2024 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON) | 979-8-3315-3202-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/SIBIRCON63777.2024.10758442

Assessing the Applicability of the Multiagent Modeling Approach to the Epidemic Surveillance of COVID-19 in Russian Cities

1st Andrey I. Korzin
ITMO University
Saint Petersburg, Russia
corzin.an@gmail.com

2nd Timofey I. Kaparulin
ITMO University
Saint Petersburg, Russia
kaparulinti@mail.ru

3rd Vasiliy N. Leonenko
ITMO University
Saint Petersburg, Russia
vnleonenko@itmo.ru

***Abstract*—This study presents a qualitative and quantitative evaluation of multiagent modeling approach as a tool for COVID-19 epidemic surveillance. For that purpose, we regard a multi-agent model based on authors' earlier research and an open source COVASIM framework, comparing their differences in structural complexity, accuracy and computational performance. The mentioned models are calibrated using incidence data from COVID-19 outbreak in Saint Petersburg, Russia, in January-April 2022. The simulation results demonstrate that both models can be applied for retrospective analysis and prediction of COVID-19 incidence. At the same time, there exist important aspects of their usage related to employed demographic data, which should be taken into account when applying them for the tasks of epidemiological surveillance.**

***Index Terms*—mathematical epidemiology, COVID-19, multi-agent model, synthetic population, COVASIM**

I. Introduction

The COVID-19 pandemic has demonstrated the need for health authorities to quickly respond to a dynamically changing epidemic situation. Mathematical models constitute an important tool that allows tracking the quantitative and spatial distribution of cases, as well as assessing the impact of government restrictions and vaccination. The most widely studied and actively used epidemic models are compartmental models, which first appeared at the beginning of the 20th century. Their advantages are simplicity, high speed of calculations and ease of interpretation of the output. At the same time, compartmental models are ill–suited for describing the spatial patterns of disease propagation. That drawback is solved in multiagent models, which are capable of analyzing in detail the heterogeneous effects of the spread of morbidity based on synthetic populations, as well as displaying the infection process on a map. On the other hand, multiagent models have their own limitations. Firstly, those models require detailed input data, particularly, a synthetic population based on demographic and geographic data. In addition to collecting the demographic data, it is necessary to distribute the individuals of the population according to their places of residence, as well as carry out appropriate verification of this distribution. The second important disadvantage is increased simulation time, which limits the performance of decision-support systems based on multiagent approach. Thus, the multiagent models which assume repetitive use as a part of epidemiological surveillance and decision support systems should have a good balance between their detail level, i.e. structural complexity, and simulation speed, which is to be found taken into account the available hardware and nomenclature of data. Ultimately, a perfect model should combine accuracy, simulation speed, and modeling detail.

The primary approach which is considered by the authors of this study is to perform COVID-19 multiagent modeling based on our earlier developed models for the spread of influenza in Russian cities [1]. However, we cannot neglect the fact, that since 2020, there was created a big number of promising COVID-19 modeling frameworks. Particularly, the COVASIM modeling framework [2], released as open source in 2021, has been a significant contribution to epidemic modeling. However, using this package includes several peculiarities:

- Using numerous parameters in COVASIM enables customizing a wide range of processes, such as detailed transitions between infectious states and incorporating various government interventions. At the same time, the large number of parameters can negatively impact the quality of model calibration due to the high degree of freedom of the model.
- The synthetic populations serving as model inputs for COVASIM are necessary for detailed modeling of the spread of COVID-19 in cities. They are generated using an open-source model called SynthPops [3]. This library mainly contains populations of various counties of Washington, U.S. what limits researchers from using the library for simulation of disease propagation in other cities. That is why the use of COVASIM in different cities involves the creation of a synthetic population based on open government data [4]. Some researches use populations based on random networks or 'hybrid' approach, which combines random networks and statistical data of distribution over households, workplaces, etc. [5].

This research was supported by The Russian Science Foundation, Agreement #22-71-10067

- Model calibration is another important issue that has not yet been fully resolved in the COVASIM and another multi-agent models, as it takes a lot of time and has unpredictable quality: according to its creators, the built-in optimizer from the Optuna library does not guarantee the successful model calibration [6].

In this paper, we compare the usage of the COVASIM framework and our model for the purpose of retrospective analysis and prediction of COVID-19 dynamics. The purpose of this study is to highlight the advantages and disadvantages of their employment, focusing on their structural complexity, the quality of calibration and ease of adaptation for use in various cities of the Russian Federation.

II. Methods

A. Data

a) Synthetic population: A synthetic population is a generated dataset which is comprised of records of each inhabitant of a regarded city or country. For simulation of COVID-19 propagation, the synthetic population of Saint Petersburg was created. Gathered data contains spatial distribution of households, workplaces and schools as well as age and gender distribution of population. Data was collected using open source data: Federal State Statistics Service, Open Street Map and Yandex Auditorii API. The detailed description of creating a similar synthetic population is discussed in paper [1]. For this work, a synthetic population of St. Petersburg, Russia was constructed for the year 2023. The characteristics of synthetic population are shown in the Table I.

b) Epidemic data: The data for model calibration was taken from the repository [7] with daily dynamics of COVID-19 in St. Petersburg, Russia. The epidemic outbreak from January 2022 to April 2022 was used for calibration of model parameters. The incidence data is shown in Fig. 1.

TABLE I
SYNTHETIC POPULATION CHARACTERISTICS

Characteristic	Value
Population size	5'294'136
Number of households	1'936'991
Number of workplaces	360'353
Number of schools	810

B. Our multi-agent model

The multi-agent model was implemented for simulation of dynamics of COVID-19 propagation. For the sake of clarity, the description of model is represented below using ODD protocol [8].

1) Purpose
The proximate purpose of the model is to simulate the dependence of number of COVID-19 cases on time, as well as to build a spatial distribution of disease spreading. The model intends to track disease incidence down to a single individual and works with input data in the form of a synthetic population of a city.

2) Entities and state variables
The following entities are included in the model:

- Individuals who correspond to people living in a city;
- Household – entity that every person visits every day of simulation;
- Workplace – entity that every working adult from 18 to 65 visits;
- Schools – entity that every child from age 7 to 17 visits.

Each place entity has a set of parameters, presented in the Table III. Each person has several characteristics presented in the Table II.

TABLE II
VARIABLES OF PERSON ENTITY

Variable name	Type	Meaning
id	int	Unique identification number for a person
gender	bool	Gender of a person (male or female)
age	int	Age of a person
household_id	int	Unique identification number for the person's household
work_id	int	Unique identification number for the person's workplace
school_id	int	Unique identification number for the person's school
susceptible_flag	bool	Shows susceptibility status
incubation_flag	bool	Shows if a person is infected and is in incubation period
infectious_flag	bool	Shows if person is infected and is in infectious period
incubation_duration	int	Day number of the incubation period
illness_duration	int	Day number of the infectious period

TABLE III
VARIABLES OF PLACE ENTITY

Variable name	Type	Meaning
id	int	Unique identification number for a place
latitude	float	Latitude of a place
longitude	float	Longitude of a place
size	int	Maximal place capacity in individuals
c_s	float	Average number of contacts per day

3) Process overview and scheduling
The model has a discrete time step of one day. During each simulation day, three processes are performed via the corresponding submodels:

- Daily activity – visiting home, work and school;
- Infection onset – simulating possible infection transmission in each place with non-zero number of susceptible and infectious individuals;
- Infection advance – simulating incubation and infectious periods for each newly infected person.

979-8-3315-3202-4/24$31.00 ©2024 IEEE

4) Design concepts
Population data is implemented in the form of a table, where each row corresponds to a specific person, and a column is the person's feature, for example, gender or day since a start of the disease. The infection propagation is implemented in the form of sequential infectious process iteration, first across households, then through workplaces, and then through schools.
5) Initialization
We set up initial parameters, such as the initial number of infected people (I_0), duration of simulation (T_{mod}), fraction of non-immune (α) and infection transmission coefficient λ. Average daily number of contacts in different places must be defined as well. Due to the stochastic behavior of multi-agent simulation, we need to make several runs for gathering more output data, parameter N_{runs} is responsible for the number of output curves. The description of these parameters is shown in the Table 5.

TABLE IV
INITIAL PARAMETERS DESCRIPTION

Parameter	Description	Value
α	Fraction of non-immune individuals in general population	Varied
λ	Infection transmission coefficient	Varied
c_{hh}	Average daily number of contacts in households	3
c_{wp}	Average daily number of contacts in workplaces	6
c_{sch}	Average daily number of contacts in schools	7
I_0	Initial number of infected individuals	1500
T_{mod}	Modeling time, days	120
N_{runs}	Number of simulation runs	10

6) Input data
The model takes as input a synthetic population based on real data consisting of four files: *people.txt*, *households.txt*, *workplaces.txt* and *schools.txt*. While the file *people.txt* mainly contains information about each person in population (see Table II), other files contain data frames with a row for each place and columns for place features: id, latitude, longitude and size (see Table III).
7) Submodels
The model contains three submodels:
- Daily activity
Each person during the day can visit one or two places depending on his age and work status: household, household + work or household + school.
- Infection onset
Firstly, we calculate the probability for an individual to contact a random person in the place *pl*:

$$p_{cnt}^{(pl)} = \min\left\{\frac{c_s}{N^{(pl)}(t)-1}, 1\right\}.$$

After that, we calculate the probability of susceptible individual to become infected:

$$p_{\text{inf}}^{(pl)} = 1 - \prod_{i=1}^{N^{(pl)}(t)-1} (1 - \lambda \cdot p_{cnt}^{(pl)}),$$

where $N^{(pl)}(t)$ is a total number of individuals at place *pl* at simulation day with number t. Finally, we calculate the number of infected people in current place by generating random variable with a binomial distribution:

$$n_{inf}^{(pl)} \sim Bin(S^{(pl)(t)}, p_{inf}^{(pl)})$$

- Infection advance
For each infected person, we initialize the duration of the incubation and infectious periods in days using log-normal distribution with parameters according to papers [9]–[13]:

$$\tau_{inc} \sim \text{lognormal}(4.5, 1.5), \tau_{inf} \sim \text{lognormal}(8, 2),$$

where τ_{inc} and τ_{inf} is a number of days for incubation and infectious period respectively. The first parameter of the distribution stands for mean and the second stands for standard deviation. This process of infection repeats in every place during the simulation day. The number of newly infected is recorded for each simulation day.

C. COVASIM model

To simulate the spread of COVID-19 in St. Petersburg using COVASIM, we used two types of contact networks. The first one was based on transforming demographic and geographic data to the data conforming to COVASIM input standard. The contact network with households, workplaces and schools was created with python library `synthpops`. One simulation run time increased from 135 seconds with random network population to 6084 seconds with synthetic population based on St. Petersburg data. The second network was based on COVASIM random networks as synthetic population data. To decrease the simulation time, we took a population of 500'000 individuals, which is approximately 10 times less than the real population of St. Petersburg. The infection transmission coefficient, the fraction of non-immune individuals, the number of initial infectious and duration of simulation were varied, while other parameters were set to COVASIM default values. The default parameters and variables of COVASIM framework can be found in [2].

D. Performance metrics

To assess the suitability of the model for use in epidemic surveillance, we will calculate three main model metrics:
- Model accuracy: R^2;
- Simulation performance: average time T_{avg} of one simulation run;
- Structural complexity: the number of free parameters k. A larger coefficient k indicates a higher structural complexity of the model, it is also widely used to

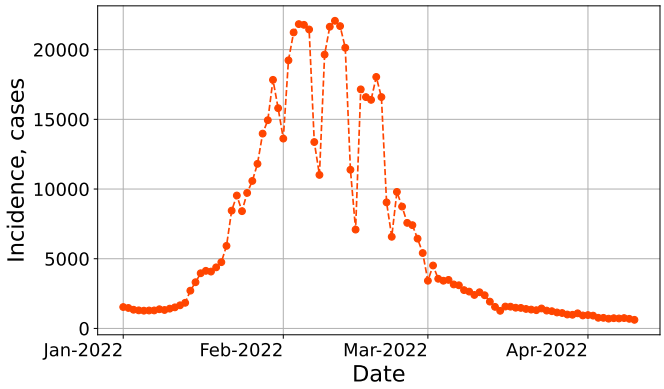

Fig. 1. COVID-19 incidence data for epidemic outbreak in 2022 Saint Petersburg, Russia.

assess model calibration accuracy with regards to their complexity based on Akaike criterion (e.g., see [14]).

III. Experiments

A. Calibration

Automatic calibration in COVASIM framework is made by Optuna library. Several experiments with automatic calibration was made and it has not given satisfactory results. In all cases we observed $R^2 \leqslant 0$. To address this issue, we made a parameter sweep — an iterative process in which simulations are run repeatedly using different values of the parameter(s) of choice. This process enables the modeler to determine a parameter's "best" value (or range of values), or even where in parameter space the model produces desirable (or non-desirable) behaviors. As a result, we obtained a comprehensive review of possible incidence curves. The parameter sweep was conducted with α and λ values ranging from 0 to 1. In COVASIM these parameters are called `frac_susceptible` and `beta` respectively. The trajectories of incidence from the parameter sweep were compared to the observed epidemic data. This allowed us to trace whether there is a model curve that approximates the real incidence data. By exploring the range of possible outcomes, we gained insights into the behavior of the model to changes in the α and λ parameters. The resulting trajectories and epidemic incidence data for COVASIM and our model are shown in Fig. 2. As a result of each parameter sweep, we detected the curve with the best R^2, these curves are shown in Fig. 3.

B. Forecast

To generate a forecast, the curves derived from a parameter sweep were utilized. It was assumed that epidemic data would be available within $t \leqslant 28$, where t is the number of day of simulation. The epidemic data, forecast curves and other simulation curves are shown in Fig. 4 for our model and in Fig. 5 for COVASIM framework. The forecast is not provided for the COVASIM simulation with contact network based on inferred patterns of daily activity, as all the epidemic curves obtained from it demonstrated negative R^2 values. The

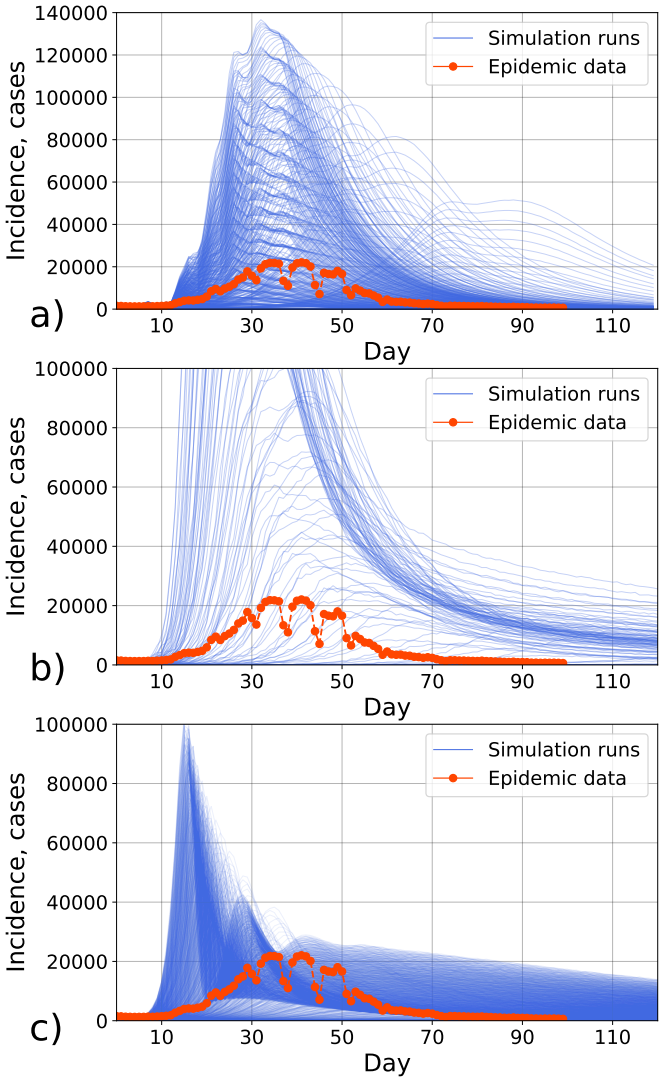

Fig. 2. Trajectories of incidence for parameter sweep simulations and epidemic data of COVID-19 propagation, Saint Petersburg, Russia. Parameters α and λ were changed with step Δ. a) our model, $\Delta = 0.05$ b) COVASIM with contact network based on inferred patterns of daily activity, $\Delta = 0.05$, c) COVASIM with random contact network with 500'000 population size, $\Delta = 0.02$.

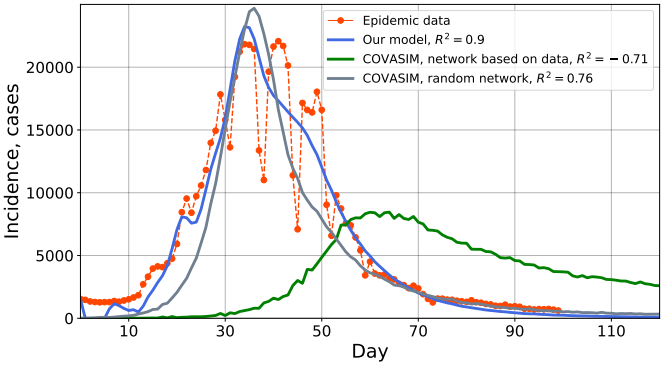

Fig. 3. Curves with best R^2 value from different parameter sweep simulations.

979-8-3315-3202-4/24$31.00 ©2024 IEEE

incidence curves for the first 28 days are represented with a gradient from red to green; a red hue indicates a decrease in R^2, while a green hue signifies an increase. For $t > 28$ only curves with R^2 greater than threshold value were retained. Grey curves stand for simulation runs with $R^2 > 0$. For our model, threshold value of R^2 is 0.9 and for COVASIM is 0.7. Prediction interval shows area between minimum and maximum incidence value in the forecast.

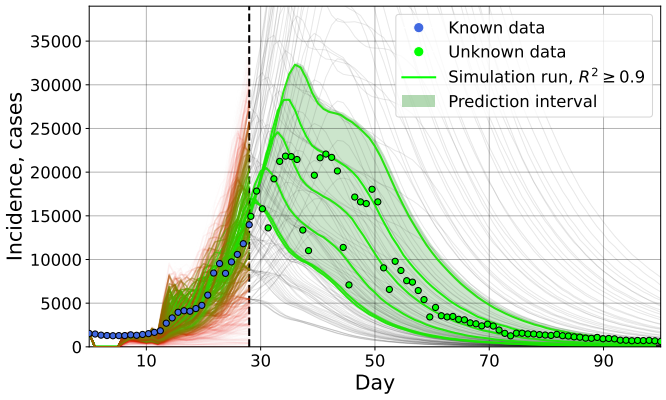

Fig. 4. Incidence forecast using our model and the synthetic population based on inferred patterns of daily activity.

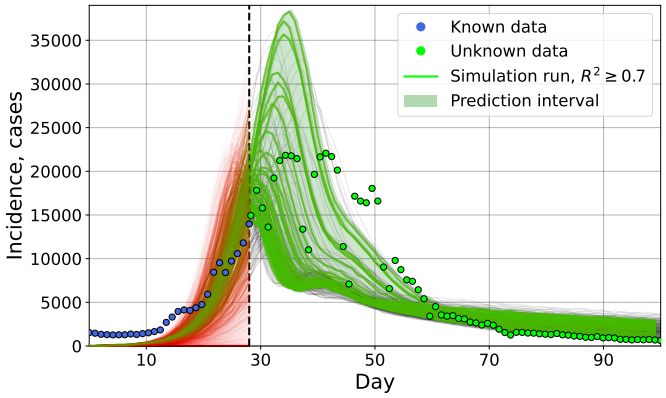

Fig. 5. Incidence forecast using COVASIM with random contact network.

To understand the behavior of the R^2 in the parameter space of our model, we constructed a plot, shown in Fig. 6. This graph demonstrates the presence of a clearly defined maximum of the R^2 coefficient in retrospective analysis (Fig. 6 a)). The point corresponding to highest R^2 has $\alpha = 0.4$ and $\lambda = 0.3$ parameter values. The prediction graph has greater uncertainty in determining the calibration parameters (Fig. 6 b)). It shows that high values of R^2 can be obtained with high λ and low α and with high α and low λ values.

C. Comparing models

The comparison of three computational experiments provides valuable insights into modeling the spread of COVID-19.

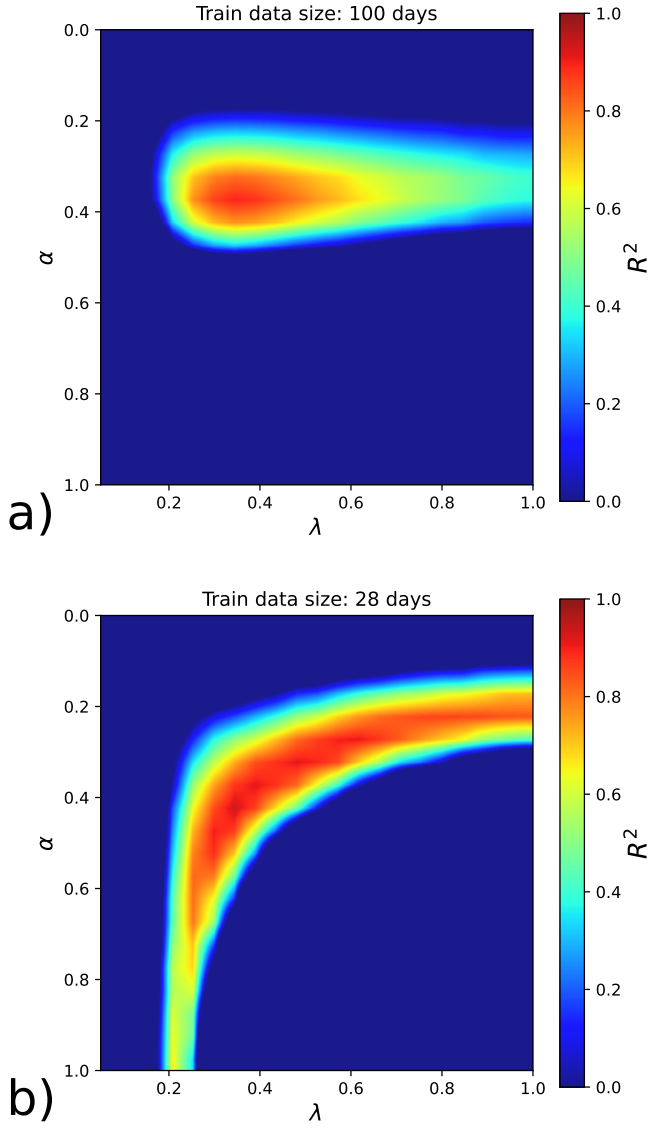

Fig. 6. R^2 for curves from parameter sweep for our model: a) retrospective analysis, train data size is 100 days; b) prediction, train data size is 28 days.

- The first experiment involved simulating the disease spread using a model which utilized the contact network based on inferred patterns of daily activity. The model showed a relatively short average operating time, as well as good R^2 values for forecasting incidence.
- In the second experiment, we employed the COVASIM framework to model the spread of COVID-19, again using the contact network based on inferred patterns of daily activity. This experiment showed extremely long simulation time, as well as poor R^2 values. Moreover, it took more than 24 hours and more than 100 GB of RAM to create the synthetic population compatible with COVASIM with 5 million agents. The change of population to a standard one for COVASIM based on random contact networks slightly improved the situation

both in terms of algorithm performance and in terms of calibration accuracy, but the results were still not good enough.

- The third experiment utilized the COVASIM framework with random contact networks and a population reduced to the size of 500'000. This experiment showed the smallest simulation time and allowed us to make a detailed parameter sweep with step $\Delta = 0.02$ along each of the parameters, resulting in satisfactory R^2. The interpretation of why the epidemic process in the population which is 10 times reduced better reflects real disease incidence compared to the simulation in the original population remains questionable. One of the possible explanations is underreporting coupled with a big fraction of non-symptomatic disease carriers, which alters the observable incidence making it lower than expected.

Comparative characteristics of the experiments are given in Table V.

TABLE V
SIMULATION COMPARISON

Model and population type	R^2	Avg. time of one simulation run, sec.	k
Our model with contact network based on inferred patterns of daily activity	**0.9**	279	**22**
COVASIM with contact network based on inferred patterns of daily activity	-0.71	6084	122
COVASIM random contact networks, population size 500'000	0.76	**39**	122

IV. DISCUSSION

In this study, we carried out a comprehensive assessment of the usage of the COVID-19 propagation models described. While both models showed decent results, we noted important features that should be taken into account when using these models in epidemic surveillance. Our model achieved the best R^2 value of 0.9 and has low structural complexity according to indicator k. On the other hand, the model has 279 seconds of average execution time of one simulation run, which is higher than default COVASIM simulation time of 39 seconds. The detail of our model is limited by the detail of synthetic population compiled from open data. In the current version of our model we only consider the movement of people between households, workplaces and schools. This simplification may restrict the model's ability to capture the nuances and complexities of real-world social interactions. Nevertheless, our model can be enhanced by incorporating new places and control measures, which would improve its accuracy to the desired level.

Using COVASIM framework with contact networks based on inferred patterns of daily activity requires significant computational resources. The creation of the contact network take more than 24 hours and more than 100 GB of RAM. Moreover, the average simulation time of one run has increased from 39 seconds with default random networks to 6084 seconds. Due to the high simulation time of the model, we were unable to find the curve that fits the incidence data. The population reduced by 10 times allowed us to decrease time of simulation run to 39 seconds and achieve R^2 of 0.76. However, the question of why reducing the population size yields accurate results, whereas the simulation on original population failed to deliver them, remains open for further investigation.

The choice between using COVASIM with synthetic populations or our model will depend on the specific research objectives, data availability, and the level of detail required to address the problem at hand. Both approaches offer unique advantages and can contribute to a better understanding of COVID-19 dynamics and the effectiveness of interventions.

REFERENCES

[1] V. Leonenko, S. Arzamastsev, and G. Bobashev, "Contact patterns and influenza outbreaks in Russian cities: A proof-of-concept study via agent-based modeling," *Journal of Computational Science*, vol. 44, p. 101156, 2020.

[2] C. C. Kerr, R. M. Stuart, D. Mistry, R. G. Abeysuriya, K. Rosenfeld, G. R. Hart, R. C. Núñez, J. A. Cohen, P. Selvaraj, B. Hagedorn *et al.*, "Covasim: an agent-based model of COVID-19 dynamics and interventions," *PLOS Computational Biology*, vol. 17, no. 7, p. e1009149, 2021.

[3] "Synthpops," https://github.com/institutefordiseasemodeling/synthpops/, (accessed October 11, 2024).

[4] R. Latkowski and B. Dunin-Kplicz, "An agent-based COVID-19 simulator: extending Covasim to the Polish context," *Procedia Computer Science*, vol. 192, pp. 3607–3616, 2021.

[5] J. Panovska-Griffiths, B. Swallow, R. Hinch, J. Cohen, K. Rosenfeld, R. M. Stuart, L. Ferretti, F. Di Lauro, C. Wymant, A. Izzo *et al.*, "Statistical and agent-based modelling of the transmissibility of different SARS-CoV-2 variants in England and impact of different interventions," *Philosophical Transactions of the Royal Society A*, vol. 380, no. 2233, p. 20210315, 2022.

[6] "Covasim docs," https://docs.idmod.org/projects/covasim/en/latest/, (accessed October 11, 2024).

[7] A. Kouprianov, "Monitoring COVID-19 epidemic in St. Petersburg, Russia: Data and scripts (2021)," https://github.com/alexei-kouprianov/COVID-19.SPb.monitoring, (accessed October 11, 2024).

[8] V. Grimm, S. F. Railsback, C. E. Vincenot, U. Berger, C. Gallagher, D. L. DeAngelis, B. Edmonds, J. Ge, J. Giske, J. Groeneveld *et al.*, "The odd protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism," *Journal of Artificial Societies and Social Simulation*, vol. 23, no. 2, 2020.

[9] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, and J. Lessler, "The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application," *Annals of internal medicine*, vol. 172, no. 9, pp. 577–582, 2020.

[10] Z. Du, X. Xu, Y. Wu, L. Wang, B. J. Cowling, and L. A. Meyers, "Serial interval of COVID-19 among publicly reported confirmed cases," *Emerging infectious diseases*, vol. 26, no. 6, p. 1341, 2020.

[11] H. Nishiura, N. M. Linton, and A. R. Akhmetzhanov, "Serial interval of novel coronavirus (COVID-19) infections," *International journal of infectious diseases*, vol. 93, pp. 284–286, 2020.

[12] R. Pung, C. J. Chiew, B. E. Young, S. Chin, M. I. Chen, H. E. Clapham, A. R. Cook, S. Maurer-Stroh, M. P. Toh, C. Poh *et al.*, "Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures," *The Lancet*, vol. 395, no. 10229, pp. 1039–1046, 2020.

[13] R. Wölfel, V. M. Corman, W. Guggemos, M. Seilmaier, S. Zange, M. A. Müller, D. Niemeyer, T. C. Jones, P. Vollmar, C. Rothe *et al.*, "Virological assessment of hospitalized patients with COVID-2019," *Nature*, vol. 581, no. 7809, pp. 465–469, 2020.

[14] I. Huaman and V. Leonenko, "Does complex mean accurate: Comparing COVID-19 propagation models with different structural complexity," in *International Conference on Computational Science*. Springer, 2023, pp. 270–277.