Journal of Computational Science 69 (2023) 101999

Contents lists available at ScienceDirect

Journal of Computational Science

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

# Scale matters: Variations in spatial and temporal patterns of epidemic outbreaks in agent-based models

Sarah Wise [a,*], Sveta Milusheva [b], Sophie Ayling [a], Robert Manning Smith [a]

[a] Centre for Advanced Spatial Analysis (CASA), University College London, London, United Kingdom
[b] World Bank Group, Washington, D.C., United States of America

ARTICLE INFO

*Keywords:*
Agent-based modelling
Scaling
Synthetic population
Simulation

ABSTRACT

Agent-based modellers frequently make use of techniques to render simulated populations more computationally tractable on actionable timescales. Many generate a relatively small number of "representative" agents, each of which is "scaled up" to represent some larger number of individuals involved in the system being studied. The degree to which this "scaling" has implications for model forecasts is an underdeveloped field of study; in particular, there has been little known research on the spatial implications of such techniques. This work presents a case study of the impact of the simulated population size, using a model of the spread of COVID-19 among districts in Zimbabwe for the underlying system being studied. The impact of the relative scale of the population is explored in conjunction with the spatial setup, and crucial model parameters are varied to highlight where scaled down populations can be safely used and where modellers should be cautious. The results imply that in particular, different geographical dynamics of the spread of disease are associated with varying population sizes, with implications for researchers seeking to use scaled populations in their research. This article is an extension on work previously presented as part of the International Conference on Computational Science 2022 (Wise et al., 2022)[1].

## 1. Introduction

Agent Based Models (ABMs) are often designed and built to model complex behaviours among large populations. However, the combination of complex behaviours and large populations can require extensive memory or CPU usage, and simulations can quickly become too computationally intensive to run efficiently outside of specialised environments. As the use of such models flourished during the global COVID-19 pandemic, the desirability of models which could be explored quickly and frequently became especially pressing. Researchers have long known that "more is different" - the scale of the simulated population matters in complex systems (see, for example, [2]). The solution has traditionally been to push technological and methodological boundaries to ensure models can be run at scale. While different methodologies have arisen to deal with the challenge of large-scale agent-based simulations (see Bithell and Parry [3] for a few of the most common), during the COVID-19 pandemic, the tremendous pressure in terms of both time and resources challenged the trade-off between model fidelity and run-time. The importance of this work has been highlighted as researchers have sought to help with the time-sensitive problem of disease forecasting (see for example [4,5], or [6]).

Briefly, some modellers approach this problem by reducing the level of complexity of the model or the scope of the simulated population in order to enable their model to run [7]. Others revert to equation-based modelling or hybrid approaches to reduce some of the burden in terms of computational intensity [6]. Still others have the option to simply increase computational power through either computer hardware or parallelisation (see, for example, [8]). Some researchers restructure their model to enable "super individuals" to represent multiple agents, which risks the dynamics of a larger population not being reflected beyond a certain point — both spatially and temporally.[9]

None of these approaches are without drawbacks, and it is important to understand the trade-off decisions researchers are making. Modellers who lack access to distributed computing, the platform-specific expertise to take advantage of it, or the funding to support such technical capacity must make do with more mass-market resources. Local government officials or humanitarian teams are increasingly interested in this kind of simulation; at the same time, they rarely have access to flexible budgets, highly specialised staff, or the longer timescales academic researchers enjoy. Further, those working with sensitive data may have concerns about making use of commercial platforms like Amazon Web Services; this is an especially alarming

* Corresponding author.
*E-mail address:* s.wise@ucl.ac.uk (S. Wise).

https://doi.org/10.1016/j.jocs.2023.101999
Received 14 October 2022; Received in revised form 10 March 2023; Accepted 14 March 2023
Available online 21 March 2023

prospect for teams working with health data. As such, many new agent-based modellers have turned to the use of "super individuals", an idea originally developed by Scheffer in 1995 [10]. In such models, researchers simulate only a fraction of the target population and scale up their output with the assumption that the trends are broadly representative.

In light of the renewed importance of this technique, we question how the assumptions researchers make about super individuals translate into different model outcomes. In particular, we study how using a smaller set of agents to represent the dynamics of the full population affects the trajectory of the model. This work provides an illustrative example of the impact of using different sizes of population samples in an ABM simulating the spread of SARS-CoV-2, an extension upon previous work (see Wise et al., 2022)[1]. The model explores the spread of disease through a representative sampled population in Zimbabwe during the first wave of the pandemic, which began in March 2020.

In Section 2, we present an overview of the background of this problem and how researchers have approached it in the past. We also highlight instances of practice whereby researchers have applied these techniques to highlight the need for the research being presented here. In Section 3 we introduce the model being used in the case study, and in Section 4 the data that underpins it. Section 5 breaks down our experiments regarding the variation in results produced by the differing proportion of population being simulated, exploring in particular how this changes as space is added to the simulation and as the infection rate changes. Section 6 further explains the results and their implications, while 7 concludes the article with our suggestions to other researchers regarding the trade offs of higher model fidelity versus quicker runtimes.

## 2. Literature review

The COVID-19 pandemic has prompted an explosion in the publication of disease models which use agent-based approaches. It has simultaneously highlighted the fact that computationally intensive models, including ABMs, may struggle to provide outputs on a timeframe useful for policymakers making time-sensitive decisions. During this period, when modellers were put under unprecedented pressure to produce outputs at short notice, simpler equation-based approaches often held a comparative advantage, despite not being able to provide equally fine grained answers. Given the importance of variation in susceptibility, access to vaccines, contact rates, and many other factors, many researchers instead turned to model simplification to render the problems computationally tractable. When models are simplified, an important aspect to consider and account for is how the population is scaled and the impacts this scaling could have on results and policy implications. This paper will focus on these implications.

First, we provide a review of how the question of scaling more broadly is approached in the literature. The remainder of this section will review a subset of ABMs, both relating to the pandemic and more broadly, that have used some of the approaches to scaling outlined by Bithell and Parry [3]. The first set of approaches opts to address computational needs by simply increasing computational power. However, due to both expense and technical expertise needed, such approaches can be out of the reach of all but the most high capacity, well funded stakeholders. The context of a global pandemic revealed the demand for modelling approaches optimised for lower capacity contexts also. We will address both approaches in this literature review, and explore points of caution when applying scaling.

### 2.1. Computing power solutions

One way of speeding up computationally intensive ABMs is to simply increase the computational power available to the simulation. Typically, this is achieved either through specialised hardware or through carefully fine-tuned parallelisation approaches. For example, Hoertel et al. [5] sought to model the potential impact of post-lockdown measures in France. Simulating a demographically data-rich population, their model uses Intel TBB libraries for multi-threading (and thus parallelisation) to improve performance. Another model developed by Aleman et al. [11] consists of almost five million individuals in 1.8 million households in the Greater Toronto Area. They were able to run 1000 simulations in approximately 13 min when using their parallel computing infrastructure. Parker & Epstein [12] built a platform of disease transmission called the Global-Scale Agent Based Model (GSAM), which enables the simulation of a population of several billion agents in Java and application to any disease; this platform makes use of multi-threading and increasingly available multi-core machines to deliver such speedups, although the code used to achieve this has not been made open source.

In addition to parallelisation, other methods that are CPU oriented (e.g. [13]) use General-Purpose Computing on Graphic Units [14,15], or use dynamic load balancing across HPCs [16] where possible. In each of these cases, either the code was not made available online for others to use (a common problem; see [17]) or the modellers made use of specialised hardware in order to make the models tractable. As such, this puts these solutions outside the reach of many interested users.

### 2.2. Complexity reduction

A more cost effective, accessible and less computationally intensive alternative to increased computational power is the simplification of complexity in order to reduce the computational load. Some studies dealing with particularly complex behaviour in space choose to work with a small scale population. For example, in the case of a different study by Perez & Dragicevic [7] the researchers were unable to simulate the entire population of the city for computational reasons. The model implementation instead used 1000 individuals involved in a measles epidemic and interacting at a city scale. Arguably, this is not a realistic representation of the dynamics that would transpire were the entire city population taken into account.

### 2.3. Hybrid approaches

One way to simplify complexity and thus speed up individual based models is to adopt a hybrid approach. Hunter and MacNamee [6] adopt such an approach by combining equation based and agent based modelling. After modelling a disease outbreak in both a small town and a county in Ireland, they allow the model to switch between the ABM and the equation based model (EBM) depending on the number of agents infected in an area. Once the threshold of the number infected is passed, the model will automatically switch from agent-based to equation-based. They argue that the advantage of the hybrid approach is to allow for further scaling and modelling of a larger population while still keeping a heterogeneous population. They point out that it is important to decide in which parts of the model the fidelity can be reduced and made equation-based, lest the model lose performance. While an equation-based disease model can save time and computing power, it misses capturing individual agents' actions and the importance of contacts and different contact patterns between agents in the spread of a disease. Further, it makes it impossible to explore how adoption of non-pharmaceutical interventions might vary by group or place, robbing the ABM of one of its great advantages.

### 2.4. "Super agents"

A compromise approach that tries to balance the twin demands of model fidelity and tractability is to make use of the aforementioned "super agents". This allows modellers to both preserve a degree of low-level heterogeneity and to avoid troubleshooting the challenges of tailoring a simulation for multi-threading or specialised hardware. The Covasim model by Kerr et al. [4] uses the idea of dynamic scaling in

conjunction with the super agents. When the epidemic is small, one agent corresponds to one person. Then, when a certain threshold is reached, the non-susceptible agents are down-sampled and a corresponding scaling factor is introduced. As the epidemic expands, the process is repeated iteratively until the scale factor reaches its upper limit. They state that "this enables arbitrarily large populations to be modelled, starting from a single infection and maintaining a constant level of precision and computational time throughout"[4]. The scaling factors themselves can be set by the user. This technique can also be used outside of epidemic contexts — see for example [18], where super agents stand in as the "leaders" and decision-makers for larger groups of individuals.

This "super agent" approach has proved popular, and may be a good solution to the challenges outlined, especially in a context where the data, software, hardware or team capacity may not be available. However, it assumes that the scale of the input population each "super agent" represents does not matter. Ben-Dor et al. [9] used a traffic model to illustrate that scaling could in fact make a substantial difference to results from the full-scale model. When testing simulations of the Sioux falls test case road network with MATSim, they note that within the 10%–25% interval, downscaling could become unstable for some statistics. For samples below 10%, they illustrate that statistics can vary substantially from the full scale model. This paper is one of a limited selection that explores the importance of scale in the super-agent approach, and this is precisely the motivation behind the contribution of this paper.

### 2.5. The unspecified route

Although a few approaches to scaling have been described in this section, many models fail to document the population sizes or scaling approaches used in the model at all. Systematic reviews of the literature by Hazelbag and Willem have documented this [19,20]. In a full text review of 265 articles, Hazelbag notes that the size of the simulated population was explicitly mentioned in only 54 (64%) of the 84 included articles; for another 9 (11%) articles it was possible to deduce this number from either text or figures. The remaining 21 (25%) articles either provided incomplete information (3/21) or no information at all (18/21). For the 63 (75%) articles for which they obtained a number, the median population size was 78,000 (range: 250–47,000,000). This goes to show how some models attempt to avoid the challenges around scaling up altogether [19].

In this paper, we aim to illustrate the potential bias that can arise when scaling is not explicitly addressed, showing how this bias may also vary with the complexity of the model. It is known that nonlinearities in complex network interactions can lead smaller populations to produce qualitatively different outputs than those of larger populations; we quantify this in the context of a spatial network.

## 3. Methodology

In this paper we use the example of an ABM simulating the spread of the COVID-19 pandemic in Zimbabwe to illustrate the role of super individuals and the relative sample scale on the model's outputs. We will refer to the number of individuals modelled out of the full population as the **sample size**, where the number of people being represented by each super individual is determined by this sample size. For example, in a sample of 5% of the population, each super individual represents 20 people; in the 50% sample, each super individual represents only two people.

We will present a very brief overview of the model and its functioning using the ODD Protocol [21]. A full description of the model will be documented in a forthcoming separate paper.

### 3.1. Overview

The **purpose** of the model is to forecast the spread of an infectious disease throughout a population of spatially distributed humans.

The **entities** in the model are either individual humans or infections. The human agents, called **Persons**, are characterised by their age and sex. Persons are assigned to households — physical locations where a consistent, designated group of Persons gather in the evenings. They move around their communities, sometimes travelling to other nearby districts, interacting with other people and potentially transmitting infections to one another. **Infections** represent a case of the given disease. An Infection must be assigned with a host Person, and may progress over time based on the age of the host.

The model **environment** represents space. Persons can be located within either household or community locations, both of which are situated within districts (larger spatial units which together make up the country of Zimbabwe). A Person who is visiting a district outside of their own home district must be out in the "community" - in their own home district, they may either be out in the community or else in their own household. Persons make decisions every 4 h and interact with others based on their location.

The **processes** represented in this model include movement and infectious behaviours. Individual Persons choose whether to go out into the community every day; they may visit the community of their own district or may travel to another district and visit the community there. If they have an Infection, they will potentially prompt the Persons with which they interact to contract their own Infections. Infections develop over time, developing from the initial exposure all the way to either the recovery or death of the host. In advanced stages, the Infection may render the host immobile, preventing them from moving.

### 3.2. Design

The design of the model allows for the emergence of local outbreaks and hotspots within target districts. Interactions between agents give rise to the spread of disease, and the movement of Persons among districts allows for the disease to spread between otherwise relatively closed communities.

### 3.3. Details

The **initialisation** of the model is significant because after the populations have been generated in their target households and districts, a set number of Infections are generated in hosts in the target districts. The hosts are randomly chosen for each instantiation of the simulation.

The input data, being of particular interest in this paper, has been specified in its own section, Section 4. The submodels are simply the movement module and the infection module. In the former, Persons choose whether to leave the house with a pre-determined probability; if they choose to leave, they will select a target destination based on the movement matrix described in Section 4. Their movement pattern will also depend on the model, to be described in **Model Versions and Population Scales**. If they move to a community, they will interact with some set number of individuals present in the same community, potentially prompting an Infection in them. At the end of 8 h in the community, they will return home and interact with those in their household.

The infection behaviour sees the Infection transitioning from the state of initial exposure to being, potentially, either symptomatic or asymptomatic. In the asymptomatic case, the Infection will resolve into recovery after a stochastic period of time; in the symptomatic case, the Infection may either resolve into recovery or transition to a mild case. It may continue along this path, at each stage either recovering or worsening into a severe case, a critical case, and ultimately death. An individual Person who has recovered becomes susceptible to new Infections. If the Infection progresses beyond the stage of being mild, the Person is set to be immobile and restricted to their household.

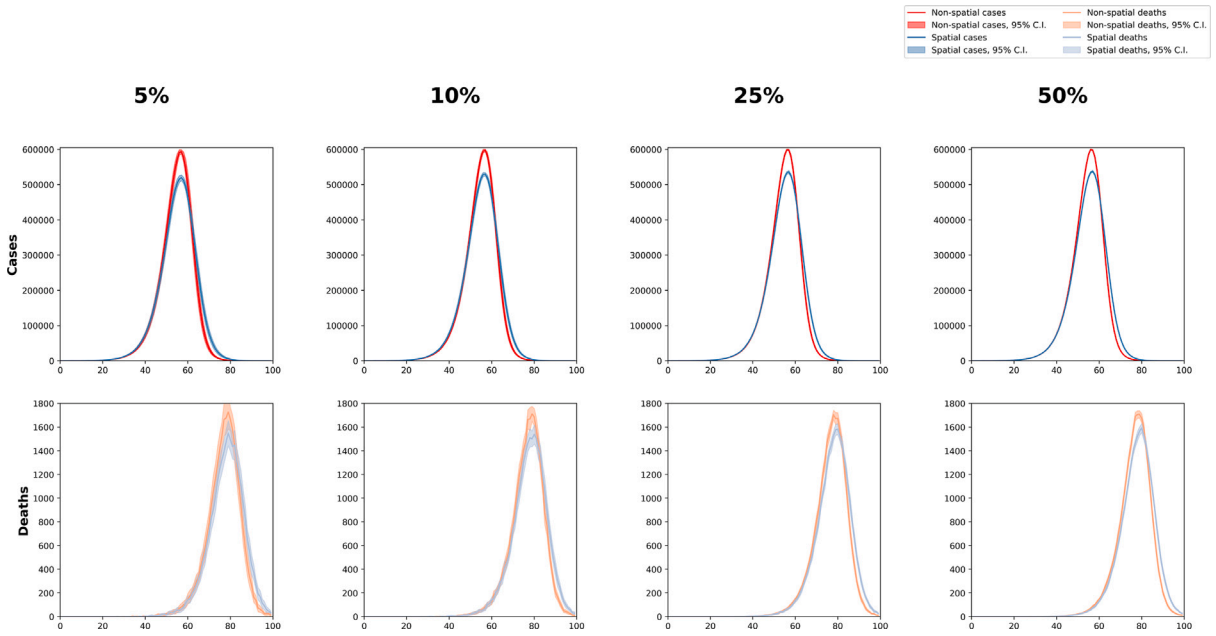


**Fig. 1.** Each model's predicted number of cases and deaths at different population scales. This figure shows the predicted number of cases, deaths and corresponding 95% confidence intervals for the non-spatial and spatial models. The first row illustrates the number of cases and the second the number of deaths. Here we see that as the model population size increases, the variability in the estimated number of cases and deaths decreases.

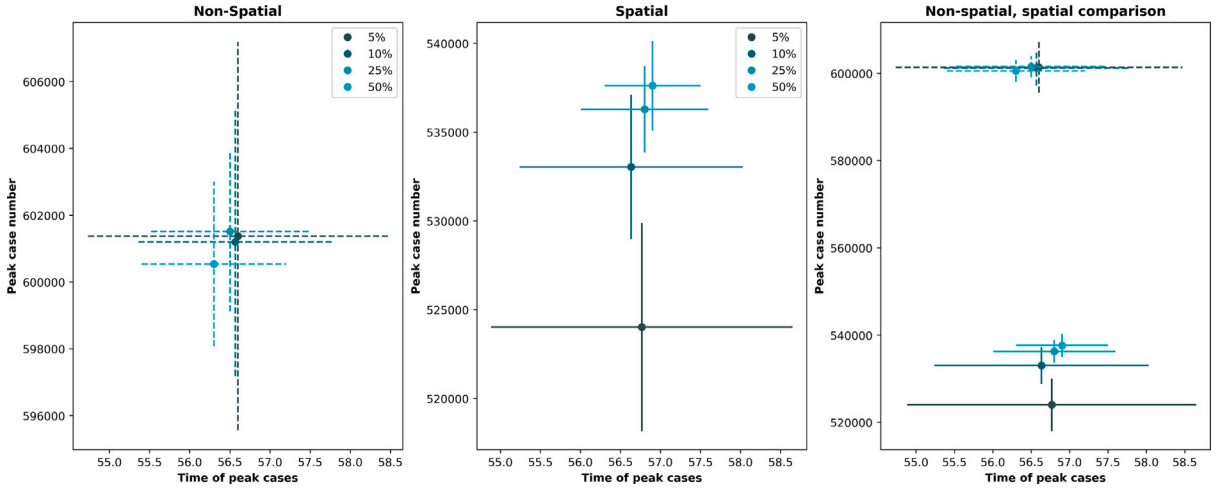


**Fig. 2.** The mean peak epidemic size and the mean time at which the peak occurred for the various sample sizes, along with their respective 95% confidence intervals. The non-spatial form of the model shows no significant difference among the peak number of cases, whereas the spatial form of the model predicts a significantly smaller maximum case number in the 50% compared to the 10% and 5% samples, but not the 25% sample. In both the spatial and non-spatial forms of the model, the time at which the peak epidemic size occurred did not differ at a statistically significant level.

### *3.4. Model versions and population scales*

In order to assess how the complexity of the model interacts with the size of the sample, we generate two versions of the model described above, then test each one with 4 sample sizes, i.e. 8 combined variations in total.

- **A Non-Spatial Model** - individuals have representative ages and live in households of an appropriate size. Everyone exists in the same geographic location, so when in the community, every Person has an equal probability of interacting with every other Person.
- **A Spatial Model** - in the same way as the Non-Spatial Model, individuals have ages and household sizes drawn from real data. However, in the Spatial Model, households are assigned to sixty individual districts (with population sizes equivalent to the 60 districts of Zimbabwe) to create a spatially reasonable distribution of population across the country.

For each model, we create a 5% sample, a 10% sample, a 25% sample, and a 50% sample, where the percentage corresponds to the share of the size of the total 2012 Census population of Zimbabwe. For the 5% sample we simply use the 5% census sample we have for

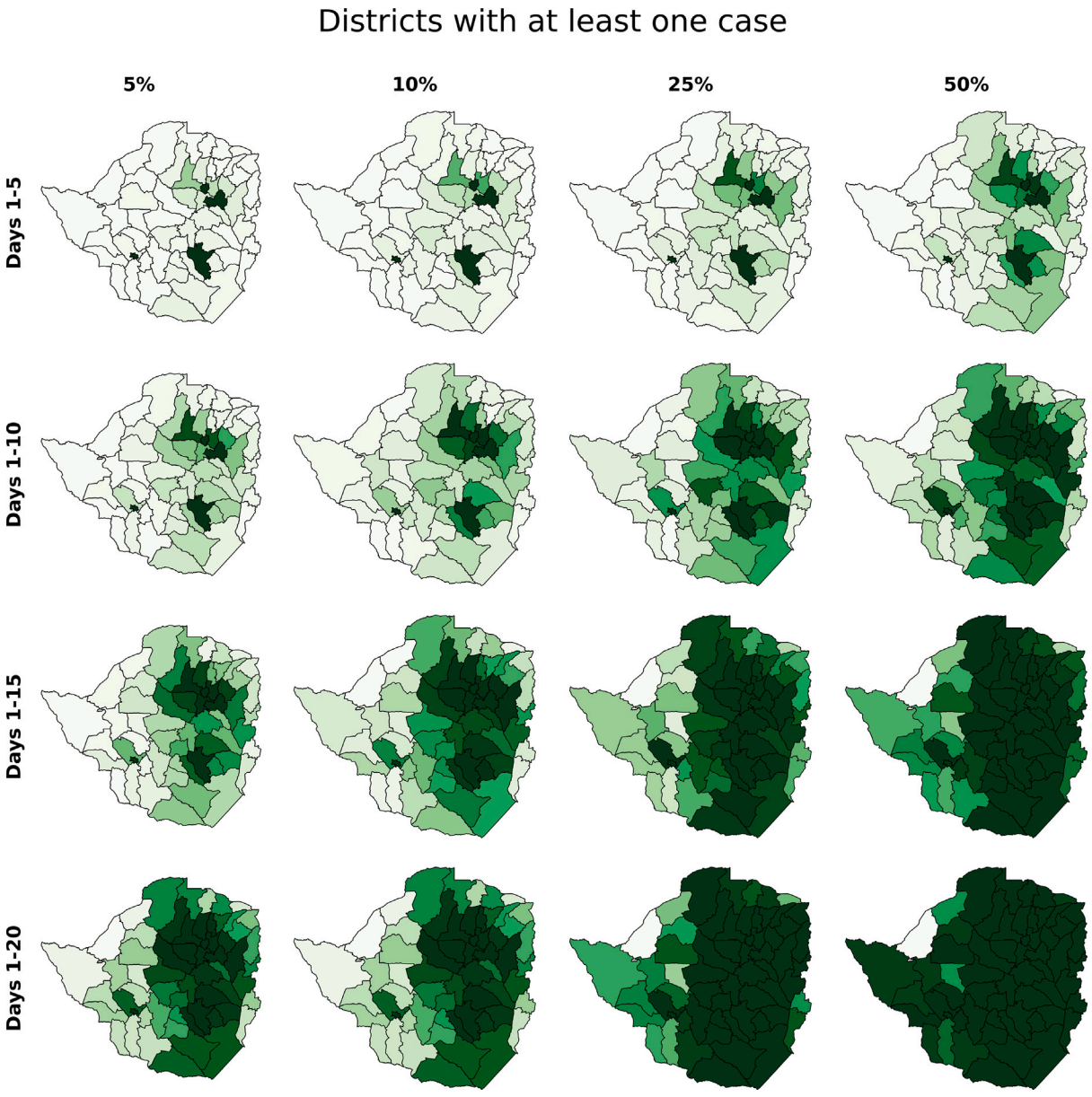


**Fig. 3.** The spread of the COVID-19 epidemic in time and space for different population sample sizes ($\beta = 0.3$). The shading of each district refers to the fraction of runs in which districts had cases after 5, 10 and 20 days respectively, based on 30 runs of the simulation.

the population described in **Data**, using the characteristics provided. This includes 654,688 agents in 160,728 households. To get the other percentages, we expand the original 5% sample by generating identical replicas of each household. For example, in the 10% sample, every agent from the 5% sample is duplicated, so that there are two people with the exact same characteristics, for a total of 1.3 million agents over 321,456 households. The 25% sample includes 5 replicas of each agent, and so on. Therefore, differences that arise between the simulation results can be more easily attributed to the increased size of the sample. When increasing the sample size, there is no additional information added into the model since the extra agents are clones of those in the 5% sample. This is the same concept as that of "super-agents" as described in the literature review, but since we are adding these replicated individuals into the simulation, it allows the agents that are otherwise exactly the same to end up in different situations. Evidently, these are somewhat contrived populations but this should be understood as a mechanism for testing rather than a realistic census distribution, enabling us to run a cleaner experiment.

## 4. Data

The model makes use of a number of different forms of data to motivate and contextualise the simulation, including:

- ***Census Data*** The 2012 Zimbabwe Census was taken from IPUMS International. The data that is available is a 5% sample of the original census of the 13 million individuals who lived in Zimbabwe at that time. This data contains information on the age, sex, economic status, household ID, and district of origin of every person within this 5% sample (a total of 654,688 individuals).[1]
- ***Mobility Data*** Mobility data was calculated from approximately 8.1 billion Call Detail Records (CDRs) and reflect the levels of mobility as monitored from and to each of Zimbabwe's 60 districts.

[1] The population has since grown to approximately 14.86 million, according to most recent estimates. Therefore if we were to scale up to the full population of 2022, the original sample would in fact have to be expanded 23 times.

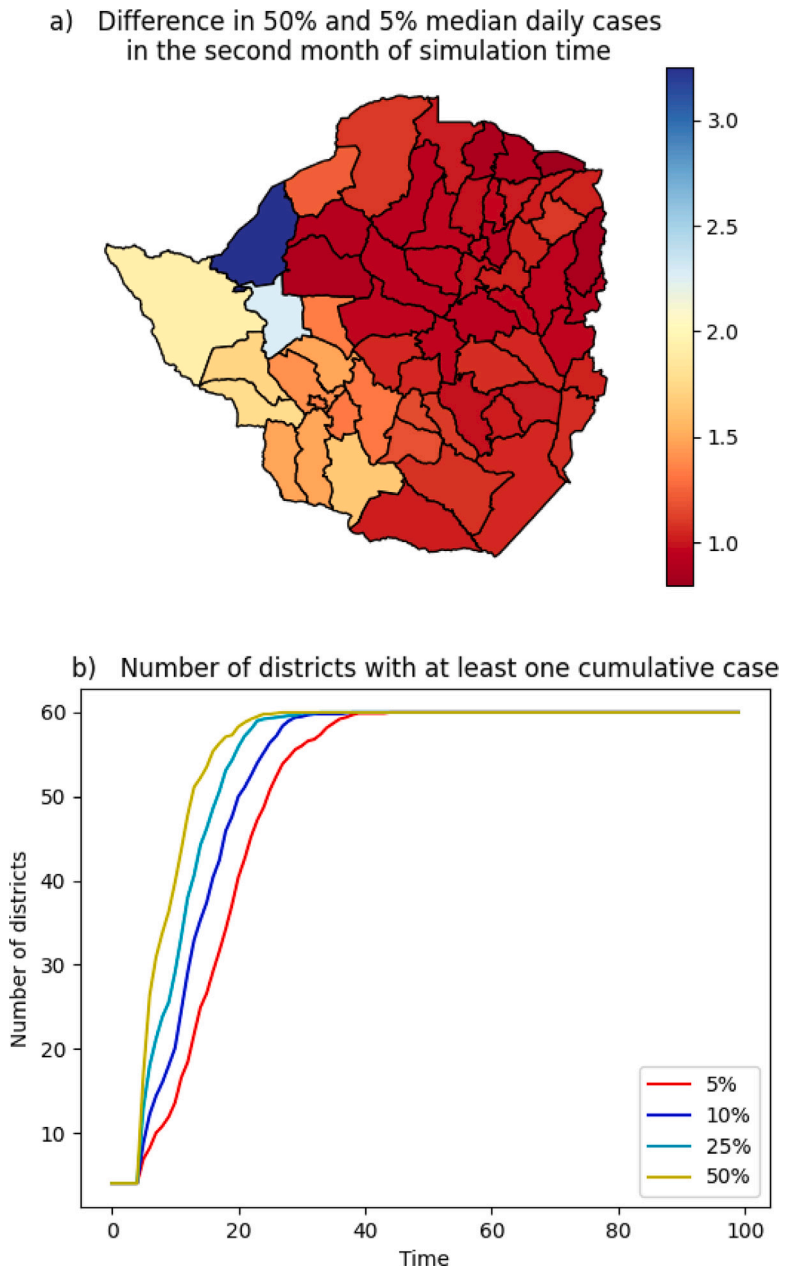


**Fig. 4.** The temporal spread of the simulated epidemic varies with respect to sample size. Subplot a) shows the average spatial distribution of the ratio in median daily cases between the 5% and 50% population sizes for days 30–60 in the simulation. Subplot b) shows the cumulative number of districts with COVID-19 cases for each population size. Larger model population sizes consistently resulted in a faster spread in the number of districts with COVID-19 cases ($\beta = 0.3$).

The detailed data were aggregated by a telecom operator into daily origin and destination matrices using code developed by the research team (see [22]). Only the aggregated, fully anonymised output was shared by the telecom company with the research team. The daily data was further aggregated into an Origin Destination Matrix showing trips between two districts, relative to day of the week by averaging values across dates falling on the same day of the week. This was used to produce an average probability that someone in a given district travels to each of the other districts in the country on a given day of the week.

- ***Epidemiological Data*** A series of parameters to define the characteristics of infection were input into the model to establish the infection dynamics. These include age-specific susceptibility to transmission, hospitalisation, critical cases, and death. The characteristics of SARS-CoV-2 such as the incubation period, infectious period, and recovery times were also included and taken from the Covasim model which in turn were taken from Ferguson et al. (2020) [23] and Verity et al. (2020) [4,24].
- ***Case data*** The aggregate district level case numbers from March 2020 provided by the Ministry of Health of Zimbabwe to the World Bank representing the number of cases. These are used to inform seeding of cases in the districts in the Spatial Model presented here. In the 5% population sample, we seeded 12 cases across four districts (districts Bulawayo, Harare, Marondera and Masvingo). The number of cases in each district was then scaled up to its corresponding input population size (for a total of 24 cases in the 10% sample, 60 in the 25% sample, and 120 in the 50% sample). The overall number of cases was equivalent for the Non-Spatial Model.

These data sources were combined and synthesised in order to support our modelling efforts.

## 5. Results

To understand how the size of the simulated population impacts the output of the model, we began by exploring the epidemic curves of a range of population sizes, in terms of both daily new cases and daily new deaths. Each of the two models is run with each of the synthetic populations 30 times. The outputs are scaled up directly to the full population (e.g. for the 5% sample we use a factor of 20 and for the 25% sample we use a factor of 4) to facilitate comparison. Unless otherwise mentioned, the $\beta$ value used for the simulated infection is 0.3 - kept purposefully high for easier comparison.

Fig. 1 shows that, for both the Non-Spatial and Spatial models, the model produces an epidemic curve as expected. Smaller populations inflate the impact of stochasticity and therefore have slightly more varied results, while the larger, higher fidelity models show greater precision. Deaths, being more rare than cases, exaggerate this effect; they also lag cases by an appropriate time period. The height and shape of these curves, across population sizes, is very consistent. These are all in keeping with our expectations and support the conclusion that the model is correctly capturing the dynamics of an epidemic outbreak.

### 5.1. The epidemic peak

While both versions of the model recreate the appropriate epidemic curve, the Spatial model has a noticeably lower peak than that of the Non-Spatial model. Further, this peak appears to occur very slightly later in the Spatial model than it does in the Non-Spatial model. To more precisely investigate the impact of spatiality on the time to peak case numbers, we calculated the mean and standard deviation of time to the epidemic peak and the mean peak case number respectively for each sample size, Fig. 2. We find that in the Non-Spatial model, the maximum case number and time at which it occurs is not significantly influenced by sample size. In contrast, the maximum case numbers of the Spatial model are significantly different between the 50% and both the 10% and 5% samples. The time at which the peak infections occurs is not significantly different for different sample sizes.

### 5.2. Spatial distribution of cases

It is worthwhile to consider also the spatial distribution of cases. Fig. 3 reflects how the finer-grained populations capture the spread of disease to a wider range of districts. While the aggregate case counts can seem quite similar, the spatial distribution of these cases may be very different. Fig. 4 shows a comparison of these parameters, with Fig. 4a highlighting the fact that in the more remote areas, the smaller simulated populations consistently underpredict the presence of cases in remote districts during the early period of a pandemic. Fig. 4b gives a temporal sense of this delay.

### 5.3. Varying $\beta$

Another consideration was whether the importance of population size varied relative to the $\beta$ of the infection. Fig. 5 shows the spread of disease through each of the synthetic populations at different $\beta$ values. Because smaller values of $\beta$ so often lead to epidemics apparently dying out, the impact of population is most immediately obvious in the $\beta = 0.3$ case we have highlighted previously. Note the presence of cases in unseeded districts for the lower values of $\beta$ with the 50% population, in particular the extremely noticeable difference between final results for $\beta = 0.03$. Thus, the lower the $\beta$, the more population fidelity matters.This point is further illustrated in Fig. 6, which shows the average number of districts with COVID-19 cases after 20 days for different values of $\beta$ and sample sizes. As sample size increases, the average number of districts with COVID-19 cases also increases for the same $\beta$ value. Note the relatively log-linear increase associated with the 5% population, in comparison to the increasingly log-superlinear trendlines of the larger samples.

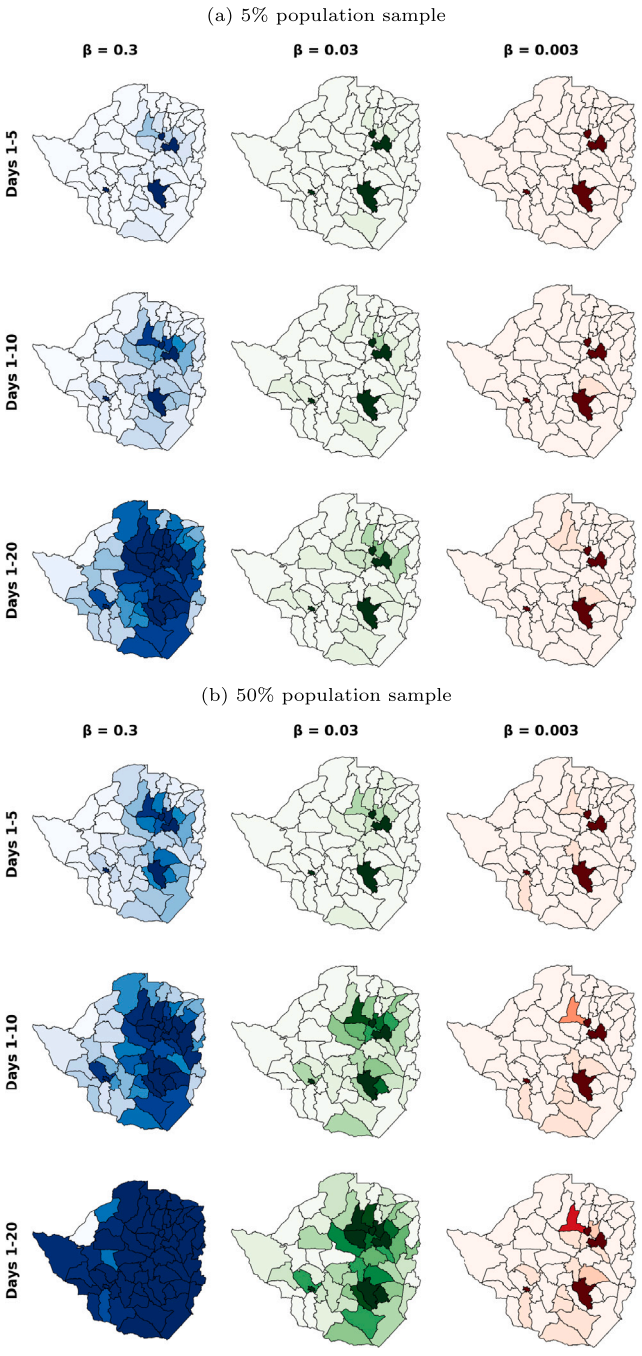


**Fig. 5.** The average spatio-temporal spread of COVID-19 for varying values of $\beta$ and differing population sizes. The shading of each district refers to the fraction of runs where districts had cases after 5, 10 and 20 days respectively, based on 30 runs of the simulation.

## 6. Discussion

The question of why the Spatial and Non-Spatial models differ in terms of epidemic peak size is worth discussing in greater detail. Intuitively, we know that the number of new infections depends on the ability of infected persons to interact with uninfected persons. The greater the number of uninfected persons an infected person can access, the greater the potential increase will be and the higher the epidemic peak. The spread of the disease is constrained only by time or by a lack of new hosts.

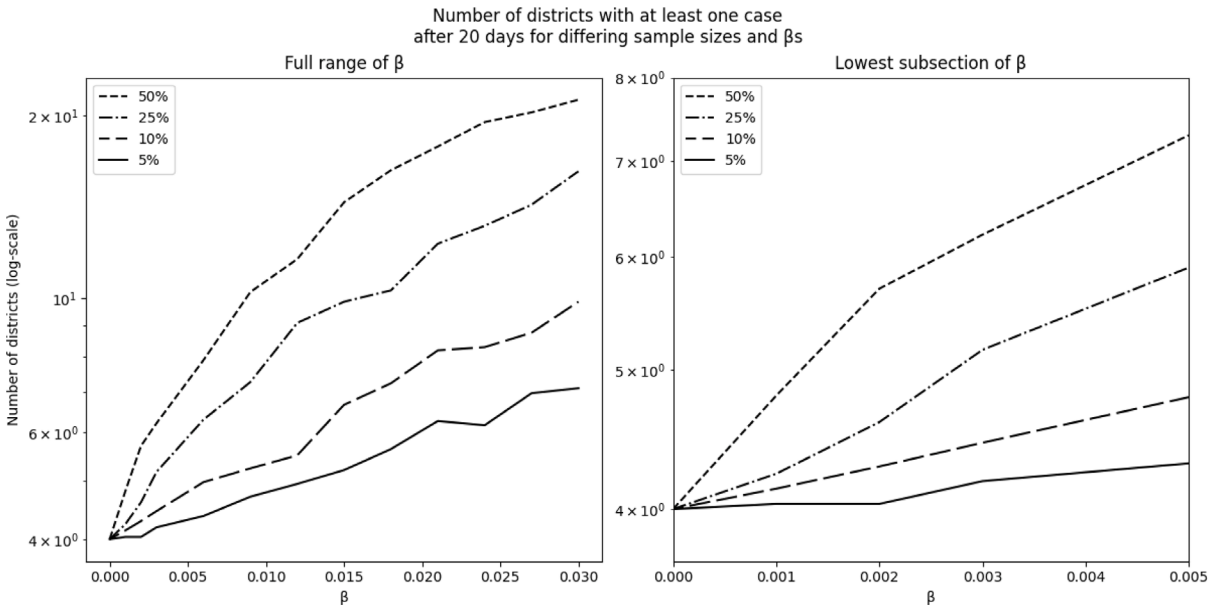


**Fig. 6.** The average number of districts with COVID-19 cases after 20 days for differing population sample sizes. As sample size increases, the average number of districts with COVID-19 cases increases for the same value of $\beta$. Note that the only value of $\beta$ where we do not see a difference in outcome is $\beta = 0$, where the infection cannot spread and only the initial infections are present.

In the Spatial version of the model, individuals interact with a proper subset of the overall population. Thus, every set of candidate interactions the agent can possibly draw in the Spatial model is also possible for the candidate to draw in the Non-Spatial model. There are no "outbreak events" possible in the Spatial model that are impossible in the Non-Spatial model. There is no way for an agent in the Spatial model to reach some number of new potential hosts that are inaccessible to the same agent in the Non-Spatial model, while the opposite is untrue. Thus, the spread of the disease is unconstrained in the Non-Spatial model, and the epidemic continues unimpeded. In the Spatial model, agents may eventually move to locations where they are exposed to the disease, but barring the vanishingly unlikely case in which every agent travels to the same district at once, the immediate availability of new hosts will always be smaller in the Spatial case. This limits the fastest increases in growth seen in the Non-Spatial version, resulting in the relatively lower, flatter curve of the Spatial model.

Of particular interest, too, is the intersection between the spatial spread of the disease and the $\beta$ value, relative to the actual implications of the model results. Smaller sample populations might lead researchers to conclude that certain areas are highly unlikely to have cases — leading them to wrongly conclude that some other aspect of their model is incorrectly calibrated, should the validation data report cases in the areas. It would be easy for a researcher to assume that their estimate of $\beta$ was too low and to adjust it. More sophisticated statistical tests could potentially be deployed to avoid this kind of error, but it is important modellers recognise these risks.

## 7. Conclusion

Agent-based modellers have historically lacked firm references or benchmarking showing how population size, the inclusion of space, or the setting of parameters can combine to impact their modelled outcomes. This work attempts to begin a conversation about the need for this, especially as more and more simulators attempt to apply this methodology to time-sensitive health questions.

We demonstrate that researchers must take care in selecting the scale of the population sample in their models, particularly if there is interest in understanding the initial phases of a pandemic when case counts and death numbers are low. Importantly, when there is a spatial component to the model, the ultimate size of the epidemic is smaller — but more widely spread. Running the model more times or with a bigger population will not correct for this difference; researchers should be aware of this when they design their work or try to compare it with the findings of others.

Small samples also obviously lead to higher uncertainty. This is not only in the early stages, but for the entire curve. Given that there is already a high level of uncertainty in these epidemiological models due to the large number of assumptions that are made, the added uncertainty from a small sample size may reduce the reliability of such a model for policy planning. For low values of $\beta$, they may change the nature of the dynamic altogether, missing out on situations which have the potential to become low-level epidemics.

From a policy perspective, there is interest in understanding when a disease might spread to a new geographic area. This geographic analysis can be one of the important advantages of an agent based model, which allows for the simulation of how different agents might move between areas. Yet, we see that if the sample used is very small, it may not accurately portray the timing of when a disease will spread to a new area. This is especially true for diseases with lower $\beta$ values. This is important from a policy perspective since identifying when a disease might first enter an area is important for mitigation strategies and planning. On the other hand, finer grained populations allow the larger populations to spread more widely, meaning that a low-fidelity model might fail to identify locations where cases might reach. The presence of cases in the gold standard or verification dataset might conflict with such simulated spreads, and the researchers conclude that finding a case in one district means that we have wrongly set our $\beta$ lower than we should, when in fact it may instead be that our simulated super individuals are being asked to represent too many of their peers.

We have laid out these variations in light of the simplest possible epidemic model, which includes no non-pharmaceutical interventions of any kind or specialised movement patterns for individuals (e.g. long-range truckers or flight attendants). Future work should explore how these vary in light of travel restrictions, variable adoption of hand-washing or masking, targeted vaccination campaigns, and so forth. Given that resources and cultural practices vary noticeably by location, the ability of agent-based models to capture the heterogeneity of human behaviour may prove crucial to understanding epidemics.

Overall, trade-offs between model speed and fidelity are increasingly relevant to researchers and we hope this work can contribute to both discussion and awareness of them.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Almost all data will be made available, including a dummy dataset for the data which we are not permitted to share. These will be available on the project GitHub openly.

## Acknowledgements

This work was supported by the UKRI Medical Research Council MR/T02075X/1. The project benefited from financial contributions from the United Kingdom Foreign, Commonwealth & Development Office (FCDO) through the ieConnect for Impact Program, the Trust Fund for Statistical Capacity Building III (TFSCB-III), and the Research Support Budget in the Development Economics Vice-Presidency.

The findings, interpretations and conclusions expressed in this paper do not necessarily reflect the views of the World Bank, the Executive Directors of the World Bank or the governments whom they represent. The World Bank does not guarantee the accuracy of the data included in this work.

## References

[1] Sarah Wise, Sveta Milusheva, Sophie Ayling, The importance of scaling for an agent based model: An illustrative case study with COVID-19 in Zimbabwe, in: Computational Science – ICCS 2022: 22nd International Conference, Springer, 2022.

[2] John H. Miller, Scott E. Page, Complex Adaptive Systems, Princeton University Press.

[3] Hazel R. Parry, Mike Bithell, Large scale agent-based modelling: A review and guidelines for model scaling, in: Alison J. Heppenstall, Andrew T. Crooks, Linda M. See, Michael Batty (Eds.), Agent-Based Models of Geographical Systems, Springer Netherlands, Dordrecht, 2012, pp. 271–308.

[4] C.C. Kerr, Robyn M. Stuart, Dina Mistry, Romesh G. Abeysuriya, Katherine Rosenfeld, Gregory R. Hart, Rafael C. Núñez, Jamie A. Cohen, Prashanth Selvaraj, Brittany Hagedorn, et al., Covasim: an agent-based model of COVID-19 dynamics and interventions, PLoS Comput. Biol. 17 (7) (2021).

[5] N. Hoertel, Martin Blachier, Carlos Blanco, Mark Olfson, Marc Massetti, Marina Sánchez Rico, Frédéric Limosin, Henri Leleu, A stochastic agent-based model of the SARS-CoV-2 epidemic in France, Nat. Med. 26 (9) (2020).

[6] E. Hunter, B. Mac Namee, J. Kelleher, A hybrid agent-based and equation based model for the spread of infectious diseases, J. Artif. Soc. Soc. Simul. 23 (4) (2020).

[7] L. Perez, Suzana Dragicevic, An agent-based approach for modeling dynamics of contagious disease spread, Int. J. Health Geogr. 8 (1) (2009).

[8] R. Minson, Georgios K. Theodoropoulos, Distributing RePast agent-based simulations with HLA, Concurr. Comput.: Pract. Exper. 20 (10) (2008).

[9] G. Ben-Dor, Eran Ben-Elia, Itzhak Benenson, Population downscaling in multi-agent transportation simulations: A review and case study, Simul. Model. Pract. Theory 108 (2021).

[10] M. Scheffer, J.M. Baveco, D.L. DeAngelis, Kenneth A. Rose, E.H.s. van Nes, Super-individuals a simple solution for modelling large populations on an individual basis, Ecol. Model. 80 (2–3) (1995) 161–170.

[11] D.M. Aleman, Theodorus G. Wibisono, Brian Schwartz, A nonhomogeneous agent-based simulation approach to modeling the spread of disease in a pandemic outbreak, Interfaces 41 (3) (2011).

[12] J. Parker, Joshua M. Epstein, A distributed platform for global-scale agent-based models of disease transmission, ACM Trans. Model. Comput. Simul. 22 (1) (2011).

[13] N. Fachada, Agostinho C. Rosa, Assessing the feasibility of OpenCL CPU implementations for agent-based simulations, in: Proceedings of the 5th International Workshop on OpenCL, 2017.

[14] E. Hermellin, F. Michel, GPU delegation: Toward a generic approach for developing MABS using GPU programming, in: AAMAS: Autonomous Agents and Multiagent Systems, ACM, 2016, pp. 1249–1258.

[15] W. Tang, Meijuan Jia, Global sensitivity analysis of a large agent-based model of spatial opinion exchange: A heterogeneous multi-GPU acceleration approach, Ann. Assoc. Am. Geogr. 104 (3) (2014).

[16] C. Márquez, Eduardo César, Joan Sorribes, Graph-based automatic dynamic load balancing for HPC agent-based simulations, in: European Conference on Parallel Processing, Springer, 2015.

[17] Karel Mls, Milan Kořínek, Kamila Štekerová, Petr Tučník, Vladimír Bureš, Pavel Čech, Martina Husáková, Peter Mikulecký, Tomáš Nacházel, Daniela Ponce, Marek Zanker, František Babič, Ioanna Triantafyllou, Agent-based models of human response to natural hazards: systematic review of tsunami evacuation, Nat. Hazards (ISSN: 0921-030X) (2022) http://dx.doi.org/10.1007/s11069-022-05643-x, URL https://link.springer.com/10.1007/s11069-022-05643-x.

[18] A. Sharpanskykh, Jan Treur, Group abstraction for large-scale agent-based social diffusion models with unaffected agents, in: International Conference on Principles and Practice of Multi-Agent Systems, Springer, 2011.

[19] C.M. Hazelbag, Jonathan Dushoff, Emanuel M. Dominic, Zinhle E. Mthombothi, Wim Delva, Calibration of individual-based models to epidemiological data: A systematic review, PLoS Comput. Biol. 16 (5) (2020).

[20] L. Willem, Frederik Verelst, Joke Bilcke, Niel Hens, Philippe Beutels, Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006–2015), BMC Infect. Dis. 17 (1) (2017).

[21] V. Grimm, S.F. Railsback, C.E. Vincenot, U. Berger, C. Gallagher, D.L. Deangelis, B. Edmonds, J. Ge, J. Giske, J. Groeneveld, A.S.A. Johnston, A. Milles, J. Nabe-Nielsen, J.G. Polhill, V. Radchuk, M.S. Rohwäder, R.A. Stillman, J.C. Thiele, D. Ayllón, The ODD protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism, J. Artif. Soc. Soc. Simul. 23 (2) (2020).

[22] Sveta Milusheva, Anat Lewin, Tania Begazo Gomez, Dunstan Matekenya, Kyla Reid, Challenges and opportunities in accessing mobile phone data for COVID-19 response in developing countries, Data Policy 3 (2021) e20.

[23] N.M. Ferguson, Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma Cucunubá, Gina Cuomo-Dannenburg, et al., Report 9: Impact of Non-Pharmaceutical Interventions (NPIs) to Reduce COVID19 Mortality and Healthcare Demand, Imperial College London, 2020.

[24] R. Verity, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick GT Walker, Han Fu, et al., Estimates of the severity of coronavirus disease 2019: a model-based analysis, Lancet. Infect. Dis. 20 (6) (2020).