Calibration in a Data Sparse Environment: How Many Cases Did We Miss?

Robert Manning Smith
The Bartlett Centre for Advanced Spatial Analysis, University College London, UK

Sarah Wise
The Bartlett Centre for Advanced Spatial Analysis, University College London, UK

Sophie Ayling
The Bartlett Centre for Advanced Spatial Analysis, University College London, UK

Abstract

Reported case numbers in the COVID-19 pandemic are assumed in many countries to have underestimated the true prevalence of the disease. Deficits in reporting may have been particularly great in countries with limited testing capability and restrictive testing policies. Simultaneously, some models have been accused of over-reporting the scale of the pandemic. At a time when modeling consortia around the world are turning to the lessons learnt from pandemic modelling, we present an example of simulating testing as well as the spread of disease. In particular, we factor in the amount and nature of testing that was carried out in the first wave of the COVID-19 pandemic (March - September 2020), calibrating our spatial Agent Based Model (ABM) model to the reported case numbers in Zimbabwe.

2012 ACM Subject Classification Computing methodologies → Modeling methodologies

Keywords and phrases Agent Based Modelling, Infectious Disease Modelling, COVID-19, Zimbabwe, SARS-CoV-2, calibration

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.50

Category Short Paper

Funding *Robert Manning Smith*: UKRI Grant MR/T02075X/1.
Sarah Wise: UKRI Grant MR/T02075X/1.
Sophie Ayling: UBEL-Doctoral Training Partnership ES/P000592/1.

1 Introduction

From the early stages of the COVID-19 pandemic, there have been initiatives to estimate the true scale and impact of the epidemic in terms of cases, hospitalizations and deaths across different countries around the world. Starting with the World Health Organization [23], a number of other data trackers sprung up (e.g. [15, 21, 12] or the more policy-focused [3]). These trackers fed into disease models which sought to predict the future spread of disease. Agent-based models (ABMs) became popular, especially as researchers sought more granular dimensions to population characteristics and scenario modelling (see [4, 16, 10]).

During the pandemic, criticisms were levelled at modellers in the public eye that the model forecasts did not reflect the number of cases that were reported in the media [2]. Certain studies suggested that the cases detected and reported were substantially under-reporting the true magnitude of the epidemic. In different contexts, researchers estimated that true case numbers might outstrip reported case numbers by a factor of between 5 and 20 ([19]. What accounts for this discrepancy?

In this paper, we attempt to recreate these "hidden" cases, taking as a case study Zimbabwe. We endeavour to replicate the true reported case numbers by layering a simulated testing process on top of our existing model of disease. The work presented in this paper

incorporates the available data on Zimbabwe's pandemic response policy, testing, and reported cases. The following sections will address some relevant background for this question (Section 2) before presenting the modelling framework and data used to inform it (Section 3). The results of the applied model will be presented (Section 4) and contextualised (Section 5).

2 Background

This section will present motivating context for understanding reported cases of disease as well as Zimbabwe's handling of the COVID-19.

2.1 Understanding reported cases

ABMs experienced an explosion in popularity as a result of the COVID-19 pandemic. The ways in which researchers sought to understand how their simulations related to reported cases varied. Some modelers have made efforts to either a-priori include an understanding of testing, resulting in only a proportion of cases being detected, or to somehow back-calibrate to reported data. For example, the US based Institute of Disease Modelling's model Covasim [16] added a parameter to incorporate testing. Others tried to compare actual and simulated hospital admissions [17] or to calibrate their models on diagnosis versus morality rates [14].

In many Low and Middle Income Country (LMICs) contexts, where testing capacities were often more limited, these underestimates on reported case numbers are likely to have been at least as high as those in High Income Countries (HICs). Many have proactively attempted to mitigate this: for example, in Kenya, researchers used a combination of serological and PCR test data to calibrate their work for this reason [20]. Research seems to support the idea that true cases were undercounted: in Kazakhstan, researchers used death and the Case Fatality Ratios (CFR) to attempt to backcast true case numbers from July 2020 to May 2021 of the pandemic in that country [22]. The authors of the study asserted that official cases reported undercounted the number of infections by at least 60%. A similar situation was reported in various African countries [6], where serological surveys also retrospectively appeared to reveal a much higher prevalence of those who had developed SARS-CoV-2 antibodies in the population than the reported case statistics would appear to show. For example across 3 high density suburbs in Harare, Zimbabwe researchers found that the seroprevalence was at 19% in 2020 and 53% in 2021, with almost half of the participants who tested positive reporting no symptoms in the preceding six months [11]. With this background, it is useful to explore further the specific case of Zimbabwe.

2.2 The case of Zimbabwe

Zimbabwean authorities acted very quickly after the first case was detected in their country on 20^{th} March 2020 [5]. They launched the country's Preparedness and Response Plan for Coronavirus the very next day.However, during this initial period testing was very limited. Large scale rapid diagnostic testing did not become available till September 11^{th}, 2020 [13]. As of 27^{th} June 2020, Zimbabwe had 567 confirmed SARS-CoV-2 cases [21]. Eighty-two percent of these were returning residents and 18% were the result of local transmission. The testing was heavily skewed towards returnees despite a comprehensive testing strategy [18]. For those tests that were conducted, there were also logistical issues in transporting samples to the few available testing centers (see [7]) further confounding the picture. Thus, despite proactive measures by leadership, it is likely that cases in Zimbabwe were substantially underreported.

With this understanding of the need for simulation which can calibrate against systematically underreported data, we proceed to a description of the method we adopt in the rest of this paper.

3 Methodology

This model is an extension of work documented in [24], based on simulation available as an open-source project available online[1]. To briefly review the simulation framework, we constructed a spatial agent-based model (ABM) simulating the spread of SARS-CoV-2 in Zimbabwe with district level dis-aggregation in movement patterns for individual agents in the model. Default model values are taken from [16], which in turn draws upon [10].

In this paper, we introduce the incorporation of a testing regime into the model to enable us to measure both cases that *exist* and cases that have been *detected* in the population.

3.1 The testing regime

The modelled testing regime sits on top of our existing simulation of the spread of the virus. In the testing regime, a number of tests are distributed amongst the population each day. Individuals who exhibit symptoms of SARS-CoV-2 are eligible for testing. The symptoms of SARS-CoV-2 - such as a continuous cough or fever - are common to many other infections; thus we take into account that people without SARS-CoV-2 will present for testing. To simulate the allocation of tests to those without the infection, we generate a number of people with "spurious" SARS-CoV-2 symptoms. These symptoms will last for 7 days before subsiding. A person will seek a test only once. This process is based upon the work of numerous contextual studies (see [7, 6, 13, 8, 9, 5]).

Two factors will necessarily influence the number of detected cases beyond the actual underlying number of cases: the number of tests administered per day and the number of people with SARS-CoV-2 infections who are tested. The number of tests given out each day is a set number taken from the government's reported numbers [21]. Because the number of tests distributed daily was not available to us, we calculated the number of tests performed each day from the reported number of cases and the percent of tests that were positive as per [21]. The total number of tests administered each day were then scaled to match the models population size. The number of people with SARS-CoV-2 who are tested remains an unknown; false positives and negatives make it impossible to objectively determine this. Thus, we explore different possibilities in the results section.

3.2 Movement

One key feature of the model is the movement of individuals between districts. As we wanted to compare our test results to real reported case data, it was important to ensure that lockdowns and their consequent lower mobility levels were incorporated into the simulation.

The model calculates the likelihood of any agent moving between districts based on a number of different factors: their economic status, the day of the week, and the baseline likelihood of moving between their current district and another. That last factor is represented in the model by an origin-destination (OD) matrix, which draws from Call Detail Records (CDR) provided by the largest mobile phone service provider in the country. The raw data

[1] see `https://github.com/dime-worldbank/Disease-Modelling-SSA`

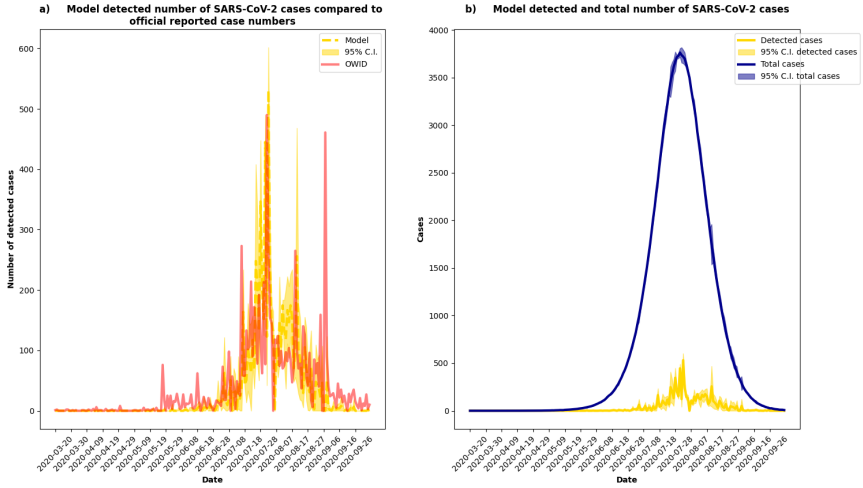

Figure 1 a) The model's number of SARS-CoV-2 cases detected through the testing regime compared to the official reported case number, taken from Our World in Data. b) The model's detected number of SARS-CoV-2 and the total number of the model's predicted cases.

(to which this study did not have access) covered the period February 1-June 30, 2020. At the dis-aggregated level, it contains data on 1900 towers to include 8.1 billion observations across each of the country´s 60 districts. The World Bank research team which handled this data partitioned it into two periods: the first from February 2 to March 14 (prior to the first Level 4 lockdown), and the second from March 15 - June 2020. By extracting the inter-district movements for these two time periods into separate OD matrices, they created patterns of travel representative of both normal and lockdown conditions.

Thus, in order to ensure that our simulated individuals were moving correctly, we applied a "lockdown" in the simulation by drawing the movement of individuals from a distribution defined by either the pre- or lockdown OD matrices. The simulation imposes a level 4 lockdown on the 30^{th} of March, with reduced movement; we then revert back to the pre-lockdown levels of interdistrict travel on the 17^{th} of May, when the imposed restrictions on intercity travel were removed as part of Level 2 measures (as per [5]).

4 Results and Discussion

Each instantiation of the model was run for 200 simulated days; our model start date and testing routine coincides with the start of the case reporting from Zimbabwe from the 20^{th} of March 2020. The simulated population is based on a 5% sample of the 2012 Zimbabwe Census was taken from IPUMS International [1], allowing us to incorporate realistic distributions of age, sex, economic status, and household composition.

We performed a parameter grid search to calibrate the models' number of detected cases to those reported. We paired combinations of the infection transmission parameter, β, to the rate in which a person will develop spurious symptoms, γ. The total error in the number of detected cases in each parameter combination was assessed and models were selected to minimize the total error. Initially, SARS-CoV-2 testing in Zimbabwe was limited to points of entry (functionally, districts with an official boarder crossing, airport or train station).

Within our parameter grid search, the parameter combination which resulted in model runs that most closely fit the true reported case data came when $\beta = 0.128$ and $\gamma = 0.0875$. The simulated reported cases are shown in Figure 1a, with a 95% confidence interval indicating the variation among runs. Figure 1b demonstrates the total number of simulated cases in the same model, demonstrating the significant number of cases missed as a result of a limited testing regime. During the simulation period, the model's daily detected number of cases peaked at 531, whereas the peak number of both undetected and detected cases was 3763.

Our methodology of filtering the model's simulated cases through a simulated testing regime allowed us to closely match the reported case numbers. Over the course of the simulation, the model generated a total of 153,807 cases, yet the simulated testing documented only 6892 cases. Thus, only 5% of the model's "true" cases were discovered by the testing regime. Other modelling studies have found similar discrepancies in the detected and total cumulative number of cases estimated (see for example [19]).

5 Conclusion

The results of this paper are dependent on the outcome of the model's calibration and a number of assumptions made. For example, one relevant assumption is the number of cases distributed in the population at the beginning of the simulation. Initially, we created a single infection in a 25% scale size population (equivalent to four initial cases, once scaling is taken into account). A single initial infection was chosen to represent the single initial case reported on the 20^{th} of March. It may be that more cases existed in Zimbabwe at the time; however, in hindsight it would be impossible to establish the exact number. Seeding more infections initially would result in an increased number of cases overall. Future work might explore the sensitivity of the epidemic to the number of initial cases as well as the parameters β and γ.

Broadly, this work contributes to the discussion around disease forecasting and prediction. As described above, many people were skeptical of the apparent "overprediction" of cases of SARS-CoV-2 cases. Our results show a clear example of how the results of such simulations might track well with the reality of testing. The fit between our simulated testing data and real testing data in our chosen case study suggests the model is capturing the true epidemic peak - and also of reflecting the impact of a testing regime. Exploration of different testing regimes represents a promising future direction for research. Regardless, researchers should ensure that modelled results distinguish between cases and *reported* cases, and should seek to document the statistical process which mediates the relationship between these. Reported case numbers will paint only a partial picture of the full situation, but through simulation we may begin to better understand the underlying reality.

References

1 National Statistics Agency. Zimbabwe Population Census 2012. https://international.ipums.org/international-action/sample_details/country/zw, 2012.

2 Adam T Biggs and Lanny F Littlejohn. Revisiting the initial COVID-19 pandemic projections. *The Lancet Microbe*, 2(3):e91–e92, March 2021. doi:10.1016/S2666-5247(21)00029-X.

3 BSG. Oxford University Gortnment Response Tracker. https://www.bsg.ox.ac.uk/research/covid-19-government-response-tracker, 2020.

4 Sheryl L. Chang, Nathan Harding, Cameron Zachreson, Oliver M. Cliff, and Mikhail Prokopenko. Modelling transmission and control of the COVID-19 pandemic in Australia. *Nature Communications*, 11(1):5710, November 2020. doi:10.1038/s41467-020-19393-6.

5 Itai Chitungo, Tafadzwa Dzinamarira, Nigel Tungwarara, Munashe Chimene, Solomon Mukwenha, Edward Kunonga, Godfrey Musuka, and Grant Murewanhema. COVID-19 Response in Zimbabwe: The Need for a Paradigm Shift? *COVID*, 2(7):895–906, June 2022. doi:10.3390/covid2070065.

6 Tafadzwa Dzinamarira, Mathias Dzobo, and Itai Chitungo. COVID-19: A perspective on Africa's capacity and response. *Journal of Medical Virology*, 92(11):2465–2472, November 2020. doi:10.1002/jmv.26159.

7 Tafadzwa Dzinamarira, Munyaradzi P. Mapingure, Gallican N. Rwibasira, Solomon Mukwenha, and Godfrey Musuka. COVID-19: Comparison of the Response in Rwanda, South Africa and Zimbabwe. *MEDICC Review*, July 2021. doi:10.37757/MR2021.V23.N3.4.

8 Tafadzwa Dzinamarira, Solomon Mukwenha, Rouzeh Eghtessadi, Diego F Cuadros, Gibson Mhlanga, and Godfrey Musuka. Coronavirus Disease 2019 (COVID-19) Response in Zimbabwe: A Call for Urgent Scale-up of Testing to meet National Capacity. *Clinical Infectious Diseases*, 72(10):e667–e674, May 2021. doi:10.1093/cid/ciaa1301.

9 Federal Research Centre for Cultivated Plants. Official Ports of Entry for Zimbabwe.

10 N Ferguson, D Laydon, G Nedjati Gilani, N Imai, K Ainslie, M Baguelin, S Bhatia, A Boonyasiri, ZULMA Cucunuba Perez, G Cuomo-Dannenburg, A Dighe, I Dorigatti, H Fu, K Gaythorpe, W Green, A Hamlet, W Hinsley, L Okell, S Van Elsland, H Thompson, R Verity, E Volz, H Wang, Y Wang, P Walker, P Winskill, C Whittaker, C Donnelly, S Riley, and A Ghani. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. Technical report, Imperial College London, March 2020. doi:10.25561/77482.

11 Arun Fryatt, Victoria Simms, Tsitsi Bandason, Nicol Redzo, Ioana D. Olaru, Chiratidzo E Ndhlovu, Hilda Mujuru, Simbarashe Rusakaniko, Michael Hoelscher, Raquel Rubio-Acero, Ivana Paunovic, Andreas Wieser, Prosper Chonzi, Kudzai Masunda, Rashida A Ferrand, and Katharina Kranzer. Community SARS-CoV-2 seroprevalence before and after the second wave of SARS-CoV-2 infection in Harare, Zimbabwe. *EClinicalMedicine*, 41:101172, November 2021. doi:10.1016/j.eclinm.2021.101172.

12 FT. Financial Times Covid Tracker. https://www.ft.com/content/a2901ce8-5eb7-4633-b89c-cbdf5b386938, 2020.

13 Muchaneta Gudza-Mugabe, Kenny Sithole, Lucia Sisya, Sibongile Zimuto, Lincoln S. Charimari, Anderson Chimusoro, Raiva Simbi, and Alex Gasasira. Zimbabwe's emergency response to COVID-19: Enhancing access and accelerating COVID-19 testing as the first line of defense against the COVID-19 pandemic. *Frontiers in Public Health*, 10:871567, July 2022. doi:10.3389/fpubh.2022.871567.

14 Nicolas Hoertel, Martin Blachier, Carlos Blanco, Mark Olfson, Marc Massetti, Marina Sánchez Rico, Frédéric Limosin, and Henri Leleu. A stochastic agent-based model of the SARS-CoV-2 epidemic in France. *Nature Medicine*, 26(9):1417–1421, September 2020. doi:10.1038/s41591-020-1001-6.

15 JHU. John Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html, 2020.

16 Cliff 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, Lauren George, Michał Jastrzębski, Amanda Izzo, Greer Fowler, Anna Palmer, Dominic Delport, Nick Scott, Sherrie Kelly, Carrie Bennette, Bradley Wagner, Stewart Chang, Assaf P. Oron, Edward Wenger, Jasmina Panovska-Griffiths, Michael Famulare, and Daniel J. Klein. Covasim: An agent-based model of COVID-19 dynamics and interventions. Preprint, Epidemiology, May 2020. doi:10.1101/2020.05.10.20097469.

17 Imran Mahmood, Hamid Arabnejad, Diana Suleimenova, Isabel Sassoon, Alaa Marshan, Alan Serrano-Rico, Panos Louvieris, Anastasia Anagnostou, Simon J E Taylor, David Bell, and Derek Groen. FACS: A geospatial agent-based simulator for analysing COVID-19 spread and

public health measures on local regions. *Journal of Simulation*, 16(4):355–373, July 2022. doi:10.1080/17477778.2020.1800422.

18 Grant Murewanhema, Trouble Burukai, Dennis Mazingi, Fabian Maunganidze, Jacob Mufunda, Davison Munodawafa, and William Pote. A descriptive study of the trends of COVID-19 in Zimbabwe from March - June 2020: Policy and strategy implications. *Pan African Medical Journal*, 37, 2020. doi:10.11604/pamj.supp.2020.37.1.25835.

19 Jungsik Noh and Gaudenz Danuser. Estimation of the fraction of COVID-19 infected people in U.S. states and countries worldwide. *PLOS ONE*, 16(2):e0246772, February 2021. doi:10.1371/journal.pone.0246772.

20 John Ojal, Samuel P. C. Brand, Vincent Were, Emelda A. Okiro, Ivy K. Kombe, Caroline Mburu, Rabia Aziza, Morris Ogero, Ambrose Agweyu, George M. Warimwe, Sophie Uyoga, Ifedayo M. O. Adetifa, J. Anthony G. Scott, Edward Otieno, Lynette I. Ochola-Oyier, Charles N. Agoti, Kadondi Kasera, Patrick Amoth, Mercy Mwangangi, Rashid Aman, Wangari Ng'ang'a, Benjamin Tsofa, Philip Bejon, Edwine Barasa, Matt J. Keeling, and D. James Nokes. Revealing the extent of the first wave of the COVID-19 pandemic in Kenya based on serological and PCR-test data. *Wellcome Open Research*, 6:127, February 2022. doi:10.12688/wellcomeopenres.16748.2.

21 OWID. Our World in Data. https://ourworldindata.org/coronavirus, 2020.

22 Antonio Sarría-Santamera, Nurlan Abdukadyrov, Natalya Glushkova, David Russell Peck, Paolo Colet, Alua Yeskendir, Angel Asúnsolo, and Miguel A. Ortega. Towards an Accurate Estimation of COVID-19 Cases in Kazakhstan: Back-Casting and Capture–Recapture Approaches. *Medicina*, 58(2):253, February 2022. doi:10.3390/medicina58020253.

23 WHO. World Health Organization's Coronavirus Tracker. https://covid19.who.int/, 2020.

24 Sarah Wise, Sveta Milusheva, Sophie Ayling, and Robert Manning Smith. Scale matters: Variations in spatial and temporal patterns of epidemic outbreaks in agent-based models. *Journal of Computational Science*, 69:101999, May 2023. doi:10.1016/j.jocs.2023.101999.