A spatio-temporal Covid-19 surveillance tool for Scotland using telehealth data

Duncan Lee

School of Mathematics and Statistics, University of Glasgow, Scotland

Chris Robertson

Department of Mathematics and Statistics, University of Strathclyde, Scotland, and Public Health Scotland

Diogo Marques

Public Health Scotland

July 22, 2020

Abstract

Covid-19 is the biggest global health crisis in a generation, having been responsible for over 584,000 deaths worldwide as of $16th$ July 2020. Modelling the spatio-temporal spread of the disease and predicting the likely locations for future outbreaks is therefore of major public health importance for health agencies worldwide, as it allows them to target their efforts to subdue the virus where they are most needed. This paper presents a spatio-temporal surveillance tool for Covid-19 developed for Scotland, which uses data from the country's national telehealth phone service NHS 24. We develop a novel multivariate spatio-temporal model for estimating the proportions of calls to NHS 24 that were classified as being Covid-19 or having related symptoms. We applied the model to the most recent data available up to $6th$ July 2020, and show that the trends in the proportions of calls to NHS 24 due to Covid-19 do indeed provide an early warning system of similar trends in the numbers of cases and deaths.

Keywords: Covid-19; Spatio-temporal disease surveilance; Telehealth data.

1 Introduction

Covid-19 represents the biggest public health challenge in decades, and was declared a global pandemic by the World Health Organisation (WHO) on 11th March 2020. The disease originated in the city of Wuhan in the People's

Republic of China in December 2019, and reached the USA and Europe towards the end of January 2020. The first European epicentre for Covid-19 was in northern Italy in February, and in Scotland, the focus of this paper, the first confirmed case occurred on the 2nd March (Public Health Scotland, `https://www.opendata.nhs.scot/dataset/covid-19-in-scotland`). Since then Covid-19 has spread across the world causing global health and economic devastation, and as of $21st$ July, there were over 14.7 million cases worldwide, and over 610,000 people have sadly died from the disease (Johns Hopkins Coronavirus Resource Centre, `https://coronavirus.jhu.edu/map.html`).

Unsurprisingly, modelling the spread and dynamics of the Covid-19 pandemic has become a research priority, and there is a quickly growing research literature in this area. This literature has focused on a range of important epidemiological topics, including: (i) predicting the spread of the pandemic and its impacts on healthcare systems (Remuzzi and Remuzzi, 2020); (ii) identifying the factors that make people more at risk of displaying severe symptoms of covid-19 (Conticini et al., 2020 and Wu et al., 2020); (iii) developing real-time surveillance systems for covid-19 tracking (Dong et al., 2020); and (iv) identifying the wider health impacts of Covid-19 (Douglas et al., 2020).

The aim of this paper is to develop a population-level spatio-temporal surveillance system for Covid-19 in Scotland, using data from the Scottish National Health Service's (NHS) NHS 24 telehealth system. NHS 24 is the national telehealth service for Scotland, which people can call anytime of day or night if they require non-emergency medical attention. National telehealth data are ideal for building a surveillance system, because an increase in calls for Covid-19 related symptoms in an area will likely be an early warning of an increase in the numbers of cases, hospitalisations and deaths in that area in the near future. Our surveillance system estimates the spatio-temporal trends in the proportions of calls to NHS 24 about Covid-19 or related symptoms, including the identification of: (i) local hotspots of high proportions; and (ii) areas where the proportions in the final week of the data are increasing / decreasing compared to the previous week. This allows health professionals to monitor the evolution of the Covid-19 epidemic in each local area, allowing them to target public health interventions appropriately at areas exhibiting increased proportions. Thus the surveillance model can be run in near real time each week as new data become available, allowing the national health agency Public Health Scotland (PHS) to adapt to the changing state of the Covid-19 epidemic.

The surveillance system we develop is based on a multivariate binomial spatio-temporal random effects model, with inference in a Bayesian setting using Markov chain Monte Carlo (MCMC) simulation. It jointly models the spatio-temporal variation in the proportions of calls to NHS 24 directly categorised as Covid-19, as well as those calls categorised with related symptoms such as fever and difficulty breathing, the latter ensuring that potential local epidemics are not missed due to misclassification of calls. In developing this model the key methodological challenge we address is the complex multivariate spatio-temporal structure of the data, which means we need to capture spatial, temporal and between call type correlations. While multivariate spatial (e.g. Gelfand and Vounatsou, 2003, Martinez-Beneito, 2013 and Gomez-Rubio and Palmi-Perales, 2019) and univariate spatio-temporal (e.g. Knorr-Held, 2000, Lee and Lawson, 2016 and Aregay et al., 2017) models for disease risk are commonplace, the development of multivariate space-time models are in their infancy in disease risk modelling, with two recent examples being Quick et al. (2017) and Jack et al. (2019). While Quick et al. (2017) proposed a fully multivariate spatio-temporal (MVST) Gaussian Markov Random Field (GMRF, Rue and Held, 2005) process to model these correlations, Jack et al. (2019) combined separate simpler multivariate spatial and multivariate temporal processes.

Here, we propose a fully MVST GMRF model based on either a multivariate first or second order autoregressive process, and provide software to allow others to fit the model and hence build their own surveillance system. Our temporally driven conditional model specification differs from Quick et al. (2017) who defined their model in terms of a spatial multivariate conditional autoregressive process (MCAR), because in their application space was the dominant dimension (more spatial areas then time periods). For our surveillance application time is the dominant dimension in terms of the dynamics in the data that we aim to capture, which is the motivation for our temporally conditioned specification. The NHS 24 telehealth data motivating the development of our model are described in Section 2, while the novel multivariate spatio-temporal model is presented in Section 3. Our surveillance model is applied to the Scottish telehealth data in Section 4, which includes a comparison of how the estimated trends compare to the numbers of cases and deaths. Finally, Section 5 concludes the paper.

2 Telehealth data in Scotland

NHS 24 (`https://www.nhs24.scot/`) is Scotland's national telehealth service, and gives the public phone access to non-emergency medical advice 24 hours a day and 7 days a week when their regular primary health care providers are closed. NHS 24 deals with around 1.5 million calls per year and serves a population of around 5.4 million people, and at peak demand answers around 14,500 calls over the course of a weekend.

Data were obtained from Public Health Scotland (`https://publichealthscotland.scot/`) on the weekly numbers of calls to NHS 24 for Covid-19 and other similar conditions from the week beginning $2nd$ March 2020 to the week beginning $6th$ July 2020 inclusive, totalling $N = 19$ weeks. The classification for Covid-19 was only initially available from $14th$ April on-wards, but was back predicted to $2nd$ March using a prediction model developed by PHS to allow trends to be modelled over the peak of the epidemic. The prediction model was developed using NHS 24 call data from mid April to the end of May relating to respiratory and gastro instestinal syndromes plus the patients age. The prediction performance of this model had a specificity of 96% and a sensitivity of 75% with an AUC of 0.88. Therefore to ensure the Covid-19 series covers the peak of the epidemic, we treat these predictions as observed counts.

These data are available for the 444 postcode districts (PD) in Scotland, and a shapefile containing the spatial boundary information for these PDs was obtained from the National Records for Scotland (`https://www.nrscotland.gov.uk`). This spatial boundary information did not include 8 of the PDs in the data set, but as these PDs only accounted for 41 NHS 24 calls out of a total of 457,961 calls for all conditions they were removed from the data. After removing these PDs there were 882 instances (PD and week combinations) with no NHS 24 calls at all, which were spread relatively evenly across the 13 weeks with between 34 and 56 instances per week. Therefore, to ensure a rectangular data set for analysis, only the $K = 329$ PDs having at least 1 NHS 24 call (about any illness) per week were retained in the data. The PDs removed from the data only accounted for 0.66% of the total calls to NHS 24, and were mostly sparsely populated rural or industrial / commercial areas.

For the kth PD and tth week the data comprise the following counts of the numbers of calls to NHS24: (i) N_{kt} - the total number of calls to NHS 24; (ii) Y_{kt1} - the number of calls classified as Covid-19; and (iii) Y_{kt2} - the number of calls classified as `Simple Estimate 1` (hereafter `SE1`), which is a set of symptoms potentially related to Covid-19 including cold, flu, coughs, fever and difficulty breathing. The latter is modelled here to ensure that potential local epidemics are not missed due to a misclassification of calls. Each NHS 24 call can have multiple classifications however, and as expected there is substantial overlap in the calls classified as Covid-19 and SE1. Additionally, the total number of calls classified as Covid-19 or SE1 is sometimes greater than the total number of calls, i.e $Y_{kt1} + Y_{kt2} > N_{kt}$, particularly where N_{kt} is small. Thus these data cannot be modelled as a multinomial distribution, and instead we model them as a correlated multivariate binomial process.

The correlations between the proportions of calls, $\hat{\theta}_{ktj} = Y_{ktj}/N_{kt}$, classified as Covid-19 ($j = 1$) and SE1 ($j = 2$) across all PDs for each week range between 0.61 and 0.94, suggesting there is a strong relationship between them. This is further evidenced by Figure 1, which displays the temporal trends in these raw proportions. In the figure jittering has been added to the week beginning (horizontal) dimension to improve the visibility of the points, and the proportions for Covid-19 are in red while those for SE1 are in blue. The trend line in each case has been estimated using LOESS smoothing. The figure shows a number of key points, the first of which is large amounts of noise in the data arising from small numbers of calls in some PDs, with sample proportions equal to 0 or 1 in 6.1% (Covid-19) and 6.8% (SE1) of week and PD combinations. Secondly, the temporal trends are broadly similar for Covid-19 and SE1, showing a rise in the proportions from the $2nd$ March, a peak around $23rd$ March, a decrease until $1st$ June, and a general steady state since then. Thirdly, the figure shows that the dominant classification seems to change around the week beginning $6th$ April, with more calls classified as SE1 before that date and more Covid-19 calls after that date. This may be an artifact of the prediction model used to predict the Covid-19 classification before $14th$ April, or alternatively it may be that as the epidemic became more prevalent from late March onwards people might be more likely to mention Covid-19 directly when they called NHS 24.

The median lag-1 temporal autocorrelation coefficient across the $K = 329$ PDs are respectively 0.47 (Covid-19) and 0.66 (SE1), which suggests these data are likely to exhibit temporal autocorrelation as expected. The raw proportions also exhibit some spatial autocorrelation, which was quantified for each week and call classification using Moran's I (Moran, 1950) statistics and a corresponding Monte-Carlo p-value to test the null hypothesis of no spatial autocorrelation. From these tests 32% (Covid-19) and 32% (SE1) of these weekly p-values were significant at the 5% level, suggesting that despite the noise in these raw proportions, spatial autocorrelation is likely to be present in the data. Thus as the data exhibit spatio-temporal and between call type correlations contaminated by noise due to small numbers, a multivariate spatio-temporal smoothing model is proposed in the next section to estimate the underlying trends in these data.

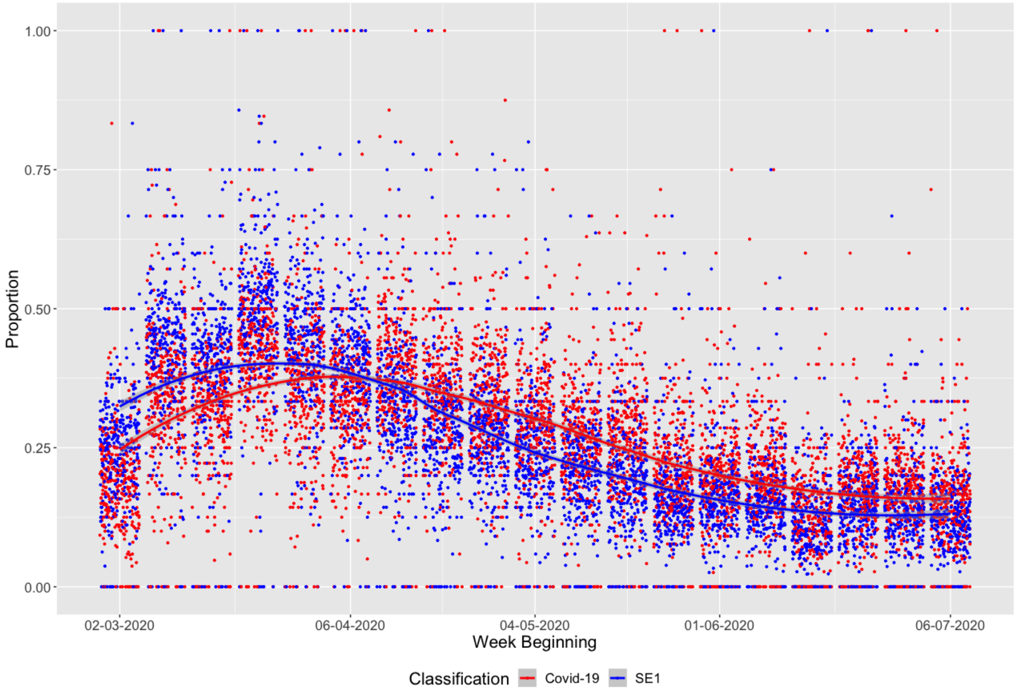

Figure 1: Scatterplot showing the temporal trends in the sample proportions of calls to NHS 24 that were related to Covid-19 (red) and SE1 (blue) for all PDs as points, with LOESS trend lines superimposed. The points have been jittered in the Week Beginning (horizontal) direction to improve their visibility.

Specifically, our 3 underlying goals when modelling these data are to:

1. Estimate the spatio-temporal trends in the proportions of calls due to Covid-19 and related conditions (SE1), to better understand the dynamics and spread of the epidemic in Scotland.
2. Compare the estimated spatio-temporal trends in the telehealth data to the same trends in data relating to confirmed cases and deaths, to assess the extent to which our modelling results can be used as an early warning surveillance system to identify future peaks in the epidemic.
3. Produce informative visualisations about the current dynamics of the epidemic, in particular where and when new peaks might occur, to support public health decision making in controlling the virus.

3 Methodology

This section proposes a new multivariate spatio-temporal (MVST) surveillance model for estimating the spatio-temporal trends in the proportions of NHS 24 calls due to Covid-19 or related symptoms (SE1). The model is fitted in a Bayesian setting using MCMC simulation, using a combination of Gibbs sampling and Metropolis-Hastings steps. Software to implement the model in `R` is available at `https://github.com/duncanplee/Covid-19-model`, which allows others to build surveillance models for their own data.

3.1 Level 1 - Data likelihood model

Let Y_{ktj} denote the number of calls to NHS 24 in the kth PD ($k = 1, \ldots, K$) during the tth week ($t = 1, \ldots, T$) for the jth outcome ($j = 1, \ldots, J$), where for our data $j = 1$ is Covid-19 and $j = 2$ is SE1. Additionally, let N_{kt} denote the total number of NHS 24 calls in the kth PD and tth week. Then as the two outcomes (call classifications) are not disjoint as described in Section 2, a multinomial model is not appropriate for these data. Instead, we model these data as conditionally independent binomial distributions, where the spatio-temporal and between outcome (auto) correlations are modelled by random effects at the second level in the model hierarchy. The first level of the hierarchical model is given by:

$$\begin{aligned} Y_{ktj} &\sim \text{Binomial}(N_{kt}, \theta_{ktj}) \qquad (1) \\ \ln\left(\frac{\theta_{ktj}}{1-\theta_{ktj}}\right) &= \beta_j + \phi_{ktj}. \end{aligned}$$

Here, θ_{ktj} is the estimated proportion of calls (or probability that a single call) to NHS 24 in PD k during week t that are due to outcome j, and an increase in θ_{ktj} in the last week of the data provides an early warning signal about a possible new outbreak of cases in the kth PD. We do not include any covariates in our model for two reasons, the first of which is that our aim is to estimate the spatio-temporal trends in $\{\theta_{ktj}\}$ via the random effects $\{\phi_{ktj}\}$, rather than explaining what factors are associated with these trends. Secondly, up-to-date temporally varying covariate information is not available on a weekly basis in a timely fashion, meaning that it would not be available to include in the model in near real time. The intercept terms for each outcome are assigned weakly informative independent Gaussian prior distributions given by $\beta_j \sim \mathrm{N}(0, 100000)$, which allows the data to play the dominant role in estimating their values.

3.2 Level 2 - Multivariate spatio-temporal random effects model

The remaining term in (1) $\{\phi_{ktj}\}$ are random effects, which are the mechanism for estimating the smooth multivariate spatio-temporal trends in $\{\theta_{ktj}\}$ for all outcomes. As such, the prior distribution for these random effects must induce (auto)correlations in time, space and between outcomes. The entire set of random effects are denoted by $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N)$, where $\boldsymbol{\phi}_t = (\boldsymbol{\phi}_{1t}, \ldots, \boldsymbol{\phi}_{Kt})$ denotes the set of $K \times J$ random effects at time t, while $\boldsymbol{\phi}_{kt} = (\phi_{kt1}, \ldots, \phi_{tkJ})$ denotes the subset of effects at the kth PD for all J outcomes. As mentioned earlier MVST models are in their infancy for areal unit data, and we follow Quick et al. (2017) and propose a zero-mean multivariate Gaussian Markov random field (GMRF, Rue and Held, 2005) model for $\boldsymbol{\phi}$. The general form of our model is given by

$$\boldsymbol{\phi} \quad \sim \quad \mathrm{N}\left(\mathbf{0}, \left[\mathbf{D}(\boldsymbol{\alpha}) \otimes \mathbf{Q}(\mathbf{W}, \rho) \otimes \boldsymbol{\Sigma}^{-1}\right]^{-1}\right), \tag{2}$$

where $\otimes$ denotes a Kronecker product. The precision matrix is $\mathbf{P}(\boldsymbol{\alpha}, \rho, \boldsymbol{\Sigma}) = \mathbf{D}(\boldsymbol{\alpha}) \otimes \mathbf{Q}(\mathbf{W}, \rho) \otimes \boldsymbol{\Sigma}^{-1}$, where $\mathbf{D}(\boldsymbol{\alpha})_{N \times N}$ controls temporal autocorrelations, $\mathbf{Q}(\mathbf{W}, \rho)_{K \times K}$ controls spatial autocorrelations and $\boldsymbol{\Sigma}_{J \times J}$ captures between outcome correlations. The precision matrix $\mathbf{P}(\boldsymbol{\alpha}, \rho, \boldsymbol{\Sigma})$ is sparse because $[\mathbf{D}(\boldsymbol{\alpha}), \mathbf{Q}(\mathbf{W}, \rho)]$ are both built from specific cases of GMRFs described below, which enables computationally efficient Bayesian inference by making use of their triplet form representation. The model is defined in terms of its precision matrix $\mathbf{P}(\boldsymbol{\alpha}, \rho, \boldsymbol{\Sigma})$ rather than its covariance matrix, which means that multivariate Gaussian theory gives the following partial (auto)correlations for (ϕ_{ktj}, ϕ_{rsi}) conditional on the remaining random effects $\boldsymbol{\phi}_{-ktj,rsi}$:

$$\mathrm{Corr}(\phi_{ktj}, \phi_{rsi} | \boldsymbol{\phi}_{-ktj,rsi}) \; = \; \frac{-\mathbf{D}(\boldsymbol{\alpha})_{ts} \mathbf{Q}(\mathbf{W}, \rho)_{kr} \left(\boldsymbol{\Sigma}^{-1}\right)_{ji}}{\sqrt{\left(\mathbf{D}(\boldsymbol{\alpha})_{tt} \mathbf{Q}(\mathbf{W}, \rho)_{kk}\right) \left(\boldsymbol{\Sigma}^{-1}\right)_{jj} \left(\mathbf{D}(\boldsymbol{\alpha})_{ss} \mathbf{Q}(\mathbf{W}, \rho)_{rr} \left(\boldsymbol{\Sigma}^{-1}\right)_{ii}\right)}}. \tag{3}$$

The between outcome covariance matrix $\boldsymbol{\Sigma}$ is not assigned a specific structure, and is instead assigned the following conjugate Inverse-Wishart prior distribution

$$\boldsymbol{\Sigma} \sim \text{Inverse-Wishart}(d, \boldsymbol{\Omega}). \tag{4}$$

The hyperparameters are set at $(d = J+1, \boldsymbol{\Omega} = 0.01\mathbf{I})$ where $\mathbf{I}$ is the identity matrix, and are chosen to ensure it is only weakly informative. In contrast, spatial autocorrelation is modelled by the conditional autoregressive (CAR) prior proposed by Leroux et al. (2000), which corresponds to the following spatial precision matrix

$$\begin{aligned}\mathbf{Q}(\mathbf{W},\rho) &= \rho(\text{diag}[\mathbf{W}\mathbf{1}]-\mathbf{W})+(1-\rho)\mathbf{I} \\ \rho &\sim \text{Uniform}(0,1).\end{aligned} \tag{5}$$

Here $(\mathbf{1},\mathbf{I})$ are a $K\times 1$ vector of ones and the $K\times K$ identity matrix respectively, while diag[.] denotes a diagonal matrix. The spatial autocorrelation structure assumed by this matrix is determined by the $K\times K$ neighbourhood or adjacency matrix $\mathbf{W}=(w_{kr})$, which denotes whether each pair of PDs are close together. Here we adopt the binary specification that $w_{kr}=1$ if the rth PD is one of the 5 nearest PDs to the kth PD, and $w_{kr}=0$ otherwise. This leads to an asymmetric $\mathbf{W}$ matrix, which is made symmetric for the purposes of fitting the model by if $w_{kr}=1$ and $w_{rk}=0$ then we set $w_{rk}=1$. We note that we did not construct $\mathbf{W}$ based on the commonly used border sharing approach because the study region has islands, which thus don't share any borders with other areas. This specification models (ϕ_{ktj},ϕ_{rtj}) as partially spatially autocorrelated if $w_{kr}=1$ and conditionally independent if $w_{kr}=0$, which can be seen from (3) and the fact that for $k\neq r$ $\mathbf{Q}(\mathbf{W},\rho)_{kr}=-\rho w_{kr}$. This also illustrates that ρ is a global spatial dependence parameter, with a value of 0 corresponding to spatial independence while a value of 1 corresponds to strong spatial autocorrelation, specifically the intrinsic CAR model proposed by Besag et al., 1991). We model temporal autocorrelation using either first order or second order autoregressive processes, and the joint distribution for $\boldsymbol{\phi}$ from (2) in each case can be decomposed as described below.

3.2.1 First-order autoregressive process

For a first-order autoregressive process the joint prior distribution $f(\boldsymbol{\phi})$ can be decomposed as

$$\begin{aligned}f(\boldsymbol{\phi}) &= f(\boldsymbol{\phi}_1)\prod_{t=2}^{N}f(\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1}) \\ &= \text{N}\left(\boldsymbol{\phi}_1\middle|\mathbf{0},\left[\mathbf{Q}(\mathbf{W},\rho)\otimes\boldsymbol{\Sigma}^{-1}\right]^{-1}\right)\prod_{t=2}^{N}\text{N}\left(\boldsymbol{\phi}_t\middle|\alpha\boldsymbol{\phi}_{t-1},\left[\mathbf{Q}(\mathbf{W},\rho)\otimes\boldsymbol{\Sigma}^{-1}\right]^{-1}\right)\end{aligned} \tag{6}$$

which is combined with an improper non-informative prior $f(\alpha)\propto 1$. This specification corresponds to a tridiagonal matrix for $\mathbf{D}(\alpha)$ with entries

$$\begin{aligned}\mathbf{D}(\boldsymbol{\alpha})_{t,t} &= \left\{\begin{array}{ll}1+\alpha^2 & \text{for } t=1,\ldots,N-1 \\ 1 & \text{for } t=N\end{array}\right., \\ \mathbf{D}(\boldsymbol{\alpha})_{t,t-1} &= -\alpha \quad \text{for } t=2,\ldots,N.\end{aligned}$$

Thus from (3) it is clear that (ϕ_{ktj},ϕ_{ksj}) are conditionally independent if $s\notin\{t-1,t,t+1\}$.

3.2.2 Second-order autoregressive process

For a second-order autoregressive process the joint prior distribution $f(\boldsymbol{\phi})$ can be decomposed as

$$
\begin{aligned}
f(\boldsymbol{\phi}) &= f(\boldsymbol{\phi}_1)f(\boldsymbol{\phi}_2)\prod_{t=3}^{N} f(\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1},\boldsymbol{\phi}_{t-2}) \\
&= \mathrm{N}\left(\boldsymbol{\phi}_1\middle|\mathbf{0},\left[\mathbf{Q}(\mathbf{W},\rho)\otimes\boldsymbol{\Sigma}^{-1}\right]^{-1}\right)\mathrm{N}\left(\boldsymbol{\phi}_2\middle|\mathbf{0},\left[\mathbf{Q}(\mathbf{W},\rho)\otimes\boldsymbol{\Sigma}^{-1}\right]^{-1}\right) \\
&\times \prod_{t=3}^{N}\mathrm{N}\left(\boldsymbol{\phi}_t\middle|\alpha_1\boldsymbol{\phi}_{t-1}+\alpha_2\boldsymbol{\phi}_{t-2},\left[\mathbf{Q}(\mathbf{W},\rho)\otimes\boldsymbol{\Sigma}^{-1}\right]^{-1}\right), \qquad (7)
\end{aligned}
$$

which is combined with an improper non-informative prior $f(\alpha_1,\alpha_2)\propto 1$. This specification corresponds to the following sparse matrix for $\mathbf{D}(\boldsymbol{\alpha})$ with non-zero entries

$$
\begin{aligned}
\mathbf{D}(\boldsymbol{\alpha})_{t,t} &= \begin{cases} 1+\alpha_2^2 & \text{for } t=1 \\ 1+\alpha_1^2+\alpha_2^2 & \text{for } t=2,\ldots,N-2 \\ 1+\alpha_1^2 & \text{for } t=N-1 \\ 1 & \text{for } t=N \end{cases}, \\
\mathbf{D}(\boldsymbol{\alpha})_{t,t-1} &= \begin{cases} \alpha_1\alpha_2 & \text{for } t=2 \\ \alpha_1\alpha_2-\alpha_1 & \text{for } t=3,\ldots,N-1 \\ -\alpha_1 & \text{for } t=N \end{cases}, \\
\mathbf{D}(\boldsymbol{\alpha})_{t,t-2} &= -\alpha_2 \quad \text{for } t=3,\ldots,N.
\end{aligned}
$$

Thus from (3) it is clear that (ϕ_{ktj},ϕ_{ksj}) are conditionally independent if $s\notin\{t-2,t-1,t,t+1,t+2\}$.

4 Covid-19 surveillance in Scotland

This section presents the results of the Covid-19 surveillance model in Scotland, focusing on its estimation of the spatio-temporal trends in the telehelath data and how well it tracks the epidemic of confirmed cases of, and deaths due to, Covid-19.

4.1 Model fitting

Both the AR(1) and AR(2) models outlined in Section 3.2 are fitted to the data, and inference from each model is based on 3,000 MCMC samples generated from 3 independent Markov chains. Each chain was burnt in for 50,000 samples by which time convergence was assessed to have been reached, and then run for a further 300,000 samples which were thinned by 300 to greatly reduce their autocorrelation. Convergence was visually assessed using traceplots and numerically assessed using the Gelman-Rubin diagnostic, and for the latter all values of $\hat{R}$ were less than 1.1 as suggested as a convergence criteria by Gelman et al. (2013).

A summary of each model is presented in Table 1, which includes posterior medians and 95% credible intervals for the covariance parameters $(\rho,\boldsymbol{\alpha},\boldsymbol{\Sigma})$, as

Table 1: Summary of the two models fitted to the data, including overall model fit via the DIC and the effective number of independent parameters (p.d), and the posterior medians and 95% credible intervals for the covariance parameters.

	Model	
	AR(1)	**AR(2)**
$\boldsymbol{\Sigma}_{11}$	0.153 (0.133, 0.174)	0.173 (0.153, 0.195)
$\boldsymbol{\Sigma}_{22}$	0.156 (0.136, 0.180)	0.181 (0.160, 0.203)
$\boldsymbol{\Sigma}_{12}/\sqrt{\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{22}}$	0.995 (0.993, 0.997)	0.995 (0.993, 0.997)
ρ	1.00 (0.999, 1.000)	1.000 (0.999, 1.000)
$\boldsymbol{\alpha}$	α 0.680 (0.630, 0.725)	α_1 0.441 (0.377, 0.507) α_2 0.287 (0.218, 0.355)
DIC	59,693	59,609
p.d	2,311	2,415

well as the overall model fit measure the deviance information criterion (DIC, Spiegelhalter et al., 2002) and the effective number of independent parameters (p.d). The table shows that the estimated proportions of calls classified as Covid-19 and SE1 have similar levels of spatio-temporal variation, as the posterior medians of $(\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22})$ are similar for both models. The table also shows substantial spatio-temporal and between outcome (call classification) dependencies, with the between outcome correlation $(\boldsymbol{\Sigma}_{12}/\sqrt{\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{22}})$ being very close to one in both cases. The spatial dependence is also high, because the posterior median for ρ equals 1 for both models which corresponds to the intrinsic CAR model for strong spatial dependence proposed by Besag et al. (1991). Substantial temporal dependence is also present in these data, because in the AR(1) and AR(2) models the respective 95% credible intervals for α and (α_1, α_2) are not close to zero which represents temporal independence. Finally, the DIC for the second order autoregressive process is lower than that for the first order autoregressive process suggesting it is the best fitting model considered here, and thus all subsequent inference is based on this model.

4.2 Spatio-temporal trends in the telehealth data

The temporal trends in the estimated (posterior medians) proportions of calls $\{\hat{\theta}_{kt1}, \hat{\theta}_{kt2}\}$ to NHS 24 classified as Covid-19 and SE1 are displayed in Figure 2, which has the same format as Figure 1, with Covid-19 in red and SE1 in blue. The estimated proportions exhibit much less noise than the raw proportions due to the spatio-temporal smoothing applied by the model, and the peak in the average proportions is 0.42 for Covid-19 and 0.49 for SE1 in the week beginning $23rd$ March. The trends in the estimated proportions are shown by LOESS curves, and the curve for SE1 has a steeper ascent and descent compared to that of Covid-19 which is more gradual. The average estimated proportions for the week beginning $6th$ July (the last week of the data) are similar to (Covid-19) or lower than (SE1) the estimated proportions for $2nd$ March (the first week of the data), suggesting the first wave of the epidemic is starting to come to an end.

The spatial patterns in the estimated proportions are presented in Figure 3 for the week beginning $23rd$ March, with this week being chosen as it

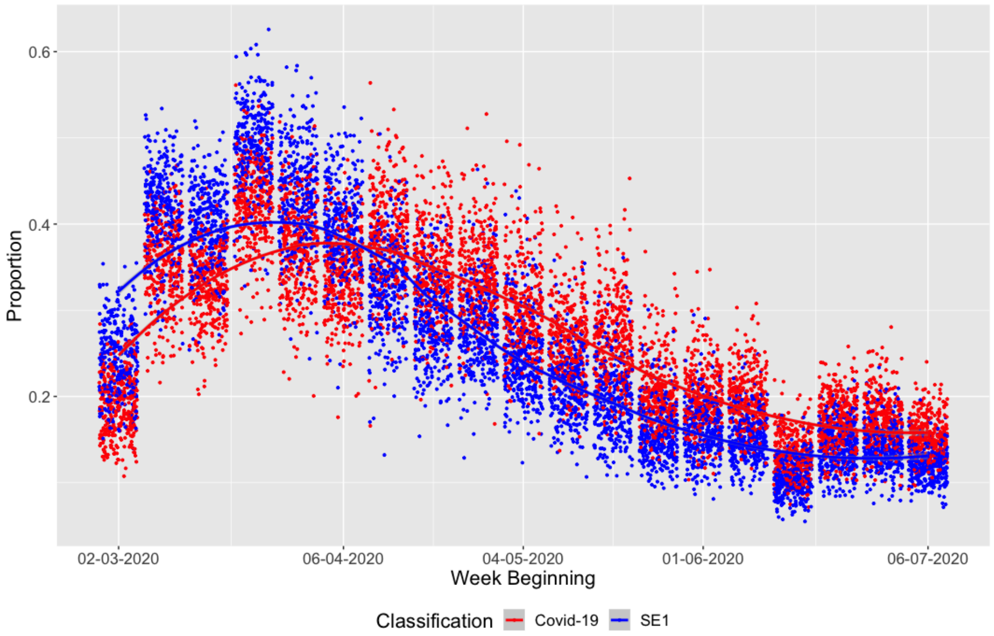

Figure 2: Scatterplot showing the temporal trends in the estimated proportions of calls to NHS 24 that were related to Covid-19 (red) and SE1 (blue) for all PDs as points, with LOESS trend lines superimposed. The points have been jittered in the Week Beginning (horizontal) direction to improve their visibility.

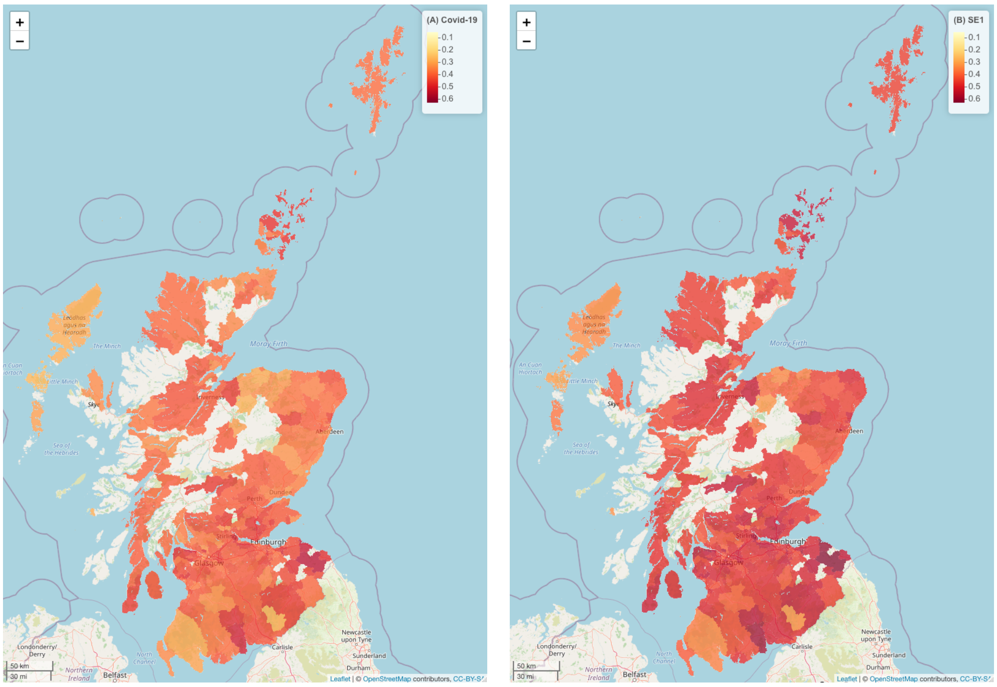

Figure 3: Maps displaying the spatial trends in the estimated proportions of NHS 24 calls due to (A) Covid-19 (left) and (B) SE1 (right) in the week beginning $23rd$ March.

has the highest average proportions. Additionally, an interactive Rshiny visualisation tool displaying the spatial patterns in these estimated proportions for Covid-19 for all weeks is available at `https://github.com/duncanplee/Covid-19-model`. The figure shows similar spatially smooth trends in the proportions for both outcomes, with the proportions generally being larger for SE1 compared to Covid-19. The highest proportions for this week are in the far south east of Scotland and along the southern bank of the river Clyde just west of Glasgow, while the lowest proportion areas are the Western Isles and parts of north east and south west of Scotland.

4.3 Assessing the use of telehealth data in predicting the epidemic

We assess the utility of the telehealth data in predicting the spatio-temporal dynamics of the epidemic, by comparing the estimated proportions $\{\hat{\theta}_{kt1}, \hat{\theta}_{kt2}\}$ with Covid-19 data summarising the numbers of confirmed cases and deaths. These latter data are only available publicly by Health Board, which are large administrative units in Scotland that are regionally autonomous bodies responsible for the organisation and delivery of health care in Scotland. Additionally, as the numbers of confirmed cases and deaths are much smaller than the numbers of calls to NHS 24, using the PD spatial scale for this comparison would result in noisy data with small counts obscuring the trends we wish to observe. The data on the numbers of cases each week has a sharp discontinuity between $14th$ and $15th$ June, because before this date only cases from tests carried out in NHS Scotland labs were included. However, after this date all historical cases identified from tests carried out in UK government labs over the entire pandemic to date were added to the data for this week. This discontinuity is therefore artificial, and to account for this the numbers of cases in the week beginning $8th$ June was imputed by averaging the numbers of new cases from the 2 weeks either side in time.

The PDs are nested exactly within Health Boards, with between 3 and 57 PDs being in each HB. Here we focus on the Greater Glasgow and Clyde, and Lothian Health Boards because they contain the two largest cities of Glasgow and Edinburgh respectively, with similar analysis for the remaining 12 Health Boards being available via an Rshiny visualisation tool at `https://github.com/duncanplee/Covid-19-model`. For each Health Board Figure 4 displays the temporal trends in: (i) the estimated telehealth proportions $\{\hat{\theta}_{kt1}, \hat{\theta}_{kt2}\}$; (ii) the numbers of confirmed cases of Covid-19; and (iii) the numbers of deaths due to Covid-19. For the telehealth data the trends in Figure 4 relate to the average of all $\{\hat{\theta}_{kt1}, \hat{\theta}_{kt2}\}$ within the Health Board in question. A common vertical axis for the different data series (which are on different scales) has been created by dividing the value for the current week by the maximum value of that series, giving a range of possible values for each series between 0 and 1.

The figure shows a good visual connection between the temporal trends for both the NHS 24 call types and the numbers of confirmed cases and deaths, which is consistently observed for both Health Boards. The rises and falls in both NHS 24 trends (Covid-19 and SE1) are followed by similar rises and falls in the confirmed cases and deaths, with lags of around 2 to 3 weeks between NHS calls and confirmed cases, and around 4 weeks between NHS 24 calls and deaths. The fall in these trends is more gradual for the Covid-19 call type compared to SE1, and the latter visually appears to better approximate the shape of the trends in the cases and deaths. Thus the telehealth data do appear to provide an early warning system for possible spikes in cases of and deaths due to Covid-19 between 2 to 4 weeks in the future.

4.4 Visualising the changing state of the epidemic

Our final goal in modelling these data is to visualise the spatial variation in the changing state of the epidemic, particularly focusing on identifying areas with higher and / or increasing proportions of calls to NHS 24 due to Covid-19 or

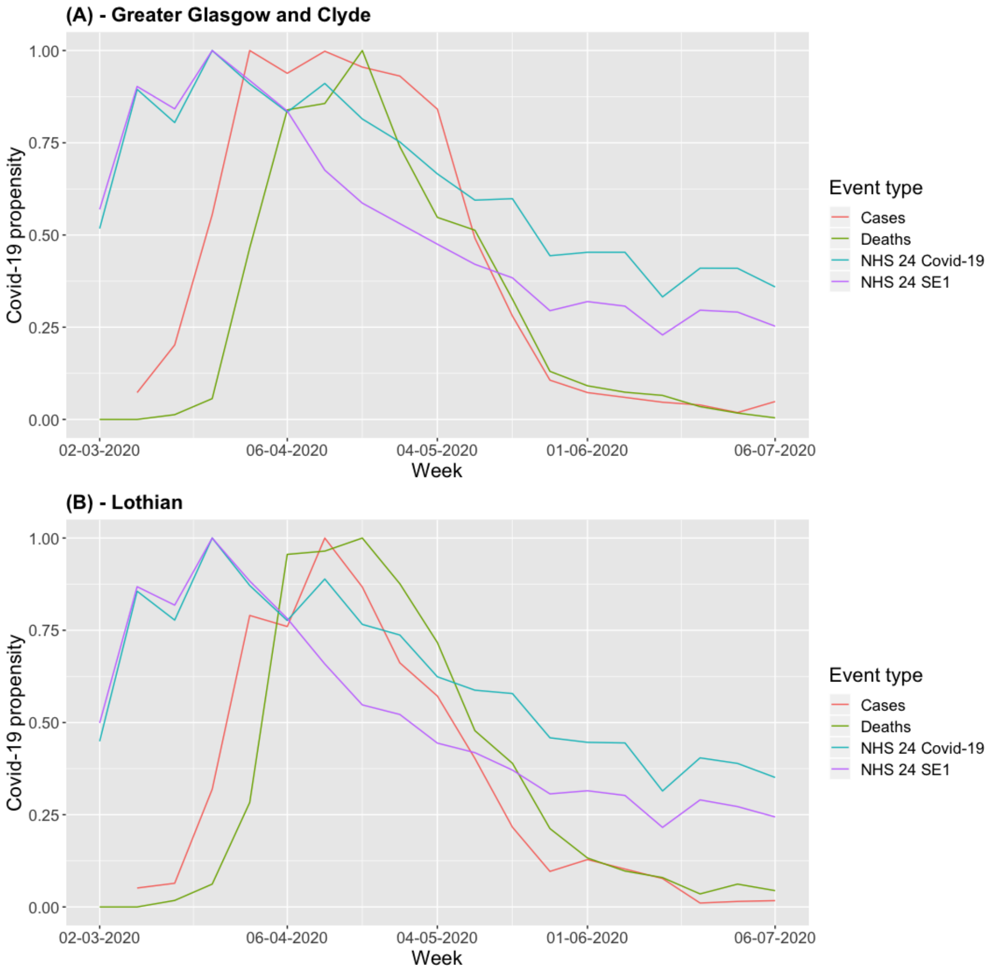

Figure 4: Temporal trends in the Covid-19 epidemic as measured by telehealth, case and death data for (A) Greater Glasgow and Clyde and (B) Lothian Health Boards. The vertical axis has been unified by dividing each series by its maximum value.

related symptoms (SE1). The preceding section illustrated that these areas are likely to see a spike in the numbers of cases in future weeks. We do this by computing for each PD k and week t the posterior distribution of

$$\{\delta_{ktj} = \theta_{ktj} - \theta_{k(t-1)j}\},$$

the change in the proportions of NHS 24 calls being about Covid-19 or related symptoms (SE1) between weeks $t-1$ and t. If $\delta_{ktj} > 0$ then the proportions of calls classified as Covid-19 or SE1 have increased compared to the previous week, suggesting there could be an increase in the numbers of cases in this area in the near future. In contrast, if $\delta_{ktj} < 0$ the numbers of cases is likely to reduce. We visualise $\{\delta_{ktj}\}$ for the week of $6th$ July, the final week in our data, by mapping: (i) the posterior median of δ_{ktj}; and (ii) the posterior probability that $\delta_{ktj} > 0$. These visualisations are presented in Figure 5 for calls classified as Covid-19, and the corresponding maps for SE1 are similar and are thus not shown for brevity. These maps for Covid-19 are also available via the Rshiny visualisation app at `https://github.com/duncanplee/Covid-19-model`.

The figure shows that the proportions of calls due to Covid-19 in the week beginning $6th$ May were generally lower than the previous week, with an average reduction across Scotland of 0.06. However, 8 (2%) PDs exhibited an increased risk, which can be clearly seen by the dark red colours in both the maps of the estimated differences (left) and the probabilities of an increase in the proportions (right). The areas that exhibit these likely increases are mainly in the far north of mainland Scotland and the Western, Orkney and Shetland islands, with only a small number of other areas showing evidence of an increase. The remainder of Scotland exhibits substantial evidence that the proportions of calls are decreasing, again suggesting a reduction in the epidemic during this period.

5 Discussion

This paper has developed a novel multivariate spatio-temporal surveillance model for Covid-19 in Scotland, using data from the national telehealth service NHS 24. The model estimates the joint spatio-temporal trends in the proportions of calls to NHS 24 classified as either Covid-19 directly or as having related symptoms (called SE1), and will be run on a weekly basis by Public Health Scotland as new data become available to monitor the changing dynamics of the epidemic. The multivariate spatio-temporal model developed here is fitted in a Bayesian paradigm using MCMC simulation, and code to fit the model as well as an Rshiny visualisation tool for the data analysed here are available at `https://github.com/duncanplee/Covid-19-model`.

The paper has presented two main findings from our data analysis, the first being that spatio-temporal trends in the proportions of NHS 24 calls classified as Covid-19 or SE1 do appear to provide a good early warning system for confirmed Covid-19 cases and deaths. This can clearly be seen from Figure 4, and the lag between NHS 24 calls and confirmed cases appears to be around 2 to 3 weeks, while the lag to deaths is a further week. The temporal trends in the proportions of calls to NHS 24 due to Covid-19 exhibit less variation than those for SE1, with shallower ascents and descents (see Figure 2). The shallower ascent may be an artifact of the Covid-19 classification coming from a prediction model before $14th$ April. The slower decline in the proportions for Covid-19 after the epidemic

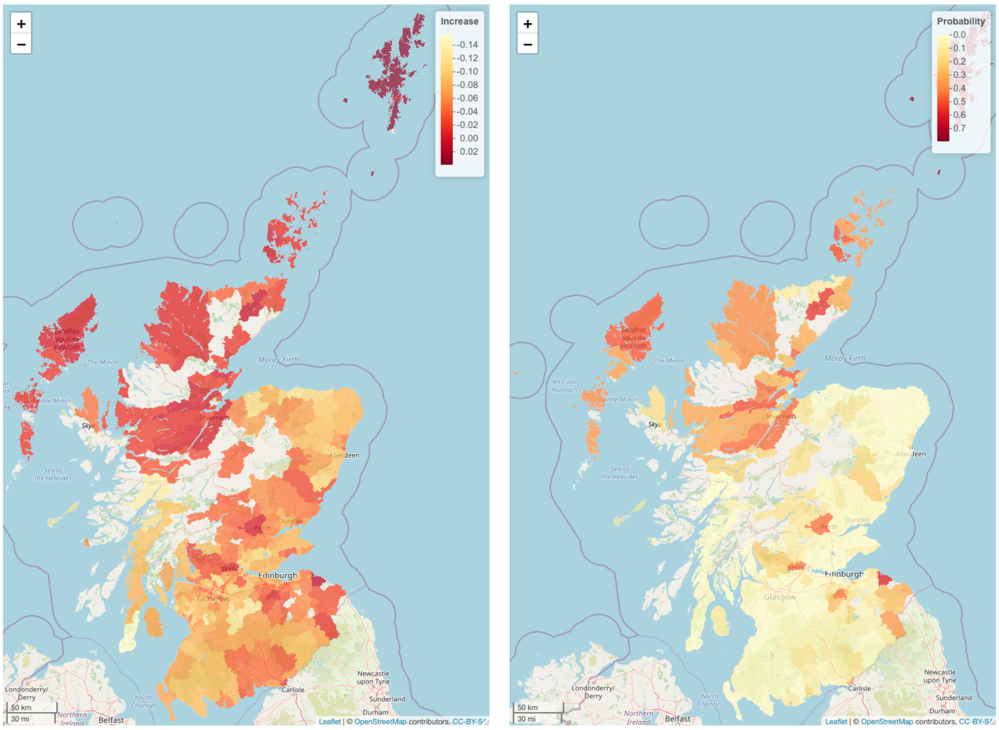

Figure 5: Maps displaying the posterior median difference in the proportions of calls classified as Covid-19 from the previous week δ_{kt} for $6th$ July (left) and the posterior probability that this difference is positive (right).

had peaked may be because Covid-19 is still at the forefront of people's minds, and calls are therefore more likely to be classified as Covid-19 than SE1. Thus the trends in SE1 visually appear to be a slightly better early warning predictor of future confirmed cases than calls classified as Covid-19, as its steeper descent more closely matches that of the confirmed cases and deaths.

Our second main finding concerns the estimated spatio-temporal dynamics of Covid-19 in Scotland, which peaked in terms of the telehealth calls around the week beginning $25th$ March, with confirmed cases and deaths peaking a few weeks later. Since then the proportions of calls classified as both Covid-19 and SE1 have fallen, suggesting that the first wave of the epidemic might be coming to an end. Finally, the development of such an early warning system using telehealth data has clear future applications to other diseases such as ordinary seasonal flu and outbreaks of norovirus, which would give the NHS better information on the likely prevalence of these diseases and where and when outbreaks are likely to occur.

Acknowledgements

The NHS 24 data were provided by Public Health Scotland.

References

Aregay, M., Lawson, A., Faes, C., and Kirby, R. (2017). Bayesian multi-scale modeling for aggregated disease mapping data. *Statistical Methods in Medical Research* **26,** 2726–2742.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics* **43,** 1–59.

Conticini, E., Frediani, B., and Caro, D. (2020). Can atmospheric pollution be considered a co-factor in extremely high level of sars-cov-2 lethality in northern italy? *Environmental Pollution* **261,** 114465.

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases* **20,** 533–534.

Douglas, M., Katikireddi, S., Taulbut, M., McKee, M., and McCartney, G. (2020). Mitigating the wider health effects of covid-19 pandemic response. *BMJ* **369,**.

Gelfand, A. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* **4,** 11–15.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis.* Chapman and Hall / CRC, 3rd edition.

Gomez-Rubio, V. and Palmi-Perales, F. (2019). Multivariate posterior inference for spatial models with the integrated nested laplace approximation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **68,** 199–215.

Jack, E., Lee, D., and Dean, N. (2019). Estimating the changing nature of scotland's health inequalities by using a multivariate spatiotemporal model. *Journal of the Royal Statistical Society Series A* page to appear.

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

Lee, D. and Lawson, A. (2016). Quantifying the spatial inequality and temporal trends in maternal smoking rates in Glasgow. *The Annals of Applied Statistics* **10,** 1427–1446.

Leroux, B., Lei, X., and Breslow, N. (2000). *Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence*, chapter Statistical Models in Epidemiology, the Environment and Clinical Trials, Halloran, M and Berry, D (eds), pages 135–178. Springer-Verlag, New York.

Martinez-Beneito, M. (2013). A general modelling framework for multivariate disease mapping. *Biometrika* **100,** 539–553.

Moran, P. (1950). Notes on continuous stochastic phenomena. *Biometrika* **37,** 17–23.

Quick, H., Waller, L., and Casper, M. (2017). Multivariate spatiotemporal modeling of age-specific stroke mortality. *Annals of Applied Statistics* **11,** 2165–2177.

Remuzzi, A. and Remuzzi, G. (2020). Covid-19 and italy: what next? *The Lancet* **395,** 1225–1228.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall / CRC.

Spiegelhalter, D., Best, N., Carlin, B., and Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B* **64,** 583–639.

Wu, X., Nethery, R., Sabath, B., Braun, D., and Dominici, F. (2020). Exposure to air pollution and covid-19 mortality in the united states: A nationwide cross-sectional study. *medRxiv preprint* page https://doi.org/10.1101/2020.04.05.20054502.