ARTICLE IN PRESS

Epidemics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Contagion! The BBC Four Pandemic – The model behind the documentary

Petra Klepac[a,b,*], Stephen Kissler[a], Julia Gog[a]

[a] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK
[b] Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Public Health, London School of Hygiene and Tropical Medicine, London, UK

ARTICLE INFO

Keywords:
Influenza dynamics
Fine-level spatial simulation
Human mobility
Contact data
Mobile phone data

ABSTRACT

To mark the centenary of the 1918 influenza pandemic, the broadcasting network BBC have put together a 75-min documentary called 'Contagion! The BBC Four Pandemic'. Central to the documentary is a nationwide citizen science experiment, during which volunteers in the United Kingdom could download and use a custom mobile phone app called *BBC Pandemic*, and contribute their movement and contact data for a day.

As the 'maths team', we were asked to use the data from the app to build and run a model of how a pandemic would spread in the UK. The headline results are presented in the TV programme. Here, we document in detail how the model works, and how we shaped it according the incredibly rich data coming from the *BBC Pandemic* app.

We have barely scratched the depth of the volunteer data available from the app. The work presented in this article had the sole purpose of generating a single detailed simulation of a pandemic influenza-like outbreak in the UK. When the *BBC Pandemic* app has completed its collection period, the vast dataset will be made available to the scientific community (expected early 2019). It will take much more time and input from a broad range of researchers to fully exploit all that this dataset has to offer. But here at least we were able to harness some of the power of the *BBC Pandemic* data to contribute something which we hope will capture the interest and engagement of a broad audience.

1. Introduction

In a nationwide citizen science experiment, 360 Production, commissioned by the British Broadcasting Corporation (BBC), launched an app called *BBC Pandemic* that was available for download to smartphones via App Store or Google Play. Using the app, the volunteers could participate in two studies: (1) one focusing on Haslemere, a town in Surrey, where there was a campaign to enroll a considerable number of people and volunteers' mobile phone locations were simultaneously tracked with permission over three consecutive days, and (2) a bigger study for users across the United Kingdom that, with permission, recorded volunteers' hourly locations to the nearest square kilometre over 24-h period chosen by the volunteer. At the end of each of the study periods volunteers were asked to input whom they encountered during that period. Here we focus exclusively on the national dataset, consisting of recorded movement data and self-reported contact data. We were tasked with using this data to develop a mathematical model for the spread of influenza, and thereby to simulate how a pandemic-like strain of influenza might spread through the United Kingdom. This virtual outbreak was to start in Haslemere, a town in Surrey, in the south of England, to follow the programme's narrative, with the documentary's presenter acting as a hypothetical index case. Detailed data from Haslemere formed the basis for an individual based model (detailed in the companion paper by Kissler et al., 2018b) used to simulate an outbreak in Haslemere that was to seed the virtual national outbreak.

To meet the tight production and filming deadlines, we had to make quick decisions, often responding to requested outputs and changes in under a day. The bulk of our work here took place in three weeks: starting from the maths team finishing modelling and filming for the previous part of the programme (the part on Haslemere outbreak Kissler et al., 2018b) and getting the main part of the data on which we could start to investigate and make decisions on the model for the national simulations. There were many challenges here, chiefly associated with working with very large datasets which have never been used before. We had no specification imposed on the model structure, indeed the detail of how it worked was not included in the programme, but we were aiming for an output that would give a detailed geographic picture of pandemic spread.

We were able to make extensive use of the new and very promising dataset to develop, parameterise and run a detailed national simulation, all in time for the required schedule. We are presenting in this

* Corresponding author at: Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Public Health, London School of Hygiene and Tropical Medicine, London, UK.
E-mail addresses: pklepac@alum.mit.edu (P. Klepac), jrg20@cam.ac.uk (J. Gog).

https://doi.org/10.1016/j.epidem.2018.03.003
Received 9 March 2018; Received in revised form 14 March 2018; Accepted 14 March 2018
1755-4365/

Please cite this article as: Klepac, P., Epidemics (2018), https://doi.org/10.1016/j.epidem.2018.03.003

document what we actually did, starting with building from the data, to the model construction, and resulting simulation outputs. With the luxury of even a few more weeks, we would certainly have investigated alternative approaches and explored robustness to model choices and parameters: we comment further on this in the discussion section.

All of the results presented here and in the TV Programme make use of the app data collected only up to October 30, 2017 (i.e. about a month's worth of data as the app launched on September 27, 2017). As we write this document, the app is still collecting data and will continue to do so during the rest of 2018, and the final dataset will be published with a separate paper (expected early 2019). Until then, the results from the datasets here should be treated as preliminary only. The app data consists of three interlinked data streams: (i) user profiles, (ii) location logs, and (iii) user encounter data. The user profiles gives us a brief information about the user, and the key thing used here is their age. The location logs give GPS position of the user (to square kilometre resolution, one record per hour) and we used this to extract detailed movement profiles. The self-reported encounter data gives a list of basic information on people they met in that day, and we use this to build age-structured contact matrices below. In total, the data used to shape the model comes from information contributed by 28,947 users.

2. Data analysis and preparation

2.1. BBC data – from location logs to movement patterns

2.1.1. Data extraction

From available location logs and user profiles data we create a single file with user characteristics (age, gender, max-distance-travelled), and location coordinates for each hour during a 24-h period starting with the time of the first location log. To calculate travel distances we take the first recorded location as a reference location, and calculate distance between reference and destination coordinates using Haversine distance with radius set to Earth's radius in Haslemere (R = 6,365,295 m). We are assuming that the reference location is usually ‘home’ or somewhere nearby, and we eliminate all entries whose reference location is not in the UK. To later run the model, we assign each reference location to one of 9370 model patches (defined in Section 2.3) using function `over()` from `sp` R-package.

2.1.2. Distance travelled – within 100 km

We abstracted from this data the (time-weighted) distribution of distance from reference location on the scale of kilometres. Distances were binned into one kilometre ranges (so 0–1000 m, 1000–2000 m, etc.). Then a tally was made of all of these, summing over all users and all recordings for each user. For this part of the analysis, we consider only distances up to 100 km and discard the rest (but see below for long distance jumps). This was all done separately for users whose home locations were within urban areas and for those within rural areas (shown in Fig. 1).

An interesting way to look at these counts was as cumulative density, in other words: what proportion of the time do users spend more than distance X away from home. Both the raw counts and the cumulative densities are shown in Fig. 2. From the cumulative density plots, it can be seen that the rural users typically spent more time far away from home.

To go from movement patterns accumulated from many individuals to the ‘right’ kernel in a gravity-like patch model is an important and interesting open question. We believe this warrants much further attention from researchers, as we move into an era where such data is becoming available (*BBC Pandemic* data will be widely available). Here, we were limited by availability of good methods, and not enough time to develop and test anything sophisticated. We took the best approach we could (described below), but we still feel this point deserves much further careful work.

The distribution of distances for our recordings gives a simple

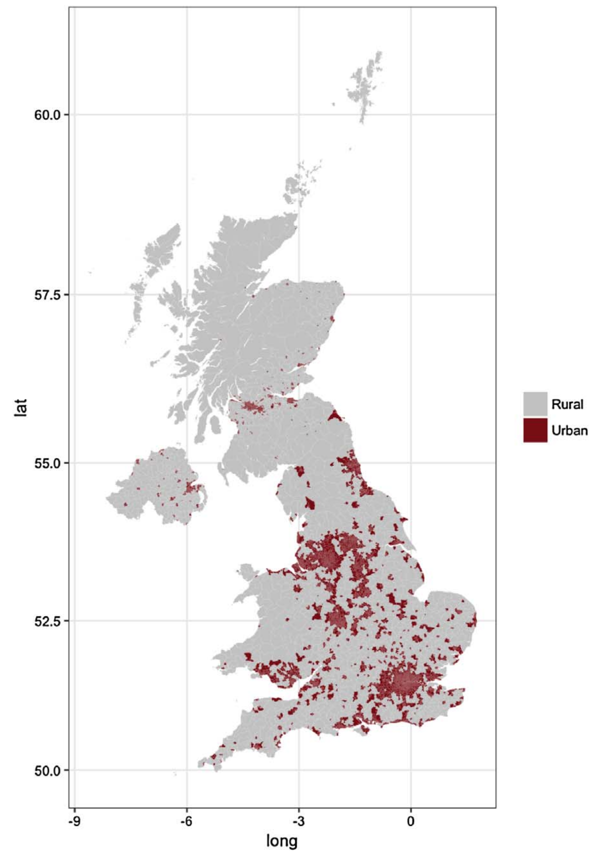

Fig. 1. Distribution of rural and urban mid-layer areas in the UK.

measure of how much time a user spends a given distance away from ‘home’. For our purposes, we were interested in transmission between model patches (typically several or many kilometres apart). The bulk of recordings are within 1000 m of home, plus the app resolution is only of the order of 1000 m. So to make a kernel for between-patch movement, we used only the values for over 1000 m away (effectively dropping the first bin count and replacing it with a duplicate of the second count to represent movement to very nearby other patches). Then the counts were normalised, to give a distribution of where people are, given they are away from home. At this point we have two lists of length 100 to give the proportion of time away from home that is spent at each kilometre binned distance. Denote these as $F_u(i)$ for urban and $F_r(i)$ and rural (for i = 1–100).

We also explored differences in movement patterns with respect to many other factors, including the participants' age and gender, illustrated in Fig. 3. The difference by gender is interesting, particularly over the mid-range of distance, and deserves further attention, but we decided not to pursue it for inclusion in the model here. The split by age group is even more intriguing, especially given different age pattern observed in a smaller dataset of self-reported distances and contacts from southern China (Read et al., 2014), where elderly age groups appear to move the least. Again, we did not use this distribution directly here, but return to it in the discussion.

2.1.3. Distance travelled – long jumps

A model based purely on density of movement from above would have transmission rates tailing off sharply over tens of kilometres. An epidemic simulation of the UK would be strongly wave-like, and jumps across the sea to Northern Ireland would be rare, and epidemic travel would be very slow indeed across less densely population regions (such as around the England–Scotland border). The epidemic would then effectively get stuck, politely waiting some time for infection to