A spatio-temporal Covid-19 syndromic 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

August 10, 2020

Abstract

Predicting the locations of future Covid-19 outbreaks is of major public health importance, because it allows health agencies to target their efforts to subdue the virus where they are most needed. This paper presents a novel multivariate spatio-temporal syndromic surveillance model for Covid-19 in Scotland, which was developed in partnership with the national health agency Public Health Scotland. The model uses data from the country's national telehealth phone service NHS 24, and models the proportions of calls that are classified as being Covid-19 or having related symptoms. The model is applied to data up to the week beginning $27th$ July 2020, and is used to predict the likely locations for future disease outbreaks using posterior exceedance probabilities.

Keywords: Covid-19; Spatio-temporal modelling, Syndromic 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 $10th$ August, there were over 19.8 million cases worldwide, and over 731,000 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 (Conticini et al., 2020 and Wu et al., 2020); (iii) identifying the wider health impacts of the pandemic (Douglas et al., 2020); and (iv) developing real-time surveillance systems for identifying increases in disease prevalence (Dong et al., 2020). The development of real-time surveillance systems are a vital tool in the fight against the virus, because they allow public health agencies to monitor its spread and identify emerging hotspots of cases, thus allowing interventions to be taken such as the enforcement of local lockdowns.

Disease surveillance is a well researched topic (see for example Unkel et al., 2012 and Lawson, 2018), and surveillance systems can either directly model data relating to confirmed cases (e.g. Gu et al., 2017 and Bauer and Wakefield, 2018), or model proxy measures of disease prevalence that are available before confirmed diagnoses (e.g. Kavanagh et al., 2012 and Martin et al. (2019)). The latter is known as syndromic surveillance (Fricker, 2008), and while the data are of a lower quality in that they do not directly relate to confirmed cases of the disease, they are more timely and thus provide an early warning of where future confirmed cases are likely to occur. Numerous statistical approaches have been proposed for disease surveillance systems, including spatio-temporal models to identify trends in disease risk (e.g. Bauer and Wakefield, 2018 and Hale et al., 2019), and hypothesis testing via scan statistics to identify a limited number of hotspots (e.g. Wakefield and Kim (2013) and Tang et al. (2017)).

The aim of this paper is to develop a population-level spatio-temporal syndromic surveillance system for Covid-19 in Scotland, which can provide an early warning of areas that are likely to exhibit outbreaks of disease in the near future. Our model is applied to data from the Scottish National Health Service's (NHS) telehealth system NHS 24, and models the spatio-temporal trends in the proportions of calls categorised as Covid-19 or related symptoms. National telehealth data are ideal for building a syndromic 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. This is because people exhibiting symptoms of Covid-19 are likely to call NHS 24 for medical advice, and may be subsequently advised to be tested for the virus, hence inducing a lag between the NHS 24 call and the confirmed case. Additionally, NHS 24 data have been used previously in a simpler temporal syndromic surveillance setting for the 2009 H1N1 swine influenza pandemic in the UK (Kavanagh et al., 2012). Our surveillance system

is being run in real time by analysts in Public Health Scotland (PHS) to monitor the evolution of the Covid-19 epidemic in different areas of Scotland, allowing them to target public health interventions appropriately at areas at greatest risk from the virus.

Our surveillance system 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 from our telehealh model 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 (PHS, `https://publichealthscotland.scot/`) on the weekly numbers of calls to NHS 24 for Covid-19 and other similar

conditions, with a weekly temporal scale being used as it smooths out noise in the daily data caused by small numbers of calls and known day of the week effects (e.g. there are more calls during the weekends when doctors surgeries are closed). We have weekly data from the week beginning $2nd$ March 2020 to the week beginning $27th$ July 2020 inclusive, totalling $N = 22$ 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 area under the curve (AUC) of 0.88. Therefore to ensure the Covid-19 series covers the peak of the epidemic, we treat these predictions as observed data.

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 44 NHS 24 calls out of a total of 524,036 calls for all conditions they were removed from the data. After removing these PDs there were 1005 instances (PD and week combinations) with no NHS 24 calls at all, which were spread relatively evenly across the 22 weeks with between 34 and 56 instances per week. Therefore, to ensure a rectangular data set for analysis, only the $K = 328$ 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 0.7% 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 0.60 and 0.94, suggesting there is a strong relationship between them. This is further evidenced by the top panel (A) of 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 6.4% (Covid-19) and 7.4% (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 generally 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 coefficients across the $K = 328$ PDs are respectively 0.54 (Covid-19) and 0.70 (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 41% (Covid-19) and 23% (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. 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 pandemic 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 syndromic surveillance system can provide an early warning of future peaks in infection rates.

3. Construct and visualise disease surveillance metrics for identifying likely hotspots of future outbreaks, to support public health decision making in controlling the virus.

3 Methodology

This section proposes a new multivariate spatio-temporal (MVST) syndromic 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.

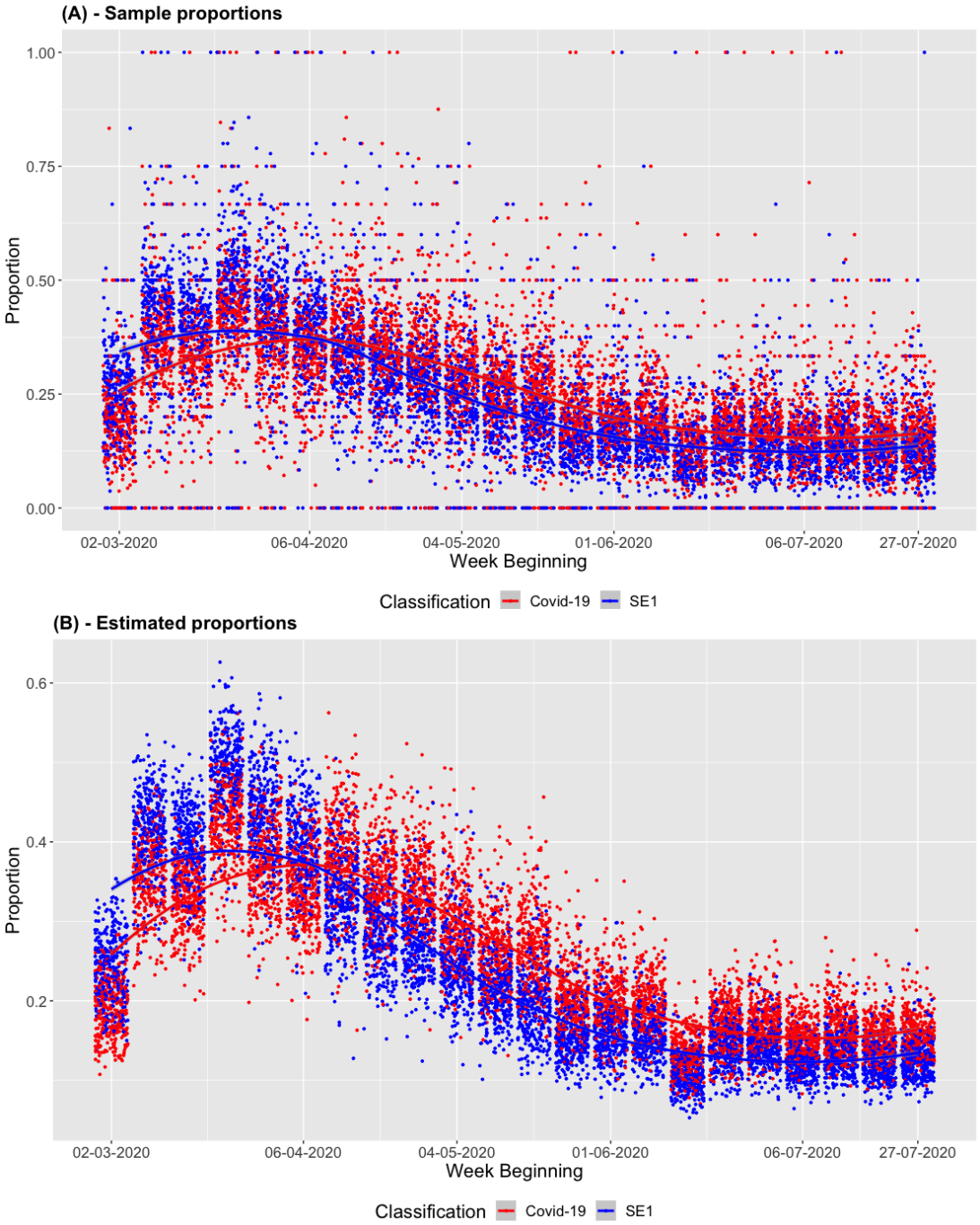

Figure 1: Scatterplots showing the temporal trends in the proportions of calls to NHS 24 that were related to Covid-19 (red) and SE1 (blue) for all PDs as points, with LOESS trend lines superimposed. The points have been jittered in the Week Beginning (horizontal) direction to improve their visibility.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

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 of the model hierarchy. The first level of the hierarchical model is given by:

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

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 $\beta_j \sim \text{N}(0, 100000)$, 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 (Rue and Held, 2005) model for $\boldsymbol{\phi}$. The general form of our model is given by

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

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 $\mathbf{\Sigma}_{J\times J}$ captures between outcome correlations. The precision matrix $\mathbf{P}(\boldsymbol{\alpha},\rho,\mathbf{\Sigma})$ 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 $\mathbf{P}(\boldsymbol{\alpha},\rho,\mathbf{\Sigma})$ 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 $\boldsymbol{\phi}_{-ktj,rsi}$:

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

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

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

The hyperparameters are set at $(d = J+1, \mathbf{\Omega} = 0.01\mathbf{I})$ 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

$$\begin{aligned} \mathbf{Q}(\mathbf{W},\rho) &= \rho(\text{diag}[\mathbf{W1}] - \mathbf{W}) + (1-\rho)\mathbf{I} \\ \rho &\sim \text{Uniform}(0,1). \end{aligned} \tag{5}$$

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

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

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

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

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

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

which is combined with the 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 syndromic surveillance in Scotland

This section presents the results of the Covid-19 syndromic surveillance model in Scotland, focusing on: (a) its estimation of the temporal trends in the telehelath data; (b) how well it tracks the epidemic of confirmed cases of, and deaths due to, Covid-19; and (c) the presentation of surveillance metrics for identifying the most likely locations of future outbreaks.

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 none of the values of $\hat{R}$ were above 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 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, albeit slightly larger for SE1 calls. 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 for both models. 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 would represent 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 (a) 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 the bottom panel of Figure 1, which has the same format as the top panel of the same figure, 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.

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)**
$\boldsymbol{\Sigma}_{11}$	0.152 (0.132, 0.172)	0.172 (0.153, 0.192)
$\boldsymbol{\Sigma}_{22}$	0.158 (0.138, 0.181)	0.183 (0.163, 0.205)
$\boldsymbol{\Sigma}_{12}/\sqrt{\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{22}}$	0.995 (0.993, 0.997)	0.995 (0.993, 0.997)
ρ	1.000 (0.999, 1.000)	1.000 (0.999, 1.000)
$\boldsymbol{\alpha}$	α 0.686 (0.640, 0.731)	α_1 0.421 (0.364, 0.481) α_2 0.324 (0.260, 0.385)
DIC	68,177	68,059
p.d	2,610	2,723

The average estimated proportions for the week beginning $27th$ 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 has come to an end.

4.3 (b) Exploratory assessment of the syndromic ability of telehealth data as an early warning system

We visually assess the extent to which the estimated trends $\{\hat{\theta}_{kt1}, \hat{\theta}_{kt2}\}$ provide an early warning of similar trends in the proportions of positive test results (representing confirmed cases) and the numbers of deaths. We make this comparison at the Health Board scale, because data on positive tests and deaths are only available publicly at this much larger spatial scale. Health Boards 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 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 2 displays the temporal trends in: (i) the estimated telehealth proportions $\{\hat{\theta}_{kt1}, \hat{\theta}_{kt2}\}$; (ii) the proportion of tests for Covid-19 that were positive; and (iii) the numbers of deaths due to Covid-19. For the telehealth data the trends in Figure 2 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 NHS 24 call types and positive tests and deaths, which is consistently observed for

both Health Boards. Simple correlation analysis using Pearson's correlation coefficients shows that the highest correlations between NHS 24 calls classified as Covid-19 and the proportion of positive tests occurred at a 1-week lag, with values of 0.88 (Greater Glasgow and Clyde) and 0.90 (Lothian) respectively. Similarly, the highest correlations between NHS 24 calls classified as Covid-19 and deaths occurred at a lag of 3 weeks, with correlation coefficients of 0.89 (Greater Glasgow and Clyde) and 0.87 (Lothian) respectively. However, while the telehealth data do appear to provide an early warning system for possible spikes in cases of and deaths due to Covid-19, the assessment of the lag times should only be seen as exploratory. This is because there have been changes to the Covid-19 testing regime throughout the pandemic, which complicates this comparison. For example, more testing and more targeted testing has been carried out in the later weeks of the data, due to the introduction of the contact tracing 'Test and Protect' system (`https://www.nhsinform.scot/campaigns/test-and-protect`) on $28th$ May.

4.4 (c) Surveillance of possible future outbreaks

The main goal of our analysis is prospective Covid-19 surveillance, by using the model to identify the locations of possible future outbreaks of confirmed cases. The most obvious starting point for this are maps of the estimated proportions of the NHS 24 calls classified as Covid-19 or related symptoms (SE1) for the final week of the data, namely the week beginning $27th$ July 2020. Figure 3 displays these maps for both classifications, while 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`. Figure 4 shows that the two spatial surfaces are highly correlated (Pearson's correlation coefficient is 0.999), with the Covid-19 classification always exhibiting higher estimated proportions than SE1. A number of postcode districts exhibit increased proportions of around 0.2 to 0.25 of NHS 24 calls classified as Covid-19, most notably in Aberdeenshire in the north east and parts of Ayrshire in the south west just south of Glasgow.

In addition to these estimated proportions, existing disease surveillance systems use a variety of metrics to signal where outbreaks are likely to occur, and examples are given in Kavanagh et al. (2012), Unkel et al. (2012), Lawson (2018) and Hale et al. (2019). A number of these metrics are based on posterior exceedance probabilities (PEP), which in our context is the posterior probability that the proportion $\{\theta_{ktj}\}$ exceeds a certain threshold. Here we consider the following relative and absolute exceedance probabilities

$$\pi_{ktj}^{rel} = \mathbb{P}\left(\theta_{ktj} > \bar{\theta}_{.tj} = \frac{\sum_{k=1}^{K} Y_{ktj}}{\sum_{k=1}^{K} N_{kt}} \middle| \mathbf{Y}\right) \quad \text{and} \quad \pi_{ktj}^{abs} = \mathbb{P}(\theta_{ktj} > \epsilon = 0.2|\mathbf{Y}), \tag{8}$$

where π_{ktj}^{rel} is an exceedance relative to the Scottish national average proportion $\bar{\theta}_{.tj} = \sum_{k=1}^{K} Y_{ktj} / \sum_{k=1}^{K} N_{kt}$ for the week in question. This relative PEP is produced weekly by Public Health Scotland to monitor the changing state of the pandemic. In contrast π_{ktj}^{abs} is an absolute exceedance probability relative to a fixed threshold ϵ, whose specification is always going to be somewhat arbitrary. Here we choose $\epsilon = 0.2$ because for the final week of the data

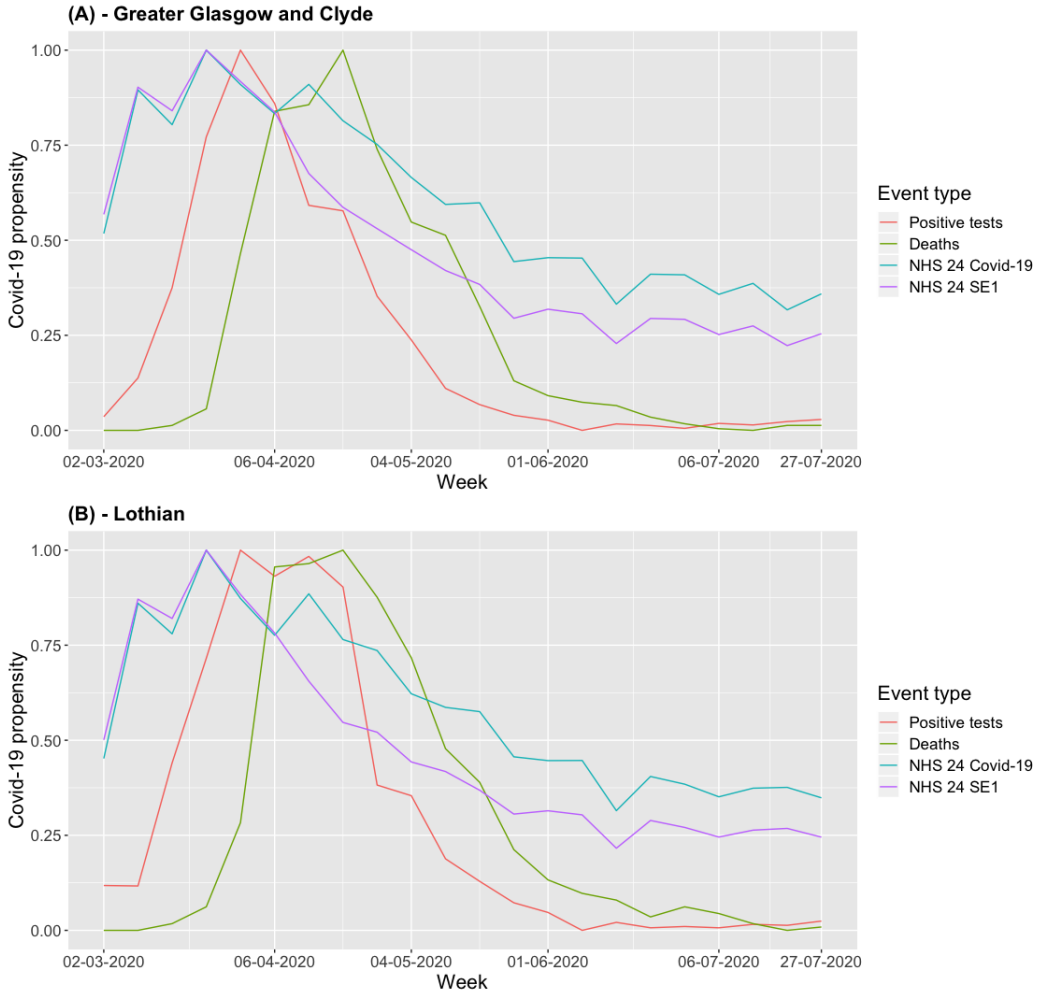

Figure 2: Temporal trends in the Covid-19 pandemic as measured by telehealth, positive tests (cases) and death data for (A) Greater Glasgow and Clyde and (B) Lothian Health Boards. The vertical axis has been unified by dividing each series by its maximum value.

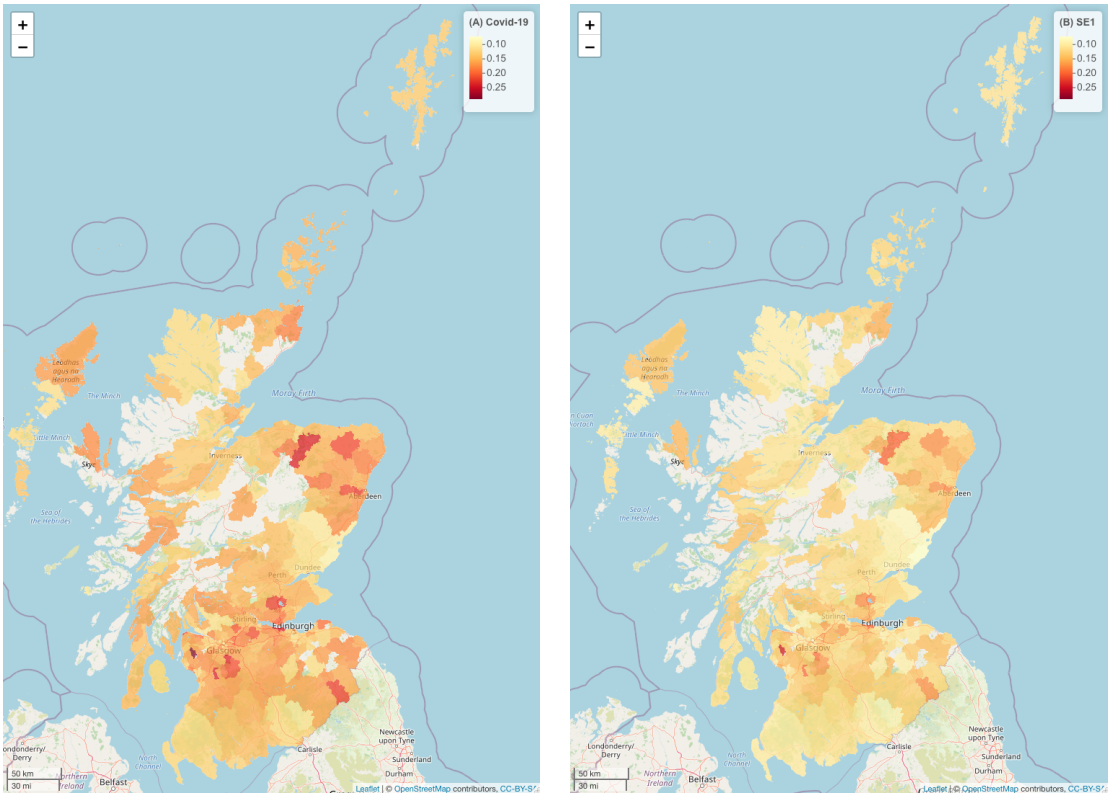

Figure 3: Maps displaying the spatial trends in the estimated proportions of NHS 24 calls due to (A) Covid-19 (left) and (B) SE1 (right) in the week beginning 27*th* July.

$\bar{\theta}_{.t1} = 0.16$ for the Covid-19 classification, so this absolute threshold should signal a small number of areas that warrant further investigation by Public Health Scotland. Maps of these two exceedance probabilities for the Covid-19 call type are displayed in Figure 4 for the week beginning $27th$ July. The figure shows there are a sizeable number of PDs with high relative PEP values, which is because it compares each PD to the Scottish average. Thus in this example the absolute PEP measure which is based on a higher threshold of $\epsilon = 0.2$ might be preferable, because it only signals those PDs that have the highest estimated proportions for further investigation. One of these locations is the city of Aberdeen and the surrounding area, which subsequently saw an outbreak of cases in early August a few days after the time frame of our data. Interactive maps displaying these PEPs are available via an Rshiny visualisation tool at `https://github.com/duncanplee/Covid-19-model`.

5 Discussion

This paper has developed a novel multivariate spatio-temporal syndromic 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 a simplification of the model using only the Covid-19 classification is 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 2, and the lag between NHS 24 calls and confirmed cases appears to be around 1 week, while the lag to deaths is around a further 2 weeks. 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 1). The shallower ascent may be an artifact of the Covid-19 classification coming from a prediction model before $14th$ April, or alternatively it may be that as the epidemic became more prevalent from late March on-wards people might be more likely to mention Covid-19 directly when they called NHS 24. Similarly, the slower decline in the proportions for Covid-19 after the epidemic 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

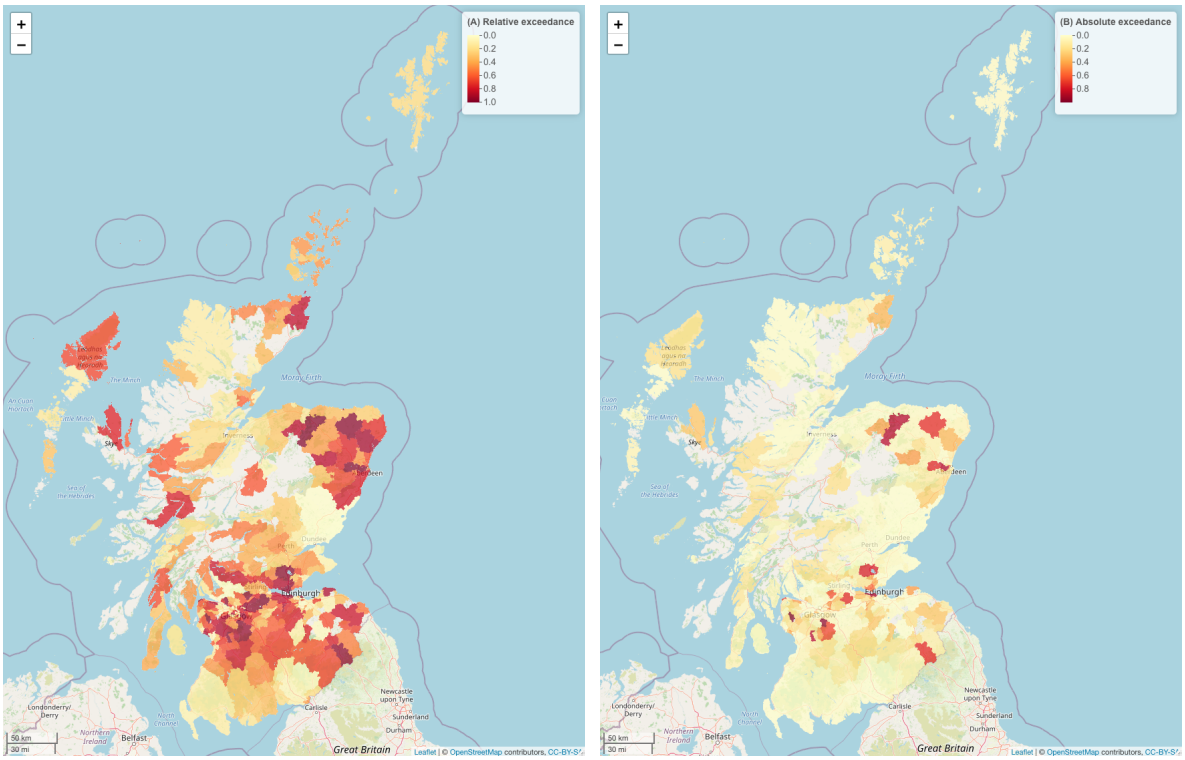

Figure 4: Maps displaying the posterior exceedance probabilities for the: (A) relative threshold; and (B) absolute threshold; given by (8) for the week beginning $27th$ July.

SE1 have fallen, suggesting that the first wave of the epidemic might be coming to an end. However, Figure 4 shows that Covid-19 has not disappeared from Scotland, because it highlights local areas that are likely to exhibit renewed outbreaks in the near future. Indeed, the area highlighted around Aberdeen exhibited a cluster of cases shortly after the timeframe of the data analysed here. 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., A. Lawson, C. Faes, and R. Kirby (2017). Bayesian multi-scale modeling for aggregated disease mapping data. *Statistical Methods in Medical Research 26*, 2726–2742.

Bauer, C. and J. Wakefield (2018). Stratified space–time infectious disease modelling, with an application to hand, foot and mouth disease in china. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 67*, 1379–1398.

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

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

Douglas, M., S. Katikireddi, M. Taulbut, M. McKee, and G. McCartney (2020). Mitigating the wider health effects of covid-19 pandemic response. *BMJ 369*, m1557.

Fricker, D. (2008). *Encyclopedia of Quantitative Risk Analysis and Assessment*, Chapter Syndromic Surveillance, pp. 1743–1752. John Wiley and Sons.

Gelfand, A. and P. Vounatsou (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics 4*(1), 11–15.

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

Gomez-Rubio, V. and F. Palmi-Perales (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.

Gu, H., W. Fan, K. Liu, S. Qin, X. Li, J. Jiang, E. Chen, Y. Zhou, and Q. Jiang (2017). Spatio-temporal variations of typhoid and paratyphoid fevers in zhejiang province, china from 2005 to 2015. *Scientific Reports 7*, 5780.

Hale, A., F. Sanchez-Vizcaino, B. Rowlingson, A. Radford, E. Giorgi, S. O'Brien, and P. Diggle (2019). A real-time spatio-surveillance system with application to small companion animals. *Scientific Reports 9*, 17738.

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

Kavanagh, K., C. Robertson, H. Murdoch, G. Crooks, and J. McMenamin (2012). Syndromic surveillance of influenza-like illness in scotland during the influenza a h1n1v pandemic and beyond. *Journal of the Royal Statistical Society: Series A 175*, 939–958.

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

Lawson, A. (2018). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, Third Edition*. Chapman and Hall/CRC.

Lee, D. and A. Lawson (2016). Quantifying the spatial inequality and temporal trends in maternal smoking rates in Glasgow. *The Annals of Applied Statistics 10*(3), 1427–1446.

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

Martin, L., H. Dong, Q. Liu, J. Talbot, W. Qiu, and Y. Yasui (2019). Predicting influenza-like illness-related emergency department visits by modelling spatiotemporal syndromic surveillance data. *Epidemiology and Infection 147*, e312.

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

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

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

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

Tang, X., A. Geater, E. McNeil, Q. Deng, A. Dong, and G. Zhong (2017). Spatial, temporal and spatio-temporal clusters of measles incidence at the county level in guangxi, china during 2004–2014: flexibly shaped scan statistics. *BMC Infectious Diseases 17*, 243.

Unkel, S., C. Farrington, P. Garthwaite, C. Robertson, and N. Andrews (2012). Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A 175*, 49–82.

Wakefield, J. and A. Kim (2013). A Bayesian model for cluster detection. *Biostatistics 14*, 752–765.

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