Real-time spatial health surveillance: mapping the UK COVID-19 epidemic

Richard Fry\textsuperscript{1,3,*}, Joe Hollinghurst\textsuperscript{1,3}, Helen R Stagg\textsuperscript{4}, Daniel A Thompson\textsuperscript{1,3}, Claudio Fronterre\textsuperscript{2}, Chris Orton\textsuperscript{1,3}, Ronan A Lyons\textsuperscript{1,3}, David V Ford\textsuperscript{1,3}, Aziz Sheikh\textsuperscript{1,4}, and Peter J Diggle\textsuperscript{1,2}

\textsuperscript{1}Health Data Research, UK
\textsuperscript{2}Medical School, Lancaster University
\textsuperscript{3}Swansea University Medical School
\textsuperscript{4}Usher Institute, University of Edinburgh

*Corresponding author: Richard Fry, r.j.fry@swansea.ac.uk

Abstract The COVID-19 pandemic has highlighted the need for robust data linkage systems and methods for identifying outbreaks of disease in near real-time. Using self-reported app data and the Secure Anonymised Information Linkage (SAIL) Databank, we demonstrate the use of sophisticated spatial modelling for near-real-time prediction of COVID-19 prevalence at small-area resolution to inform strategic government policy areas. A pre-requisite to an effective control strategy is that predictions need to be accompanied by estimates of their precision, to guard against over-reaction to potentially spurious features of ‘best guess’ predictions. In the UK, important emerging risk-factors such as social deprivation or ethnicity vary over small distances, hence risk needs to be modelled at fine spatial resolution to avoid

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
aggregation bias. We demonstrate that existing geospatial statistical methods originally developed for global health applications are well-suited to this task and can be used in an anonymised databank environment, thus preserving the privacy of the individuals who contribute their data.

1 Introduction

On 11th March 2020, the World Health Organization declared a pandemic of COVID-19 caused by the SARS-CoV-2 coronavirus [1]. By this date, the UK had reported 373 confirmed COVID-19 cases and six deaths [2]. Up to 15th July 2020, these figures had risen to 291,911 and 45,053 [3]. Lockdowns governing the movement of the population and opening of shops and other facilities, initially imposed across the entire country on 23rd March 2020 [4], have been a key tool in the government’s response to COVID-19. Since that date, detection of local variations in infection rates has been critical for controlling the spread of SARS-CoV-2 [5], including ascertaining the level of required local public health response across the UK. A key example of this was the implementation of the first ‘local lockdown’ in Leicester on 30th June 2020, in response to a cluster of COVID-19 accounting for approximately one in ten of all new disease cases across the country in the preceding week [6].

The COVID Symptom Study app (Zoe Global Limited, King’s College London) was released publicly on 24th March 2020 [7], the day after the UK-wide lockdown rules were first imposed. The app collects postcode of residence at the time of registration and daily updates on self-reported COVID-19-associated symptoms. The Secure Anonymised Information Linkage (SAIL) Databank facilitates robust secure storage and use of anonymised person-based data for research to improve health, well-being and services [8, 9]. During the pandemic, SAIL has been receiving daily updates of the COVID Symptom Study app data, facilitating near real-time health surveillance of COVID-19 across the UK.
To help understand the localised spread and flare up of the disease we adapted existing statistical methodology [10] for the analysis of geo-referenced health outcome data to map, at Lower-layer Super Output Area (LSOA) resolution (Datazone in Scotland and Super Output Area in Northern Ireland), the prevalence of positive symptom reports amongst app users over a rolling 14-day period, together with associated limits of statistical uncertainty. Notwithstanding the limitations of this self-reported health outcome, these maps provide the first fine-scale, UK-wide assessment of the geographical distribution of probable COVID-19 infections, and have been used by the devolved administrations in each country for pandemic planning [11][12].

2 Methods

2.1 Population Sampling

Across the UK, as of 15th July 2020, 4,002,397 individuals had registered on the COVID Symptom Study app. Use of the app is voluntary and thus the population sample is non-random. Users of the app had to have access to an internet-enabled telephone, although reporting for multiple individuals in the same household was instigated on 1st May 2020 for those people who could not access or use the app [13]. At the time of registration, individuals report baseline demographic and clinical information (e.g. underlying health conditions), as well as postcode of residence. Self-reported data on COVID-19-associated symptoms, including fever and persistent cough, are recorded for any day on which an individual reports. For a full metadata summary see the HDR Gateway deposit [14].

The SAIL Databank acts as a secure gateway to the ZOE app data for the whole of the UK. Data are made available daily via a secure data transfer and processed into a SQL DB2 database. Access to the data is via a secure remote desktop login following approval for a project via an application to the SAIL Information Governance Review Panel (IGRP). In light
of the COVID-19 crisis, IGRP applications were typically approved within 24-hours. Prior to transfer of the data to the SAIL Databank, postcode data were aggregated to LSOA level using the Office for National Statistics (ONS) postcode lookup directory [15] to maintain an app user’s privacy.

2.2 Informatics

Data were extracted from the SAIL Databank using SQL and processed to generate suitable inputs for the geospatial modelling. Where an individual reported their symptoms more than once in a day, the last record was taken. Likely instances of COVID-19 were calculated through either a) the presence of high fever and persistent cough (‘classic symptoms’) or b) an algorithm developed by the King’s College team, which used an array of symptoms and other characteristics (persistent cough, skipping meals, loss of smell, gender, age and fatigue; ‘multi-symptom algorithm’), [16]. The denominator of users in each LSOA for each analysis was calculated using a 14-day retrospective window i.e. the number of individuals who had reported data to the app at any time during that period. The numbers of app users and cases were then aggregated to the LSOA level. The resulting data for each LSOA consisted of its population-weighted centroid, \( x \), the number of people who used the app at least once over the time-period in question, \( n \), and the number of those who were predicted to have COVID-19 at least once within the time-period, \( y \).

2.3 Geospatial statistical model and inference

Our statistical model is a geospatial extension of the logistic regression model for binomial (numerator/denominator) data, in which the log-odds of the probability, \( P(x) \), of at least one positive symptom report is the unobserved realisation of a spatially correlated stochastic process and, conditional on \( P(x) \), the corresponding numerator \( y \) follows a binomial distribution.
with denominator $n$. The model has three parameters that determine the mean and variance of $P(x)$ and the rate at which the correlation between the values of $P(x)$ at two different locations decays with increasing distance between them.

We estimated the parameters using Monte Carlo maximum likelihood and used the fitted model to draw samples from the joint predictive distribution of $P(x)$ over all LSOA population weighted centroids. Parameters were re-estimated separately for each of the UK’s constituent countries, in each case using data aggregated over a rolling 14-day time-period. In the Supplementary Material we describe in detail how we developed and fitted the particular model that we used for our application to the COVID-19 app data.

The complete prediction for each LSOA is a probability distribution for its underlying prevalence. This distribution can be summarised as the user wishes. We chose to map four summaries: the mean, a point prediction of prevalence; the 5% and 95% quantiles, which together measure the uncertainty associated with each point prediction; and the probability that the prevalence in the LSOA in question is greater than the country-wide average for each devolved nation, with mapped values close to 1 or 0 indicating “hot-spots” and “cold-spots” respectively. If the primary aim of the mapping is hot-spot detection, additional useful summaries would be the probabilities that local prevalence exceeds each of a set of thresholds representing increasing multiples of the country-wide average. Patches of mapped probabilities close to 1 would then indicate both the geographical extent and magnitude of local hot-spots. Predicted prevalence data were summarised for the whole of the UK, and for each of its four constituent countries.
3 Results

3.1 App users

Table 1 summarises the users of the app for the UK as of 14th July 2020. Across the UK users were predominately female, white, between 30 and 55 and lived in the least deprived areas. It is also worth noting that the 20% of users did not provide a full postcode thereby limiting the utility of these users data provision for the purposes of high-resolution spatial modelling.

3.2 App usage over time

There was no national government requirement for members of the public to use the app. In each country, the number of people registering to use the system increased rapidly in the early weeks of its availability, and more slowly thereafter (Figure 1 top panel). In England, Scotland and Wales the number of active users (people who recorded one or more app submission in the preceding 14 days) also increased between mid-April and early May, but declined thereafter, from a peak of around 60% in early May to around 45% in mid July. (Figure 1 bottom panel). In Northern Ireland, where the app is, in effect, competing with Northern Ireland’s own app the percentage of active users peaked at around 50% in early May and had declined to about 30% by mid July.

3.3 Predicted prevalence of COVID-19 over time

Predicted disease prevalence over time, weighted for population size within each LSOA, was plotted using both classic symptoms and the multi-symptom algorithm. Both provided similar patterns of predicted disease prevalence for the first two weeks of data collection (Figure 2), the figures from the multi-symptom algorithm were higher than those using classic symptoms (classic symptoms 3.6% to 4.3% across the four countries at their peak, multi-symptom...
# Table 1: Summary of registered ZOE Symptom Study App users as of 14th July 2020

<table>
<thead>
<tr>
<th>Registered App Users</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>3840824</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Median (IQR)</strong></td>
<td>40 (30,55)</td>
<td>100%</td>
</tr>
<tr>
<td>0-10</td>
<td>154091</td>
<td>4.0%</td>
</tr>
<tr>
<td>15-20</td>
<td>346591</td>
<td>9.0%</td>
</tr>
<tr>
<td>25-30</td>
<td>692117</td>
<td>18.0%</td>
</tr>
<tr>
<td>35-40</td>
<td>714001</td>
<td>18.6%</td>
</tr>
<tr>
<td>45-50</td>
<td>682677</td>
<td>17.8%</td>
</tr>
<tr>
<td>55-60</td>
<td>536308</td>
<td>14.0%</td>
</tr>
<tr>
<td>65-70</td>
<td>321197</td>
<td>8.4%</td>
</tr>
<tr>
<td>75-80</td>
<td>85531</td>
<td>2.2%</td>
</tr>
<tr>
<td>85+</td>
<td>21204</td>
<td>0.6%</td>
</tr>
<tr>
<td>Not Recorded</td>
<td>307360</td>
<td>8.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Female</strong></td>
<td>2150324</td>
<td>56.0%</td>
</tr>
<tr>
<td>Male</td>
<td>1380425</td>
<td>35.9%</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>2458</td>
<td>0.1%</td>
</tr>
<tr>
<td>intersex</td>
<td>256</td>
<td>0.1%</td>
</tr>
<tr>
<td>Not Recorded</td>
<td>307361</td>
<td>8.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic</td>
<td>52</td>
<td>0.1%</td>
</tr>
<tr>
<td>Other</td>
<td>11883</td>
<td>0.3%</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>10322</td>
<td>0.3%</td>
</tr>
<tr>
<td>Asian (UK)</td>
<td>56140</td>
<td>1.5%</td>
</tr>
<tr>
<td>Black (UK)</td>
<td>15935</td>
<td>0.4%</td>
</tr>
<tr>
<td>Chinese</td>
<td>9134</td>
<td>0.2%</td>
</tr>
<tr>
<td>Middle Eastern</td>
<td>10154</td>
<td>0.3%</td>
</tr>
<tr>
<td>Mixed (other)</td>
<td>30974</td>
<td>0.8%</td>
</tr>
<tr>
<td>Mixed (White/Black)</td>
<td>16051</td>
<td>0.4%</td>
</tr>
<tr>
<td>White (UK)</td>
<td>2448755</td>
<td>63.8%</td>
</tr>
<tr>
<td>US Residents</td>
<td>1151</td>
<td>0.1%</td>
</tr>
<tr>
<td>Not Recorded</td>
<td>1230273</td>
<td>32.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Deprivation</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Townsend Quintile</td>
<td>3067197</td>
<td>79.9%</td>
</tr>
<tr>
<td>1. Least deprived</td>
<td>814557</td>
<td>26.6%</td>
</tr>
<tr>
<td>2</td>
<td>741201</td>
<td>24.2%</td>
</tr>
<tr>
<td>3</td>
<td>629180</td>
<td>20.5%</td>
</tr>
<tr>
<td>4</td>
<td>490686</td>
<td>16.0%</td>
</tr>
<tr>
<td>5. Most deprived</td>
<td>391573</td>
<td>12.8%</td>
</tr>
</tbody>
</table>
algorithm 6.0% to 7.5%) but diverged thereafter. Using classic symptoms, predicted prevalence slowly declined to mid-May and then was approximately stable at slightly more than 2% in Wales, slightly less than 2% in England, Scotland and Northern Ireland (Figure 2, upper panel). Using the multi-symptom algorithm, the decline was more rapid, stabilising around two weeks earlier at approximately 0.4% in all four countries. (Figure 2, lower panel)
Figure 2: Longitudinal trajectories of symptom-prevalence (percentage of active users reporting symptoms) for each of the four UK countries up to 14th July 2020, based on classic symptoms (upper panel) and multi-symptom algorithm (lower panel).

### 3.4 Predictive mapping

Predictive mapping at LSOA-level geography based on inputs derived using the prevalence algorithm described in Menni et al. [16] revealed small-scale spatial variation in disease prevalence, which varied over time. Figures 3 and 4 show UK-wide LSOA-level maps of predicted prevalence and predictive probability that each LSOA exceeded the national average prevalence, over the pandemic to 15th July 2020. Most hot-spots (bright yellow areas in Figure 4) were located in or close to major cities, with Aberdeen and Bristol as notable exceptions.
Prevalence and exceedance probability maps need to be interpreted in combination. The prevalence exceedance (hot-spot) maps focus attention on areas that show statistically significant deviations above the national averages. However, all other things being equal, these were likely to occur in areas that have high population density and, consequently, deliver more precise local predictions. The three-panel format of Figure 5 facilitates this combined interpretation by allowing the reader to check whether areas indicated in Figure 4 as hot-spots are also areas whose predicted prevalence is markedly high, and conversely, whilst the left-hand and right-hand panels act as a guard against over-interpretation of imprecise point predictions. For example, the centre-panel of Figure 5 shows that the largely rural area of west Cumbria had relatively high prevalence over the 14-day period ending 28th June, but the associated probability limits were wide, and the bottom row, second-left column of Figure 4 does not indicate any part of west Cumbria to have been a hot-spot. Conversely, over this same period, prevalence levels in London were no longer among the highest in the country (Figure 5, centre panel), but were nevertheless almost certainly above the English national average (Figure 4, bottom row, second-left column).
Figure 3: LSOA Level COVID-19 prevalence predictions using a 14-day window 9\textsuperscript{th} March - 14\textsuperscript{th} July 2020. Purples represent low values, green mid range and yellow high values.
Figure 4: Predictive probabilities for LSOA-level prevalence to exceed the UK-wide average using a 14-day window 29th March - 14th July 2020. Purples represent low probabilities, green mid range and yellow high probabilities.
Figure 5: Predicted (centre panel (b)), lower 5% (left-hand panel (a)) and upper 95% (right-hand panel (c)) limits for the probability that LSOA-level prevalence to exceed the UK-wide average over the 14-day period ending 28th June 2020. Purples represent low values, green midrange and yellow high values.

4 Discussion

The COVID-19 pandemic has shown how a combination of skills (Health Informatics, Statistics and Geography) can provide insights to inform local and national government policy at a UK level. Building on the HDR UK ‘One Institute’ principles, we have prototyped and delivered data infrastructures and analysis pipelines capable of delivering timely and insightful analytics to all levels of government. These have formed a cornerstone of the COVID-19 response in Scotland and Wales in particular, and illustrate how fine-grained spatio-temporal inferential mapping tools for near-real-time crowd-sourced georeferenced health outcome data are critical to inform rapid public health responses.

The platform established within the SAIL system should not be seen as uniquely applicable to data from the COVID Symptom Study app, or indeed to COVID-19. Such tools can be
applied to any source of georeferenced health data, including when comparing data sourced via
different detection and testing platforms, each of which may have a different target population,
sensitivity and specificity. We note the particular relevance to newly emerging infectious
conditions – where spatiotemporal mapping of spread is critical – and thus the need to maintain
such platforms so that they can be activated as each pandemic arises, e.g. for influenza\textsuperscript{18}.

The required geospatial methodology for these tools has been available since the early years of
the century\textsuperscript{19}. Diggle, Rowlingson and Su\textsuperscript{20} report on a real-time surveillance system for
calls to the now-defunct NHS Direct on-line triage service\textsuperscript{21} for which the primary reason
for the call was recorded as non-specific gastro-intestinal illness. This system was developed
in collaboration with the Southampton Public Health Service and ran in prototype form over
the years 2001 to 2003.

There has been a proliferation of mapped outputs related to the COVID-19 pandemic with
policy makers and the general public equally concerned with tracking the pandemic. However,
most of the published maps have been at a regional level\textsuperscript{7, 5, 22}. Although valuable, these
can mask localised hotspots as exemplified by the Leicester, UK outbreak. In situations where
expedient decisions are required, mapping statistical outputs to granular geographies, with
supporting information relating to confidence intervals, gives policy makers better information
to take necessary action. Also, by recognising and exploiting spatial correlation in the
underlying prevalence surface, geospatial statistical methods can deliver substantially more
precise estimation of local prevalence than classical methods that implicitly assume independence
of outcomes in different spatial units\textsuperscript{23}.

We acknowledge the limitations that come with using self-reported symptom data from an
app used voluntarily. Firstly, confirmation that self-reported symptoms did indeed represent
COVID-19 disease was not possible at a UK level, although the multi-symptom algorithm
utilised was generated using predictive regression modelling comparing symptoms to self-reported
reverse transcription polymerase chain reaction SARS-CoV-2 test results\textsuperscript{16}. Secondly, the
individuals included in the studied population are not a random sample of the UK population,
potentially presenting a source of collider bias due to the link between age and app usage\cite{24},
nor are they necessarily representative with respect to other factors that are either known,
or thought likely, to affect susceptibility; for example, gender or ethnicity. The inclusion in
the model of LSOA-level covariate information is a potential route to controlling for these
at LSOA level, although not at individual level. For example, in the current COVID-19
context information on the age distribution of app users would allow adjustment for the
potentially non-representative sub-population of active app users. For environmentally driven
health outcomes, such as asthma symptom exacerbation in relation to air quality, covariate
adjustment could also materially improve predictive precision.

This combination of real-time data sources and rapid analytical tools using readily adaptable
methodologies is a powerful one in the control of pathogens where evolving spatio-temporal
patterns of incidence are of public health concern. Our response to COVID-19 has much to
Teach us about preparedness for the next pandemic\cite{18, 25}.

5 Conclusions

In conclusion, we have demonstrated the value of a real-time spatio-temporal inferential
mapping platform for public health efforts during the emergence and spread of infectious
diseases. The work has been conducted in the confines of the privacy-protecting SAIL
Databank to produce statistically robust results at a spatially granular level whilst ensuring
that no individual contributor to the ZOE Symptom Study app can be identified. Such tools
are not only essential to produce population-weighed estimates of disease prevalence, but they
provide a unique insight into the geographical distribution of the disease, thus informing local
and national control efforts.
6 Acknowledgements

This work uses data provided by participants of the COVID-19 Symptoms Study, developed by ZOE Global Limited with scientific and clinical input from King’s College London. We would also like to acknowledge all data providers who made anonymised data available for research.

We wish to acknowledge the collaborative partnership that enabled acquisition and access to the de-identified data, which led to this output. The collaboration was led by BREATHE – The Health Data Research Hub for Respiratory Health, in partnership with SAIL Databank at Swansea University, the Health Data Research UK Swansea University site team and the Usher Institute at the University of Edinburgh. We wish to acknowledge the input of ZOE Global Limited and King’s College London in their development and sharing of the data, and their input into the understanding and contextualisation of data for COVID-19 research. All research conducted has been completed under the permission and approval of SAIL independent Information Governance Review Panel (IGRP) project number 1078. HRS is supported by the Medical Research Council [MR/R008345/1].

References


[22] The Path to Zero: Key Metrics For COVID Suppression – Pandemics Explained. URL: https://globalepidemics.org/key-metrics-for-covid-suppression/.


Supplementary material: Geospatial statistical model and inference

S.1 Model formulation

The standard way to relate prevalence – the probability of a positive outcome – to one or more variable risk-factors is a logistic regression equation. For example, writing $P(x)$ for the probability of a positive outcome at a location $x$ within a specified time-interval, if geographical variation in prevalence were thought to be primarily associated with deprivation, a candidate logistic regression model would be

$$\log\left[\frac{P(x)}{1 - P(x)}\right] = \alpha + \beta \times \text{deprivation score} \quad (1)$$

In a geospatial model, equation (1) is extended to

$$\log\left[\frac{P(x)}{1 - P(x)}\right] = \alpha + \beta \times \text{deprivation score} + S(x), \quad (2)$$

where now $S(x)$ is an unobserved, spatially correlated stochastic process. Its role is to account for that part of the geographical variation in prevalence that is not explained by available covariates. Its inclusion in the model has two potentially important consequences: it guards against invalid inferences on covariate effects by acknowledging that data from different locations may not be statistically independent; more importantly in the current context, it delivers more precise estimates of the underlying prevalence surface $P(x)$ by transferring some of what would otherwise have been treated as unpredictable random variation into predictable, spatially correlated variation. The extent to which these potential benefits are realised in practice is context-specific – the stronger the spatial correlation structure, the greater the benefits. This makes it essential that the spatial correlation structure is estimated from the data, rather than being pre-specified.

S.2 Exploratory analysis
In preliminary analyses of data from different parts of the UK we did not find a robust association of app-reported symptom prevalence with LSOA-level deprivation or population density. We therefore fitted a version of the geospatial model \(^{(2)}\) without covariate-adjustment, i.e. all geographical variation in app-reported symptom prevalence is ascribed to the stochastic term \(S(x)\).

To explore the spatial correlation structure of the data, we converted the positive symptom report count, \(y\), and the number of active users, \(n\), for each LSOA in a given 14-day time-window into empirical logits, \(el = \log\{(y + 0.5)/(n - y + 0.5)\}\) and calculated their variogram, as follows. Associate each empirical logit, \(el_i\), with the population-weighted centroid, \(x_i\), of the corresponding LSOA. Calculate the quantities \(v_{ij} = \frac{1}{2}(el_i - el_j)^2\) and \(d_{ij} = ||x_i - x_j||\), the distance between \(x_i\) and \(x_j\). Group the \(d_{ij}\) in a set of distance bins centred on \(u_k : k = 1, ..., m\) and calculate \(\bar{v}_k\) as the average of all of the \(v_{ij}\) whose corresponding \(d_{ij}\) fall within the \(k\)th distance bin. A plot of \(\bar{v}_k\) against \(u_k\), called the empirical variogram, estimates the quantity \(V(u_k) = \tau^2 + \sigma^2\{1 - \rho(u)\}\), where \(\sigma^2\) is the variance of \(S(x)\), \(\rho(u)\) is the correlation between pairs \(S(x)\) and \(S(x')\) at locations a distance \(u\) apart, and \(\tau^2 > 0\) represents the sampling variation in the individual \(el_i\). Figure 6 shows the result using square-root-transformed empirical logits from Scotland’s data over the 14-day time-window 1 to 14 April 2020.

Based on Figure 6 and similar results from other parts of the UK we used a standard specification for \(S(x)\) as a Gaussian process with twice-differentiable Matérn correlation function (Diggle and Giorgi, 2019, p32), giving a model with three parameters: the intercept, \(\alpha\), in \(^{(2)}\), the variance, \(\sigma^2\), of the stochastic process \(S(x)\), which represents how variable are the log-odds of prevalence from place to place; and a correlation parameter, \(\phi\), which determines the rate at which the correlation between log-odds of prevalence at two locations decays as the distance between them increases.

**S.3 Inference**

We estimated the model parameters by Monte Carlo maximum likelihood, as implemented in
Figure 6: Empirical variogram for empirical logits of positive symptom reports in Scotland over the 14-day time-window 1 to 14 April 2020.

This method ensures statistical efficiency, but requires the tuning of a Markov chain Monte Carlo algorithm. We ran the chain with 11,000 samples, from which the first 1,000 were discarded and the remainder sampled every tenth iteration, leaving a sample of size 1,000 from the joint predictive distribution of the prevalences $P(x)$ for each LSOA. Initial values of the model parameters were taken from visual inspection of the empirical variogram shown in Figure 6 but optimised separately for each time-window of data in each of the UK’s constituent countries.

Summary Points

What is known:

- COVID-19 has highlighted the need for robust methods for identifying outbreaks of disease and local levels
- Most mapping efforts have so far been restricted to regional level estimates - there are very few local level estimates of COVID-19 prevalence
- Self reported app data is currently being contributed by 4 million people in the UK

What we are adding:

- We demonstrate the use of sophisticated spatial modelling for near-real-time prediction of COVID-19 prevalence at small-area resolution to inform strategic government policy areas
- We provide estimates of their precision, to guard against over-reaction to potentially spurious features of 'best guess' predictions
- We demonstrate that adapting existing geospatial statistical methods, originally developed for global health applications, can be used in an anonymised databank environment, thus preserving the privacy of the individuals who contribute their data.