

Journal Pre-proof

A dynamic microsimulation model for epidemics

Fiona Spooner, Jesse F. Abrams, Karyn Morrissey, Gavin Shaddick, Michael Batty, Richard Milton, Adam Dennett, Nik Lomax, Nick Malleon, Natalie Nelissen, Alex Coleman, Jamil Nur, Ying Jin, Rory Greig, Charlie Shenton, Mark Birkin



PII: S0277-9536(21)00793-0

DOI: <https://doi.org/10.1016/j.socscimed.2021.114461>

Reference: SSM 114461

To appear in: *Social Science & Medicine*

Received Date: 19 May 2021

Revised Date: 25 August 2021

Accepted Date: 5 October 2021

Please cite this article as: Spooner, F., Abrams, J.F., Morrissey, K., Shaddick, G., Batty, M., Milton, R., Dennett, A., Lomax, N., Malleon, N., Nelissen, N., Coleman, A., Nur, J., Jin, Y., Greig, R., Shenton, C., Birkin, M., A dynamic microsimulation model for epidemics, *Social Science & Medicine* (2021), doi: <https://doi.org/10.1016/j.socscimed.2021.114461>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier Ltd.

Credit Author Statement

Fiona Spooner	Methodology; Investigation; Software; Writing, reviewing and editing
Jesse Abrams	Methodology; Investigation; Software; Writing, reviewing and editing
Karyn Morrissey	Methodology; Investigation; Writing, reviewing and editing
Gavin Shaddick	Methodology; Writing, reviewing and editing
Michael Batty	Methodology; Writing, reviewing and editing
Richard Milton	Methodology; Investigation; Software; Writing, reviewing and editing
Adam Dennett	Methodology; Writing, reviewing and editing
Nik Lomax	Methodology; Writing, reviewing and editing
Nick Malleson	Methodology; Investigation; Software; Writing, reviewing and editing
Natalie Nelissen	Investigation; Software
Alex Coleman	Software
Jamil Nur	Methodology; Writing, reviewing and editing
Ying Jin	Methodology; Writing, reviewing and editing
Rory Greig	Methodology; Software; Writing, reviewing and editing
Charlie Shenton	Methodology; Software
Mark Birkin	Methodology; Writing, reviewing and editing

Title: A Dynamic Microsimulation Model for Epidemics

Fiona Spooner^{a,b,1}, Jesse F. Abrams^{b,c,1}, Karyn Morrissey^a, Gavin Shaddick^{c,g}, Michael Batty^d, Richard Milton^d, Adam Dennett^d, Nik Lomax^{e,g}, Nick Malleson^{e,g}, Natalie Nelissen^{e,g}, Alex Coleman^h, Jamil Nur^f, Ying Jin^f, Rory Greigⁱ, Charlie Shentonⁱ, Mark Birkin^{e,g*}

¹ Joint first authors

* Corresponding Author: m.h.birkin@leeds.ac.uk

^a European Centre for Environment and Human Health, University of Exeter, Knowledge Spa, Truro, UK

^b Institute for Data Science and Artificial Intelligence & Global Systems Institute, University of Exeter, UK

^c Joint Centre for Excellence in Environmental Intelligence, Exeter, UK

^d Bartlett Centre for Advanced Spatial Analysis, University College London, London, UK

^e School of Geography and Leeds Institute for Data Analytics, University of Leeds, Leeds, UK

^f Martin Centre for Architectural and Urban Studies, University of Cambridge, 1 Scroope Terrace, Cambridge, UK

^g Alan Turing Institute, London, UK

^h Research Computing, University of Leeds, Leeds, UK

ⁱ Improbable, London, UK

A Dynamic Microsimulation Model for Epidemics

A large evidence base demonstrates that the outcomes of COVID-19 and national and local interventions are not distributed equally across different communities. The need to inform policies and mitigation measures aimed at reducing the spread of COVID-19 highlights the need to understand the complex links between our daily activities and COVID-19 transmission that reflect the characteristics of British society. As a result of a partnership between academic and private sector researchers, we introduce a novel data driven modelling framework together with a computationally efficient approach to running complex simulation models of this type. We demonstrate the power and spatial flexibility of the framework to assess the effects of different interventions in a case study where the effects of the first UK national lockdown are estimated for the county of Devon. Here we find that an earlier lockdown is estimated to result in a lower peak in COVID-19 cases and 47% fewer infections overall during the initial COVID-19 outbreak. The framework we outline here will be crucial in gaining a greater understanding of the effects of policy interventions in different areas and within different populations.

Keywords: Coronavirus, COVID-19, microsimulation, SEIR, spatial-interaction, dynamics

1. Introduction

Across the world, governments have introduced non-pharmaceutical interventions (NPI) to try and control the spread of COVID-19 through a reduction in the number of contacts between susceptible members of the population and those with the disease (Desvars-Larrive et al., 2020). Those interventions include social distancing, isolation, wearing face masks and lockdowns at national, regional and local scales. In the UK, each policy has been underpinned by much speculation surrounding its timeliness, extent and subsequent effectiveness. However, what has become clear is that pre-existing systemic health

inequalities (Daras et al., 2021, Kontopantelis et al., 2021, McNamara et al., 2020) have meant that regardless of NPI, certain communities have been disproportionately impacted in terms of COVID-19 cases, hospitalisations and mortality outcomes. There is evidence of markedly different impacts on health across various domains, including: geographical region (Kontopantelis et al., 2021); level of deprivation (Cabinet Office, 2017, Office for National Statistics, 2021); race and ethnicity (Mathur et al., 2020, Race Disparity Unit Cabinet Office, 2020). The causes behind these patterns are complex and interlinked (Bibby et al., 2020, Zhang et al., 2021). Such factors include economic circumstances whereby people in more disadvantaged communities are less able to comply with requirements to work from home due to their occupation. Additionally, some communities are less inclined to comply with restrictions due to mistrust of authorities (Daras et al., 2021, Harris, 2020, Zhang et al., 2021).

The risk factors leading to COVID-19 cases, hospitalisation, and mortality exist not only at the individual level; neighbourhood-level factors and their interactions with individual-level factors are also responsible for the observed disparities (Daras et al., 2021, KC et al., 2020). Lack of access to health care, unemployment, occupation type, level of education, and housing conditions significantly increase the risk of COVID-19 infection (Bilal et al., 2021, KC et al., 2020, Shah et al., 2020). The varying levels of vulnerability between people and places has been increasingly shown to have important consequences for individual and community responses to the pandemic (Daras et al., 2021, Harris, 2020). Given these complexities, it is increasingly clear that to understand the effectiveness of government policies we require detailed data that reflects the everyday lives of the British population.

Since the onset of the pandemic, researchers across a variety of disciplines have come together to understand the transmission of COVID-19 at the population level.

Compartmental models, specifically the Susceptible – Exposed – Infection – Removed (SEIR; Rvachev and Longini (1985)) have formed the bedrock of this research. However, with the partial exception of a number of models that allow for the effect of population age structure (Keeling et al., 2020, van Leeuwen and Sandmann, 2020) or specific behaviour changes in response to public health interventions and seasonal change (Dureau et al., 2013, Ferguson et al., 2006, Kucharski et al., 2020) through stochastic model extensions, most of this work has largely failed to embed and replicate the complex space and time dynamics that underline the spread of COVID-19 across different populations and communities within their models.

In this paper we outline an enhancement of the traditional SEIR model of infectious disease transmission through adoption of a spatial microsimulation modelling framework that brings together epidemiological modelling, urban analytics, spatial analysis and data integration. Specifically, we combine the power of well-established methods within the social and behavioural sciences, namely spatial microsimulation and spatial interaction models, within a dynamic SEIR to offer the best approximation of (i) the daily, individual-level mobilities that characterise many of the interactions which lead to COVID-19 transmission and (ii) the impact of different NPI based on the complex health, socio-economic and behavioural attributes of the British population. This framework provides the much-needed ability to assess the effects of past interventions and simulate the effects of future policy decisions on different population groups at a variety of spatial scales.

The modelling framework proposed here is based on synthetic georeferenced population which has been enriched with additional socio-economic, demographic, activity and health attributes required to understand individuals' typical mobility patterns and likelihood of being severely impacted by the disease. In each simulated day, the common daily behaviours of the synthetic individuals – currently shopping, schooling and working – are simulated and then, if they have the disease, the individuals impart a hazard to the locations that they visit. Disease-free individuals who also visit these locations receive some exposure which, when combined with their individual vulnerability, may lead to them contracting the disease themselves. The model runs for a user-defined number of simulated days and, on completion, outputs aggregate disease statistics.

The remainder of this paper is organised as follows. Section 2 describes the risk modelling framework including how hazards and exposures are estimated and integrated within a compartmental epidemiological risk model. This section also contains details on the generation of a synthetic population (Section 2.2), how health, socio-demographics, and activity information are incorporated into that population (Section 2.2.1) and how individuals are assigned to appropriate locations (e.g. school, home, work) for their activities (Section 2.3). In Section 3 the result of a case study in Devon is presented, showing the effects of the lockdown that started on 23rd March 2020 compared to those predicted if the lockdown had started a week earlier. Finally, Section 4 provides a concluding discussion and ideas for future developments and applications.

2. Methods and Materials

2.1 Disease Modelling

Compartmental models have been used widely in infectious disease epidemiology with many based on the Susceptible – Infection – Removed (SIR) model introduced by Kermack and McKendrick (1927) or the Susceptible – Exposed – Infection – Removed (SEIR) model introduced by Rvachev and Longini (1985). They have been used extensively for modelling COVID-19 with examples including Arcede et al. (2020), Dureau et al. (2013), Keeling et al. (2020), Kucharski et al. (2020), van Leeuwen and Sandmann (2020).

As an important characteristic of COVID-19 is the possibility of transmission when individuals are unknowingly infectious, i.e. in the pre-symptomatic and asymptomatic phases (Arcede et al., 2020). The SEIR model used here has a further breakdown of the infectious and removed components (Fig. 1). The additional compartments provide enhanced additional resolution in the disease status of individuals that is important to determine individual behaviour and transmission probabilities (He et al., 2020).

[Figure 1 about here.]

Individuals within the model may progress between compartments based on a probabilistic approach to determine the progression from one compartment (phase of infection) to the next (See Section 2.1.3). SEIR models have been combined with high resolution social interaction networks to explore COVID-19 transmission pathways at local scales (Aleta et al., 2020, Firth et al., 2020) and metapopulation models have been used to capture broad scale COVID transmission dynamics with an SEIR model used within each electoral ward (Danon et al., 2020). Here, a dynamic microsimulation modelling framework is used to calculate those the probabilities of transmission are calculated for each individual within a given population, based on their movements across time and space

according to their demographic and socioeconomic characteristics, and hence their exposure to the disease according to the different locations they regularly visit, i.e. shops, schools and workplaces.

The dynamic simulation framework consists of three, interlinked, components:

1. Stage 1, Hazard allocation - individuals with the disease impart *hazard* to the locations they visit. See Section 2.1.1
2. Stage 2, Risk estimation - as individuals without the disease visit different locations with increased hazards their risk of contracting the disease will increase. See Section 2.1.2
3. Stage 3, Disease status - individuals that are exposed to the disease will either contract the disease whilst those with the disease may recover. Each day, the disease status (Susceptible, Exposed, Infectious, or Removed) is updated probabilistically. See Section 2.1.3

This daily update is illustrated in Figure 2. Before simulating daily dynamics, the model estimates an initial disease status for each individual. This initialization is only performed once and, in effect, seeds the disease into the population. After this initial step, in each iteration of the model synthetic individuals spend time at some locations; current locations are their homes, shops, schools, and workplaces. If an individual is infected then they impart some of this infection risk on to the location that will then form the basis of the risk of disease for others at those locations.

[Figure 2 about here.]

2.1.1 Hazard Allocation

In each iteration the synthetic individuals spend time in four possible locations; these are currently homes, shops, schools, and workplaces. If an individual is infected then they impart some of this infection risk on to the location, denoted location *hazard*, H . The overall hazard, H , associated with a location, l is calculated by summing the individual hazards, h , imparted by each agent/individual, a , from a total population of N agents, as they visit location, l :

$$H_l = \sum_{a=0}^N h_{a,l} \quad (1)$$

If an individual, a , does not visit location l , or if they are not infected, then $h_{a,l} = 0$. If the individual is infected, then the individual hazard is proportional to the amount of time per day that the individual spends doing that activity, t , and the probability that the individual will visit that particular location l . Individuals have a probability of visiting a number of different school, work, and retail locations, so the time spent doing a particular activity is distributed among the possible locations that they might visit – denoted by p :

$$h_{a,l} = \begin{cases} 0 & \text{if } a \text{ is not infected} \\ t \cdot p & \text{if } a \text{ is infected \& symptomatic} \\ t \cdot p \cdot \mu & \text{if } a \text{ is infected \& asymptomatic} \end{cases} \quad (2)$$

Symptomatic individuals impart ‘full’ hazard on a location, while asymptomatic individuals will impart a reduced amount of hazard due to reduced transmission rates (Koh et al., 2020, Madewell et al., 2020, Qiu et al., 2021). We can scale the transmission asymptomatic individuals by using the μ parameter. If an infected, symptomatic individual spends 18

hours per day at home and 6 hours per day shopping in two possible shops, each with a 50% probability of being visited, then the individual hazard assigned to those locations from that individual are:

$$h_{shop1} = h_{shop2} = \frac{\text{proportion of time spent shopping}}{0.25} * \frac{0.5}{\text{probability of visiting the shop}} = 0.125$$

and

$$h_{home} = \frac{\text{proportion of time spent at home}}{0.75} * \frac{1.0}{\text{probability of visiting the single home location}} = 0.75$$

The derivation of time spent performing an activity (t) and the possible locations of that activity (p) are outlined in Sections 2.2 and 2.3 respectively.

2.1.2 Exposure and Risk Estimation

In the second stage of each iteration, individuals may receive some exposure to the disease based on the locations they visit. The exposure, ϵ , that an individual, a , receives per day, is the summation of the hazard, H , of all the locations that they visit, L , proportioned by the amount of time they spend there, t , and the proportion of visits to that particular location that they make, p :

$$\epsilon_a = \sum_{l=0}^L \frac{\text{total hazard at location } l}{\hat{H}_l} \frac{t_l}{\text{proportion of time spent at } l} \frac{\text{probability of visiting } l}{\hat{p}_l} \quad (3)$$

Hence if an individual spends 24 hours per day in a location that has a hazard score of 1.0, then their exposure will be 1.0.

An individual's exposure is then combined with their vulnerability (V) to give the risk (probability) of infection on that day:

$$r_a = 1.0 - e^{-(V_a * \epsilon_a) \Delta t}, \quad (4)$$

where $\Delta t = 1$ day and, for the simulations reported here, V is set to 1 for all individuals. In future work this mechanism can be used to describe which individuals are more likely to be infected.

2.1.3 Disease Status

As disease-free individuals are exposed to the disease through visiting locations with increased hazards. For any given day they will contract the disease with probability $p_a = r_a$ from Equation (4) where a represents the effects of personal characteristics for each individual that determine their behaviour and where they spend their time - the key components of calculating their individual risk of contracting the disease. The Bernoulli distribution is used to assign each individual either a zero (doesn't get exposed) or a one (does get exposed) based on the principle of a coin-flip with the weight of the coin (i.e. the chance of being exposed) being determined by the probability p_a . The higher the probability, p_a , the more likely the random number drawn from the Bernoulli distribution will be a one, and the more likely they are to transition from susceptible (**S**) to exposed (**E**). This process is repeated for every individual in the population at each (daily) time step.

When an individual is exposed, they are assigned an exposed duration transition time and a pre-symptomatic duration and a symptomatic/asymptomatic duration. Following approaches commonly used in the literature (see for example, Li et al. (2020), Linton et al. (2020), Wei et al. (2020)). The first two of these are realisations of Weibull distributions (i.e. non-negative, flexible and allow for long-tails / extended durations) and the latter from a log-normal distribution (non-negative and right-skewed). Details of parameters used for the different stages, together with references of their sources, can be found in Supplementary Information.

Once in the Exposed (**E**) state an individual will next move into the Infectious (**I**) state. This can mean moving into the asymptomatic or the pre-symptomatic and then symptomatic stage. This will be influenced by an individual's age and BMI, with older and overweight individuals less likely to be asymptotically infected (Table 1) according to:

$$(\mathbf{E} \rightarrow \mathbf{I})_a = \text{Bernoulli}(\theta_{\mathbf{I},a}) \quad (5)$$

where $\theta_{\mathbf{I},a}$ is determined by the symptomatic probabilities outlined in Table 1.

Lastly, individuals will move from the Infectious (**I**) state to the Removed (**R**) state. All asymptotically infected individuals will recover. Symptomatically infected individuals will either recover or die based upon their age and BMI (Table 1). Older and more overweight individuals are less likely to recover (Table 1). This transition is described by the following:

$$(\mathbf{I} \rightarrow \mathbf{R})_a = \text{Bernoulli}(\gamma_{\mathbf{R},a}) \quad (6)$$

where $\gamma_{\mathbf{R},a}$ is determined by the mortality probabilities outlined in Table 1.

[Table 1 about here.]

2.2 *Generating a Synthetic Population*

The underlying population used in the dynamic simulation model comes from a spatial microsimulation model, SPENSER (Synthetic Population Estimation and Scenario Projection Model), developed to provide timely georeferenced population forecasts at a high resolution (individual and household level) for scenario projections (Lomax and Smith, 2017, Smith and Russell, 2018). SPENSER uses Iterative Proportional Fitting (IPF) techniques (Lovelace et al., 2015) to reweight microdata and area level counts from the 2011 Census of Population for England and Wales to create a micro-level synthetic dataset for the entire population. Spatial microsimulation has been widely employed in support of financial and economic policy analysis across Europe and North America (Tanton, 2018). Over the last two decades, spatial microsimulation techniques have been used increasingly to examine health and health inequalities (Morrissey et al., 2015).

The SPENSER model comprises four steps: (1) estimate the individual population from 2011 Census Data; (2) estimate the household population from 2011 Census data; (3) simulate the baseline population and households forward to the jump off year 2020, needed for input to the dynamic model; and (4) assign individuals to households to provide consistency between files. Synthetic individuals are placed in households and are attributed demographic (age and sex for each individual), socioeconomic (based on the socioeconomic status of the household's reference person) and housing condition variables according to the individual and household estimates from the 2011 Census. The individual and household characteristics of relevance to this work can be seen in Table 2, along with

additional health and time-use variables that are included through the use of Propensity Score Matching (discussed below).

[Table 2 about here.]

In the output from SPENSER, each individual is assigned to a Middle Layer Super Output Area (MSOA) while in the household output, individual households are assigned to a Lower Super Output Area (LSOA). This is due to differences in the constraint tables used to construct the synthetic population, where household constraints variables are available with higher levels of disaggregation for smaller areas than population constraint variables. As individuals are assigned to a household, combining the two files means that information for individuals can ultimately be derived at LSOA scale. MSOA is a census geography in which each area represents a mean population in the order of 7,200 individuals, and LSOA is a finer geography in the order of 1,500 individuals.

2.2.1 Enriching the Synthetic Population

Following work by Morrissey et al. (2015), propensity score matching (PSM) using a kernel density algorithm was used to allow each individual simulated by the SPENSER model to be matched to an individual in two external datasets based on the similarity of their demographic, socioeconomic and spatial characteristics. Using a kernel density algorithm, PSM was used to enrich the baseline SPENSER dataset to include data from the United Kingdom Time Use Survey, 2014/2015 (UKTUS) and the Health Survey of England, 2019 (HSE). UKTUS is a large-scale household survey that provides data on how people aged eight years and over in the UK spend their time. The survey instrument is a time diary instrument in which respondents record their daily activities over two weeks. The UKTUS

provides the richest source data on how people spend their time, their location throughout the day, and who they spend their time with. The UKTUS also has detailed employment information as part of its core set of questions including information on employment status, and industrial sector and occupation category for those in employment or previously in employment (i.e. they are now retired). Including employment data and the occupation and industrial sector in which individuals are employed in is important as it allows the identification of key workers in the dataset. The HSE is an annual survey that provides health and care information on adults aged 16 and over and children aged 0 to 15. The HSE survey is used to monitor the rate of obesity and to estimate the proportion of people in England who have certain health conditions and the prevalence of risk factors and health related behaviours, such as smoking and drinking alcohol. The additional variables matched to the outputs from SPENSER can be seen in Table 2. Following the approach used in Morrissey et al. (2015) validation of the matching process was performed to assess whether the resulting enriched dataset could be considered unbiased conditional on the observed characteristics (the conditional independence assumption). Frequencies and distributions of both matching variables (used in the PSM) and non-matching variables were compared. One would expect the matching variable to show good agreement across the population as this variable was used in the PSM process. However, it is also important to understand if the distribution for key variables of subsequent interest not included in the PSM process are captured. Table 3 shows an example of this evaluation: the distributions of the National Statistics Socio-economic Classification (NS-SEC), one of the matching variables in the PSM, and health status, a non-matching variable. The proportions in each category in the enriched SPENSER dataset are compared to corresponding Office of

National Statistics data and HSE for Devon (Census, 2011) and both the matching and non-matching variables show good agreement. Figure 3 presents a snapshot of the augmented SPENSER data, empirically demonstrating a number of key variables for the MSOAs in the five Local Authority Districts that comprise the case study area.

[Table 3 about here.]

[Figure 3 about here.]

2.3. Estimating Interaction with Locations of Disease Transmission

Currently three activities, other than spending time at home, are simulated in the model: working, attending school and shopping. Having estimated the amount of time that individuals spend doing these activities (Section 2.2), this section outlines a general method for estimating the probabilities that individuals will visit particular sites of disease transmission. For example, given that an individual might spend an hour per day shopping, which shops are they most likely to visit?

2.3.1. Supermarket and School Probabilities

The following provides an illustrative example based on trips to supermarkets and schools, but the principle is the same for sending individuals to any point destinations including those not explicitly considered currently such as pubs and restaurants. Workplaces are an exception, as discussed in Section 2.3.2.

The probabilities of individuals visiting specific locations are calculated using spatial interaction models (SIMs; O'Kelly, 2009). SIMS estimate the aggregate flows of a population from origin zones (neighbourhoods where the synthetic populations live) to

destination locations. SIMs are analogous to a Newtonian model of gravity where the strength of interaction (in our case the flows of people or the money they spend) is proportional to the mass of the origin and destination locations (represented by the size of the residential population or the attractiveness of the destination) and inversely proportional to the cost of this interaction (frequently represented by travel distance or time). Where information about aspects of the system is known such as the total number of residents at an origin or pupils on a school roll, constraints can be applied such that estimated interactions correspond to this known information. Where data on aspects of the interaction are available such as known flows or travel times, parameters of the model can be calibrated to improve the estimates produced. The locations of schools (both primary and secondary) and shops have been established from Department for Education (<https://get-information-schools.service.gov.uk/>) and the Geolytix retail point open data (<https://www.geolytix.co.uk/#!geodata>), respectively. The 'attractiveness' of each location is estimated using the school capacity and the approximate retail floorspace (augmented with retail turnover) respectively.

[Figure 4 about here.]

A cost matrix is used to compute flows between origins and destinations (i.e. trip probabilities) based on that used in the QUANT project (Batty and Milton, 2021). QUANT is a spatial analysis system which calculates shortest paths between every pair of zones in the model, using a network containing all roads in England, Scotland and Wales. As the model contains 8,436 MSOA and Intermediate Zones, this equates to 71 million shortest paths on an 8 million node road network; it is computationally intensive. Hence, the pre-built

QUANT costs matrix is used to calculate costs between 8,436 model zones and 14,227 retail point locations. This is achieved by taking the origin zone cost to the destination zone nearest to the retail point and then adding an additional term reflecting straight line distance from the destination zone to retail point term. These values are available in the files generated by the software. This process is repeated for the primary and secondary schools.

Having assembled the data for the origin, destination and costs of travel between zones, a spatial interaction model is used to calculate trip probabilities. Details of these models can be found in the Supplementary Materials (Section 7). Figure 5 shows the trip probabilities for South West England region using flow lines.

[Figure 5 about here.]

2.3.2. Workplace Probabilities

Workplace flows would ideally be estimated through a spatial interaction model similar to that employed in the estimation of flows to schools and shops. However, the problem with journey to work is significantly more difficult because: (i) there are vastly more workplaces than shops or schools; (ii) there is no definitive list of workplace locations; (iii) even if workplace locations are known, there is no clear link between a synthetic individual's employment category and equivalent workplace categories.

To address this issue, we initially adopt a stylized approach constructing 'virtual workplaces' which rely on the 2011 UK Census commuting origin-destination tables at the

MSOA level for individuals with a fixed workplace. The UKTUS data includes a Standard Industry Classification (SIC) code for everyone in the dataset. Matching data from the UKTUS to SPENSER baseline data via the PSM process and the UKTUS we were able to assign to each of our synthetic resident workers an employer industry among the 21 divisions from the Standard Industry Classification (SIC) 2007. We assume that all workers have an equal ex-ante probability to commute to all destinations independently from the SIC to which they belong. We build the set of possible destinations by multiplying the number of MSOAs in the study area, $M = 107$, to that of the SIC divisions, $S = 21$, obtaining 2,247 options. We then populate these virtual workplaces with synthetic workers based on their reference SIC and their Census relative probability to commute from M_i to any M_j , with $j = 1 \dots i \dots J$, thus including the MSOA in which the worker resides.

3. Case Study: UK Lockdown, March 2020

The first confirmed case of the novel coronavirus in the UK was documented on 21st January, 2020. This was followed by the first confirmed COVID death in the UK on 5th March. On 16th March the Prime Minister encouraged social distancing, telling people in the UK that they should stop all non-essential contact. Although they could remain open, people were asked not to visit pubs, clubs and theatres. Workers were asked to work from home if they could and households were asked to isolate for two weeks if any member had symptoms. On the day of the announcement of these measures the death toll of people in the UK with COVID-19 listed as the cause of death reached 55. One week later, on 23rd March, 2020, the Prime Minister announced a UK wide lockdown in which he ordered

people to only leave the house to shop for basic necessities “as infrequently as possible” and encouraged them to perform no more than one form of exercise a day.

In the following, we provide a case study on the potential reduction in cases and subsequently deaths that implementation of the lockdown one week earlier may have had in Devon County, England. Devon is a county in the Southwest of England that extends from the Bristol Channel in the north to the English Channel in the south and is bounded by Cornwall to the west, Somerset to the north-east and Dorset to the east. Devon is a sparsely populated, predominantly rural county with a total population of about 700,000.

3.1 *Simulating the lockdown*

The simulation of cases during the first lockdown is based on the temporal distribution of cases recorded by Public Health England (PHE; coronavirus.data.gov.uk). During this period, the Royal Devon & Exeter NHS Foundation Trust and Northern Devon Healthcare Trust estimate that the prevalence of COVID-19 was 2% (*personal communication*). This equates to ca. 14,000 individuals compared with 790 cases recorded by PHE for the Unitary Authority of Devon over the first 70 days, due to limited testing at the beginning of the outbreak. We smoothed the PHE cases using a negative binomial generalised additive model, $s(\text{cases}_t)$, and applied a multiplying factor to give the expected number of cases on day t as $ec_t = \left(\frac{\text{population} * \text{prevalence}}{s(\text{cases}_t)} \right)$.

The model was ‘seeded’ by constraining the number of infections in the first 10 days to be equal to ec_t after which the number of new daily infections are generated by the model, unrestricted, for a further 60 days. In order to impose lockdown on the simulated

population, the amount of time individuals spent outside their home was scaled according to data from the Google Community Mobility Reports. As Google Community Mobility Reports are available at a regional scale, we used data specific to Devon. These data provide aggregated estimates for the proportion of time, on average, a population spends in six types of locations relative to a baseline of the median value for the corresponding day of the week, during the 5-week period 3 Jan–6 Feb, 2020. The six locations are: retail & recreation, grocery & pharmacy, parks, transit stations, workplaces and residential. It is assumed that the residential component refers to individuals spending time in their own homes and therefore an individual's baseline is equivalent to the estimated amount of time individuals spend at home from the UKTUS (as discussed in Section 2.2).

The values from the Google Community Mobility data were smoothed for time spent in residential locations using a 14-day moving average (g_t). Using this in conjunction with the average proportion of time spent at home (\bar{p}_h) and outside the home (\bar{p}_o) from the individuals in the population, we created time-series of daily lockdown multipliers (l_t , Figure 6). As can be seen from Figure 6, the values for proportion of time outside the home from March to June 2020 are all less than 1. For any given day, the amount of time that any individual spends at a location outside the home is reduced in proportion to the lockdown multiplier. Time no longer spent on activities outside the home will be added on to time spent at home for each individual. The only condition under which the lockdown multiplier does not apply is if an individual is in the symptomatic disease status. Here we assume they reduce their activities outside the home by 90% to reflect self-isolation behaviour.

Lockdown restrictions are applied universally across the population so that, for example, there is no differentiation for enhanced mobility of key workers or to allow for variations

between business sectors (Batty and Milton, 2021), which would be a possible avenue for future refinement of the model.

$$l_t = \frac{1-(g_t*\bar{p}_h)}{\bar{p}_o} \quad [7]$$

[Figure 6 about here.]

3.2 Results: Lockdown restrictions imposed one week earlier

Other countries went into lockdown earlier than the UK and here the effects of implementation of a UK-wide lockdown one week earlier than it occurred are simulated. To explore the effect of official lockdown occurring earlier, the time-series of lockdown multipliers (Figure 6) is shifted to be one week earlier. For the purpose of comparing scenarios the lockdown scenario as it happened is referred to as the ‘baseline’ scenario, while the scenario in which lockdown is imposed one week earlier will be called the ‘experimental’ scenario.

The model simulation in the baseline scenario produced a good fit to the known daily cases of COVID-19 according to PHE data. The total infection count in Devon county at the end of the 70 day simulation is summarised by age group in Table 4. As expected, the model suggests that an earlier lockdown would have significantly reduced the spread of the disease. For the baseline scenario daily infections peaks at 763 (266-1047, 95% CI) people per day, while the experimental (i.e. lockdown one week earlier) scenario shows a peak of 556 (137-718, 95% CI) people per day (Figure 7).

Being able to explore heterogeneity in the transmission of the disease in different groups within the population and over different spatial aggregations and periods of time is one of the key features of the microsimulation approach. The outputs of the model are at the individual level and it is straightforward to aggregate the results from the simulations to any specified groupings. As an example, Table 4 shows the results by age groups and Figure 7 the number of cases over time. Another feature of the model is being able to extract information for individuals within the population according to their disease status at any point in time and this information can be cross-tabulated with other variables to assess heterogeneity in disease status across different groups (over time). As an example, Figure 8 shows the number of people with different disease status by age group, together with the reduction in cases associated with lockdown being a week earlier. This shows a clear difference between age groups with a higher proportion of asymptomatic cases in younger age groups.

[Table 4 about here.]

[Figure 7 about here.]

The model is spatially explicit, allowing us to explore the geographical distribution of COVID-19 infections in our scenarios. Figure 9 shows that the baseline scenario leads to some distinct hot-spots located around more densely populated MSOAs, such as those in Exeter, which is one of the largest cities in Devon county. In the baseline scenario as much as 6% of the population of an MSA becomes infected. In the experimental scenario, we see a similar spatial distribution as that seen in the baseline scenario, with hot-spots located

around larger cities with denser populations. However, the maximum infection rate is reduced to under 4% in the experimental scenario.

[Figure 8 about here.]

[Figure 9 about here.]

4. Discussion and Conclusions

This paper presents a novel, data-driven modelling framework that reflects the complexities of the British population to model the transmission of COVID-19 within communities and to assess the effect of policy interventions. The framework brings together a wide variety of data driven approaches, including epidemiological disease modelling, urban analytics and spatial analysis, as well academic and private sector researchers to develop a computationally efficient framework for its implementation. This enables questions related to the geographical transmission, diffusion, acceleration and the regulation in the incidence of cases to be traced through physical interactions between the many components that determine the way entire populations move and interact with one another in their daily lives. The power and spatial flexibility of the framework to assess the effects of different interventions is demonstrated within the case study where the effects of the first UK national lockdown are estimated for the county of Devon. Here we find that an earlier lockdown is estimated to result in a lower peak in daily infections and 47% fewer infections overall.

As outlined in this paper, the framework is based on a spatial microsimulation model, SPENSER, that reproduces data on household and its constituent population across the whole of Great Britain. The data produced by the spatial microsimulation model replicates the structure and behaviour of the real population in terms of demographic, socioeconomic and health characteristics, along with detailed time use data. Spatial Interaction models 'mobilise' this data according to the profile of each individual via a series of spatial allocations for each individual into a series of real-world physical locations in which the transmission of coronavirus could take place. Data from a variety of third-party sources are introduced to allow calibration of the models to reproduce existing patterns of movement and spatial interaction. In the case study demonstration for Devon, shops, schools and hospitals are included as destination locations. These models are being extended to embrace the key activity of the journey to work which is an essential component of the balance between working from home and place of work.

The model is calibrated against a variety of data sources including public health records, mobility data, measures of retail activity, employment and educational participation, and the socio-demographic composition of small areas. The benefits to wider exploitation and sharing of such sources has been widely noted (von Borzyskowski et al., 2021, Science Academies of the Group of Seven, 2021). With such resources at our disposal, the development of dynamic microsimulation models could provide a step change in the ability of national governments to prepare and respond to the threat of future pandemics.

The flexibility of the modelling framework presented here allows the parameters and distributions within the individual components to be updated to reflect updated scientific

understanding and factors such as increased levels of transmission associated with multiple variants. It offers a multitude of opportunities for future scenario development, including exploring the effects of alternative lockdown scenarios both at an aggregate level, but also across different sub-populations, and the ramifications of the vaccination roll-out. In the case of the former, it will be possible to consider variations in the timing of movements between different mitigation/adaptation strategies on the number and distribution of cases, and the capacity of local health services to meet the associated need. More refined options such as the restriction of specific types of employment type or activity, e.g. schools, restaurants or retail outlets, or the variation of controls across more disaggregate geographies than local authority areas can also be considered. For the latter, scenarios could be designed that explore the nature of long-term equilibrium dynamics, e.g. in a progression towards herd immunity or seasonal cycles of infection, with the model creating projections of future infections, by local area, for example, in relation to efficacy, uptake, compliance, and availability of the vaccines across social and demographic groups. The dynamic simulation model was developed using a combination of R and Python. After the initial development, it was refactored using OpenCL, a framework for parallel programming. OpenCL allows the simulation to be executed on a CPU or GPU, depending on the available hardware, and leads to a significant speedup due to multi-threaded execution. The OpenCL implementation is able to run the simulation for 100 timesteps for the whole population of Devon in around a second, which is in the order of 10,000 times faster than the original implementation. This improved computational speed is crucial if models such as this are going to be used by policy-makers within real decision-making environments. In addition, an interactive Graphical User Interface (GUI) was built (see Figure 10). The GUI

allows the user to explore scenarios while they are executing by interactively starting, stopping, stepping and resetting the model. The GUI also allows the values of model parameters to be modified and the model to be re-run with updated parameter values. This allows rapid exploration of the model output and how it changes with different parameter values.

[Figure 10 about here.]

The importance of reflecting the real-life behaviours of individuals given their health, demographic and socioeconomic circumstances is reflected in the large evidence base that demonstrates that the outcomes of COVID-19 are not distributed equally across sub-populations and space. This is linked to a variety of factors including occupational profile, housing circumstances and transportation options. To date, COVID-19 transmission models have failed to capture the necessary data to capture the inequality in outcomes across different sub-groups. This paper extends the growing number of COVID-19 transmission models by developing a dynamic SEIR model underpinned by a 'digital twin' British population. The digital twin underpinning the dynamic SEIR model represents the complex health, socio-economic and behavioural attributes, as well as mobility patterns required to understand the transmission of COVID-19 within the community and the impact of different interventions. Importantly, the synthetic modelling approach is reproducible in any country for which small area demographic counts are available, along with nationally representative health and time use data.

References

A. Desvars-Larrive, E. Dervic, N. Haug, T. Niederkrotenthaler, J. Chen, A. Di Natale, J. Lasser, D. S. Gliga, A. Roux, J. Sorger, et al., A structured open dataset of government interventions in response to covid-19, *Scientific Data* 7 (2020) 1–9.

K. Daras, A. Alexiou, T. C. Rose, I. Buchan, D. Taylor-Robinson, B. Barr, How does vulnerability to COVID-19 vary between communities in England? Developing a Small Area Vulnerability Index (SAVI), *Journal of Epidemiology and Community Health* (2021) jech-2020-215227. doi:10.1136/jech-2020-215227.

E. Kontopantelis, M. A. Mamas, J. Deanfield, M. Asaria, T. Doran, Excess mortality in England and Wales during the first wave of the COVID-19 pandemic, *Journal of Epidemiology and Community Health* 75 (2021) 213–223. doi:10.1136/jech-2020-214764.

S. McNamara, J. Holmes, A. K. Stevely, A. Tsuchiya, How averse are the UK general public to inequalities in health between socioeconomic groups? A systematic review, *European Journal of Health Economics* 21 (2020) 275–285. URL: <https://doi.org/10.1007/s10198-019-01126-2>. doi:10.1007/s10198-019-01126-2.

Cabinet Office, Race Disparity Audit Summary Findings from the Ethnicity Facts and Figures website, Technical Report, 2017.

Office for National Statistics, Monthly mortality analysis, England and Wales: January 2021, Technical Report, 2021.

R. Mathur, C. T. Rentsch, C. E. Morton, W. J. Hulme, A. Schultze, B. MacKenna, R. M. Eggo, K. Bhaskaran, A. Y. Wong, E. J. Williamson, H. Forbes, K. Wing, H. I. McDonald, C. Bates, S. Bacon,

A. J. Walker, D. Evans, P. Inglesby, A. Mehrkar, H. J. Curtis, N. J. DeVito, R. Croker, H. Drysdale, J. Cockburn, J. Parry, F. Hester, S. Harper, I. J. Douglas, L. Tomlinson, S. J. Evans, R. Grieve, D. Harrison, K. Rowan, K. Khunti, N. Chaturvedi, L. Smeeth, B. Goldacre, Ethnic differences in COVID-19 infection, hospitalisation, and mortality: An OpenSAFELY analysis of 17 million adults in England, medRxiv (2020) 1–19. doi:10.1101/2020.09.22.20198754.

Race Disparity Unit Cabinet Office, Quarterly report on progress to address COVID-19 health inequalities, Technical Report October, 2020. URL: <http://www.nationalarchives.gov.uk/doc/open-government-licence/>.

J. Bibby, G. Everest, I. Abbs, Will COVID-19 be a watershed moment for health inequalities?, 2020.

X. Zhang, G. Owen, M. Green, I. Buchan, B. Barr, Evaluating the impacts of tiered restrictions introduced in England, during October and December 2020 on COVID-19 cases: A synthetic control study, medRxiv (2021) 1–20. URL: <http://medrxiv.org/content/early/2021/03/12/2021.03.09.21253165.abstract>.

R. Harris, Exploring the neighbourhood-level correlates of Covid-19 deaths in London using a difference across spatial boundaries method, Health and Place 66 (2020). URL: <https://doi.org/10.1016/j.healthplace.2020.102446>.
doi:10.1016/j.healthplace.2020.102446.

M. KC, E. Oral, S. Straif-Bourgeois, A. L. Rung, E. S. Peters, The effect of area deprivation on COVID-19 risk in Louisiana, PLoS One 15 (2020) 1–12. URL: <http://dx.doi.org/10.1371/journal.pone.0243028>. doi:10.1371/journal.pone.0243028.

- U. Bilal, L. P. Tabb, S. Barber, A. V. Diex Roux, Spatial Inequities in COVID-19 Testing , Positivity, Confirmed Cases, and Mortality in 3 U.S. Cities, *Annals of Internal Medicine* (2021). doi:10.7326/M20-3936.
- G. H. Shah, P. Shankar, J. S. Schwind, V. Sittaramane, The detrimental impact of the COVID-19 crisis on health equity and social determinants of health, *Journal of Public Health Management and Practice* 26 (2020) 317–319. doi:10.1097/PHH.0000000000001200.
- L. Rvachev, I. M. Longini, A mathematical model for the global spread of influenza, *Mathematical Biosciences* 75 (1985) 3–22. doi:10.1016/0025-5564(85)90064-1.
- M. Keeling, E. Hill, E. Gorsich, B. Penman, G. Guyver-Fletcher, A. Holmes, T. Leng, H. McKimm, M. Tamborrino, L. Dyson, M. Tildesley, Predictions of COVID-19 dynamics in the UK: short-term forecasting and analysis of potential exit strategies, *medRxiv* (2020). doi:10.1101/2020.05.10.20083683.
- E. van Leeuwen, F. Sandmann, Augmenting contact matrices with time-use data for fine-grained intervention modelling of disease dynamics: A modelling analysis, *medRxiv* (2020). doi:10.1101/2020.06.03.20067793.
- J. Dureau, K. Kalogeropoulos, M. Baguelin, Capturing the time-varying drivers of an epidemic using stochastic dynamical systems, *Biostatistics* 14 (2013) 541–555. doi:10.1093/biostatistics/kxs052. arXiv:1203.5950.
- N. M. Ferguson, D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, D. S. Burke, Strategies for mitigating an influenza pandemic, *Nature* 442 (2006) 448–452. doi:10.1038/nature04795.

A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday, N. Davies, A. Gimma, K. van Zandvoort, H. Gibbs, J. Hellewell, C. I. Jarvis, S. Clifford, B. J. Quilty, N. I. Bosse, S. Abbott, P. Klepac, S. Flasche, Early dynamics of transmission and control of COVID-19: a mathematical modelling study, *The Lancet Infectious Diseases* 20 (2020) 553–558. doi:10.1016/S1473-3099(20)30144-4.

W. O. Kermack, A. G. McKendrick, A contribution to the mathematical theory of epidemics, *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115 (1927) 700–721.

J. P. Arcede, R. L. Caga-Anan, C. Q. Mentuda, Y. Mammeri, Accounting for symptomatic and asymptomatic in a seir-type model of covid-19, *Mathematical Modelling of Natural Phenomena* 15 (2020) 34.

X. He, E. H. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. C. Lau, J. Y. Wong, Y. Guan, X. Tan, X. Mo, Y. Chen, B. Liao, W. Chen, F. Hu, Q. Zhang, M. Zhong, Y. Wu, L. Zhao, F. Zhang, B. J. Cowling, F. Li, G. M. Leung, Temporal dynamics in viral shedding and transmissibility of COVID-19, *Nature Medicine* 26 (2020) 672–675. doi:10.1038/s41591-020-0869-5.

A. Aleta, D. Martin-Corral, A. P. y Piontti, M. Ajelli, M. Litvinova, M. Chinazzi, N. E. Dean, M. E. Halloran, I. M. Longini Jr, S. Merler, et al., Modelling the impact of testing, contact tracing and household quarantine on second waves of covid-19, *Nature Human Behaviour* 4 (2020) 964–971.

J. A. Firth, J. Hellewell, P. Klepac, S. Kissler, A. J. Kucharski, L. G. Spurgin, Using a real-world network to model localized COVID-19 control strategies, *Nature Medicine* 26 (2020) 1616–1622.

L. Danon, E. Brooks-Pollock, M. Bailey, M. Keeling, A spatial model of CoVID-19 transmission in England and Wales: early spread and peak timing, *medRxiv* (2020) 1–10. doi:10.1101/2020.02.12.20022566.

X. Qiu, A. I. Nergiz, A. E. Maraolo, I. I. Bogoch, N. Low, M. Cevik, The role of asymptomatic and pre-symptomatic infection in SARS-CoV-2 transmission—a living systematic review, *Clinical Microbiology and Infection* 27 (2021) 511–519. URL: <https://doi.org/10.1016/j.cmi.2021.01.011>. doi:10.1016/j.cmi.2021.01.011.

Z. J. Madewell, Y. Yang, I. M. Longini, M. E. Halloran, N. E. Dean, Household Transmission of SARSCoV-2: A Systematic Review and Meta-analysis, *JAMA Network Open* 3 (2020) e2031756. doi:10.1001/jamanetworkopen.2020.31756.

W. C. Koh, L. Naing, L. Chaw, M. A. Rosledzana, M. F. Alikhan, S. A. Jamaludin, F. Amin, A. Omar, A. Shazli, M. Griffith, R. Pastore, J. Wong, What do we know about SARS-CoV-2 transmission? A systematic review and meta-analysis of the secondary attack rate and associated risk factors, *PLoS One* 15 (2020) 1–23. URL: <http://dx.doi.org/10.1371/journal.pone.0240205>. doi:10.1371/journal.pone.0240205.

Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong, et al., Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia, *New England Journal of Medicine* (2020).

N. M. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A. R. Akhmetzhanov, S. M. Jung, B. Yuan, R. Kinoshita, H. Nishiura, Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data, medRxiv (2020). doi:10.1101/2020.01.26.20018754.

W. E. Wei, Z. Li, C. J. Chiew, S. E. Yong, M. P. Toh, V. J. Lee, Presymptomatic Transmission of SARS550 CoV-2-Singapore, Morbidity and Mortality Weekly Report 69 (2020) 411–415.

N. Lomax, A. P. Smith, An introduction to microsimulation for demography, Australian Population Studies 1 (2017) 73–85. URL:
<http://www.australianpopulationstudies.org/index.php/aps/article/view/14>.

A. P. Smith, T. Russell, ukpopulation: unified national and subnational population estimates and projections, including variants, Journal of Open Source Software 3 (2018) 803.
doi:10.21105/joss.00803.

R. Lovelace, M. Birkin, D. Ballas, E. Van Leeuwen, Evaluating the performance of iterative proportional fitting for spatial microsimulation: New tests for an established technique, JASSS 18 (2015) 1–15. doi:10.18564/jasss.2768.

R. Tanton, Spatial microsimulation: Developments and potential future directions, International Journal of Microsimulation 11 (2018) 143–161. doi:10.34196/ijm.00176.

K. Morrissey, G. Clarke, P. Williamson, A. Daly, C. O'Donoghue, Mental illness in Ireland: Simulating its geographical prevalence and the role of access to services, Environment and Planning B: Planning and Design 42 (2015) 338–353. doi:10.1068/b130054p.

M. O'Kelly, Spatial interaction models, in: R. Kitchin, N. Thrift (Eds.), *International Encyclopedia of Human Geography*, Elsevier, Oxford, 2009, pp. 365–368. URL: <https://www.sciencedirect.com/science/article/pii/B9780080449104005290>. doi:<https://doi.org/10.1016/B978-008044910-4.00529-0>.

M. Batty, R. Milton, A new framework for very large-scale urban modelling, *Urban Studies* 0 (2021) 0042098020982252. doi:10.1177/0042098020982252.

S. N. Wood, N. Pya, B. Saßken, Smoothing parameter and model selection for general smooth models, *Journal of the American Statistical Association* 111 (2016) 1548–1563.

Google LLC, Google COVID-19 Community Mobility Reports, 2020. URL: <https://www.google.com/covid19/mobility/>.

I. von Borzyskowski, A. Mazumder, B. Mateen, M. Wooldridge, Data science and AI in the age of COVID-19: Reflections on the response of the UK's data science and AI community to the COVID-19 pandemic, Technical Report, London, England, 2021.

Science Academies of the Group of Seven, Data for international health emergencies: governance, operations and skills, <https://royalsociety.org/-/media/about-us/international/g-science-statements/G7-data-for-international-health-emergencies-31-03-2021.pdf>, 2021.

N. F. Brazeau, R. Verity, S. Jenks, H. Fu, C. Whittaker, P. Winskill, I. Dorigatti, P. Walker, S. Riley, R. P. Schnekenberg, H. Hoeltgebaum, T. A. Mellan, S. Mishra, J. T. Unwin, O. J. Watson, Z. M. Cucunuba', M. Baguelin, L. Whittles, S. Bhatt, A. C. Ghani, N. M. Ferguson, L. C. Okell, Report 34 - COVID-19 infection fatality ratio estimates from seroprevalence, Imperial

College London (2020) 1–18. URL: <https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-34-ifr/%0Ahttps://doi.org/10.25561/83545>.

B. M. Popkin, S. Du, W. D. Green, M. A. Beck, T. Algaith, C. H. Herbst, R. F. Alsukait, M. Alluhidan, N. Alazemi, M. Shekar, Individuals with obesity and COVID-19: A global perspective on the epidemiology and biological relationships, *Obesity Reviews* 21 (2020) 1–17. doi:10.1111/obr.13128.

N. G. Davies, P. Klepac, Y. Liu, K. Prem, M. Jit, C. A. Pearson, B. J. Quilty, A. J. Kucharski, H. Gibbs, S. Clifford, A. Gimma, K. van Zandvoort, J. D. Munday, C. Diamond, W. J. Edmunds, R. M. Houben, J. Hellewell, T. W. Russell, S. Abbott, S. Funk, N. I. Bosse, Y. F. Sun, S. Flasche, A. Rosello, C. I. Jarvis, R. M. Eggo, Age-dependent effects in the transmission and control of COVID-19 epidemics, *Nature Medicine* 26 (2020) 1205–1211. doi:10.1038/s41591-020-0962-9.

Table 1: The symptomatic and mortality rates of COVID-19 infections based on age and BMI (Brazeau et al., 2020, Popkin et al., 2020, Davies et al., 2020). The base symptomatic and mortality rates are taken from Davies et al. (2020) and Brazeau et al. (2020) respectively. We multiply the symptomatic rate by 1.46 for overweight individuals (BMI _ 25) and multiply the mortality rate by 1.48 for obese individuals (BMI _ 30) based on the findings by Popkin et al. (2020).

Age Group (Years)	Probability of Symptomatic Infection	Probability of Symptomatic Infection if BMI _ 25	Probability of Mortality if Infected	Probability of Mortality if Infected and BMI _ 30
0-4	0.21	0.21	0.0000	0.0000
5-9	0.21	0.21	0.0001	0.0001
10-14	0.21	0.21	0.0001	0.0001
15-19	0.21	0.21	0.0002	0.0002
20-24	0.45	0.66	0.0003	0.0004
25-29	0.45	0.66	0.0004	0.0006
30-34	0.45	0.66	0.0006	0.0009
35-39	0.45	0.66	0.0010	0.0015
40-44	0.45	0.66	0.0010	0.0024
45-49	0.45	0.66	0.0024	0.0036
50-54	0.45	0.66	0.0038	0.0056
55-59	0.45	0.66	0.0060	0.0089
60-64	0.45	0.66	0.0094	0.0139
65-69	0.45	0.66	0.0147	0.0218
70-74	0.69	0.96	0.0231	0.0342
75-79	0.69	0.96	0.0361	0.0534
80-84	0.69	0.96	0.0566	0.0838
85-89	0.69	0.96	0.0886	0.1311
90+	0.69	0.96	0.1737	0.2571

Table 2: Attributes for the synthetic population from SPENSER and propensity score matching

Variable	SPENSER	Time Use Survey	Health Survey of England
<i>Individual</i>			
Sex	X	X	X
Age	X		X
Ethnicity	X	X	X
National Statistics Socio-economic Status (NS-SEC) of household reference person	X	X	X
Number in household	X		
Time use data (proportion of time doing different activities)		X	
(CVD, high blood pressure, diabetes, COPD, BMI>40)			X
In-work status		X	
Standard Industrial Classification of economic activities (SIC)		X	
<i>Household</i>			
Type of dwelling inhabited (e.g. semi-detached house)	X		
Tenure (e.g. rented, mortgaged)	X		
Household Composition (e.g. cohabiting couple)	X		
Number of occupants	X		
Number of rooms	X		
Presence of central heating	X		
Type of dwelling	X		
Number of cars in household	X		

Table 3: Evaluation of propensity score matching: frequencies of a matching and non-matching variables (National Statistics Socio-economic Classification (NS-SEC) and health status) in the enriched SPENSER dataset and from the Office of National Statistics and Health Survey for England

PSM Matching Status	Validation	Devon Census	Enriched Dataset
Matched variable	NS-SEC 1	34%	37%
	NS-SEC 2	10%	11%
	NS-SEC 3	12%	17%
	NS-SEC 4	6%	8%
	NS-SEC 5	21%	24%
Non-matching Variable	Very good health	46%	45%
	Good health	35%	33%
	Fair health	14%	11%
	Bad health	4%	8%
	Very bad health	1%	3%

Table 4: The number of infections (medians from 1000 simulations) in Devon county in each age-group between the baseline scenario and the experimental scenario under which lockdown started was a week earlier.

Age Group	Baseline	Experimental	Percentage Decrease
0-18	4967	4076	18%
19-30	2662	1045	61%
31-65	8017	2190	73%
66+	1757	569	68%
Total	17221	7880	54%

Table S1: Model parameters.

Disease status	Length of time (days)	Source
Exposed Period	2.56 \pm 0.72 (SD)	Peirlinck et al. (2020)
Pre-symptomatic (infectious)	Contagious 1-3 days before symptom onset	Wei et al. (2020)
Pre-symptomatic (infectious)	2.3 days (95% CI, 0.8–3.0 days) before symptom onset	He et al. (2020)
Symptomatic	17 days = The median number of days from symptoms onset until the successive negative detection of SARS-CoV-2 RNA [IQR] 12–21 days	Qi et al. (2020)
Symptomatic	21 = Maximum number of days since symptom onset with detectable virus	He et al. (2020)
Asymptomatic	Viral load lower for asymptomatic cases but duration of shedding similar	Zhou et al. (2020)
Asymptomatic	mean 5.2 (CI 95% 4.1-7.0)	Li et al. (2020)
Incubation period (exposed + pre-symptomatic)	mean 6.4 (CI 95% 5.6-7.7)	Backer et al. (2020)
Incubation period (exposed + pre-symptomatic)	median 5 (CI 95% 4.0-5.8)	Backer et al. (2020)
Incubation period (exposed + pre-symptomatic)	mean 5 (CI 95% 4.0-5.8)	Linton et al. (2020)
Incubation period (exposed + pre-symptomatic)	mean 4.8 (CI 95% 2 - 11)	Liu et al. (2020)
Incubation period (exposed + pre-symptomatic)	median 5.1 (CI 95% 4.5 -5.8)	Lauer et al. (2020)
Incubation period (exposed + pre-symptomatic)	97.5% of those that get symptoms will do so within 11.5 days (CI, 8.2 to 15.6 days)	Lauer et al. (2020)
Infectious (pre-symptomatic + symptomatic or asymptomatic)	17.82 \pm 2.95 (SD)	Peirlinck et al. (2020)

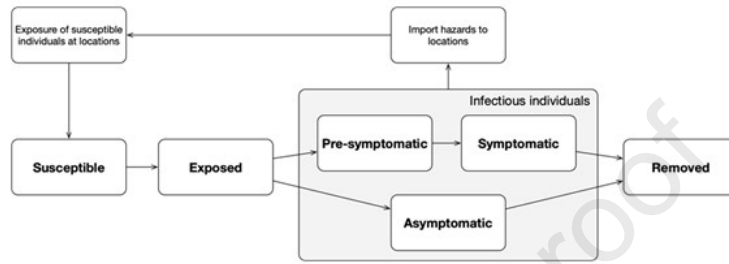


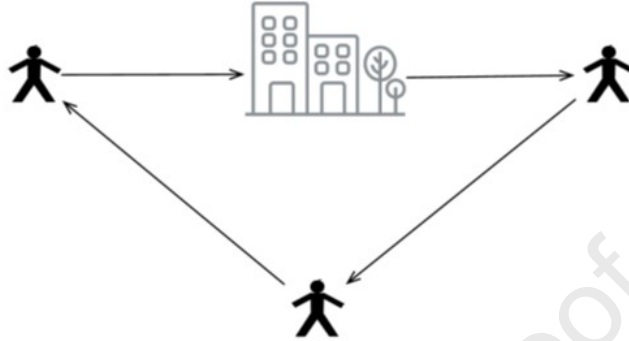
Figure 1: Schematic showing the structural components of the SEIR model.

Stage 1. Hazard Allocation

Individuals visit different locations (homes, schools, shops, workplaces.). If they are infected they contribute to the *hazard* in the locations they visit.

Stage 2. Exposure Estimation

Individuals are exposed to a hazard from the locations that they visit. These exposures cumulate so contribute to their overall exposure score

**Stage 3. Disease Status Estimation**

Exposure scores are used, amongst other attributes, to estimate the new disease status for all individuals

Figure 2: The process of disease spread using the dynamic microsimulation model.

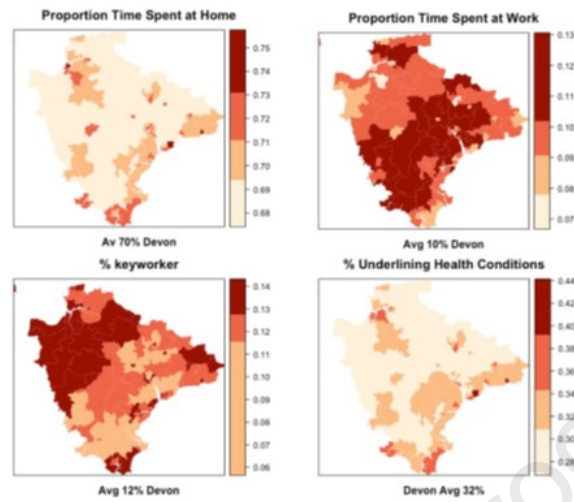


Figure 3: Example output from augmented SPENSER dataset proportion of time spent at home, proportion of time spent at work, the percentage of key workers and the percentage of individuals with underlying health conditions (doctor diagnosed CVD, high blood pressure, diabetes, COPD and a BMI greater than 40) for the MSOAs in the five Local Authority Districts that comprise Devon.



Figure 4: Geolytix retail point location coverage in the UK. The colours on the points represent one of four different floorspace bands. Flow lines connect origin MSOA zones to Geolytix retail points according to the modelled trip probabilities.



Figure 5: Trip probabilities from MSOA origin zones to Geolytix retail points (black dots). The colour of the flow line represents the magnitude of the trip probability from <0.0094 (blue) to ≥ 0.0226 (red). The map on the left contains the Geolytix retail points, which are omitted for clarity in the map on the right.

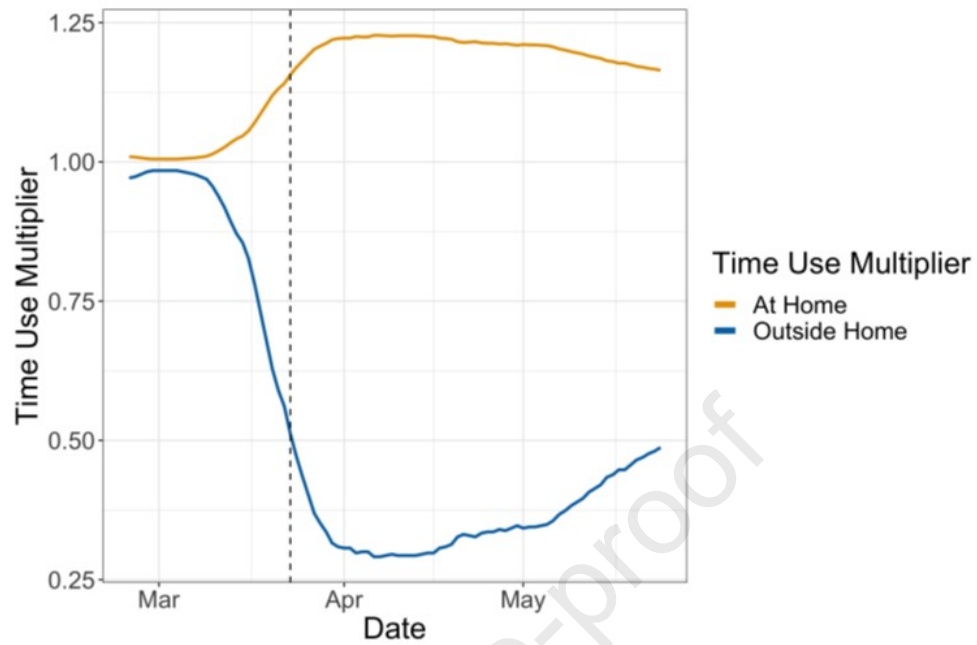


Figure 6: Relative to a baseline of Jan 3–Feb 6 2020 the proportion of time spent outside home (blue) and at home (orange). The official lockdown date 23rd March 2020 is shown with the dashed black line.

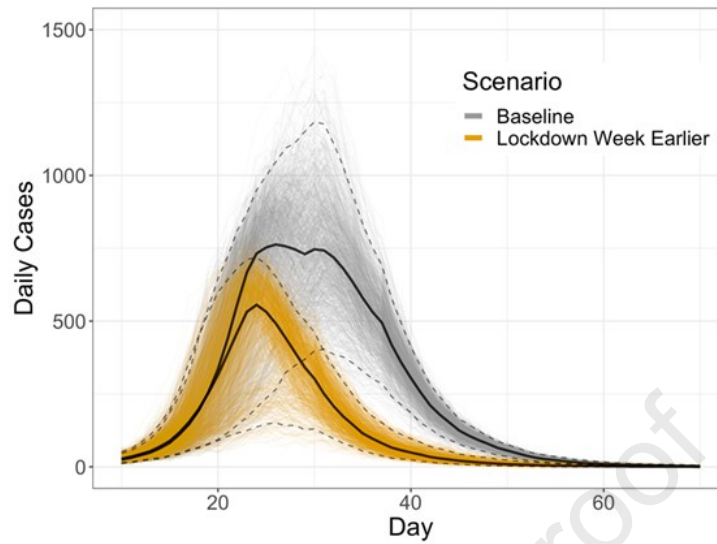


Figure 7: Predicted daily number of COVID cases under a baseline scenario and a scenario where lockdown effects were a week earlier, by age group. Each of the scenarios were run 1,000 times to capture variation in the stochastic elements of the model. The daily number of cases from each of these runs is shown as a transparent line. The median value of all of the runs of each scenario is shown in bold. The dashed lines show the 95% uncertainty intervals for each scenario. For the baseline scenario Google Mobility data is used to invoke lockdown effects on reducing individual time outside their home, for the experimental scenario where lockdown is a week earlier we shift the Google Mobility data a week earlier.

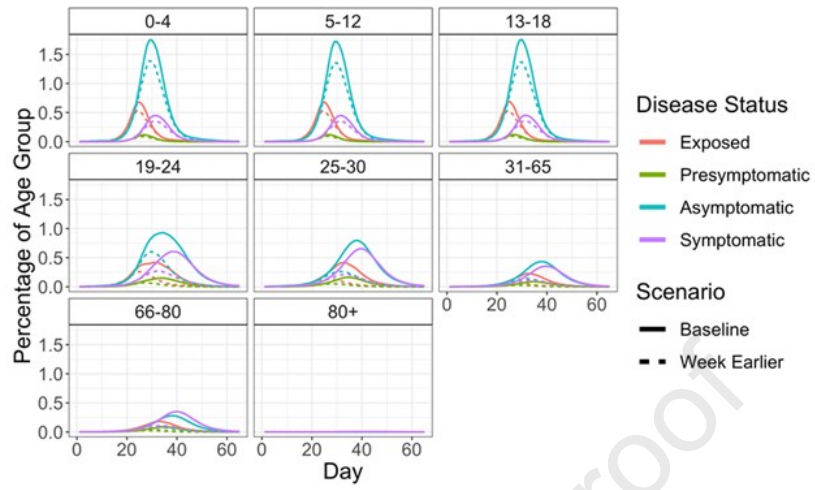


Figure 8: Predicted proportion of individuals in each age group, colours represent each disease status.

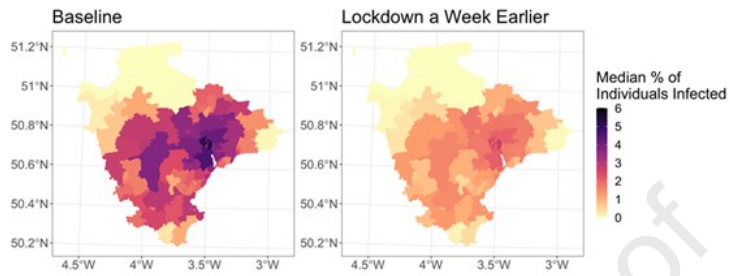


Figure 9: Predicted percentage of each MSA that was infected with COVID during the first 70 days of the outbreak under two scenarios, the baseline scenario and an experimental scenario where lockdown was shifted a week earlier.

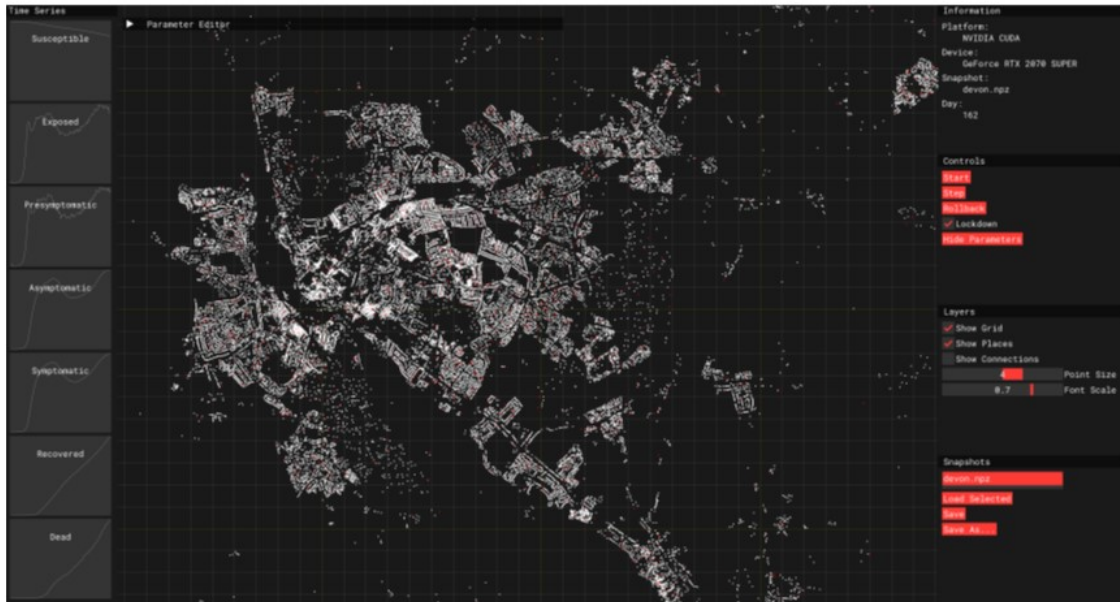


Figure 10: Screenshot of the GUI, showing point cloud visualisation and time-series charts.

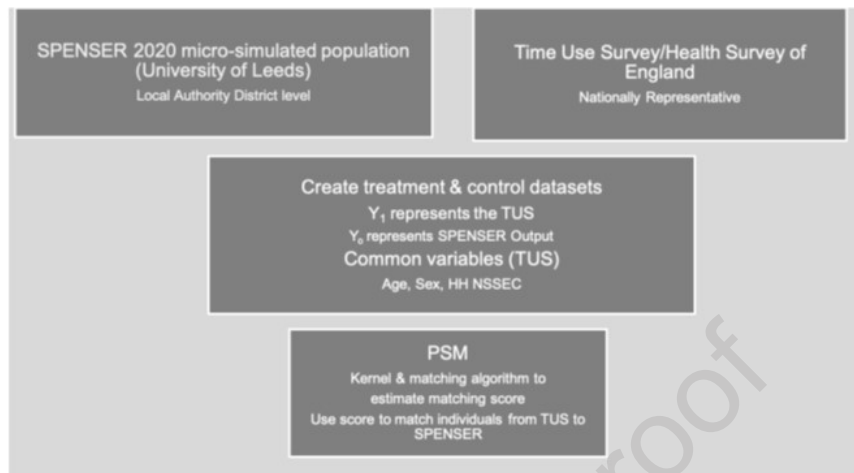


Figure S1: Overview of Propensity Score Matching Methodology.

Highlights

- A microsimulation framework for modelling disease transmission within communities.
- Integration of data from multiple, diverse sources to estimate exposure and risk.
- Computationally efficient implementation allows rapid assessment of scenarios.
- Allows the effects of different non-pharmaceutical interventions to be assessed.
- Results show substantial reduction in infections associated with earlier UK lockdown.

Journal Pre-proof