Ordnance Survey COVID-19 Response: Predicting the Geospatial Spread of Disease using Spatial Interaction Modelling with Gridded Data

Ben Attwood BSc, Arjan Dhaliwal MMath, Kyriaki Dionysopoulou PhD, Freja Hunt PhD, Josh Pooley PhD, Jacob Rainbow MMath, and Jonathan Hughes MMath
Rapid Prototyping Team
Ordnance Survey Ltd.

Kirk Harland PhD*
Actionable Intelligence

Martyn Fyles MSci†
The Alan Turing Institute

David Martin PhD‡
Geography and Environmental Science
University of Southampton

(Office for National Statistics and Population247 Collaboration)
(Dated: July 6, 2020)

To support the scientific community in understanding the dynamic behaviour of COVID-19, we present a framework designed to facilitate the investigation of geospatial factors influencing disease spread, with example outputs to demonstrate how it can be used. Here, we focus on COVID-19, but the framework is applicable to any transmissible disease. We aim to help address two questions: (1) what are the geospatial variations in virus spread and impact, and (2) how does the geospatial spread and impact of the virus change if population mobility is altered? We approach these questions by using a network to represent the geographical interactions within and between populations in nodes, combined with a Susceptible-Exposed-Infected-Recovered-Dead (SEIRD) model to determine disease transmission and impacts. The model is demonstrated for the regions of Southampton and Gosport (in the South of England), with both pre-lockdown and lockdown scenarios simulated.

I. Introduction

The emergence, and subsequent global spread, of a novel coronavirus during late 2019 has led to widespread day-to-day disruption as governments try to reduce the number of people becoming infected and minimise the growing death toll. The virus causes a disease known as `COVID-19` in humans and was confirmed as a global pandemic on 11$^{\text{th}}$ March 2020. Despite the virus now being a global phenomenon, the impact has varied dramatically by region, not only because of different measures taken by different governments, but also within areas under the same degree and type of restriction; for example, in the UK, despite identical non-pharmaceutical interventions being applied nationwide, infections initially rose fastest in London, the Midlands and the North West. This demonstrates the vital role geospatial information has to play in understanding the impact of the virus, determining how best to provide resources to support those impacted, and designing strategies to balance population movement needs with reducing virus transmission rates. Evidence from a major study early on in the pandemic [1] suggests that "[g]iven local epidemics are not perfectly synchronised, local policies are also more efficient and can achieve comparable levels of suppression to national policies while being in force for a slightly smaller proportion of the time"; this means that accounting for the influence of geospatial factors when addressing the challenges posed by this novel coronavirus could significantly improve the effectiveness of a given approach.

This paper brings together a unique combination of geospatial data scientists from Ordnance Survey (OS), epidemiologists from The Alan Turing Institute and the University of Manchester, and population dynamics experts from the University of Southampton. To support efforts by the scientific community to understand coronavirus dynamics, we present here a framework designed to facilitate the investigation of geospatial factors influencing disease spread, with example outputs to demonstrate how it can be used. We hope this framework, called `InfectioNet`, will help with targeting healthcare and community resources where they are most needed, as well as informing decisions about the way mobility restrictions are eased by demonstrating the impact of different options. Here we focus on `COVID-19`, but the

* https://www.researchgate.net/scientific-contributions/80028845_Kirk_Harland
† https://www.turing.ac.uk/people/doctoral-students/martyn-fyles
‡ https://www.southampton.ac.uk/geography/about/staff/djm1.page

framework is applicable to any transmissible disease.

We aim to help address two questions:

1. What are the geospatial variations in virus spread and impact?

2. How does the geospatial spread and impact of the virus change if population mobility is altered?

We approach these questions using a network to represent the geographical interactions within and between populations in nodes, combined with a Susceptible-Exposed-Infected-Recovered-Dead (`SEIRD`) model to determine disease transmission and impacts. The data used to define the nodes in the network is described in Section II A, while the structure of the network itself is explained in Section II B. Section II C sets out how population movements are determined and Section II E details the principles of the `SEIRD` disease model. Section II F shows how these elements are brought together to simulate a time evolving geospatial disease model, with the calibration process for specific disease attributes detailed in Section II G. Examples that demonstrate `InfectioNet` are provided in Section III, with two models set in Southampton, UK, and another two set in Gosport, UK. Lastly, in Section IV we outline some limitations of the `InfectioNet` model and discuss some ideas for improvement in the future.

II. Method

A. Data

To model the spread of the disease, we need to know the geospatial distribution of the population, since changes in this distribution (as people move during the day) contribute to opportunities for disease transmission. We use gridded population maps from `Population247` to determine the population distribution at two representative times of the day: `0200` hours and `1400` hours. Whilst it is possible for `Population247` to estimate population distributions at other times of the day, we elected to bisect each day at `0200` hours and `1400` hours as these seemed like reasonable times for the general population to be considered at home or away from home, respectively. OS's `AddressBasePlus` is used to provide a gridded count of residential dwellings (hereafter referred to as households), which helps to determine the number and type of interactions individuals may have in each location. These data sources are described in more detail below.

To determine the number of households in a location, we used the Unique Property Reference Number (UPRN) from `AddressBasePlus`. In this context, the UPRN is a proxy for a postal address, so a house may have one UPRN assigned to it, while a block of flats will have several. Households were defined to mean UPRNs that have a domestic Valuation Office Agency (VOA) or belonged to one of the following classes: *Residential*, *Residential Dwelling*, *Detached*, *Semi-detached*, *Terraced*, *Self-Contained Flat*, *Sheltered Accommodation*, or *House in Multiple Occupancy* (note, this does not include bedsits, caravans, or house boats). These addresses are aggregated into `200m` grid cells (aligning with integer multiples of 200 in both Eastings and Northings of British National Grid) and summed within that cell, comprising the households in which the local populations live. Although households and residential spaces are not synonymous (some dwellings may be empty or some spaces may have more than one 'household') they are assumed to be so for the work presented here. For specific applications or in future work it may make sense to consider more complex household / residential dwelling relationships.

The data supplied by `Population247` has an advantage over census data in this context, as the latter only relate to a particular date and are primarily based on 'night-time' household assumptions. In practice, there is no time of day when all residents can be expected to be at their primary residential address. It is, however, a reasonable approximation if we assume that the middle of the night (here, `0200` hours) corresponds to a time when the highest proportion of the population is at home. We use the distribution at `0200` hours to consider how disease is transmitted between members of the same household, while day-time distributions allow interactions with those outside of an individual's household. This is explained in more detail in Section II F.

`Population247` data is ingested as two raster (gridded) files: one for the `0200` hours population and one for the `1400` hours population. The resolution of the data is currently 200m, though there is flexibility should a finer granularity be required. There is also extra stratification available: by age bracket in years `(0-4, 5-9, 10-15, 16-17, 18-64, 65+)`, or by term-calendar (Term Time, Out-of-Term Time). We use two different configurations of the `Population247` data—one representing 'business as usual' or pre-lockdown mobility, and one representing reduced mobility during lockdown.

For the purpose of creating a representation of the pre-lockdown population distribution, all age brackets were aggregated and the Office for National Statistics (ONS) usual residence definition was implemented, which assumes that all students are at their term-time addresses: typically halls of residence or rental properties in their place of study [2]. As well as residential households, other origins of populations include communal establishments, such as prisons and care homes. In this paper, a destination is defined as a grid cell which can contain population without itself necessarily being a residential origin. Examples of destinations include workplaces, educational establishments and healthcare sites. The source data is largely from the 2011 Census (e.g. Ages [3], Student Accommodation [4], Communal Establishments [5], Workplace Population [6], Employment Status [7]), with additional inputs from NHS Digital (Attendance Fig-

ures [8]), Higher Education Statistics Agency (enrolled student numbers [9]), Government Data Portal (School Pupil Numbers [10]), and British Tourist Authority (Attractions [11]).

For the purpose of creating a lockdown version of these gridded population counts, `Population247` recreated the same data structure using updated assumptions. The resident population counts were revisited using the `2018 Small Area Population Estimates` [12]. Using the best available data in the press about the relocation of students, `Population247` assumed that 90% of students had returned to parental/out-of-term time addresses by April; this figure is considered a reasonable initial estimate, but is subject to change. Moreover, national lockdown restricted the mobility of (and access to) the care home population and prisoners. Estimates for the number of people in care homes came from 2011 Census, with the locations taken from the Care Quality Commission [13]. The number of incarcerated individuals came from the Government Prison Statistics [14]. Some of these assumptions were discussed in personal communication with `Population247`, but for some more information, refer to *Pop247Cov: Notes on building a lockdown population base*, available upon request.

For both the pre-lockdown and lockdown versions of the `Population247` datasets, the modelling also used a raster mask based on open datasets (Roads [15], Coast [16], Inland Water [17]) to identify land areas into which population could theoretically be allocated. The trunk roads were also weighted using data from the Department for Transport (Annual Average Daily Flows [18]). Using these datasets, it is possible to restrict populations to places they are most likely to be as well as admitting appropriate numbers to be in transit somewhere on the road network.

B. Network Structure

The data described in Section II A is used to define an `InfectioNet`, which is a custom mobility network for modelling disease propagation. `InfectioNet` is formed of the components described below: nodes, edges, disease parameters and population. Figure 1 is a schematic of the network demonstrating how these components interact.

a Nodes The nodes in the network represent geographic locations. In the examples presented in Section III, node locations refer to the centroids of a `200m` grid. This means each node represents the same geographic extent but the populations of each node can vary freely. Although we use gridded data here, there are no restrictions on the geographic distribution of the nodes; for example, we have also used population weighted centroids of Lower Layer Super Output Areas (LSOAs) as node locations within this framework. In contrast to a uniform grid, LSOAs can vary greatly in geographic extent but contain broadly similar total populations. Additionally, the nodes do not need to provide total geographic coverage to a region—for areas where there is a very low chance of people being present at any time of day (for the duration of the run), it is not necessary to define nodes. The structure of the network is fixed for the duration of the model but edge weights can be changed at specific epochs by using `EvolveNet` (see Section II H).

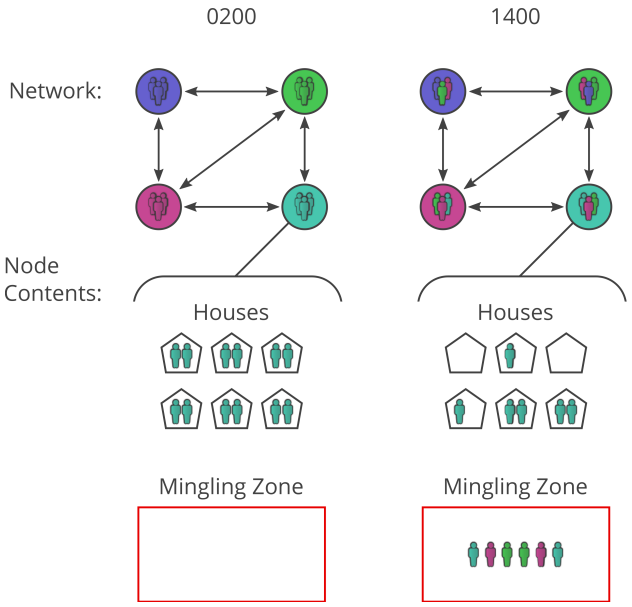

Figure 1. Schematic of the network demonstrating how the components interact and the configuration at two different times in a day (at `0200` hours left panel, and `1400` hours right panel). Each one of the nodes (coloured circles) can have many households/homes (pentagons) and contains a number of `Person` objects, which are colored by their home-node color. At `0200` hours everyone is at their home node within their house. At `1400` hours some `Person` objects will have moved out of their house and entered a Mingling Zone, either within their home node or a connected node.

Each node stores the `Person` objects assigned to that location at any given time step and a count of households within the geographic area (for more information on `Person` see part *d* of this section). The internal structure of a node consists of a single Mingling Zone and several households, if present. The Mingling Zone represents a public space within the node where disease transmission can occur from visitor to visitor, or between visitors and a sample of locals. This means that each node may contain multiple independent interaction zones: one per household for interactions between locals of the same household, and one for interactions between the population in the Mingling Zone.

Independent zones help constrain interactions to more realistic levels, rather than assuming an individual interacts with the entire population of the node. The idea behind the zone splitting is based on [19], which provides the scientific basis of why an infected person is much more likely to infect their cohabitants or work colleagues than people they interact with in public spaces.

Finally, each one of the nodes contains statistics about

the development of the disease (in this case `COVID-19`) in terms of the populations in each of the compartments of the Susceptible-Exposed-Infected-Recovered-Dead (`SEIRD`) model presented in Section II E.

b Edges Nodes are connected to each other by edges, with weights that determine the number of local people, per age bracket, that will move to the Mingling Zone of each adjoining node. Edges are directional and so exchange between connected nodes is not constrained by symmetry. The method to determine the edge weights is described in Section II C. The network is initially fully connected (i.e. every node is connected to every other node), however once the edge weights are calculated, any edge with a weight less than 0.001 is removed from the network. A weight of 0.001 means there is only a one in 1000 chance of one `Person` being moved along this edge at each epoch. Removing these edges reduces the computational load by removing connections that are unlikely to contribute to any significant population exchange over the timescales of interest.

c Disease The network is supplied with a `Disease` object that defines how quickly the disease will spread and the associated impact it will have on morbidity and mortality.

The force of infection determines how easily a given infected `Person` transmits the disease to a susceptible `Person` within the same interaction zone. The object has a separate force of infection for interactions occurring within a household vs in a Mingling Zone, with the household value usually being higher [19].

The symptom development rate controls the length of time a `Person` remains susceptible before becoming infected and is synonymous with the incubation period of a disease. Similarly, the recovery rate parameter controls the length of time a `Person` remains infectious after becoming infected. When a `Person` is no longer infectious they either recover or die. The likelihood of an outcome of death is controlled by a death probability, which can be specified per age bracket. This process is explained in more detail in Section II E.

The `Disease` object also contains information about the proportion of infectious people that are asymptomatic and, therefore, unaware that they are infectious.

d Population The population of the region to be modelled is represented in the network by a number of `Person` objects. Each `Person` is initially assigned to a node and a household within that node, which together define the home location for that `Person`.

A `Person` is able to travel to any other node directly connected to their home node; they may also leave their household to go their home node's Mingling Zone. `Person` objects in the Mingling Zone are hereafter referred to as minglers. Minglers are comprised of visitors from adjoining nodes and a random sample of 'locals' who are removed from their households and allowed to interact with the visitors. The number of local people in the Mingling Zone of each node is controlled by the mixing proportion parameter and is a fixed number per node.

Section II C explains how we determine the overall movement between each node, whilst Section II F explains how we determine which individual `Person` objects are chosen to move. Each `Person` keeps track of their current location and their home location as they move around the network. Information about their age bracket and disease status (see Section II E) are also stored.

C. Mobility Calculations

The data supplied by `Population247` shows the distribution of people in the age ranges `0-4, 5-9, 10-15, 16-17, 18-64` and `65+` at `0200` hours and `1400` hours. To determine internal mobility, we need a way to translate the distribution of people at these times into flows. By flows we mean the number of people that move from an origin to a destination. We calculate the flows using the doubly constrained version of Wilson's gravity model (see [20–23] for more details).

For the rest of this paper, we shall talk about origins and destinations—this is a way of distinguishing the two time slices, where an origin represents a 200m square at `0200` hours and a destination is a 200m square (potentially the same one) at `1400` hours. Our use of a doubly constrained model assumes that the number of people that move from an origin to a destination is proportional to the product of their populations and decays exponentially with the distance between them. This model has an extra constraint which specifies the total amount of movement allowed in the system. We estimate this parameter as the product of the total destination population and average travel distance, which we take as 5400m [24].

The mobility network is based on an implementation described in [25], where a total distance parameter is specified by the user. The implementation takes as inputs the origin and destination populations for each age bracket, and the Euclidean distances between each origin and destination pair. The first two inputs amount to a repackaging of `Population247` data after re-scaling the total origin population to ensure it is equal to that of the destination. This is needed because the model requires a closed system where the total population of the system is conserved (see Section IV). The third input, namely the distances between the nodes, is used to specify the connectivity of the network. Flows are only calculated for the origin-destination pairs passed to the model; excluding an origin-destination pair effectively sets the flow between those locations to zero.

This mobility network provides us with information about the number of people that travel between each origin-destination pair. The values obtained from the network, for example the edges' weights, are then used to constrain the flow of the populations of the nodes.

The framework outlined in this paper is agnostic to the source of mobility data. It would be possible, there-

fore, to connect nodes based on empirical data, instead of estimating flows. For instance, known public transport routes could be used to inform the edge weights between nodes, or samples gathered from mobile phone data could be used to identify nodes with a highly-populated Mingling Zone. Ancillary datasets, such as distance travelled to work statistics [26], may also provide a stronger indication of average journey lengths when calibrating the network weights. It would also be more realistic to decouple journeys into those that are likely to involve mingling (for example, a bus journey) from those that are likely to be shielded (for example, a car journey). Other modifications and theoretical improvements will be discussed in Section IV.

D. Immobilisation

A limitation of the entropy maximisation model is that, without proper tuning of the input parameters, it can result in an unexpected increase of population movement, even when the `0200` hours and `1400` hours distributions are very similar. To address this behaviour we can immobilise a portion of the population. Consider a node with `0200` hours and `1400` hours populations of 10 and 15 respectively. To implement immobilisation, we assumed that of the 15 people present at `1400` hours, 10 of them were the people present at `0200` hours. Therefore we only require 5 people to move into the node. This means we transform the populations into 0 and 5. The case where `0200` hours has a greater population is treated similarly. This approach makes the assumption that populations are not exchanged between morning and afternoon.

E. Susceptible-Exposed-Infected-Recovered-Dead Model

The epidemiological model used to calculate the disease evolution is a discrete component of the wider framework. The `SEIRD` class applies this epidemiological model to a list of `Person` objects, by using `Disease` attributes including the forces of infection (home and away), the symptom development rate, the removal rate, the proportion asymptomatic, and the likelihood of death (which may depend on the age bracket).

The population is partitioned into five compartments. The *Susceptible* compartment represents people who have not yet caught the disease and who have no immunity. The model is initialised with a near-fully susceptible population, with a very small number of individuals who have been exposed to the disease[1].

[1] A single exposed individual often failed to elicit epidemic behaviour; seeding with two exposed individuals produced more consistent results.

We have defined the *Exposed* compartment as representing the part of the population who have caught the disease but have not yet exhibited symptoms. This definition differs from canonical compartment literature where exposed individuals are typically unable to infect susceptible population. For the purposes of this paper, it has been useful to treat an *Exposed* individual as an as-yet asymptomatic (but infectious) carrier of the disease. The reaction equation linking *Susceptible* individuals to *Exposed* is proportional to the to the size of each compartment and the appropriate force of infection parameter that governs how quickly this transition occurs. Typically, the force of infection is higher at home than away, due to the potentially prolonged contact time between residents.

After a period of time, defined by the symptom development rate, *Exposed* individuals are deemed to be *Infected*. The proportion of the infected individuals who still experience no symptoms is given by the proportion asymptomatic parameter. Whilst both *Exposed* and *Infected* individuals may infect the *Susceptible* population, these compartments have been decoupled to allow different properties to be provided for each—for example, exposed people may continue to interact with the wider population, whereas any infected individuals who experience symptoms may self-isolate or quarantine themselves. To summarise the difference between the *Exposed* and *Infected* compartments: all exposed people will continue to interact with the wider population as normal, whereas some infected people may express symptoms and preferentially remain at home.

After a period of time, defined by the removal rate, infected individuals are deemed to be removed from the infection model. This is because they may either be *Recovered* and assumed to be immune, or *Dead*. The fraction that governs which compartment infected individuals terminate in is defined by the likelihood of death, which may vary depending on the age bracket of the individual. In reality, this will also depend on other factors, such as hospital capacity or presence of comorbidities, which are not accounted for in this treatment. The full dynamical system of equations is supplied in Equations 1(a)–(e):

$$\frac{dS}{dt} = -\beta\frac{(E+I)S}{N}, \tag{1a}$$

$$\frac{dE}{dt} = \beta\frac{(E+I)S}{N} - \alpha E, \tag{1b}$$

$$\frac{dI}{dt} = \alpha E - \gamma I, \tag{1c}$$

$$\frac{dR}{dt} = (1-\delta_a)\times\gamma I, \tag{1d}$$

$$\frac{dD}{dt} = \delta_a\times\gamma I, \tag{1e}$$

where S, E, I, R, D are the numbers of *Susceptible*, *Exposed*, *Infected*, *Recovered* and *Dead* individuals in a total population of N people. The parameters β, α and γ are the force of infection, symptom development rate and removal rate, respectively. The likelihood of death

is represented by the parameter, δ_a, which may simply be a scalar probability between zero and one. On the other hand, if the likelihood of death depends on the age bracket, then the *Infected* compartment in Equation (1c) also needs to be vectorised; that is to say, the infected individuals from each age bracket need to be treated separately. The model permits both of these possibilities, extending automatically depending on whether δ_a is a singleton list or contains at least two elements.

The vanilla model described here does not take into consideration deaths that are not related to `COVID-19` or births. Since deaths involving `COVID-19` are mostly amongst elderly individuals, we would expect an appreciable number of deaths to occur anyway. For this reason, we recommend the model is not run for more than 200 epochs without including an age-dependent shortcut from Equation (1a) to Equation (1e) as well as supplementing a general birth rate into Equation (1a).

F. `InfectioNet` Model Process

Combining the `SEIRD` model, outlined in Section II E, with the mobility network results in a system we call `InfectioNet`. The model operates by stepping through consecutive epochs. Each epoch (analogous to one day) consists of two time-steps. In the first time-step, all `Person` objects are at home and only interact with members of the same household. In the second time-step, a selection of `Person` objects are moved away from their house and into either a) a Mingling Zone in another node as determined by the edge weights or b) a Mingling Zone within their local node as determined by the mixing proportion. In the current model there is only one Mingling Zone per node, where all `Person` objects who are not at home interact.

At every time-step, `Person` objects in the same space (either a household or a Mingling Zone) interact and may transmit the disease. The number of new cases is determined individually for each interaction space based on the `SEIRD` model in Section II E.

In the current model, a full epoch proceeds as follows:

1. For every household in the network, a number of susceptible `Person` objects becomes exposed based on the at-home force of infection, the number of exposed and/or infected `Person` objects coinhabiting, and the total household population (see Equation. 1). [`0200` hours Update]

2. Random samples of each node's residential population are selected to move along its outgoing edges to congregate in the respective Mingling Zone of adjacent nodes (one space per node). For example, if an edge has a weight of 3 in the `18-64` age bracket, then 3 `Person` objects in this age bracket will be randomly selected to move along this edge. If the weight were 3.2, then there would be an 80 % chance of 3 people moving and a 20 % chance of 4 people moving.

3. For every household and Mingling Zone in the network, a number of susceptible `Person` objects become exposed based on the at-home or outside-home forces of infection, the number of exposed and/or infected `Person` objects in each space and the total population in each space (see Equations. 1). [`1400` hours Update]

4. All `Person` objects are sent home.

In order to investigate the effects of mobility on the propagation of a disease, `InfectioNet` should be run multiple times with various network infrastructures, but with fixed disease parameters. In our own implementations, we decided to calibrate two forces of infection (one for at-home interactions and one for Mingling Zone interactions) based on a pre-lockdown mobility network, with a target doubling-time of three days. Subsequent lockdown models could then be investigated using the same forces of infection. Details of the calibration method are given in Section II G below.

G. Disease Calibration

To ensure that the disease propagates on a realistic timescale, a calibration scheme has been devised which attempts to tune the at-home and mingling forces of infection based on an initial doubling time. The two user-specified parameters are the initial doubling time—the number of epochs it should take for the seeded infected population to double—and an at-home weighting factor—how many times more transmissible at-home interactions are in comparison to mingling interactions. The calibration starts by estimating the average number of people in an interaction space in one full epoch, $\bar{N}_{Total}$. This is calculated as a weighted average of at-home, $\bar{N}_H$, and mingling, $\bar{N}_M$, interactions:

$$\bar{N}_{Total} = \alpha_M \bar{N}_M + \alpha_H \bar{N}_H.$$

In this case, α_M and α_H parameterise the fraction of interactions that happen in Mingling Zones and at-home respectively, such that

$$\alpha_M + \alpha_H = 1.$$

This implies an average susceptible population (for one infected `Person`) in mingling spaces of

$$\bar{S}_M = \bar{N}_M - 1$$

and an average susceptible population in at-home spaces of

$$\bar{S}_H = \bar{N}_H - 1$$

then taking $\bar{S} = \bar{S}_M + \bar{S}_H$, which can be substituted into Equations. (1) from the `SEIRD` model to estimate the average number of disease transmissions per epoch as

$$d\bar{S} = -\beta_M \alpha_M I \frac{\bar{N}_M - 1}{\bar{N}_M} - \beta_H \alpha_H I \frac{\bar{N}_H - 1}{\bar{N}_H}. \quad (2)$$

For simplicity, the at-home force of infection, β_H, and the mingling force of infection, β_M, are assumed to be proportional:

$$\beta_H = \omega \beta_M,$$

where ω is a user-defined proportionality constant. Equation (2) can now be written

$$d\bar{S} = -I\beta_M \alpha_M \frac{\bar{N}_M - 1}{\bar{N}_M} - I\omega\beta_M \alpha_H \frac{\bar{N}_H - 1}{\bar{N}_H},$$

$$d\bar{S} = -I\beta_M \Big(\alpha_M \frac{\bar{N}_M - 1}{\bar{N}_M} + \omega\alpha_H \frac{\bar{N}_H - 1}{\bar{N}_H}\Big),$$

$$C = \Big(\alpha_M \frac{\bar{N}_M - 1}{\bar{N}_M} + \omega\alpha_H \frac{\bar{N}_H - 1}{\bar{N}_H}\Big)$$

$$d\bar{S} = -I\beta_M \times C.$$

Assuming an initial infected population of I_0, the infected population after n epochs is:

$$\begin{aligned} I_0 &= I_0 \\ I_1 &= I_0 + \beta_M C \\ I_2 &= I_1 + I_1 \beta_M C \\ &\vdots \\ I_n &= I_0 \Big(\beta_M C + 1\Big)^n. \end{aligned}$$

After the specified doubling time, n_d, the initial infected population should have doubled, such that

$$I_{n_d} = 2I_0 = I_0 \Big(\beta_M C + 1\Big)^{n_d}. \quad (3)$$

Rearranging Equation (3) for the mingling force of infection, we get

$$\beta_M = \frac{\beta_H}{\omega} = \frac{\sqrt[n_d]{2} - 1}{C},$$

which yields the two forces of infection needed as inputs to the `InfectioNet` model.

One of the principal assumptions that is made during this calibration scheme is that the susceptible population remains constant. In reality, the susceptible population is depleted every epoch and, therefore, the doubling rate in the `InfectioNet` model tends to decrease. On top of this, the `Person` objects that are moved throughout the network are chosen randomly each epoch, which makes precise tuning of the force of infection difficult.

H. `EvolveNet` model process

In addition to the standard `InfectioNet` model, which uses a fixed mobility network, we experimented with dynamically changing the edge weights. This was to compare the difference between an unmitigated policy and one in which lockdown is enacted on the spread of a disease. We accomplished this with an augmented version of `InfectioNet`, that we call `EvolveNet`, which has the ability to change the mobility network that `InfectioNet` is using after a predetermined number of epochs. Other than the change of network, the infection propagates as described in Section II F.

III. Examples

In the following sections, we demonstrate two possible configurations of `InfectioNet` and explore the effects of different mobility networks on disease propagation. First, we investigate disease propagation in a region containing a highly-visited central hub. Then we demonstrate `EvolveNet` by dynamically reducing mobility once a virus is established.

A. Transmission via Central Hub

To test the influence of our mobility networks on the simulated propagation of COVID-19, we look at two contrasting scenarios in Southampton, UK. In one case, peoples movements are primarily within their own neighbourhood. In the second case, peoples movements are biased towards Southampton General Hospital, drawing visitors from across the city.

Both mobility networks were calibrated using the methodology described in Section II C. In each case, the total distance parameter was determined based on an average travel distance of 5400m per person. For the neighbourhood scenario, the population distribution was given by the pre-lockdown `Population247` data within Southampton city extent, whilst the hub scenario used a lockdown distribution of the population, because it included a significant hotspot for Southampton General Hospital at `1400` hours. For the purposes of constructing these networks, any `0200` hours populations of less than 10 were disregarded. Figures 2(a) & (b) show a visualisation of the mobility network in the neighbourhood and hub scenarios, respectively.

Using these networks, we initialised two `InfectioNet` models, with the parameters given in Table I. The force of infection parameters were calibrated from the neighbourhood model, using the method given in Section II G, and applied to both models. In this case, the target doubling time, n_d, was set to 3 days and the proportionality constant, ω, was set as 2.

Both models were seeded by exposing 2 `Person` objects in the North (one in the Central-North and the other in

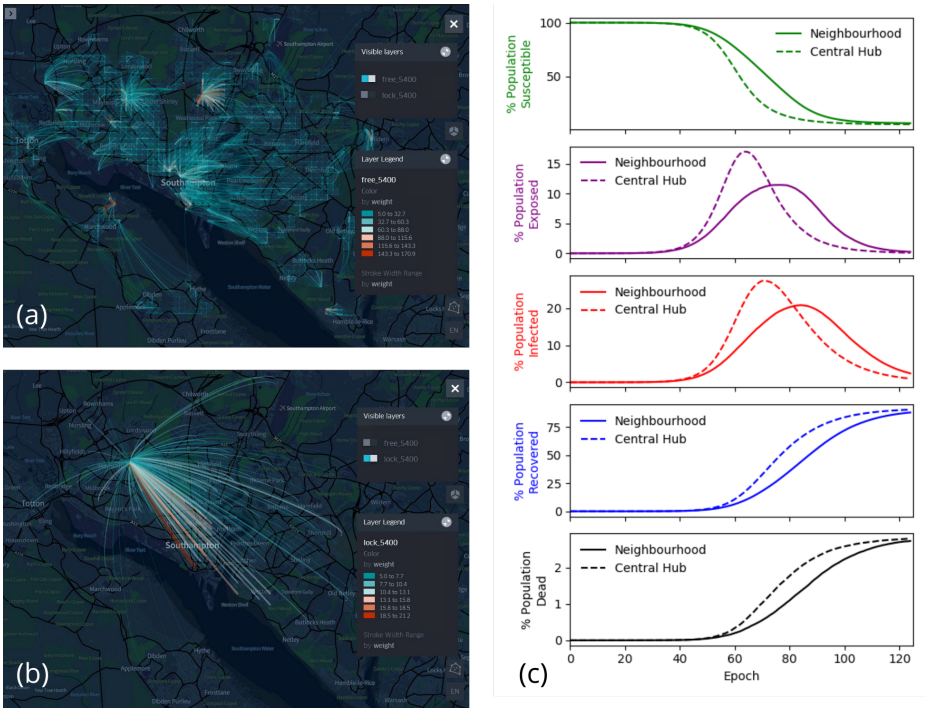

Figure 2. (a) A visualisation of the mobility network used in the neighbourhood model, Southampton, UK. (b) A visualisation of the mobility network used in the central hub model, Southampton, UK. In both cases, connections between origin-destination pairs are depicted as coloured arcs, with red arcs corresponding to pairs with higher weight values, indicating a strong connection and increased population flow. Note: the scales are not identical between the two figures. (c) A comparative line plot showing the evolution of each of the `SEIRD` disease compartments in the neighbourhood and central hub models.

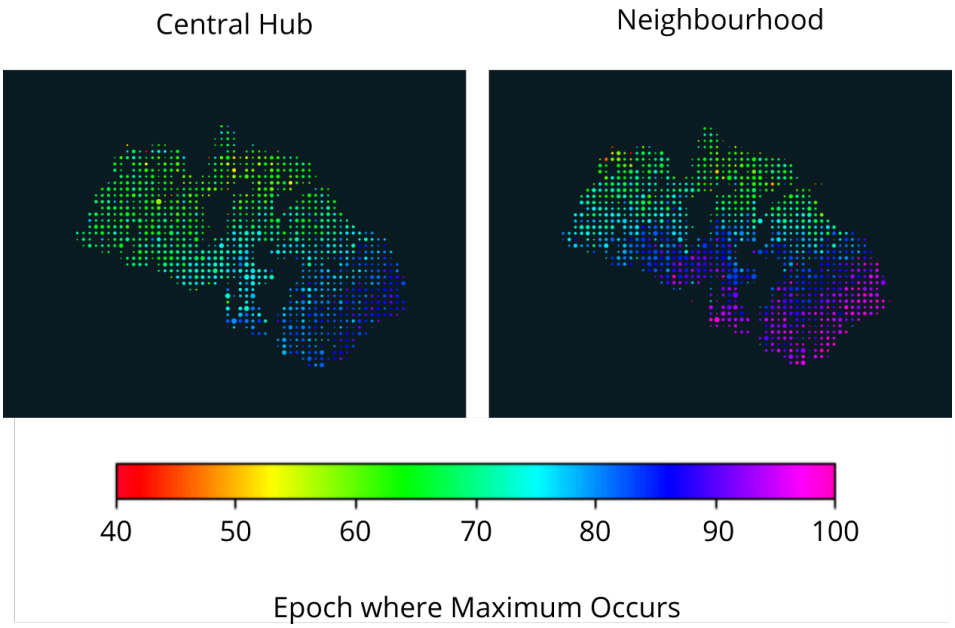

Figure 3. A comparison of the geographical disease propagation in the central hub and neighbourhood models. Notice the slight change in orientation and speed of the disease spread, despite identical infection seeding.

Table I. Disease parameters used for the neighbourhood and central hub `InfectioNet` models, as well as the restricted and unrestricted `EvolveNet` models.

Force of Infection (Mingling Zone)	**0.19**
Force of Infection (Home)	**0.38**
Symptom Development Rate	**0.2**
Proportion Asymptomatic	**0.2**
Recovery Rate	**0.1**
Probability of Death	**0.03**
Mixing Proportion	**0.2**

the North-West), and then the simulation was run for 200 epochs. The results are shown in Figure 2(c).

The development of the `SEIRD` model, in this case, is clearly affected by differences in the underlying mobility networks. In the central hub model, the infection peaks at epoch 71, with 27.4% of people infected. In comparison, the neighbourhood model peaks at epoch 84, with 20.8% of people infected. This suggests that the presence of a highly-visited central hub may accelerate the spread of infection.

The geographical evolution of the disease is shown in Figure 3, where the timing of the infection peak at each location is assigned a colour. In the central hub model, the disease spreads from North-West to South-East, primarily between epochs 55 and 85. In the neighbourhood model, the disease spreads from North to South, primarily between epochs 55 and 100. It can be seen from Figure 2(b) that the infection, in the central hub model, propagates in the direction of the strongest connections. On the other hand, in the neighbourhood model, the infection propagates outwards from the seed in this case, since the seeding occurred in the North, the wave is restricted to travelling Southwards.

B. Applying Mobility Restrictions

To test the effect of applying movement restrictions once a virus is established, we take two networks, which represent unrestricted and restricted population mobility in Gosport, UK, and combine them using an `EvolveNet` model.

The unrestricted network was constructed in the same way as the neighbourhood network in Section III A, but for the regional extent of Gosport. To produce the restricted network, the average travel distance per person was reduced to 2500m and maximal immobilisation was enforced. The resultant networks are shown in Figures 4(a) & 4(b).

In the unrestricted mobility network, there are a total of 451589 connections, whereas, in the restricted mobility network there are only 97569. Furthermore, the total number of people leaving their home node reduces from 102380 to 12379 once restrictions are applied this is a direct consequence of maximal immobilisation. Another effect of imposing stricter mobility restraints, is that longer journeys are inhibited. For example, the North-South connections across the estuary, seen in Figure 4(a), are not present in the restricted network shown in Figure 4(b).

Using these networks, we initialised an `EvolveNet` model, with the parameters given in Table I, and seeded the infection by exposing 4 people in the South. This model was run twice. In the first instance, `EvolveNet` was run for 45 epochs using the unrestricted network before dynamically switching to the restricted network. As a comparison, the same `EvolveNet` model was run with the same random seed, but where the mobility network remained fixed as the unrestricted case.

Figure 4(c) shows the impact of the mobility restrictions on disease evolution. We can see that the growth of the exposed class was immediately suppressed once restrictions were applied at epoch 45. Following these restrictions, the infection peak—and therefore the number of deaths—was reduced by approximately 50%. This also resulted in a remaining susceptible population of approximately 60000—ten times more than when mobility was left unrestricted.

IV. Discussion

The network calibration, discussed in Section II C, requires the origin and destination populations to be equal. Physically, this means that no one commutes into or out of the region being calibrated. In the case of the examples, this assumption is partially justified by the fact that Southampton and Gosport had similar `0200` hours and `1400` hours populations—only varying by up to 10%. This could be coincidental, since it is possible that some proportion of the population leaves and an equal number enters. This method will, therefore, need to be extended to be able to take account of regions which are known to have a large commuter population and/or a significant change in their `0200` hours and `1400` hours populations, such as London.

Calibrating a network over a larger region can reduce the discrepancy between `0200` hours and `1400` hours. As the size of the region is increased, a larger proportion of people are likely to travel within the region. Consider, for example, the case where one network models a region containing a block of flats, compared to one that models the UK. The first is likely to have a significant difference between `0200` hours and `1400` hours populations since many people may work outside of their flat. The network for the UK will not have this issue because relatively few people enter or leave the UK compared to its total population.

The network calibration can also suffer from edge effects. Consider a person that lives very close to the edge of a network's extent, but whose workplace is outside the network boundary. The network calibration would have no knowledge of this fact and is restricted to mov-

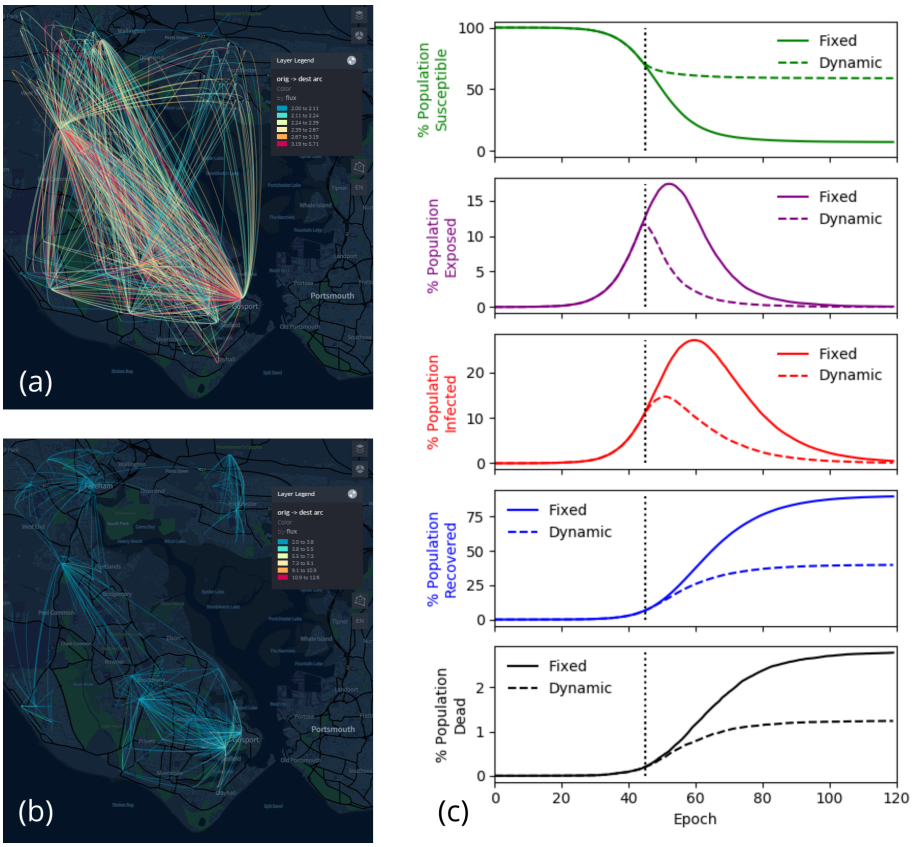

Figure 4. (a) A visualisation of the unrestricted mobility network in Gosport, UK. (b) A visualisation of the restricted mobility network in Gosport, UK. In both cases, connections between origin-destination pairs are depicted as coloured arcs, with red arcs corresponding to pairs with higher weight values, indicating a strong connection and increased population flow. Note, the scales are not identical between the two figures. (c) A comparative line plot showing the evolution of each of the `SEIRD` disease compartments when restrictions are and are not applied at epoch 45.

ing this person within the boundary. One can imagine that this effect is similar for every person who lives close to the edge, meaning that the network weights for these regions are likely to be inaccurate. To remedy this, it is advisable to select a larger region than desired, run the network calibration and then remove the additional edges around the border.

Section II D details a method for maximal immobilisation. If this is too severe, it may be appropriate to implement a more relaxed policy. For example, if 80% immobilisation is applied to `0200` hours and `1400` hours populations of 10 and 15, they would be restricted to 2 and 7, not 0 and 5 as in the case of maximal immobilisation.

Section II C explains that the network calibration finds a configuration of edges that maximise the entropy, given a total distance constraint. This is done via simulated annealing, which is not guaranteed to converge to the global maximum. It is, therefore, possible for multiple calibrations to result in different network configurations for the same population distributions.

Improvements to the mobility networks could be achieved by accounting for travel times between nodes in the network. This could be achieved by using isochrone routing along the road networks, but by taking into consideration the availability of public transport. Natural obstacles, such as rivers, would then be navigated more realistically and destination hubs that have good transport links would be more accessible throughout the network.

`InfectioNet` is a hybrid between agent based and purely deterministic epidemiological modelling: the `Person` objects act as agents and the independent `SEIRD` models within each interaction zone provide the deterministic element. As a result, the model is quite computationally intensive, with the number of edges being the biggest influence on performance. This is because each edge requires a decision about which of the origin node `Person` objects should be moved. We attempt to address this by removing edges with very small weights, where the likelihood of meaningful exchange between nodes is very small. However, a large number of edges remain; for example, after the removal of all edges with a weight less than 0.001, the networks presented in Section III A still had over two million edges. There is potential to drastically reduce processing time with parallelisation and/or

using high performance computing resources.

The number of edges is influenced by the spatial structure of the mobility network. Since the mobility network is initially fully connected, adding a single node will add an extra edge for every node already present. This rapid growth of edges requires that the choice of spatial geometry for the nodes is balanced with computational resources. Many small nodes will dramatically increase the number of edges and require many more independent `SEIRD` models, but could potentially produce more detailed results; indeed one use of `InfectioNet` might be to have each node as a single household, if this could be supported with appropriate computing infrastructure.

One possible way to reduce the number of edges is to define areas that are likely to have similar characteristics and group them together to be treated as a single node. We tried this using LSOA boundaries for aggregating the `Population247` gridded data, which allowed us to model a much larger region than would otherwise have been feasible. A similar aggregation to Middle Layer Super Output Area (MSOA) or county-level geometries could be used for national-scale modelling. We could also, for example, try using a single `Person` to represent 10 people and so reduce the data load on the model. However, preliminary tests suggest a reduction in the number of `Person` objects is likely to have a smaller impact than reducing the number of edges.

The `Population247` data we used in Section III is itself the result of a model that combines multiple data sources and assumptions about the movement of different sub-populations over the course of a day. To avoid multiple layers of modelled data, there are a range of empirical sources that could be used to provide population and mobility data to underpin `InfectioNet`, for example mobile phone data.

One of the limitations of the `InfectioNet` infrastructure is that each epoch can only be divided into two time-steps—this cannot currently be increased (for example, to four time-steps per epoch). Additional time-steps would require multiple network calibrations for various population distributions per epoch and for these networks to be updated dynamically, like in `EvolveNet`. This would be computationally expensive, but result in a more realistic pattern of movement as people will be able to travel to (and through) multiple locations in a day—including night-time interactions where, in the current model, everyone is assumed to be at home.

Furthermore, the interactions that do happen at home may be modelled more accurately if realistic household compositions are considered. This could include, for example, the number of people in each house, or the age brackets of its constituents. In the current implementation of `InfectioNet`, houses are populated without significant variation in size, for example a node with five houses and a population of ten will simply distribute two people to each house.

Interactions that occur outside of the home happen in the Mingling Zone, of which there is only one per node. In reality, peoples' interactions are likely to occur in multiple, partitioned mingling spaces such as shops or office blocks. OS's AddressBasePlus contains this information and could be added to the nodes in `InfectioNet` as attributes, e.g. number of shops, number of offices etc. There are also different types of interaction within a space, which may depend on demographics, such as age and/or the activities undertaken within it.

The only disease-affecting characteristic of the `Person` objects in the current model is age, which can be used to augment the death-probability for older members of the population. An extension could be to attribute `Person` objects with other characteristics, such as sex, ethnic group, deprivation and pre-existing health conditions. These additional characteristics could be used to alter various disease parameters (such as death-probability), if they are determined to be correlative.

These characteristics, including age, might also be useful in the `SEIRD` model to inform who interacts with whom, for example teenagers may primarily interact with people of their own age. It may constitute an area of fruitful further investigation to model the effects of different age-specific interventions, for example what is the effect of opening schools compared to opening pubs. The contact matrix of each age group may radically change depending on the intervention. For instance, if children returned to school, more adults may return to office work, thus increasing the average age of adult social interactions.

Similarly, geographical or environmental characteristics could be attributed to each node in order to further inform disease parameters. For example, air quality or distance to the nearest hospital could be used to adjust the death-probability, or the existence of a supermarket could be used to adjust the force of infection.

A relatively simple extension of the `InfectioNet` `SEIRD` model could be to add more compartments. These could include an explicit asymptomatic compartment, as well as an immunity compartment that is short-term and results in `Person` objects becoming re-susceptible.

Finally, an important aspect of this modelling that has not yet been addressed is validation. This would likely involve a Monte-Carlo approach, where the infection is seeded in multiple locations to establish a typical behaviour (if one emerges). The outcome from this analysis could then be directly compared with real-life statistics on the severity and location of the virus cases.

V. Conclusion

To support efforts by the scientific community to understand coronavirus dynamics, we present a framework, called `InfectioNet`, designed to facilitate the investigation of geospatial factors influencing disease spread. We hope this framework will help with targeting healthcare and community resources where they are most needed, as well as informing decisions about the way mobility re-

strictions are eased by demonstrating the impact of different options. We explain the framework in detail and show some examples of how it can be applied. Here we focused on `COVID-19`, but the framework is applicable to any transmissible disease.

Using a network to represent the geographical interactions within and between populations in nodes, combined with a Susceptible-Exposed-Infected-Removed (`SEIRD`) model, we show the effect of changing population mobility has on the transmission behaviour of the virus. This can vary both the pattern of geographic spread and the overall impact of the disease.

We demonstrate this using an example in Southampton, UK, where the population mobility pattern is shown to affect the spread of a disease. With a locally connected mobility network, the infection peak is lower and occurs over a longer duration than when a highly-visited central hub is present. The accelerated growth, in the central hub model, occurs because people travel longer distances to locations with lots of visitors, supporting advice that we should aim to keep travel as local as possible to reduce virus transmission.

We also present an example in Gosport, UK, using `EvolveNet`, that models a restriction being applied to mobility once the virus has taken hold (analogous to imposing 'lockdown' conditions). This results in a reduction of over 50% in the overall number of infections and therefore deaths.

There are still many things to explore using this framework and a huge range of potential applications and extensions. We have suggested some ideas in Section IV and hope that others will use, develop and adapt `InfectioNet` and `EvolveNet` in order to deepen our understanding of how geospatial factors influence disease transmission and impact.

To enquire about access to the code please contact Jacob Rainbow at Ordnance Survey (`jacob.rainbow@os.uk`).

Acknowledgments

We would like to acknowledge the support received from the Propositions & Innovation team at Ordnance Survey, as well as advice from the Technology and Design team and GeoProduction services. In particular, we would like to thank Iain Goodwin, Andrew Cooling, Jeremy Morley, Ridwan Barbhuiyan, Paul Naylor, Isabel Sargent, Mark Tabor and Paul Cruddace for their support during this work. Your input has been valuable for this paper.

[1] N. M. Ferguson and D. Laydon et al., Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand., Imperial College London (16-03-2020) 10.25561/77482 (2020).
[2] Office for National Statistics, 2011 Census: Out of term population of England and Wales - an alternative 2011 Census population base (Dataset) (2014).
[3] Office for National Statistics, Age by single year (Dataset) (2011).
[4] Office for National Statistics, Student accommodation by age (Dataset) (2011).
[5] Office for National Statistics, Communal establishment management and type - people (Dataset) (2011).
[6] Office for National Statistics, Sex by single year of age (workplace population) (Dataset) (2011).
[7] Office for National Statistics, Employment status (Dataset) (2011).
[8] National Health Service Digital, Hospital Episode Statistics Database (Dataset) (Dataset) (2019).
[9] Office for National Statistics, 2011 Census: Out of term population of England and Wales - an alternative 2011 Census population base (Dataset) (2014).
[10] Department for Education, Schools, pupils and their characteristics: January 2011 (Dataset) (2011).
[11] Visit Britain, England Research and Insights (Dataset) (2016).
[12] Office for National Statistics, Population estimates - small area based by single year of age - England and Wales (Dataset) (2020).
[13] Care Quality Commission, CQC care directory - with ratings (4 May 2020) (Dataset) (2020).
[14] H. P. S. Ministry of Justice, H. M. Prison, and P. Service, Prison population figures: 2019 (Dataset) (2019).
[15] Ordnance Survey Ltd., OS Open Map - Local, Road Network (2011).
[16] Office for National Statistics, Region (December 2015) boundaries, clipped to mean high water mark (Dataset) (2011).
[17] Ordnance Survey Ltd., OS Open Map - local, surface water (2011).
[18] Department for Transport, Annual average daily flows (2011).
[19] S. Y. Park, Y.-M. Kim, S. Yi, S. Lee, B.-J. Na, C. B. Kim, J.-I. Kim, H. S. Kim, Y. B. Kim, Y. Park, *et al.*, Coronavirus Disease Outbreak in Call Center, South Korea., Emerging Infectious Diseases **26** (2020).
[20] A. G. Wilson, A statistical theory of spatial distribution models, Transportation Research **1**, 253 (1967).
[21] A. G. Wilson, *Entropy in urban and regional modelling* (London : Pion, 1970).
[22] A. G. Wilson, A family of spatial interaction models, and associated developments, Environment and Planning A: Economy and Space **3**, 1 (1971), https://doi.org/10.1068/a030001.
[23] A. G. Wilson, *Entropy in urban and regional modelling*, Vol. 1 (Routledge, 2011).
[24] Future Foundation, Chapter 4 - shopping miles (2007).
[25] K. Harland, *Journey to learn: Geographical mobility and education provision*, Ph.D. thesis, University of Leeds (2008).
[26] Office for National Statistics, Distance travelled to work by sex by age (Dataset) (2011).