,model='iid') + # temporal unstructured component
  f(MSOA.int,model="iid", group=month.int,control.group=list(model="ar1"))

system.time(jan <- inla(formula1,family="poisson",data=data1,
                             control.compute = list(dic = TRUE,config=TRUE),
                             control.predictor = list(link = 1)))
# see model output
summary(jan)
#run model - feb
formula2 <- y ~ 1 + x1+x2+x3+x4+x5+
  f(MSOA1, model="bym", graph =adj) + # modellng structured + unstructured spatial
components
  f(month,model='ar1') + # temporal structured component
  f(month1,model='iid') + # temporal unstructured component
  f(MSOA.int,model="iid", group=month.int,control.group=list(model="ar1"))

system.time(feb <- inla(formula2,family="poisson",data=data2,
                             control.compute = list(dic = TRUE,config=TRUE),
                              control.predictor = list(link = 1)))
# see model output
summary(feb)
# run model - mar
formula3 <- y ~ 1 + x1+x2+x3+x4+x5+
  f(MSOA1, model="bym", graph =adj) + # modellng structured + unstructured spatial
components
  f(month,model='ar1') + # temporal structured component
  f(month1,model='iid') + # temporal unstructured component
  f(MSOA.int,model="iid", group=month.int,control.group=list(model="ar1"))

system.time(mar <- inla(formula3,family="poisson",data=data3,
                              control.compute = list(dic = TRUE,config=TRUE),
                              control.predictor = list(link = 1)))
# see model output
summary(mar)
# 5.0 Modify results for easier to interpret outputs & account for uncertainty
## 5.1 calculate standard deviation instead of the precision of the posterior
distribution & explained variance for the "best
# refernece @ Faraway 2018
library(brinla)
# apply function to best model
bri.hyperpar.summary(st1)
bri.hyperpar.summary(st2)
bri.hyperpar.summary(st3)
#references Blangiardo, Marta, and Michela Cameletti. Spatial and Spatio-temporal
Bayesian Models with R-INLA. John Wiley & Sons, 2015.
library(INLAOutputs)
ExplainedVariance(st1, st2, st3)
## 5.2 obtain samples from the fitted values to account for uncertainty
#  number of samples
nsample <-  1000
# define function to obtain 1000 samples of each latent observation in Jan
inla.sample <- function(x){
  # get length of each of the estimated values
  len = x$summary.fitted.values$mean
  ## reproducible results
  set.seed(1234)
  inla.seed <- as.integer(runif(1)*.Machine$integer.max)
  # apply number of samples to ar1 inla model
  x = inla.posterior.sample(nsample, x, seed=inla.seed)
  # create for loop to get the average of the latent components only for 1000 samples
  for (nr in 1:length(nsample)){
  # access all elements of the latent that are the fitted values of the linear predictor
only
  x = lapply(x, function(x) x$latent[1:length(len)])
  # get average for the same row number in each of the 1000 simulated samples
  x <- exp(colMeans(do.call(rbind,x)))
  }
  return(x)
}
# apply to all months in jan
p_ar1 <- inla.sample2(ar1)
p_rw1 <- inla.sample2(rw1)
p_st1 <- inla.sample2(st1)
p_sti1 <- inla.sample2(jan)
# apply to all months in feb
p_ar2 <- inla.sample2(ar2)
p_rw2 <- inla.sample2(rw2)
p_st2 <- inla.sample2(st2)
p_sti2 <- inla.sample2(feb)
# apply toall months in mar
p_ar3 <- inla.sample3(ar3)
p_rw3 <- inla.sample3(rw3)
p_st3 <- inla.sample3(st3)
p_sti3 <- inla.sample3(mar)

# create dataframe to compare results
par <- data.frame(y=observed_jan,fitted=ar1$summary.fitted,month=month1)
# add simulated values to fitted values to compare
par <- cbind(p_ar1,par)
# access last 662 elemeents
par <- par[7945:8606,]
# compare the results inla.posterior.sample(), fitted INLA values & INLA
sum(par$p_ar1)
sum(par$fitted.mean)
sum(par$y)
#.... for each of the months
```