The importance of Scaling for an Agent Based Model: an illustrative case study with COVID-19 in Zimbabwe

Sarah Wise[1][0000−0002−5552−4747], Sveta Milusheva[2][0000−0002−4166−5477], and Sophie Ayling[1][0000−0003−1109−3970]

[1] Centre for Advanced Spatial Analysis (CASA), University College London, London, UK s.wise, sophie.ayling.10@ucl.ac.uk
[2] World Bank Group, Washington, D.C., USA
smilusheva@worldbank.org

Abstract. Agent-based models frequently make use of scaling techniques to render the simulated samples of population more tractable. The degree to which this scaling has implications for model forecasts, however, has yet to be explored; in particular, no research on the spatial implications of this has been done. This work presents a simulation of the spread of Covid-19 among districts in Zimbabwe and assesses the extent to which results vary relative to the samples upon which they are based. It is determined that in particular, different geographical dynamics of the spread of disease are associated with varying population sizes, with implications for others seeking to use scaled populations in their research.

Keywords: Agent-based modelling · Scaling · Synthetic population · Agent-based modeling · Simulation

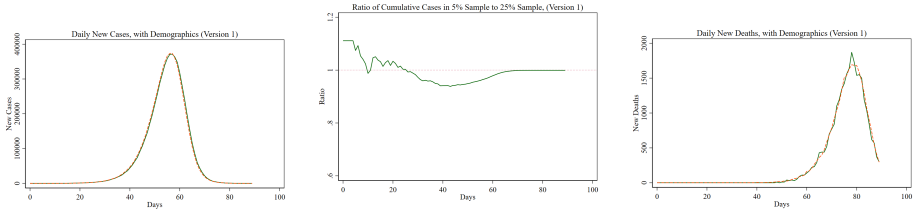
1 Introduction

Agent Based Models (ABMs) are often designed and built to model complex behaviours among large populations. However, the combination of complex behavior and a large population can require extensive memory or CPU usage and quickly become too computationally intensive to run efficiently in terms of speed or memory use. Thus, different methodologies have arisen to deal with the challenge of large-scale simulations in ABM literature (see Bithell and Parry [9] for a few of the most common). This is a point of particular interest as researchers have sought to lend a hand to the time-sensitive problem of disease forecasting (see for example [6], [4], or [5]).

Briefly, some models approach this problem by reducing the level of complexity of the model or the number of agents in order to enable it to run [10]. Others revert to equation-based modelling or hybrid approaches to reduce some of the burden in terms of computational intensity [5]. 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 each "super individual" to represent multiple agents, which risks the dynamics of a larger population not being reflected beyond a certain point - both spatially and temporally. [1]

None of these approaches are without drawbacks, and it is important to understand the trade-off decisions researchers are making. 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. 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.



(a) Daily New Cases (b) Ratio of 5% to 25% (c) Daily New Deaths

(d) Daily New Cases (e) Ratio of 5% to 25% (f) Daily New Deaths

Fig. 1: Comparison of 5% (orange dotted) and 25% (green solid) Sample for New Symptomatic Cases and Deaths Across Two Models. Note: Panels a-c use a model where individuals all live in one administrative unit. Panels d-f use a model where individuals can move between the different districts in Zimbabwe.

2 Methodology

In this paper we use the example of an agent based model simulating the spread of the COVID-19 pandemic in Zimbabwe to illustrate the role of sample size in the model outputs. We will present a very brief overview of the model and its functioning using the ODD Protocol [3].

2.1 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 hours 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 generate their own Infections. Infections develop over time, developing from being exposed all the way to either the recovery or death of the host. In advanced stages, the Infection may render the host immobile, disallowing them from moving.

2.2 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 between districts allows for the disease to spread between otherwise relatively closed communities.

2.3 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 3. The submodels are simply the movement module and the infection module. In the former, Persons choose whether to leave the house with some probability; if they choose to leave, they will select a target destination based on the movement matrix described in Section . 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 hours 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 being exposed to, potentially, either symptomatic or asymptomatic. In the latter case, after a stochastic period of time, the Infection will resolve into recovery; in the former 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.

3 Data

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

- ***Census*** The Zimbabwe Census data from 2012 was taken from IPUMS International. The data that is available is a 5% sample of the original census of 15 million individuals. This data contains information on the age, sex, economic status, household ID, and district of origin of every agent in the model.
- ***Mobility Data*** Mobility data was calculated from approximately 8.1 billion Call Detail Records (CDR) and reflect the levels of mobility as monitored from and to each of Zimbabwe's 60 districts. The detailed data were aggregated by a telecom operator into daily origin and destination matrices using code developed by the research team (see [7]. Only the aggregated, fully anonymised output was shared by the telecom company with the research team. The Origin Destination Matrix shows trips between two districts, relative to day of 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)[2] and Verity et al (2020)[11] [6].
- ***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 model version 3 (V3) presented here.

3.1 Synthetic populations

In order to assess how the complexity of the model interacts with the size of the sample, we generate two synthetic populations for the model:

- V1 - a sample of the population in which individuals have representative ages and live in households of an appropriate size. Everyone is in the same geographic location, so when in the community, every person in the sample can interact with every other person.
- V3 - as in V1, individuals have ages and household sizes drawn from real data. In V3, households are further assigned to individual districts to create a spatially reasonable distribution of population across the country.

For each version, we create a 5% sample and a 25% sample. For the 5% sample we simply use the 5% census sample we have for the population, using the characteristics provided. To get the 25% sample, we expand the original 5% sample by generating identical replicas of each household. Therefore, differences that arise between the two versions can be more easily attributed to the increased size of the sample. Obviously, it generates a somewhat contrived population - it should be understood as a mechanism for testing rather than a realistic census distribution.

4 Results

Each of the two model versions is run with the each of the synthetic populations ten times. The outputs are scaled up directly to the full population (for the 5% sample we use a factor of 20 and for the 25% sample we use a factor of 4) to facilitate comparison.

Looking across the trajectory of the disease, the numbers of new cases track fairly closely between samples (Figure 1a and d). However, it seems that adding geographic variation and mobility across districts leads to larger differences between the scaled 5% and 25% samples when it comes to other metrics (see Figure 1d-f). Note the death rates - while deaths are rarer events and their curves less smooth, the inclusion of mobility sees the number of new deaths drop consistently lower in the 25% sample than in the 5% sample.

In considering the difference between the scaled district cases across the two samples as a proportion of the scaled cases in the 25 percent sample, the differences are clear (2a). In the map, the areas in blue and green demonstrate districts where there are more cases registered in the 25% sample, as compared to orange areas where more cases were generated in the 5% sample. The 25% sample shows obvious cases where an individual has arrived in remote districts which are not reached by anyone in the 5% scenario (blue districts on the west side of the country and blue district in the north of Zimbabwe). These districts, which might be judged to be not at risk by the smaller sample, show up clearly in the 25% version of the simulation.

Additionally, if we focus on when districts have their initial case, we see that this happens much more quickly in the 25% sample, with many districts getting their first case before the 5% sample would predict (Figure 2b). In particular, while we seed initial cases in the same four districts for both the 5% and the 25% samples, we see that in the 5% sample, the disease does not spread beyond these

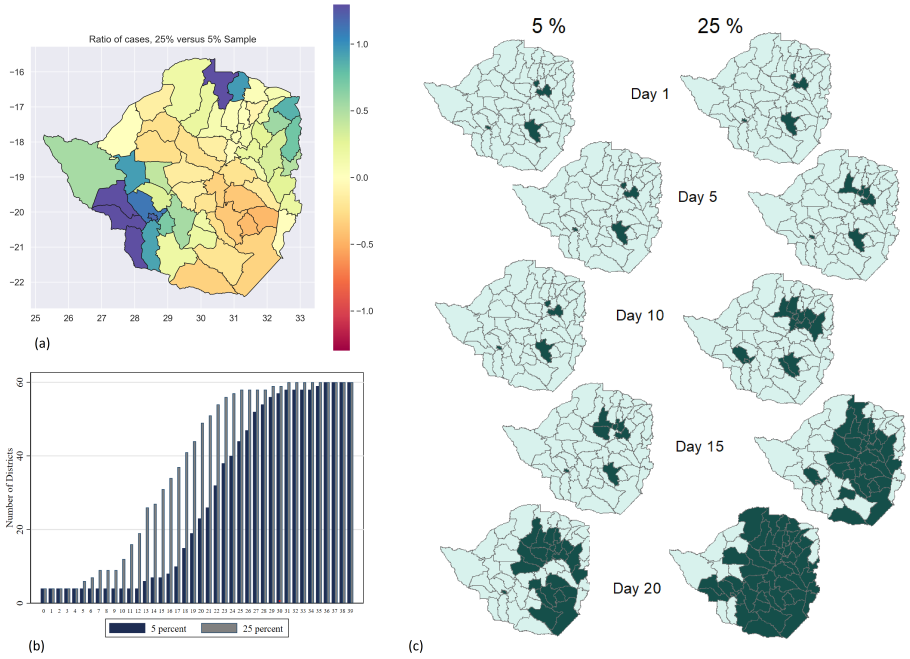


Fig. 2: Spatial differences when using different initial samples

(a) Difference between 25% and 5% median daily symptomatic case counts normalized by district cases for the second month of the V3 case (beta 0.3). (b) Number of districts with at least one cumulative case by day of outbreak. (c) Districts with at least one cumulative case by days 1,5,10 15, and 20 of the pandemic.

four districts until day 14. In contrast, with the 25% sample, the disease spreads to two new districts by day 6. This is visualized in Figure 2, which also illustrates that some of the first districts to which the disease spreads are different across the two samples.

5 Discussion and Conclusion

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 bias generated when using a small sample becomes much larger. The smaller sample leads to an overestimate of deaths across the entire time period.

Small samples also obviously lead to higher uncertainty in these models. 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.

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 important from a policy perspective since identifying when a disease might first enter an area is important for mitigation strategies and planning.

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

References

[1] G. Ben-Dor, Eran Ben-Elia, and Itzhak Benenson. "Population downscaling in multi-agent transportation simulations: A review and case study". In: *Simulation Modelling Practice and Theory* 108 (2021).

[2] N.M. Ferguson et al. "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. Imperial College London. 2020". In: (2020).

[3] V. Grimm et al. "The ODD protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism". In: *Journal of Artificial Societies and Social Simulation* 23.2 (2020).

[4] N. Hoertel et al. "A stochastic agent-based model of the SARS-CoV-2 epidemic in France". In: *Nature medicine* 26.9 (2020).

[5] E. Hunter, B. Mac Namee, and J. Kelleher. "A hybrid agent-based and equation based model for the spread of infectious diseases". In: *Journal of Artificial Societies and Social Simulation* 23.4 (2020).

[6] C.C. Kerr et al. "Covasim: an agent-based model of COVID-19 dynamics and interventions". In: *PLOS Computational Biology* 17.7 (2021).

[7] Sveta Milusheva et al. "Challenges and opportunities in accessing mobile phone data for COVID-19 response in developing countries". In: *Data Policy* 3 (2021), e20.

[8] R. Minson and Georgios K Theodoropoulos. "Distributing RePast agent-based simulations with HLA". In: *Concurrency and Computation: Practice and Experience* 20.10 (2008).

[9] Hazel R. Parry and Mike Bithell. "Large Scale Agent-Based Modelling: A Review and Guidelines for Model Scaling". In: *Agent-Based Models of Geographical Systems*. Ed. by Alison J. Heppenstall et al. Dordrecht: Springer Netherlands, 2012, pp. 271–308.

[10] L. Perez and Suzana Dragicevic. "An agent-based approach for modeling dynamics of contagious disease spread". In: *International journal of health geographics* 8.1 (2009).

[11] R. Verity et al. "Estimates of the severity of coronavirus disease 2019: a model-based analysis". In: *The Lancet infectious diseases* 20.6 (2020).