

## Article

# Extracting Self-Reported COVID-19 Symptom Tweets and Twitter Movement Mobility Origin/Destination Matrices to Inform Disease Models

Conor Rosato <sup>1,\*</sup> , Robert E. Moore <sup>1</sup>, Matthew Carter <sup>1</sup> , John Heap <sup>2</sup>, John Harris <sup>3</sup>  and Jose Storopoli <sup>4</sup>   
and Simon Maskell <sup>1</sup> 

<sup>1</sup> Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool L69 3GJ, UK

<sup>2</sup> Computational Biology Facility, University of Liverpool, Liverpool L69 3GJ, UK

<sup>3</sup> Public Health England, London NW9 5EQ, UK

<sup>4</sup> Department of Computer Science, Universidade Nove de Julho—UNINOVE, Sao Paulo 03155-000, Brazil

\* Correspondence: c.m.rosato@liverpool.ac.uk

**Abstract:** The emergence of the novel coronavirus (COVID-19) generated a need to quickly and accurately assemble up-to-date information related to its spread. In this research article, we propose two methods in which Twitter is useful when modelling the spread of COVID-19: (1) machine learning algorithms trained in English, Spanish, German, Portuguese and Italian are used to identify symptomatic individuals derived from Twitter. Using the geo-location attached to each tweet, we map users to a geographic location to produce a time-series of potential symptomatic individuals. We calibrate an extended SEIRD epidemiological model with combinations of low-latency data feeds, including the symptomatic tweets, with death data and infer the parameters of the model. We then evaluate the usefulness of the data feeds when making predictions of daily deaths in 50 US States, 16 Latin American countries, 2 European countries and 7 NHS (National Health Service) regions in the UK. We show that using symptomatic tweets can result in a 6% and 17% increase in mean squared error accuracy, on average, when predicting COVID-19 deaths in US States and the rest of the world, respectively, compared to using solely death data. (2) Origin/destination (O/D) matrices, for movements between seven NHS regions, are constructed by determining when a user has tweeted twice in a 24 h period in two different locations. We show that increasing and decreasing a social connectivity parameter within an SIR model affects the rate of spread of a disease.

**Keywords:** social media; machine learning; origin destination matrices; disease modelling; death predictions



**Citation:** Rosato, C.; Moore, R.E.; Carter, M.; Heap, J.; Harris, J.; Storopoli, J.; Maskell, S. Extracting Self-Reported COVID-19 Symptom Tweets and Twitter Movement Mobility Origin/Destination Matrices to Inform Disease Models. *Information* **2023**, *14*, 170. <https://doi.org/10.3390/info14030170>

Academic Editor: Nirmalya Thakur

Received: 28 January 2023

Revised: 3 March 2023

Accepted: 5 March 2023

Published: 7 March 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The novel coronavirus (COVID-19) has, at the time of writing, resulted in over 6.88 million deaths and 676 million confirmed cases worldwide [1]. By January 2020, new cases of COVID-19 had been seen throughout Asia, and by the time the World Health Organisation (WHO) declared a global pandemic in March 2020, the disease had spread to over 100 countries. It quickly became imperative to establish reliable data feeds relating to the pandemic, such that researchers and analysts could model the ongoing spread of the disease and inform decision-making by government and public health officials. To facilitate collaboration between researchers and allow for published results to be replicated and scrutinised, these data sets and models must be open-source. A well-used interactive dashboard collating total daily counts of confirmed cases and deaths for countries and, in some cases, regions within countries can be found in [2]. The variables presented in the platform are traditionally used to calculate metrics such as the reproduction number ( $R_t$ ). One such method for estimating  $R_t$  is by modelling how the disease spreads through a population using a Susceptible, Infected and Recovered (SIR) model [3]. This method

involves splitting the population into the unobservable SIR compartments and allowing a fraction at every timestep  $t$ , to progress to the next compartment. The model consists of three nonlinear ordinary differential equations (ODE) and a set of parameters which govern how quickly individuals progress through the compartments. The standard SIR model contains two parameters,  $\beta$  and  $\gamma$ , which are the infection and recovery rates, respectively. This metric is vital in understanding both the infection growth rate, or daily rate of new infections, and the number of people, on average, infected by a single infected person and can be calculated by

$$R_0 = \beta/\gamma. \quad (1)$$

The quality of disease metrics is heavily dependent on the model and the ingested data. In the United Kingdom (UK), up until December 2022, a joint effort was undertaken to produce estimates of the  $R_t$  number, with notable examples provided in [4]. Different data sets have been used by different institutions. Laboratory-confirmed COVID-19 diagnoses are used in [5], UK's NHS Pathways data in [6] and hospital admissions data in [7]. The statistical model developed by Moore, Rosato and Maskell [8] contributes to these estimates through the incorporation of death, hospital admission and NHS 111 call data. Aggregated 111 call counts contain the individuals that reported potential COVID-19 symptoms through the NHS Pathways telephone service.

Evaluating short-term forecasts of COVID-19 related statistics is useful to determine the accuracy of a model. A multi-model comparison of predicted deaths, hospital admissions and intensive care unit (ICU) occupancy is given in [9]; deaths, hospital admissions and ICU occupancy in [7]; daily hospital admissions in [10] and short-term forecasting of deaths in [8]. A set of scoring rules for evaluating these short-term forecasts is outlined in [11], with an application to COVID-19 deaths provided in [8].

The latency and reliability of COVID-19 related data sources can vary. Death data can be seen as reliable when compared with confirmed cases derived from positive test results; however, observations of this data are typically delayed from the initial point of infection. Delays also occur between the occurrence and reporting of deaths. The reliability of confirmed cases is limited as the sampling of those tested varies with time with the reason for testing often not recorded. In addition, hospital admissions typically occur around 1–2 weeks after infection and so may be considered outdated in relation to the time of initial infection. The extent to which these issues are problematic is likely to vary over time and between countries. For example, reliable, publicly available tests only began to become available a number of months after the outbreak and declaration of the COVID-19 pandemic. As such, information on the spread of the disease was limited and varied between countries. Twitter provides real-time data that overcome the timing limitations of the aforementioned data sources. Correlation between tweets relating to influenza and true influenza counts have been observed in [12–14]. It is possible to set up a pipeline for collecting and analysing COVID-19 tweets that can be scaled up to multiple countries in a short amount of time.

### 1.1. Related Works

Infodemiology and infoveillance [15] refer to the ability to process and analyse data, pertinent to disease outbreaks, that are created and stored digitally in real-time. The availability of these data sets, particularly at the beginning of an outbreak, could provide a noisy but accurate representation of disease dynamics. Prior to the pandemic, tweets relating to influenza-like-illness symptoms were seen to substantially improve the model's predicting capacity and to boost nowcasting accuracy by 13% in [16,17], respectively. Models allowing for early warning detection of multiple diseases are proposed in [17,18] through analysis of tweet content in real time. Many research papers use social media to gain valuable information relating to the COVID-19 pandemic. Natural language processing (NLP), in particular determining the sentiment of tweets, is a popular research area. Ref. [19] uses sentiment analysis and topic modelling to extract information from conversations relating to COVID-19. When including these data within forecasting models, they observed

a 48.83–51.38% improvement in predicting COVID-19 cases. Large databases of tweets are open-sourced [20,21]. Public sentiment relating to COVID-19 prevention measures is analysed in [22]. Depression trends among individuals were analysed in [23]. Emotion was observed to change from fear to anger during the first stages of the pandemic [24]. Misinformation and conspiracy theories propagated rapidly through the Twittersphere during the pandemic [25]. Machine learning algorithms have been used to automatically detect tweets containing self-reported symptoms mentioned by users [26], with Ref. [27] finding symptoms reported by Twitter users to be similar to those used in a clinical setting. We note that the analysis in [19,22,24–26] is conducted with the English language only. Analysis conducted in multiple languages is less common. Topic detection and sentiment analysis are conducted in the Portuguese and English language in [28] while misinformation was detected in English, Hindi and Bengali [29]. To the best of our knowledge, researchers have yet to use symptomatic tweets in multiple languages to calibrate epidemiological models.

Movement mobility patterns have been derived from anonymised cell phone data [30,31] and Twitter [32,33]. Using movement between different geographic locations has been shown to be an effective way of modelling the spread of disease [31,34–36]. During an epidemic, limiting the movement of individuals with measures, such as school closures and national lockdowns, can drive the reproduction number below 1 [37]. In Italy, when analysing mobile phone movement data, less rigid lockdown measures led to an insufficient decrease in COVID-19 cases when compared to a more rigid lockdown [38]. In this paper, we outline how origin/destination (O/D) matrices can be derived from where people tweet and show, by using an epidemiological model, that restricting movement can have an effect on the spread of a disease. To the best of our knowledge, using O/D matrices derived from Twitter movement to inform SIR disease models has yet to be explored.

### 1.2. Contribution and Structure

The contribution of this paper is as follows: first, we outline how to use machine learning to identify tweets that correspond to COVID-19 related symptoms in multiple languages. We present a comprehensive study of how these symptomatic tweets differ from other open-source data sets when calibrating the extended SEIRD model described in Section 3.1. When incorporating the surveillance data outlined in Section 2.2, the Mean Absolute Error (MAE) and Normalised Estimation Error Squared (NEES) values are calculated for 7-day death forecasts. Second, we outline a method for deriving O/D matrices from Twitter and show how these can be included to better model the spread of a disease. To the best of our knowledge, using O/D matrices derived from Twitter movement to inform SIR disease models has yet to be explored.

We now present the structure of the remainder of the paper. The methodology for extracting symptomatic tweets in real time and a description of other open-source data feeds are outlined in Section 2.2. Methods for creating the O/D matrices are outlined in Section 2.3. The extended SEIRD model for predicting deaths is outlined in Section 3.1 and the SIR model including movement between NHS regions in Section 3.2. The corresponding results are presented in Sections 4.1 and 4.2, respectively. Concluding remarks and directions for future work are described in Section 5.

## 2. Data Collection

In this section, the methods for collecting UK NHS region-specific surveillance data and symptomatic tweets are outlined in Sections 2.1 and 2.2, respectively. The O/D matrices derived from Twitter mobility are included in Section 2.3. Two Twitter API developer credentials were used for data collection, in line with our two objectives: (1) querying on COVID-19 keywords and (2) querying on geo-located tweets.

Note that testing methods and criteria for classifying deaths as COVID-19-related may differ between geographic locations. All data sets and associated code can be found on the CoDatMo GitHub repository [39].

### 2.1. United Kingdom NHS Region-Specific Surveillance Data

The methods for collecting UK NHS region-specific surveillance data are presented in the following subsections. The references from where the data were obtained are given in Table 1. The NHS regions in the UK support local systems and provide more joined up and sustainable care for patients through integrated care systems. Every individual born in the UK is entitled to use this public health system.

**Table 1.** A description of the data feeds used per geographic location, the start date used in the simulations and where they were obtained.

Geographic Location	Data Feed	Start Date	Reference
U.S States and the rest of the world	Deaths	24 March 2020	[2]
	Tests	1 March 2020	[2]
	Twitter	13 April 2020	Section 2.2
U.K NHS Regions	Deaths	24 March 2020	[40]
	Hospital admissions	19 March 2020	[41]
	Twitter	9 April 2020	Section 2.2
	Zoe app	12 May 2020	[42]
	111 calls	18 March 2020	[43]
	111 online	18 March 2020	[43]

#### 2.1.1. Deaths

The aggregated death counts contain individuals with COVID-19 as the cause of death on their death certificate or those who died within 60 days of a positive test result.

#### 2.1.2. Hospital Admissions

The aggregated admission counts contain the daily COVID-19 related hospital admissions and the total number of COVID-19 patients.

#### 2.1.3. Zoe App

The aggregated Zoe App counts contain entries of COVID-19 symptoms to a mobile App. The App was developed in 2020 to help track COVID-19. However, it has since broadened its capacity to track other health related concerns such as cancer and high blood pressure. Users can input if they have COVID-19 symptoms as well as stating whether they have been tested for COVID-19.

#### 2.1.4. 111 Calls and 111 Online

The aggregated 111 call and 111 online assessment counts contain individuals that reported potential COVID-19 symptoms through the NHS Pathways telephone and online assessment services, respectively. The telephone service allows for individuals to speak to a medical specialist regarding health concerns. The 111 online service provides information regarding where it is best to obtain help for the symptoms provided. During the COVID-19 epidemic, both services provided a method for individuals to report COVID-19 symptoms.

### 2.2. Symptomatic Tweets

The geographic locations considered when querying on keywords are:

- **US:** 50 States;
- **Rest of the world:** 2 European and 16 Latin American countries;
- **UK:** 7 NHS regions.

Table 1 provides a summary of surveillance data corresponding to each geographical location. Death and positive case data for the US States and the rest of the world (ROW)

were downloaded from the dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) [2].

### 2.2.1. Pre-Processing Tweets

Tweepy [44] is the Twitter API written in the programming language Python. The free Twitter streaming API was used for this research, limiting the number of tweets available for download to 1%. We note that the premium API would allow for a higher percentage of tweets to be collected. The API was filtered using 93 keywords in English, German, Italian, Portuguese and Spanish that align with COVID-19 symptoms from the MedDRA database [45]. The list of keywords can be found here [39]. These terms include those associated with fever, cough and anosmia. While we considered other keywords (e.g., “COVID”), we found that keywords related to symptoms gave rise to a large number of tweets that related to people experiencing symptoms. We do recognise that any choice of keywords will inevitably identify some tweets that are related to advice or general discussion of the disease. This motivated us to use machine learning to post-process the output from the keyword-based queries, as is discussed further in Section 2.2.2.

### 2.2.2. Symptom Classifier Breakdown

A multi-class support vector machine (SVM) [46] was trained with a set of annotated tweets that were vectorised using a skip-gram model. The annotated tweets were labelled according to the following classes:

1. Unrelated tweet;
2. User currently has symptoms;
3. User had symptoms in the past;
4. Someone else currently has symptoms;
5. Someone else had symptoms in the past.

The total number of tweets mentioning symptoms, given by the sum of tweets in classes 2–5, was calculated for each 24 h period. Geo-tagged tweets were mapped to their location, e.g., corresponding city, via a series of tests using country-specific shapefiles. Previous studies demonstrate that approximately 1.65% of tweets are geo-tagged [47], where the exact position of the tweeter is recorded using longitude and latitude measurements.

For non-geo-tagged tweets, the author’s profile is assessed to ascertain whether they provide an appropriate location. The server was deemed to be offline if any 15 min period within the 24 h had no recorded tweets. After checking all 96 15 min periods, the count in each geographical area was multiplied by a correction factor:

$$\text{reported tweet count} = \text{total tweet count} \cdot \frac{96}{96 - \text{downtime periods}} \quad (2)$$

To ensure the labelled tweet data sets used for training and testing were balanced, under- and over-represented classes were randomly up- and down-sampled. A subset of data was used to train the classifier before testing on the remainder. The total number of labelled tweets used for training and testing are provided in Table 2. Four metrics outlined in Table 2 were used to evaluate the classifier. These include the F1 score, accuracy, precision and recall. True positive (TP) and true negative (TN) classifications are outcomes for which the model correctly predicts positive and negative classes, respectively. Similarly, false positive (FP) and false negative (FN) classifications are outcomes for which the model incorrectly predicts positive and negative classes, respectively. Accuracy, precision, recall and the F1 score, which is the harmonic mean of precision and recall, are given as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (4)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (5)$$

$$\text{F1} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}}. \quad (6)$$

**Table 2.** Testing, training and performance measures of the machine learning classifiers in different languages.

Language	Number of Data Used		F1	Performance Measures		
	Training	Testing		Accuracy	Precision	Recall
English	1105	195	0.85	0.85	0.85	0.85
German	412	260	0.89	0.89	0.90	0.89
Italian	254	260	0.97	0.96	0.97	0.96
Portuguese	3507	619	0.77	0.77	0.78	0.80
Spanish	1530	270	0.82	0.85	0.82	0.85

### 2.2.3. Comparison of Tweets and Positive Test Results

Figure 1 shows a comparison between the classified tweets and confirmed positive test results for five US States and one South American country. Both time-series are standardised between 0 and 1 and have been converted to a 7-day rolling average to smooth out short-term fluctuations. It is evident that, at least in the context of these specific examples, the classified tweets do (by eye) follow the trend of positive test results. In some cases, such as Texas and Chile, there seems to be a lag between tweets and positive test results. We suspect there is a reporting delay in these locations. A more rigorous analyses, such as change point detection, could give a stronger indication of how well the trends in the two time-series match. We note that, for some geographic locations, tweets align much less well with the corresponding case counts: we assert that this could be caused by issues with how cases are recorded in each location or by the processing of the tweets.

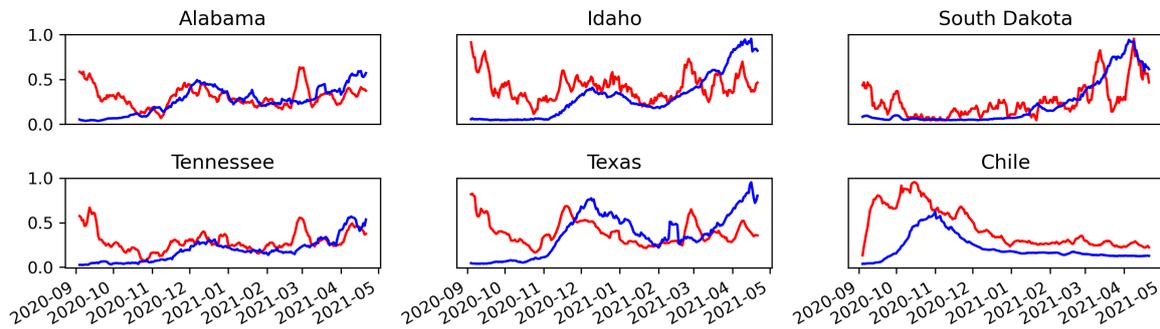
### 2.3. Twitter Mobility Origin Destination Matrices

We now present the data collection processes for the derivation of the O/D matrices.

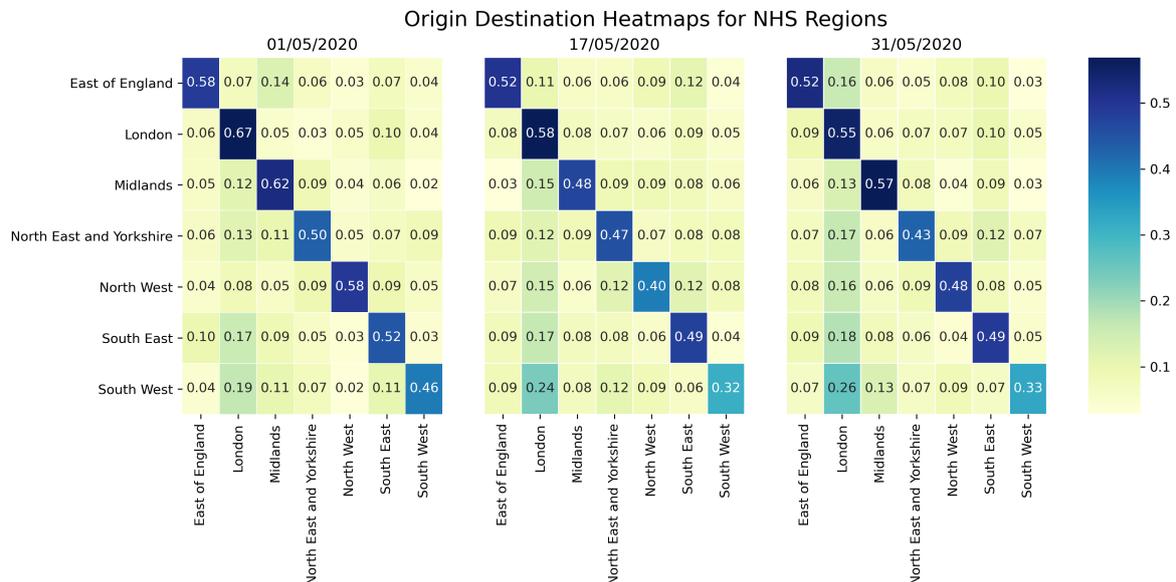
The flow of individuals travelling from one location to another can be expressed as an  $M \times M$  matrix, where  $M$  is the number of locations in the simulation area. The observation period of the data are 30 April 2020 to 31 May 2020. We divide England into the seven NHS regions, which are treated as separate locations. Tweets with the geo-location feature were collected using the same framework as described in Section 2.2.1; however, different Twitter developer API credentials used as tweets were not filtered based on keywords. To determine where an individual tweeted, a shapefile containing coordinates of the boundaries of the seven NHS regions was used.

If an individual tweets twice from two locations, for example, London (Origin) and South West (Destination), a movement is subsequently recorded. Figure 2 depicts each of these movements in the form of an O/D matrix. Locations on the  $x$ - and  $y$ -axes represent the origin and destination, respectively. Movements within regions, where an individual tweets multiple times in different locations within the same region, have also been collected. These are observed in the diagonal entries of the matrix.

Plots of Positive Cases and Twitter Counts



**Figure 1.** Plot of 7-day rolling average and standardised daily counts of positive COVID-19 cases (blue) and self-reported symptomatic tweets (red) for different US States and one South American country.



**Figure 2.** Heat-maps of origin destination matrices derived from Twitter for NHS regions. Locations on the *x*- and *y*-axes represent the origin and destination, respectively.

**3. Models**

In the following section, the model used for making inferences and death predictions when utilising different data feeds is outlined in Section 3.1. The extended SIR disease model catering for movement between different locations is described in Section 3.2.

*3.1. Model for Surveillance Data Comparison*

In this analysis, we use the statistical model developed by Moore, Rosato and Maskell [8].

The model can be described in two succinct parts. The transmission model (see Section 2(a) of [8]) is an extension of the classical SIR model outlining how individuals within the population move from being susceptible to exposed, then infected to recovered or dead. The model is implemented in the probabilistic programming language Stan [48] and uses a bespoke numerical integrator. Stan allows for statistical modelling and high-performance statistical computation by utilising the high-performance No-U-Turn Sampler (NUTS) [49]. The observation model (see Section 2(b) of [8]) outlines the relationship between the transmission model and the surveillance data feeds in Table 1 during calibration. The data are modelled via the method proposed in [8]. Daily counts of the surveillance

data feeds in Table 1 are assumed to follow a negative binomial distribution parameterised by mean  $x(t)$  and over-dispersion parameter  $\phi_x$ , such that

$$x_{\text{obs}}(t) \sim \text{NegativeBinomial}(x(t), \phi_x), \tag{7}$$

where  $x$  is data feed specific.

We refer the reader to [8] for a comprehensive description of the full model.

### 3.1.1. Computational Experiments

The time series considered begins on 17 February 2020. The start dates of each data feed follow those outlined in Table 1. The terminal time for the US States and the ROW is fixed on 1 February 2021, while, for NHS regions, the terminal time is 7 January 2021. In all cases, forecasts are considered to include seven days.

Similar to the experiments in [8], the analysis was run on the University of Liverpool’s High-Performance Computer (HPC). Each node has two Intel(R) Xeon(R) Gold 6138 CPU @ 2.00 GHz processors, a total of 40 cores and 384 GB of memory. In the following experiments, six independent Markov chains each draw 2000 samples, with the first 1000 discarded as burn-in. Run-time is dependent on the location of the data and the date at which the prediction is made. However, it typically takes 4.5 h per Markov chain for a complete run.

Initially, we only calibrate the model with death data and produce forecasts of seven daily death counts for the geographic locations described in Section 2.2 for the time periods outlined in Table 3. These forecasts are set as the baseline when comparing against forecasts incorporating low-latency data feeds.

**Table 3.** Prediction windows for the US States and the rest of the world, and NHS regions.

US States and the Rest of the World	NHS Regions
9 July 2020–16 July 2020	11 November 2020–18 November 2020
17 October 2020–24 October 2020	21 November 2020–28 November 2020
25 January 2021–1 February 2021	1 December 2020–8 December 2020
-	11 December 2020–18 December 2020
-	21 December 2020–28 December 2020
-	31 December 2020–7 January 2021

We use two metrics to determine the accuracy of the resulting forecasts. First, we calculate the MAE, which shows the average error over a set of predictions, and is given by

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |x^i - y^i|, \tag{8}$$

where  $N$  is the number of predictions and  $x_i$  and  $y_i$  are the predicted and true number of deaths on day  $i$ , respectively. The percentage difference between forecasts using only deaths ( $MAE_D$ ) and those combining deaths with low-latency data feeds ( $MAE_{DL}$ ) is calculated as follows:

$$\text{MAE \% Diff} = \frac{MAE_{DL} - MAE_D}{MAE_D}, \tag{9}$$

where a smaller percentage difference is preferred.

Secondly, we consider the uncertainties associated with the forecasts by assessing the NEES score. This is a popular method in the field of signal processing and tracking [50], recently applied to epidemiological forecasts in [8]. The metric determines whether the estimated variance of forecasts differs from the true variance. If the estimated variance is larger than the true variance, the forecast is over-cautious and if the estimated variance is smaller than the true variance, it is over-confident.

The NEES score is defined by

$$NEES = \frac{1}{N} \sum_{i=1}^N (x^i - y^i)^T C^{i-1} (x^i - y^i), \tag{10}$$

where  $C^{i-1}$  is the estimated variance at day  $i$ , as approximated using the variance of the samples for that day. If  $x^i$  is  $D$ -dimensional, then  $C^i$  should be a  $D \times D$  matrix, and the NEES score should be equivalent to  $D$  if the algorithm is consistent. As such, in assessing death forecasts, the desired NEES value is  $D \approx 1$ .

### 3.2. Model for Utilising Origin Destination Matrices

Here, we describe an extension of the discrete time approximation SIR model that includes movement between geographic locations [31,51] and is an extension of [52]. The population in location  $i$  is denoted  $P_i$ . At the beginning of the simulation,  $P_i$  is divided into three compartments: susceptible, infected and recovered, denoted  $S_{i,t}$ ,  $I_{i,t}$  and  $R_{i,t}$ , respectively, for timestep  $t$ . Location  $j$  represents the set of locations connected to location  $i$ . The origin of the pandemic is simulated at a random location, with a fraction of the susceptible compartment infected. The transmission rate in location  $i$  on day  $t$  is given by  $\beta_{i,t}$ , while  $m_{i,j}$  is the count of individuals travelling from location  $j$  to  $i$ . The global parameter  $\gamma$  describes the recovery rate.

The proportion of infected and susceptible individuals and the total population at locations  $j$  and  $i$  at time  $t$  are  $x_{j,t}$  and  $y_{i,t}$ , and  $N_j$  and  $N_i$ , respectively. The disease spreads via infected individuals travelling according to the O/D matrices in Figure 2. The full extended SIR model is described below:

$$S_{i,t+1} = S_{i,t} - \frac{\beta_{i,t} S_{i,t} I_{i,t}}{N_i} - \frac{\alpha S_{i,t} \sum_j m_{i,j}^t x_{j,t} \beta_{j,t}}{N_i + \sum_j m_{i,j}^t}, \tag{11}$$

$$I_{i,t+1} = I_{i,t} + \frac{\beta_{i,t} S_{i,t} I_{i,t}}{N_i} + \frac{\alpha S_{i,t} \sum_j m_{i,j}^t x_{j,t} \beta_{j,t}}{N_i + \sum_j m_{i,j}^t} - \gamma I_{i,t}, \tag{12}$$

$$R_{i,t+1} = R_{i,t} + \gamma I_{i,t}. \tag{13}$$

The number of infected individuals that move from all locations  $j$  to location  $i$  and transmit the disease to the susceptible population is given by

$$\sum_j m_{i,j}^t x_{j,t} \beta_{j,t}. \tag{14}$$

Uninfected individuals at location  $i$  are infected by individuals at locations  $j$  with probability

$$\frac{\alpha S_{i,t} \sum_j m_{i,j}^t x_{j,t} \beta_{j,t}}{N_i + \sum_j m_{i,j}^t}. \tag{15}$$

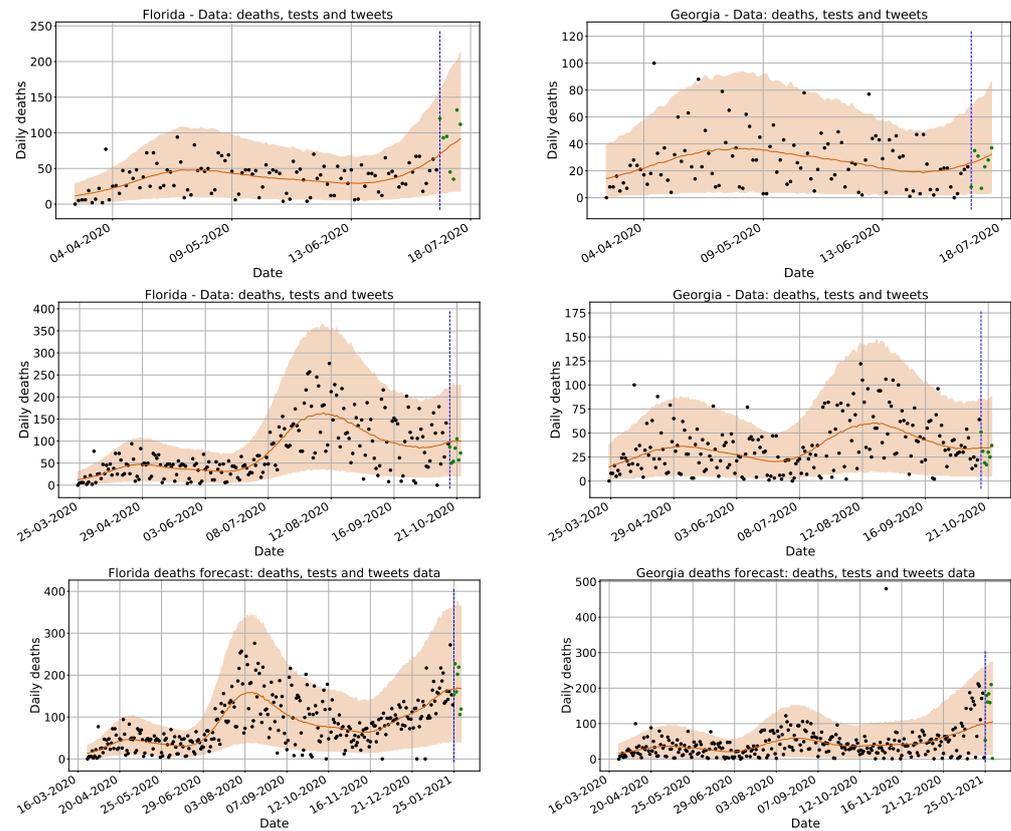
This rate is dependent on  $\alpha$ , which describes the intensity of the movement of individuals and is referred to as the social connectivity parameter.

## 4. Results

The two sets of results are now outlined. Comparison of the accuracy of death forecasts and findings on the impact of movement on the spread of a disease are presented in Sections 4.1 and 4.2, respectively.

### 4.1. Surveillance Data Comparison

The NEES value and MAE percentage difference between the baseline, ingesting solely deaths, and the incorporation of low-latency data feeds for the US States, the ROW, and NHS regions are given in Tables A1–A3, respectively. For all geographic locations, the results are averaged over the prediction windows described in Table 3. A visual representation of these predictions windows can be seen in Figure 3.



**Figure 3.** Death forecasts in Florida (left) and Georgia (right). The first, second and third prediction windows outlined in Table 3 and presented in the first, second and third rows, respectively. Confidence intervals of 1 standard deviation from the mean given by the orange ribbon, the mean sample given by the red line and the beginning of the prediction period by the vertical blue dashed line. True deaths are given by the black and green dots.

When forecasting deaths using the data in [2], calibrating the model with tests, tweets, and tests and tweets gives a 5%, 6% and 5% average increase in percentage performance, respectively, for the US States. The corresponding improvement rates for the ROW are 6%, 17% and 24%. An example of this improvement is presented in Figure 4 for death predictions in Colombia over the period 25 January 2021–1 February 2021. Considering the mean sample in the plots, outlined in red, incorporating tests and tweets follows the true death trend, outlined in green, with more accuracy when compared to only ingesting death data, for which the forecast continues to increase despite true deaths falling.

For the US States, the average NEES values are 1.696, 1.409, 1.483 and 1.269 when ingesting solely death data, tweets, tests, and tweets and tests, respectively. The corresponding results for the ROW are 0.433, 0.500, 1.198 and 0.723. As explained in Section 3.1.1, a NEES value of  $\sim 1$  is desired, with values  $< 1$  and  $> 1$  indicating that the forecast is over-cautious or over-confident, respectively. Ingesting any combination of the data feeds provides a NEES value closer to 1 than the death only forecast in both cases.

Results for NHS regions are less consistent. Ingesting hospital admissions, 111 calls and 111 online data sets provide an average increase in performance of 22%, 17% and 22%, respectively. However, tweets and Zoe App data perform less well, with decreases in performance of 2% and 124%, respectively. We perceive that this issue arises because, in these feeds, symptoms are self-diagnosed. Consequently, the counts may include relatively large numbers of people who do not have COVID-19.

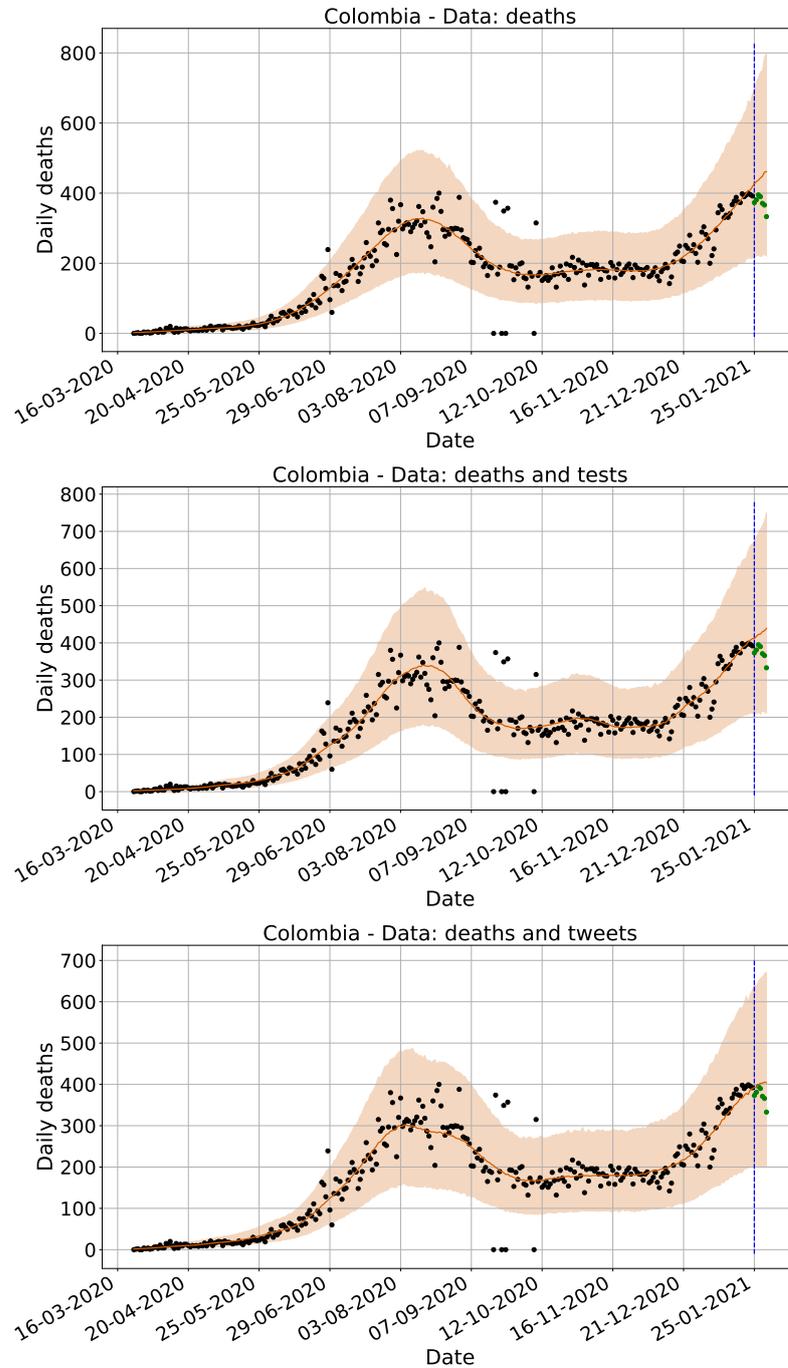
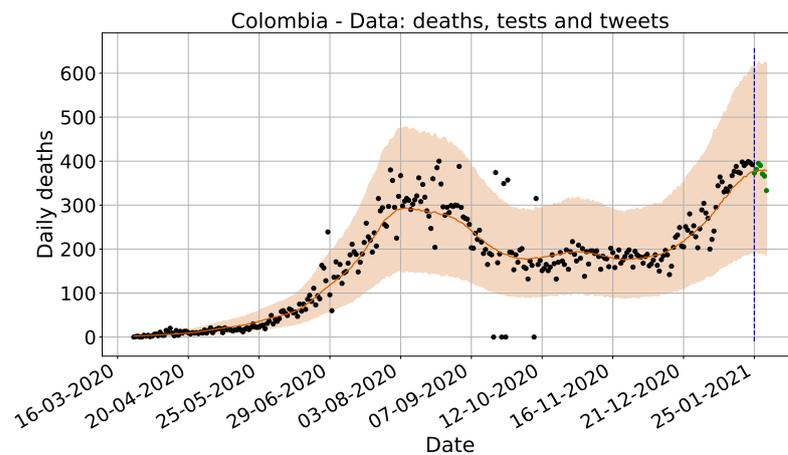
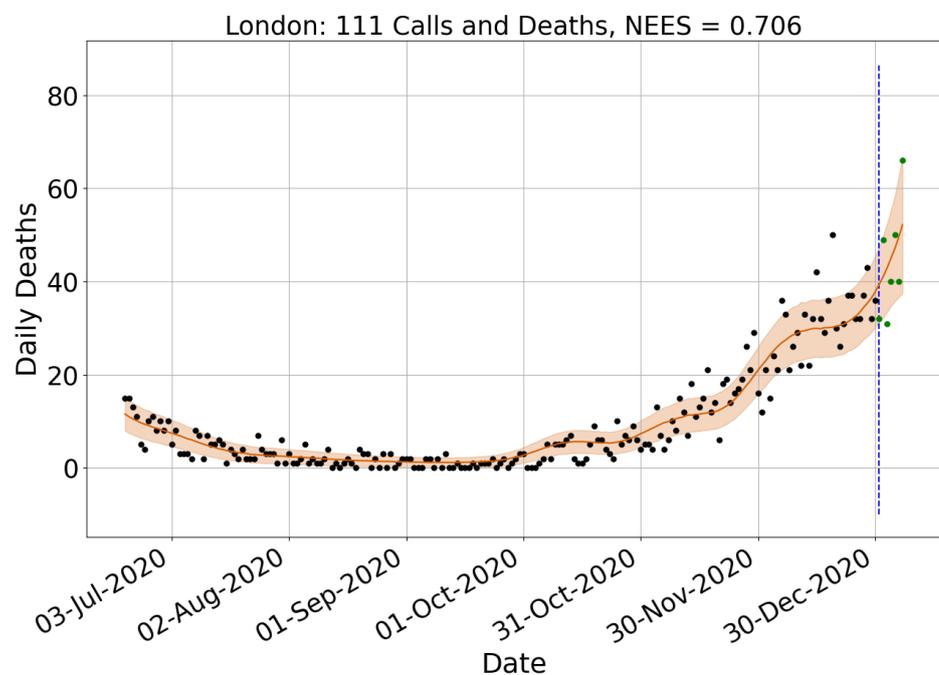


Figure 4. Cont.

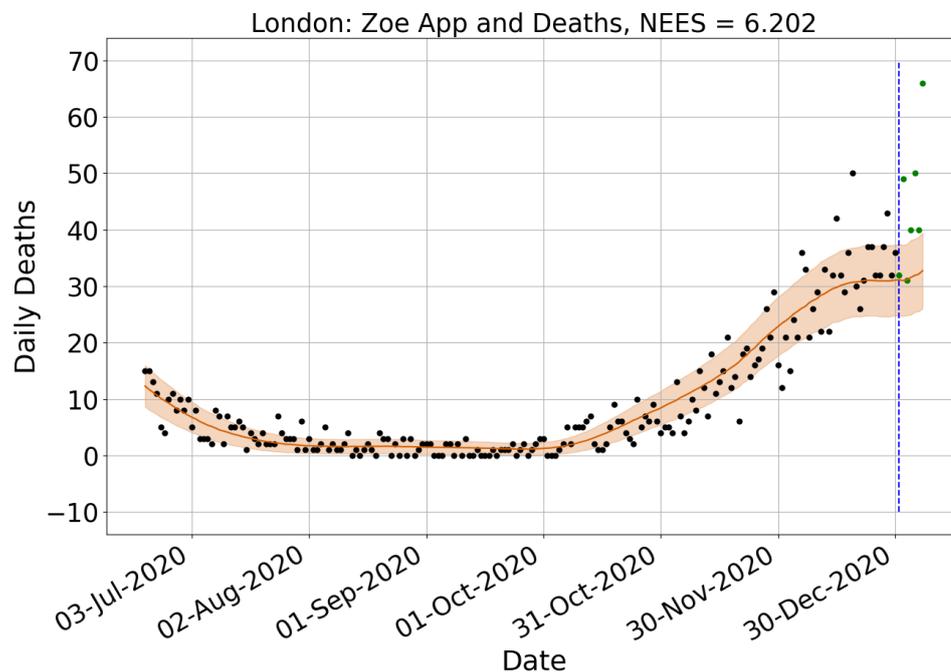


**Figure 4.** Colombian death forecasts for combinations of data sets. Confidence intervals of 1 standard deviation from the mean given by the orange ribbon, the mean sample given by the red line and the beginning of the prediction period by the vertical blue dashed line. True deaths are given by the black and green dots.

NEES values for NHS regions when ingesting solely deaths, hospital admissions, tweets, Zoe App, 111 call and 111 online data are 0.662, 0.682, 1.044, 3.160, 0.916 and 0.912, respectively. These results infer that, apart from Zoe App data where forecasts are overly-confident, ingesting all types of data feeds provides more consistent forecasts. Figure 5 exemplifies this finding. In the top image, the forecast encapsulates almost all true deaths. However, when ingesting the Zoe App data, the forecast only encapsulates two out of the seven true deaths, resulting in a NEES value of 6.202, which indicates an over-confident estimate.



**Figure 5.** Cont.



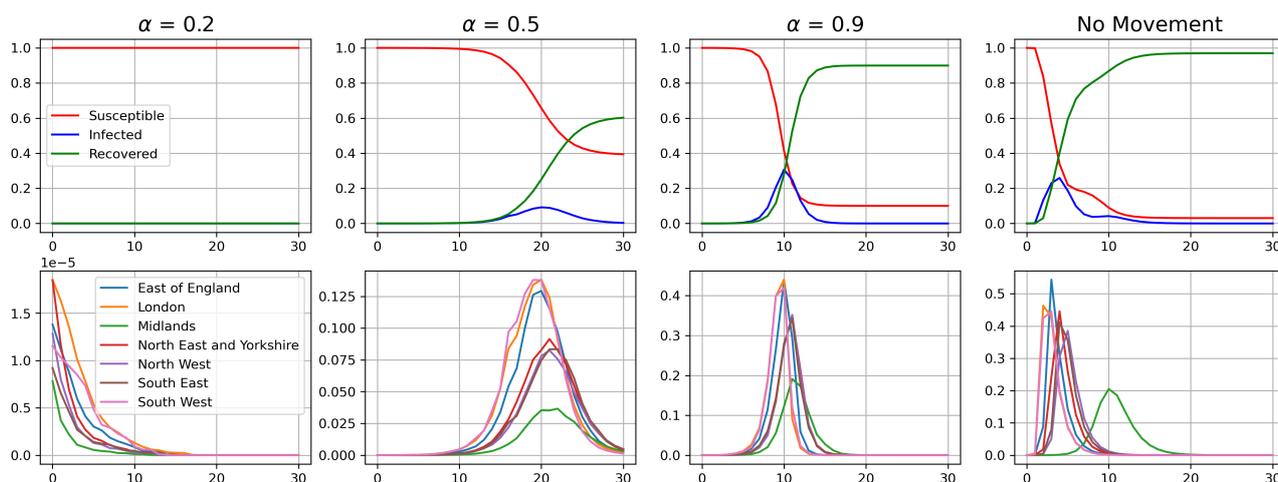
**Figure 5.** London death forecasts for death and 111 call data (**top**) and death and Zoe App data (**bottom**). Confidence intervals of 1 standard deviation from the mean given by the orange ribbon, the mean sample given by the red line and the beginning of the prediction period by the vertical blue dashed line. True deaths are given by the black and green dots.

#### 4.2. Origin Destination Matrices Analysis

As explained in Section 2.3, a movement is recorded if an individual tweets twice in one day in different locations over a 24 h period. The counts are assumed to be a percentage of the true population for the seven NHS regions. Figure 6 depicts these aggregated movements as O/D matrices.

Figure 6 shows the effect of the social connectivity parameter,  $\alpha$ , on the spread of a disease. This parameter models the level of contact individuals have with one another when travelling between locations. For example, implementing a lockdown, using a personal car or travelling via public transport will correspond to increasing values of  $\alpha$ .

Figure 6 exemplifies the role of  $\alpha$  when simulating the disease dynamics. The SIR epidemic curves for England are presented in the top row and the infected curves for each NHS region in the bottom row. Limiting contacts within the population through specification of  $\alpha = 0.2$  results in disease ceasing by day 15. For  $\alpha = 0.5$ , the peak number of infections occurs at approximately day 20 and consists of just over 0.1% of the population. In contrast, when  $\alpha = 0.9$ , the peak occurs at approximately day 10 and 0.3% of the population are infected. Simulations of the SIR curves under no movement between NHS regions are also provided in the rightmost column of Figure 6.



**Figure 6. (Top row):** Susceptible, Infected and Recovered epidemic curves for England with different values of the social connectivity parameters and with no movement between regions. **(Bottom row):** The infected curves for the different NHS regions for different social connectivity parameters and no movement between regions.

## 5. Conclusions and Future Work

In this paper, we have outlined a method for detecting symptomatic COVID-19 tweets in multiple languages. Calibrating the epidemiological model outlined in Section 3.1 with low-latency data feeds, including symptomatic tweets, provides more accurate and consistent forecasts of daily deaths when compared with using death data alone. We have also shown how to extract movement data from Twitter in the form of O/D matrices. These movement data were utilised in an extended SIR model to better represent the spread of a disease.

Incorporating symptomatic tweets for UK regions does not provide the same level of improvement as for other geographic locations. One reason for this reduced improvement could be that daily counts of tweets for NHS regions are less plentiful than for the US States or the rest of the world. It is possible to pay for a premium Twitter API that allows the user to download a higher percentage of tweets than that used in this study. A second way to potentially increase the hit rate of geo-located tweets is to use natural language processing techniques to estimate the location of the tweet user, such as those outlined in the review [53]. Another direction for future work is to train a more sophisticated classifier such as the Bidirectional Encoder Representations from Transformers (BERT) classifier [54].

Calibrating the model in Section 3.1 with movement data was not explored in this analysis due to the computational effort required. One interesting direction for future work would be to use a sequential Monte Carlo (SMC) sampler [55] in place of the MCMC sampling algorithm. An example of such sampler that uses NUTS as the proposal can be found here [56].

**Author Contributions:** Conceptualization, S.M. and J.H.A.; methodology, C.R., M.C., J.H. (John Heap) and S.M.; software, C.R., R.E.M., M.C. and J.H. (John Heap); validation, C.R., R.E.M., M.C. and J.H. (John Heap); formal analysis, C.R.; investigation, C.R.; resources, C.R., M.C., J.H. (John Heap) and J.S.; data curation, C.R., M.C., J.H. (John Heap) and J.S.; writing—original draft preparation, C.R.; writing—review and editing, C.R., R.E.M., M.C., J.H. (John Heap), J.H. (John Harris), J.S. and S.M.; visualization, C.R., M.C. and J.H. (John Heap); supervision, J.H. (John Harris), J.S. and S.M.; project administration, S.M.; funding acquisition, S.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by a Research Studentship jointly funded by the EPSRC and the ESRC Centre for Doctoral Training on Quantification and Management of Risk and Uncertainty in Complex Systems Environments Grant No. (EP/L015927/1) and an ICASE Research Studentship jointly funded by EPSRC and AWE Grant No. (EP/R512011/1), the EPSRC Centre for Doctoral

Training in Distributed Algorithms Grant No. (EP/S023445/1) and the EPSRC through the Big Hypotheses Grant No. (EP/R018537/1).

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** All data and code used in this research paper can be found at: <https://codatmo.github.io> (accessed on 6 March 2023).

**Acknowledgments:** The authors would like to thank Serban Ovidiu and Chris Hankin from Imperial College London, and Ronni Bowman and Riskaware for their support and helpful discussions of this work. We would also like to thank the team at the Universidade Nove de Julho—UNINOVE in Sao Paulo, Brazil with the help they provided in labelling the Portuguese tweets. We would also like to thank Breck Baldwin for helping to make progress with CoDatMo.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

O/D	Origin/Destination
NHS	National Health Service
WHO	World Health Organisation
$R_t$	Reproduction Number
UK	United Kingdom
ICU	Intensive Care Unit
NLP	Natural Language Processing
MAE	Mean Absolute Error
NEES	Normalised Estimation Error Squared
ROW	Rest of the World
JHU CSSE	Johns Hopkins University Center for Systems Science and Engineering
SVM	Support Vector Machine
NUTS	No-U-Turn Sampler
HPC	High-Performance Computer
BERT	Bidirectional Encoder Representations from Transformers

## Appendix A

**Table A1.** The US States: MAE and NEES when using deaths and when using deaths and different low-latency data feeds. Lower MAE diff and NEES~1 = better. Averaged over the prediction windows in Table 3. Only the English classifier was used.

Geographic Location	Deaths NEES	Tests		Twitter		Tests and Twitter	
		MAE % Diff	NEES	MAE % Diff	NEES	MAE % Diff	NEES
Alaska	0.329	−36	0.334	−29	0.301	−92	0.302
Alabama	0.684	−29	1.874	−29	1.723	−2	1.000
Arkansas	0.275	3	0.317	−1	0.288	−1	0.313
Arizona	0.337	20	0.334	18	0.344	−20	0.244
California	0.611	6	0.709	9	0.802	5	1.206
Colorado	1.886	−25	0.401	−41	0.457	10	1.278
Connecticut	13.406	−8	1.922	−2	0.875	2	11.459
Delaware	3.020	−3	0.918	16	1.046	12	0.727
Florida	0.406	−24	0.179	13	0.353	−20	0.454
Georgia	0.550	9	0.325	41	0.891	−48	0.255
Hawaii	11.459	−12	28.114	−4	24.695	17	10.149

Table A1. Cont.

Geographic Location	Deaths	Tests		Twitter		Tests and Twitter	
	NEES	MAE % Diff	NEES	MAE % Diff	NEES	MAE % Diff	NEES
Iowa	19.176	5	7.720	4	1.476	-3	1.600
Idaho	0.914	0	0.809	2	1.791	7	0.986
Illinois	0.573	9	0.350	13	0.319	-116	1.091
Indiana	0.561	-17	0.652	-40	0.781	0	0.481
Kansas	1.021	1	1.037	-2	1.835	1	0.488
Kentucky	0.355	-4	0.374	10	0.548	-15	0.214
Louisiana	0.298	-7	0.305	-2	0.341	9	0.234
Massachusetts	0.351	3	0.342	-3	0.365	14	0.409
Maryland	0.485	-3	0.619	10	0.581	31	0.313
Maine	0.488	1	0.567	-28	0.796	-9	0.952
Michigan	0.592	-6	0.445	-7	0.453	4	0.850
Minnesota	0.683	9	1.019	11	1.200	51	0.747
Missouri	0.810	-7	1.165	-27	1.609	20	0.475
Mississippi	0.683	12	0.721	2	0.997	-15	0.320
Montana	5.034	4	2.244	-1	1.538	-5	5.189
North Carolina	0.908	-1	0.453	9	0.877	-19	0.570
North Dakota	0.513	-32	0.521	-18	0.544	-8	0.661
Nebraska	0.259	5	0.253	7	0.570	5	0.286
New Hampshire	0.252	-74	0.240	-148	0.430	-36	0.288
New Jersey	0.901	-7	0.788	-6	0.926	10	3.177
New Mexico	0.832	-28	0.738	-12	0.969	0	0.489
Nevada	2.129	-24	0.353	-12	0.425	-13	1.904
New York	0.496	31	0.146	3	0.135	-17	0.418
Ohio	0.263	63	0.675	54	0.468	3	0.337
Oklahoma	0.301	-5	0.369	0	0.621	8	0.256
Oregon	0.729	0	1.032	-2	1.692	-4	0.793
Pennsylvania	0.411	-7	0.385	0	0.426	10	0.402
Rhode Island	0.609	-9	0.546	-31	0.446	-2	1.699
South Carolina	2.072	-3	2.157	-4	5.601	-39	0.429
South Dakota	1.259	14	1.080	-2	1.089	2	5.050
Tennessee	0.794	15	1.191	14	1.687	-11	0.600
Texas	0.585	6	0.784	1	0.750	-71	0.706
Utah	0.499	-98	0.716	-127	1.196	13	0.632
Virginia	0.731	-10	0.396	6	0.864	9	0.676
Vermont	0.142	59	0.300	-1	0.163	40	0.043
Washington	0.608	-8	0.561	19	1.787	-1	0.782
Wisconsin	0.842	6	1.028	25	3.921	8	0.850
West Virginia	0.650	-6	0.547	2	1.042	7	0.291
Wyoming	1.939	5	0.951	-15	1.126	25	0.395
<b>Average</b>	<b>1.696</b>	<b>-5</b>	<b>1.409</b>	<b>-6</b>	<b>1.483</b>	<b>-5</b>	<b>1.269</b>

**Table A2.** Rest of the World: MAE and NEES when using deaths and when using deaths and different low-latency data feeds. Lower MAE diff and NEES~1 = better. Averaged over the prediction windows in Table 3. Language column states which classifier was used.

Geographic Location	Language	Deaths	Tests		Twitter		Tests and Twitter	
		NEES	MAE % Diff	NEES	MAE % Diff	NEES	MAE % Diff	NEES
Argentina	Spanish	0.567	3	0.695	-17	0.904	-19	0.765
Bolivia	Spanish	0.339	-85	0.207	-117	0.182	-118	0.195
Brazil	Portuguese	0.396	-4	0.405	11	0.578	4	0.493
Chile	Spanish	0.371	15	0.439	14	0.506	10	0.425
Colombia	Spanish	0.154	17	0.243	-46	0.164	-115	0.223
Costa Rica	Spanish	0.423	6	0.583	18	3.060	2	0.786

Table A2. Cont.

Geographic Location	Language	Deaths NEES	Tests		Twitter		Tests and Twitter	
			MAE % Diff	NEES	MAE % Diff	NEES	MAE % Diff	NEES
Ecuador	Spanish	0.156	−26	0.195	−99	0.234	−69	0.234
Guatemala	Spanish	0.557	−19	0.670	−31	0.815	−31	0.713
Honduras	Spanish	0.405	−8	0.381	−27	0.915	−41	0.541
Mexico	Spanish	0.766	16	0.939	11	1.100	11	1.110
Nicaragua	Spanish	0.091	−13	0.207	−24	1.340	−22	0.364
Panama	Spanish	0.550	−20	0.421	−4	0.451	−7	0.368
Paraguay	Spanish	0.535	28	0.877	−7	2.615	8	1.473
Peru	Spanish	0.507	33	0.103	26	1.630	16	0.515
Uruguay	Spanish	0.619	11	0.742	−13	0.899	−7	0.643
Venezuela	Spanish	0.610	−14	0.713	−49	0.890	−91	0.603
Germany	German	0.379	5	0.613	15	2.131	14	1.570
Italy	Italian	0.360	17	0.557	29	3.149	34	1.991
<b>Average</b>		<b>0.433</b>	<b>−6</b>	<b>0.500</b>	<b>−17</b>	<b>1.198</b>	<b>−24</b>	<b>0.723</b>

**Table A3.** NHS Regions: MAE and NEES when using deaths and when using deaths and different low-latency data feeds. Lower MAE diff and NEES~1 = better. Averaged over the prediction windows in Table 3. Only the English classifier was used.

Geographic Location	Deaths		Hospital		Twitter		Zoe App		111 Calls		111 Online	
	NEES	MAE % Diff										
East of England	0.435	-13	0.419	-7	0.655	38	2.908	-15	0.820	-19	0.795	
London	0.878	-36	0.666	-7	1.163	131	3.150	-43	0.750	-47	0.754	
Midlands	0.635	-16	0.466	13	0.569	132	3.330	-19	0.418	-47	0.404	
North East and Yorkshire	0.753	5	1.188	-4	0.824	153	2.325	-16	0.860	-14	0.888	
North West	0.735	-1	0.756	17	1.408	129	3.285	-25	0.932	-25	0.934	
South East	0.652	-24	0.805	-3	1.255	126	4.390	8	1.018	6	0.957	
South West	0.545	-69	0.474	2	1.432	160	2.729	-8	1.617	-6	1.653	
<b>Average</b>	<b>0.662</b>	<b>-22</b>	<b>0.682</b>	<b>2</b>	<b>1.044</b>	<b>124</b>	<b>3.160</b>	<b>-17</b>	<b>0.916</b>	<b>-22</b>	<b>0.912</b>	



25. Sharma, K.; Seo, S.; Meng, C.; Rambhatla, S.; Liu, Y. COVID-19 on social media: Analyzing misinformation in twitter conversations. *arXiv* **2020**, arXiv:2003.12309.
26. Al-Garadi, M.A.; Yang, Y.C.; Lakamana, S.; Sarker, A. A Text Classification Approach for the Automatic Detection of Twitter Posts Containing Self-Reported COVID-19 Symptoms. 2020. Available online: <https://openreview.net/forum?id=xyGSIthHYO> (accessed on 6 March 2023).
27. Sarker, A.; Lakamana, S.; Hogg-Bremer, W.; Xie, A.; Al-Garadi, M.A.; Yang, Y.C. Self-reported COVID-19 symptoms on Twitter: An analysis and a research resource. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 1310–1315. [[CrossRef](#)] [[PubMed](#)]
28. Garcia, K.; Berton, L. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Appl. Soft Comput.* **2021**, *101*, 107057. [[CrossRef](#)] [[PubMed](#)]
29. Kar, D.; Bhardwaj, M.; Samanta, S.; Azad, A.P. No rumours please! A multi-indic-lingual approach for COVID fake-tweet detection. In Proceedings of the 2021 Grace Hopper Celebration India (GHCI), Bangalore, India, 18 January–3 February 2021; pp. 1–5.
30. Badr, H.S.; Du, H.; Marshall, M.; Dong, E.; Squire, M.M.; Gardner, L.M. Association between mobility patterns and COVID-19 transmission in the USA: A mathematical modelling study. *Lancet Infect. Dis.* **2020**, *20*, 1247–1254. [[CrossRef](#)] [[PubMed](#)]
31. Goel, R.; Sharma, R. Mobility based sir model for pandemics-with case study of covid-19. In Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), The Hague, The Netherlands, 7–10 December 2020; pp. 110–117.
32. Osorio-Arjona, J.; García-Palomares, J.C. Social media and urban mobility: Using twitter to calculate home-work travel matrices. *Cities* **2019**, *89*, 268–280. [[CrossRef](#)]
33. Huang, X.; Li, Z.; Jiang, Y.; Li, X.; Porter, D. Twitter reveals human mobility dynamics during the COVID-19 pandemic. *PLoS ONE* **2020**, *15*, e0241957. [[CrossRef](#)]
34. Lombardi, A.; Amoroso, N.; Monaco, A.; Tangaro, S.; Bellotti, R. Complex Network Modelling of Origin–Destination Commuting Flows for the COVID-19 Epidemic Spread Analysis in Italian Lombardy Region. *Appl. Sci.* **2021**, *11*, 4381. [[CrossRef](#)]
35. Gómez, S.; Fernández, A.; Meloni, S.; Arenas, A. Impact of origin-destination information in epidemic spreading. *Sci. Rep.* **2019**, *9*, 2315. [[CrossRef](#)]
36. Kondo, K. Simulating the impacts of interregional mobility restriction on the spatial spread of COVID-19 in Japan. *Sci. Rep.* **2021**, *11*, 18951. [[CrossRef](#)]
37. Flaxman, S.; Mishra, S.; Gandy, A.; Unwin, H.J.T.; Mellan, T.A.; Coupland, H.; Whittaker, C.; Zhu, H.; Berah, T.; Eaton, J.W.; et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **2020**, *584*, 257–261. [[CrossRef](#)]
38. Vinceti, M.; Filippini, T.; Rothman, K.J.; Ferrari, F.; Goffi, A.; Maffei, G.; Orsini, N. Lockdown timing and efficacy in controlling COVID-19 using mobile phone tracking. *EClinicalMedicine* **2020**, *25*, 100457. [[CrossRef](#)]
39. CoDatMo. 2021 Welcome to the CoDatMo Site. Available online: <https://codatmo.github.io> (accessed on 1 October 2021).
40. UK Government. 2021 Coronavirus (COVID-19) in the UK. Available online: <https://coronavirus.data.gov.uk/details/deaths> (accessed on 1 October 2021).
41. UK Government. 2021 Coronavirus (COVID-19) in the UK. Available online: <https://coronavirus.data.gov.uk/details/healthcare> (accessed on 1 October 2021).
42. Zoe App: COVID-Public-Data. Available online: <https://console.cloud.google.com/storage/browser/covid-public-data;tab=objects?prefix=&forceOnObjectsSortingFiltering=false> (accessed on 1 October 2021).
43. Potential Coronavirus (COVID-19) Symptoms Reported through NHS Pathways and 111 Online. Available online: <https://digital.nhs.uk/data-and-information/publications/statistical/mi-potential-covid-19-symptoms-reported-through-nhs-pathways-and-111-online/latest> (accessed on 1 October 2021).
44. Roesslein, J. Tweepy Documentation. 2009, Volume 5, p. 724. Available online: <http://tweepy.readthedocs.io/en/v3> (accessed on 8 May 2012).
45. COVID-19 Terms and MedDRA. Available online: <https://www.meddra.org/COVID-19-terms-and-MedDRA> (accessed on 1 October 2021).
46. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
47. Leetaru, K.; Wang, S.; Cao, G.; Padmanabhan, A.; Shook, E. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* **2013**. Available online: <https://journals.uic.edu/ojs/index.php/fm/article/view/4366> (accessed on 1 October 2021). [[CrossRef](#)]
48. Carpenter, B.; Gelman, A.; Hoffman, M.D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P.; Riddell, A. Stan: A probabilistic programming language. *J. Stat. Softw.* **2017**, *76*, 1430202. [[CrossRef](#)] [[PubMed](#)]
49. Hoffman, M.D.; Gelman, A. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **2014**, *15*, 1593–1623.
50. Chen, Z.; Heckman, C.; Julier, S.; Ahmed, N. Weak in the NEES?: Auto-tuning Kalman filters with Bayesian optimization. In Proceedings of the 2018 21st International Conference on Information Fusion (FUSION), Cambridge, UK, 10–13 July 2018; pp. 1072–1079.
51. Modelling the Coronavirus Epidemic in a City with Python. Available online: <https://towardsdatascience.com/modelling-the-coronavirus-epidemic-spreading-in-a-city-with-python-babd14d82fa2> (accessed on 24 October 2022).

52. Wesolowski, A.; zu Erbach-Schoenberg, E.; Tatem, A.J.; Lourenço, C.; Viboud, C.; Charu, V.; Eagle, N.; Engø-Monsen, K.; Qureshi, T.; Buckee, C.O.; et al. Multinational patterns of seasonal asymmetry in human movement influence infectious disease dynamics. *Nat. Commun.* **2017**, *8*, 2069. [[CrossRef](#)] [[PubMed](#)]
53. Huang, C.Y.; Tong, H.; He, J.; Maciejewski, R. Location Prediction for Tweets. *Front. Big Data* **2019**, *2*, 5. [[CrossRef](#)] [[PubMed](#)]
54. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
55. Del Moral, P.; Doucet, A.; Jasra, A. Sequential monte carlo samplers. *J. R. Stat. Soc. Ser. (Statist. Methodol.)* **2006**, *68*, 411–436. [[CrossRef](#)]
56. Devlin, L.; Horridge, P.; Green, P.L.; Maskell, S. The No-U-Turn sampler as a proposal distribution in a sequential Monte Carlo sampler with a near-optimal L-kernel. *arXiv* **2021**, arXiv:2108.02498.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.