Bayesian stochastic model-based forecasting for spatial Covid-19 risk in England
Technical Concept Note

Chris Jewell[1], Jonathan Read[1], Gareth Roberts[2], Barry Rowlingon[1], and Christopher Suter[3]

[1]CHICAS, Lancaster Medical School, Lancaster University, Lancaster, LA1 4YG, UK
[2]Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
[3]Google Research, New York

September 22, 2020

Contents

1 Introduction

This document presents a brief description of the Lancaster University "Spatial Stochastic Meta-population Model" for COVID-19 in England, designed to provide locally-resolved nowcasts and forecasts of case incidence and prevalence, and reproduction number. The approach makes use of an epidemic model built around the 317 Local Authority Districts (LADs) in England, using LAD-level population size and measures of human mobility to explain differences in positive Pillar 1 and 2 tests across the country in space and time.

Our approach is defined by its capability to formally calibrate the model to case data, using methodology we have developed during COVID-19. This builds on our experience of analogous approaches in real-time modelling of livestock diseases [e.g. 1, 2, 4], closing a capability gap in applying such techniques to national human populations.

Caveat This approach makes use of postive testing data. Results produced by this approach are therefore sensitive to fluctuations in case testing due to logistic constraints. This caveat must be heeded when interpreting results, and especially if they are used for policy planning purposes.

2 Data

This section describes the 4 sources of data we use to train our model and on which predictions are subsequently based.

Case data The model is calibrated to Public Health England (PHE) Pillar 1 and 2 postive test case reports, aggregated by day and LAD, for 3 months prior to the analysis date. The latest 4 days of data is further discarded, as we observe a consistent recording lag of 4 days[1].

Inter-LAD connectivity To establish connectivity between LADs in England, Census 2011 commuting volume data was aggregated from Middle Super Output Area (MSOA). Two pairs of LADs (Cornwall and Scilly, and City of Westminster and City of London) were merged to allow mapping of MSOAs onto LADs. This reduced the statutory number of LADs from 317 to 315.

Census 2011 provides a matrix of the number of journeys made from "Residence" to "Workplace" LADs, W of dimension 315×315. W is non-symmetric, reflecting commuting behaviour rather than the reciprocity of disease transmission. We calculated a symmetric matrix W of the daily number of journeys between each LAD

$$C = W + W^T$$

assuming that commuters return to their Residence each day, and go from their Residence to their Workplace and back at most once per day. We also set $C_{ii} = 0,$ for all i as within-LAD infection transmissibility is delegated to another part of our model (Section 3).

Traffic volume Inter-LAD connectivity data were derived from "business as usual" conditions in England. We assume that commuting is modulated during the COVID-19 outbreak by a relative measure of traffic flow provided by the UK Department for Transport (DfT). We use this information to construct a daily timeseries $\boldsymbol{w}$[1].

Population size The population size N_i for each LAD $i = 1, \ldots, 315$ is taken from ONS Population Size predictions for December 2019.

[1] Data from PHE and DfT are supplied to Lancaster University under a SPI-M-mediated data sharing agreement. These data are accessed only by Jewell, Rowlingson, and Read.

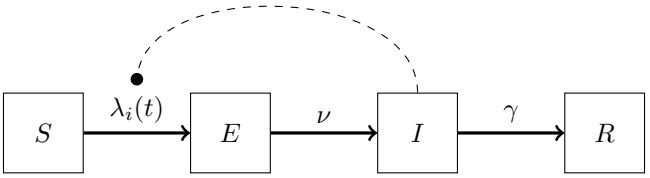

Figure 1: SEIR (susceptible, exposed, infectious, removed) model showing transition rates. In particular, the number of infectious individuals feeds back via measures of human mobility and proximity into the LAD-specific transition rate $\lambda_i(t)$.

3 Model

Within each of $m = 315$ LADs, we assume the SEIR model in Figure 1 where at any time t, individuals exist in one of 4 disease states: susceptible, exposed (i.e. infected but not yet infectious), infectious, and finally removed (either solidly immune or dead, Figure 1).

When susceptible, an individual experiences a "force of infection" ($\lambda_i(t)$, Figure 1) from all infectious individuals in its own LAD i, as well as infectious individuals in all other LADs weighted by connectivity $w_t C$:

$$\lambda_i(t) = \beta_0 e^{\xi_t} \frac{I_i(t) + \beta_1 w_t C \frac{\boldsymbol{I}(t)}{\boldsymbol{N}}}{N_i}. \tag{1}$$

where $I(t)$ is the vector of numbers of infectious individuals in each LAD, w_t is the tth element of the traffic flow timeseries $\boldsymbol{w}$, C is the mobility matrix, and N is the vector of population size in each LAD. Parameters β_0, and β_1 are assumed unknown and must be estimated from data. $\boldsymbol{\xi}$ is similarly unknown, and represents a 2-week changing baseline force of infection, measuring changes in transmission in response to unobservable changes in population behaviour. Individuals then spend $1/\nu$ days on average in the exposed state, $1/\gamma$ days on average in the infectious state before becoming removed. We assume $\nu = 0.5$ to give a 2 day latent period, and assume γ is unknown.

The model is evolved in daily discrete timesteps, using the "chain binomial" setup. For further technical information, see Appendix B.

4 Statistical model fitting

We fit our model assuming that PHE Pillar 1 and 2 positive tests represent $I \to R$ transition events. Our aim is to estimate unknown parameters $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\xi}, \beta_1, \gamma\}$ in order to subsequently make epidemic predictions forward in time. Our fitting procedure is based on our previous Bayesian approaches, allowing us to calculate unbiased estimates for the parameters, allowing for the fact that $S \to E$ and $E \to I$ events are not observed [1].

Importantly, the Bayesian approach fully quantifies uncertainty due to both statistical uncertainty in parameter estimation and random fluctuations in the epidemic process. This provides a natural way to provide uncertainty estimates, for example the probability that reproduction numbers are greater than unity. For futher details, see Appendix C.

5 Implementation

The model implementation is written in Python 3.8 using Tensorflow 2.3 (Google Brain) and Tensorflow Probability 0.14 (Google Research) to make use of GPU-accelerated high performance computing. The Python source code is freely available at `https://github.com/chrism0dwk/covid19uk`.

6 Predictions

Given our model and a fitted posterior distribution over the parameters and censored event times, we can calculate predictive distributions over a wide range of epidemic metrics of interest. Specifically, we focus on 3 different aspects:

Reproduction number R_{it} is the time-varying reproduction number R_{it} for LAD i at time t (typically the latest analysis date). It gives the expected number of further infections an average individual in that LAD will infect during its infectious period (both within its own LAD and beyond). This is a measure of the capacity of a LAD to generate more cases, should infection arrive into it, and varies as changes in population behaviour affects disease transmission.

Case incidence Case incidence λ_{it} is the expected rate of new infections in LAD i at time t. Typically, we quote this as an absolute incidence, in other words the expected number of cases in total within a LAD given a specified time-frame. We do this for resource planning purposes, in order to highlight LADs that are likely to experience a high number of cases in the short-term based on their current number of cases.

Case prevalence Case prevalence, π_{it}, gives the expected proportion of individuals in LAD i who are in *either* the Exposed or Infectious states at time t, i.e. the proportion of individuals who are infected. This metric may be used to track the progression of the outbreak in space and time, and captures an overall spatial measure of how the epidemic is unfolding.

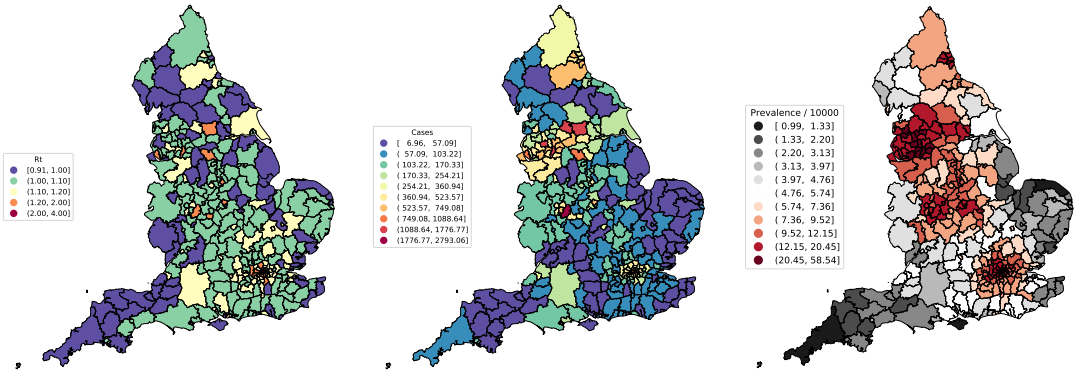

Figure 2: Analysis as of 14th September 2020: Estimated reproduction number (left); number of new COVID-19 cases by 28th September 2020 (centre); prevalence of COVID-19 on 28th September 2020 (right).

References

[1] C. P. Jewell, M. J. Keeling, and G. O. Roberts. Predicting undetected infections during the 2007 foot-and-mouth disease outbreak. *J R Soc Interface*, 6(41):1145–1151, 2009. URL `http://dx.doi.org/10.1098/rsif.2008.0433`.

[2] CP Jewell and RG Brown. Bayesian data assimilation provides rapid decision support for vector-borne diseases. *Journal of The Royal Society Interface*, 12(108):20150367, 2015.

[3] CP Jewell, T Kypraios, PJ Neal, and GO Roberts. Bayesian analysis for emerging infectious diseases. *Bayes. Anal.*, 4(3):465–496, 2009.

[4] William J. M. Probert, Chris P. Jewell, Marleen Werkman, Christopher J. Fonnesbeck, Yoshitaka Goto, Michael C. Runge, Satoshi Sekiguchi, Katriona Shea, Matt J. Keeling, Matthew J. Ferrari, and Michael J. Tildesley. Real-time decision-making during emergency disease outbreaks. *PLOS Computational Biology*, 14(7):1–18, 07 2018. doi: 10.1371/journal.pcbi.1006202. URL `https://doi.org/10.1371/journal.pcbi.1006202`.

Appendices

Appendix A Formal model description

As described above, we characterise the population within each of $m = 315$ LADs into susceptible, exposed (infected but not yet infectious), infectious, and removed states. We denote the number of individuals in each state in LAD i at time t by $S_i(t)$, $E_i(t)$, $I_i(t)$, $R_i(t)$ respectively, and write $X_i(t) = \{S_i(t), E_i(t), I_i(t), R_i(t)\}$ to denote the overall state of the meta-population. Since we have $i = 1, \ldots, m$ LADs, we omit the subscript to denote the state of the entire population, for example $X(t) = \{X_1(t), \ldots, X_m(t)\}$.

As shown in Figure 1, we assume that individuals transition from susceptible to exposed at rate $\lambda_i(t)$, from exposed to infectious at rate ν, and from infectious to removed at rate γ, all rates being in units of events per day. We assume that ν and γ are constant across both meta-populations and time. Within a LAD, we assume that the population is well-mixed, and that mixing between LADs is heterogeneous. The force of infection, $\lambda_i(t)$, is assumed to vary according to the state of the epidemic and the connectivity between LADs in a frequency-dependent (i.e. independent of population density) way such that

$$\lambda_i(t) = \beta_0 e^{\xi(t)} \frac{I_i(t) + \beta_1 w_t C \frac{\boldsymbol{I}(t)}{\boldsymbol{N}}}{N_i}. \tag{2}$$

with C the zero-diagonal commuting matrix described in Section 2, w_t the England-wide relative traffic flow on day t from Section 2, and N the vector of population sizes in each LAD. Parameters β_0, β_1, and $\boldsymbol{\xi}$ (which changes every 2 weeks) are assumed unknown.

Appendix B Data Generating Process

Given the characterisation of the SEIR model, the vector of transition rates $\boldsymbol{\lambda}(t)$, onset of infectiousness rate ν, and removal rate γ as defined in Section 3, we evolve the epidemic in daily time steps according to a discrete time "chain-binomial" Markov process. Here, the number of state transitions $Y_i^{se}(t), Y_i^{ei}(t), Y_i^{ir}(t)$ occurring on day t are drawn from a Binomial distribution conditional on the state $X_i(t)$, i.e.

$$\begin{aligned} Y_i^{se}(t) &\sim \text{Binomial}(S_i(t), p_{se}) \\ Y_i^{ei}(t) &\sim \text{Binomial}(E_i(t), p_{ei}) \\ Y_i^{ir}(t) &\sim \text{Binomial}(I_i(t), p_{ir}) \end{aligned} \tag{3}$$

where p_{se}, p_{ei}, and p_{ir} are the daily probabilities of individuals undergoing $S \to E$, $E \to I$, and $I \to R$ transitions respectively given by

$$\begin{aligned} p_{se} &= 1 - e^{-\boldsymbol{\lambda}(t)} \\ p_{ei} &= 1 - e^{-\nu} \\ p_{ir} &= 1 - e^{-\gamma}. \end{aligned}$$

Given the entire state at time t, $\boldsymbol{Y}(t)$, we propagate the epidemic state by

$$\begin{aligned} S_i(t+1) &= S_i(t) - Y^{se}(t) \\ I_i(t+1) &= I_i(t) + Y^{se}(t) - Y^{ei}(t) \\ E_i(t+1) &= E_i(t) + Y^{ei}(t) - Y^{ir}(t) \\ R_i(t+1) &= R_i(t) + Y^{ir}(t). \end{aligned} \tag{4}$$

Given an initial state $X(0)$ and parameters $\boldsymbol{\theta} = \{\beta_0, \beta_1, \nu, \gamma, \boldsymbol{\xi}\}$, the data generating process is then given by iterating Equations 3 and 4 for $t = 1, \ldots, T$ where T is the final time of interest.

Appendix C Statistical Inference

Statistical inference is performed on model parameters $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\xi}, \beta_1, \gamma\}$, assuming that $I \to R$ transitions $Y^{ir}(t)$ are observed on each day t and are synonymous with Pillar 1 and 2 case reports (with the patient assumed to fully self-isolate subsequently). For convenience we fix the $E \to I$ transition rate $\nu = 0.5$ (i.e. 2-day latent period), noting that predictive results are relatively insensitive to this parameter. Estimation is performed in a Bayesian paradigm using Markov chain Monte Carlo (MCMC), with the following priors

$$\begin{aligned} \beta_0 &\sim \text{Gamma}(1, 1) \\ \beta_1 &\sim \text{Gamma}(3, 10) \\ \gamma &\sim \text{Gamma}(100, 400) \\ \xi &\sim MVN(\mathbf{0}, \Sigma) \end{aligned}$$

where Σ is a covariance matrix with Matérn correlation function

$$\Sigma_{ij} = 0.1 \left(1 + \frac{\sqrt{3}d}{\rho}\right) \exp\left(-\frac{\sqrt{3}d}{\rho}\right)$$

with $\rho = 12$ to give a 4-week effective range.

The challenge to inference in this situation is that although a statistical likelihood function for observing events $\boldsymbol{Y}$ given the model parameters is straightforward from Equation 3, we do not observe $S \to E$ or $E \to I$ events, i.e. $Y^{se}(t)$ and $Y^{ei}(t)$ are censored data. We approach this problem by extending our previous work in MCMC data augmentation techniques for individual-level epidemic models in continuous time to the case of meta-population models in discrete time to suit our application [3, 1]. The MCMC allows us to sample from the joint posterior probability distribution $\pi(\boldsymbol{\theta}, Y^{se}(0{:}T), Y^{ei}(0{:}T) | y^{ir}(0{:}T), X(0))$ where $X(0)$ is the initial state at the beginning of our epidemic timeseries chosen by single imputation of the initial censored event times.

Crucially, both $Y^{se}(0{:}T)$ and $Y^{ei}(0{:}T)$ contain two types of censored event. Firstly, partially-censored events refer to those for which we have observed a corresponding removal (i.e. $I \to R$) event – we know that the events exist, but not precisely when they occurred. Secondly, "occult" events refer to $S \to E$ and $E \to I$ events that have happened, but that we are unaware of because of not having (yet) observed the corresponding $I \to R$ event. In the Bayesian setting, our MCMC treats these quantities as parameters to be estimated jointly with $\boldsymbol{\theta}$, with implementations of Metropolis-Hastings samplers designed to draw from their conditional posterior distributions.

The MCMC is tuned manually to obtain Metropolis-Hastings acceptance rates of approximately 23%, and run for 200,000 iterations. The initial 100,000 iterations are discarded as burn-in, and the remaining 100,000 iterations are thinned every 20 iterations for 2000 quasi-independent samples from the joint posterior. These samples are subsequently used to calculate summary statistics for the epidemic predictions.