Lightweight heterogeneous SEIR models for epidemic surveillance in Russian cities: turning synthetic populations into equations

Andrey Korzin[1] and Vasiliy Leonenko[1 [0000−0001−7070−6584]]

ITMO University, 49 Kronverksky Pr., St. Petersburg, Russia, 197101
corzin.an@gmail.com, vnleonenko@itmo.ru

Abstract. Influenza and other acute respiratory diseases pose a significant challenge to global health. The complexity of analyzing and mitigating influenza transmission is related to heterogeneity of contact network structures in modern cities. The need for effective public health strategies has driven the development of highly detailed network and agent-based models. To overcome a drawback of modeling multi-agent systems, which is their high demand for computational resources, approximate models can be employed. In our article, we present an approach that allows to convert heterogeneous synthetic populations into an input for the edge-based compartmental SEIR model. We demonstrate the method application by simulating influenza spread in a contact network of the synthetic population of Chelyabinsk, Russia. At a cost of neglecting some details in contact network structure, the proposed algorithm allows to greatly enhance simulation speed compared to multi-agent modeling, and at the same time to preserve population heterogeneity, which makes it a good choice for application in epidemic surveillance.

Keywords: epidemiology · influenza · synthetic population · complex networks · edge-based compartmental modeling

1 Introduction

Influenza, along with other acute respiratory diseases, continue to pose a significant challenge to global health. The complexity of influenza transmission, coupled with the need for effective public health strategies, has driven the development of a wide array of mathematical models. Mathematical models in epidemiology enable the prediction of epidemic dynamics and the evaluation of epidemic indicators, such as the reproductive number and the number of immune individuals. Furthermore, they allow computational experiments that facilitate the assessment of best vaccination strategies and other methods of disease control. The largest epidemics occur in major cities, where high population density and numerous contacts between individuals are prevalent. Modern cities can be considered as complex systems with heterogeneous network structure, which creates challenges for modeling the spread of diseases. Classical SIR-type compartmental models assume homogeneous mixing in the populations and are

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_13

unable to account for effects associated with heterogeneity, such as existence of super-spreaders. The two most popular approaches that come to the rescue are multi-agent and network models.

The first method, multiagent modeling (MAM) based on detailed demographic data, is a powerful tool for modeling epidemic dynamics down to the level of a single individual. MAM allow us to consider effects connected with population heterogeneity, track the spatial spread of disease and chains of transmissions. However, high model detail results in long simulation time. This aspect is critical for epidemiological surveillance purposes, as the model calibration process requires numerous simulation runs. Another drawback is the need to build a synthetic population for every city of consideration. Collecting data on demographics, schools, urban development and workplaces of the city is time-consuming, and it is often difficult to verify the resulting datasets [12].

The second method, network modeling, represent a simplified approach, where contact networks are represented by random graphs and nodes are not distinguished by individual characteristics, enabling the conservation of contact tracing without the necessity of developing detailed synthetic populations. Barabási-Albert [4] and Erdős-Rényi [15] networks are commonly utilized for constructing contact networks. This method may not represent accurately the topology of contact networks in real populations.

To overcome a drawback of modeling multi-agent systems, which is their high demand for computational resources, approximate models can be employed. Kiss et al. showed that the computationally intensive SIR network model can be effectively approximated via a range of ordinary differential equation models [6]. Namely, edge-based compartmental modelling (EBCM), being the most exact of the approaches provided, approximates well the simulation results shown by network models and at the same time provides those results much faster than them [10]. While SIR models are partially applicable for the infections with short latent period, like influenza, their applicability to diseases with long incubation period, like COVID-19, is a matter of discussion. In order to fully take into account the latent period, it is more logical to consider the SEIR (Susceptible-Exposed-Infected-Recovered) model. The SEIR EBCM models were proposed and used in [3], [13], [14]. In these articles, the networks of standard topologies were used as an input, and the usage of synthetic populations in the models was not considered.

In this work, we propose a method that unites the usage of approximate heterogeneous compartmental models, namely, SEIR EBCM, with the generation of input contact network graphs based on synthetic populations. We demonstrate the usage of the method by constructing a contact network graph from the synthetic population of Chelyabinsk, Russia, and comparing the simulation of the disease outbreak using a SEIR EBCM with the simulation via a network model. The ability of the EBCM to replicate real disease incidence is demonstrated by calibration to 2022-2023 influenza incidence in Chelyabinsk.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_13

2 Methods

2.1 Synthetic population

A synthetic population is derived from diverse data sources reflecting the actual urban population. It is formatted into text files and serves as an input for the model. We have extensively employed multiagent modeling and creating datasets for synthetic populations in our previous research [7], [8]. These datasets are compatible to the RTI synthetic populations standard [16], capturing individual attributes such as household residence, age, gender, and workplace or school identifiers. A detailed explanation of the data collection process and the synthetic population format can be found in a related publication [8]. An example of data from a synthetic population file `people.txt` describing the characteristics of individuals is given in Table 1. The variables `sp_id` and `hh_id` represent numerical identifiers of individuals and their households, while `work_id` identifies work office number. Other files that include data related to households, workplaces and schools have a similar format.

Table 1. Sample records from the file `people.txt` of a synthetic population

sp_id	age	gender	hh_id	work_id
1	25	M	784	14
2	10	F	294	83
3	74	M	33	X
⋮	⋮	⋮	⋮	⋮

Chelyabinsk, a major Russian city with a population of over one million, was chosen for this computational study. A synthetic population of the city was created using data current as of 2023. The spatial distribution of households, schools and workplaces is shown in Fig. 1. The data for workplaces is sourced from Yandex.Auditorii [2], the households are geocoded using the information from Open Street Map [1]. Workplaces are divided into offices, and households are divided into apartments. The homogeneous structure of the workplace arrangement depicted in Fig. 1 is related to the aggregated nature of office data collected by [2]. Each workplace point is associated with the aggregated information for workplaces in the corresponding area. The characteristics of the synthetic population are summarized in Table 2.

Table 2. Population statistics for Chelyabinsk

City	Population	Households	Workplaces	Schools
Chelyabinsk	1 189 000	436 000	56 800	114

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_13

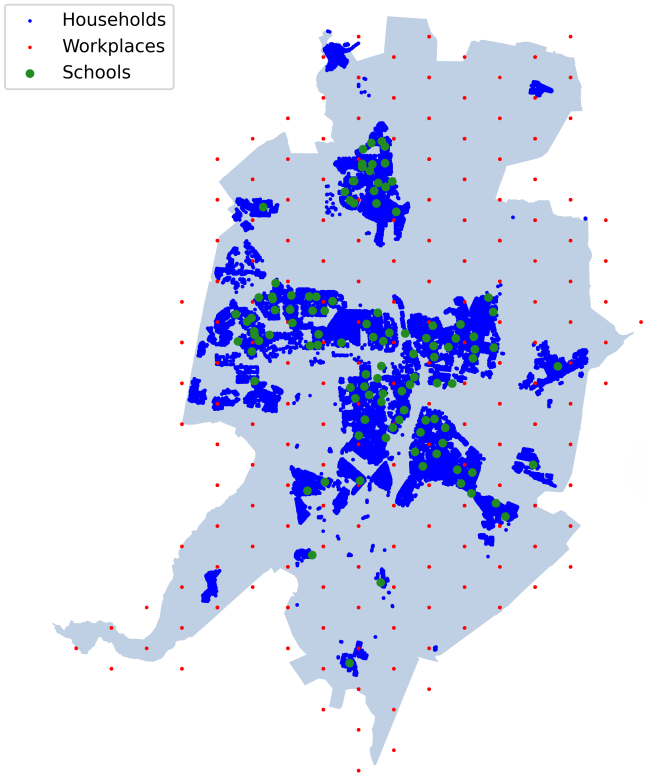

Fig. 1. Map of Chelyabinsk with spatial distribution of households, workplaces and schools from synthetic population data.

All computational experiments given in the article can be done on the full population of Chelyabinsk, however, converting a synthetic population to contact network graph of city with a population of 1 million people will require more than 30 GB of RAM and more than 2 hours of calculations (using Intel Xeon Gold 6226R 2.9 GHz). To demonstrate the possibility of reducing computational cost, a sampling algorithm was employed on a full-scale synthetic population dataset. This algorithm involves reducing the population by leaving r percent of households in each district of the city. When the household is removed, the individuals attributed to it are also removed from the synthetic population.

To generate a contact network based on the synthetic population, the following algorithm was applied:

- Create a graph with N nodes where N is the size of the population. Each node corresponds to a specific individual.
- Create an edge between each pair of individuals that have the same `hh_id`, `work_id` or `school_id`.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_13

2.2 SEIR EBCM

To explore the peculiarities of epidemic dynamics connected with unique topology of contact network and decrease the simulation time, we implemented an edge-based SEIR model (SEIR EBCM). A EBCM is an approximate model that incorporates probability generation function with degree distribution of input population network for constructing a system of equations that describes epidemic process on networks. Let the population consist of N individuals corresponding to N nodes of the network. According to works [13], [14] the system of equations for SEIR EBCM model is the following:

$$\begin{aligned}
&\dot{\theta} = -\beta\psi,\\
&\dot{\phi} = -(\alpha+\beta)\phi + (G''(\theta)\beta/G'(1))\psi,\\
&\dot{\psi} = \beta\phi - \gamma\psi + (G''(\theta)\beta/G'(1))\psi,\\
&S = G(\theta),\\
&R = \gamma I,\\
&\dot{E} = \beta SI - \alpha E.
\end{aligned}$$

Table 3 and 4 show the description of the parameters, variables and their initial conditions, taking the notations for u and v as neighbor nodes, connected by edge in network. $G(x)$ stands for probability generating function (PGF):

$$G(x) = \sum_{k=0}^{\infty} p_k x^k,$$

where p_k is the probability that a randomly chosen node degree equals k. Variables $S(t), E(t), I(t)$ and $R(t)$ in the following equations refer to fractions of susceptible, exposed, infected and recovered individuals respectively.

The system of ordinary differential equations is solved numerically using `odeint` method from `scipy` library for `Python` language. Compartment sizes are multiplied by N to find the absolute value of individuals at each state.

2.3 Data

Influenza incidence data were sourced from Research Institute of Influenza, St. Petersburg, Russia. The number of individuals infected with each strain was assessed with the help of strain-specific laboratory diagnostic data. Details of the data processing methodology are outlined in [9]. In this research, we investigated 2022-2023 epidemic season in Chelyabinsk, Russia. The incidence data is shown in Fig. 2. Some time points do not have incidence values due to a lack of data for this period. During 2022-2023, the strain A(H1N1)pdm09 was the dominant strain, accompanied by strain B, which maintained lower incidence rates, not exceeding 2,000 new cases per week. For the purpose of simplifying the analysis of the epidemic dynamics, we used the aggregate number of new cases across all strains.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_13

Table 3. Variables description for SEIR-EBCM model equations

Variable	Description	Initial value
$\theta(t)$	Probability that the node u did not transmit an infection to node v at a time step t	$1-\rho$
$\phi(t)$	Probability that the node u of an edge from u to v is exposed, and the edge did not transmit an infection at a time step t	$\varepsilon_\phi << 1$
$\psi(t)$	Probability that the node u of an edge from u to v is infectious, and the edge did not transmit an infection at a time step t	$\varepsilon_\psi << 1$
$S(t)$	Fraction of susceptible individuals in the population at time t	$1-\rho$
$E(t)$	Fraction of exposed individuals in the population at time t	0
$I(t)$	Fraction of infected individuals in the population at time t	ρ
$R(t)$	Fraction of recovered individuals in the population at time t	0

Table 4. Parameter description for SEIR EBCM equations

Variable	Description
β	Infection transmission rate over one edge
α	Rate for exposed nodes to become infectious
γ	Rate for infectious node to become recovered
ρ	Initial fraction of infectious nodes

3 Results

3.1 Synthetic population transformation

The synthetic population of Chelyabinsk was reduced using a sampling algorithm, so that only $r = 10\%$ of households were left. That resulted in approximately $N \approx 350000$ individuals compared to the full city population of 1189000. To evaluate the effect of the sampling algorithm on the contact network structure within the population, we generated histograms illustrating the distribution of households and workplaces sizes. The plots are presented in Fig. 3, showing that the distribution of household sizes moved to the right, while sizes of workplaces significantly decreased (from 8 to 2 individuals in average). Since in this study we do not calculate epidemic indicators, we consider these changes not crucial for fulfilling our goals, but in case of using the model in epidemic surveillance we consider using sampled populations with larger r, along with 'repopulating' workplaces to retain original average number of individuals in them.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_13

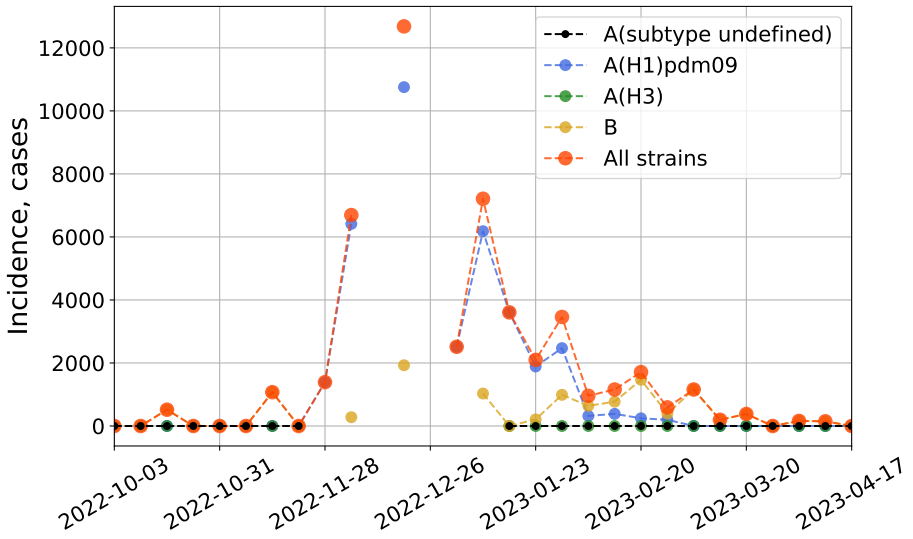

Fig. 2. Strain-specific influenza incidence data for 2022-2023 epidemic season in Chelyabinsk, Russia

As a next step, a contact network was generated from the synthetic population, according to the aforementioned algorithm, with its software implementation in Python by means of `networkx` library. The resulting contact network degree distribution is shown in Fig. 4.

To demonstrate the usage of the contact network, we performed a simulation on it using a stochastic SEIR network model. The stochastic simulation was made by Gillespie algorithm [5] from EoN (Epidemics on Networks) library [11]. In a Gillespie simulation, the timing of the subsequent event is determined by computing the rate of all possible events at the current state. A waiting time is then sampled from an exponential distribution characterized by that rate. Subsequently, an additional random number is utilized to select which specific event among the possible options will occur. In our case, initially, most nodes are susceptible, with a certain fraction ρ of them set as infected. Each infected node may transmit infection to its susceptible neighbors according to a defined transmission rate. Exposed nodes transition to infected states, and infected nodes recover based on corresponding rates. The simulation output is time series equal to those of a compartmental SEIR model, i.e. $S(t), E(t), I(t), R(t)$.

3.2 Model calibration

The second part of our method consists in feeding the created contact network to the edge-based SEIR model and launching calibration to data. The first 7 weeks on the graph with data were not considered for model calibration, as these data points represent only small fluctuations. We assume that a full-fledged epidemic

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_13

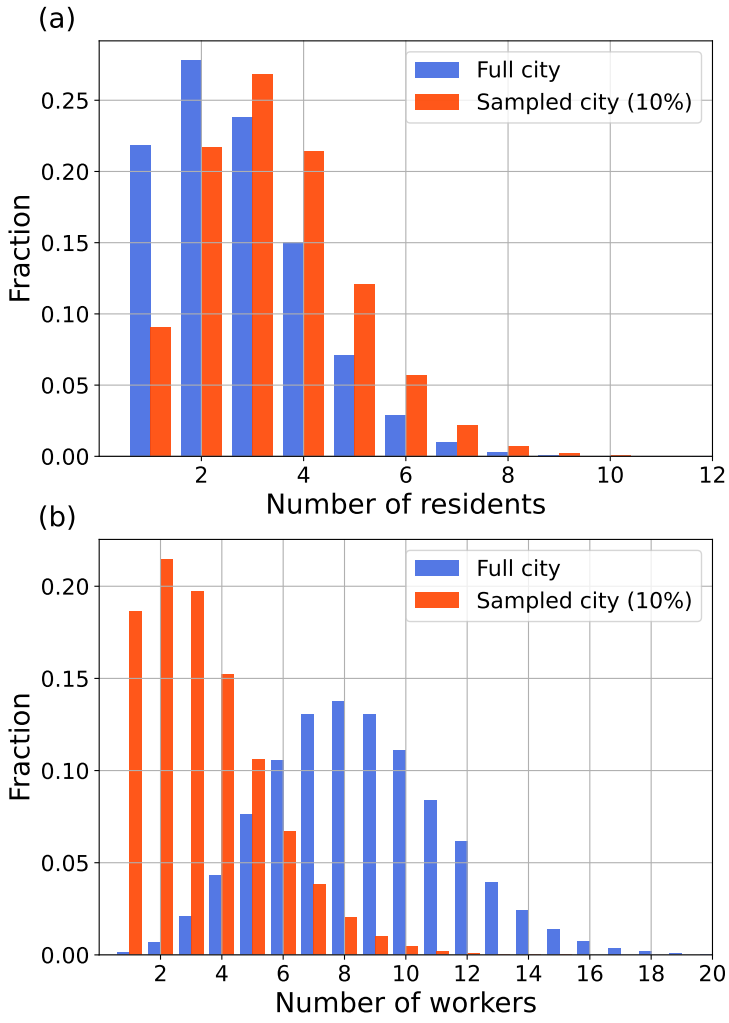

Fig. 3. Distribution of sizes of dwellings and offices according to synthetic population data

begins from week 7, when the growth of number of new cases becomes apparent. The calibration was conducted using a combination of automatic approach via simulated annealing method and a manual parameter tuning. The resulting parameter values are presented in Table 4. To measure the accuracy of calibration, the R^2 metric was used. The calibration result is shown in Figure 5.

The comparison of simulation time for different model types is presented in Table 6 and Fig. 6.

During the simulations, we have discovered an unpleasant effect which apparently takes place for some synthetic population structures. In Figure 7, we

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_13

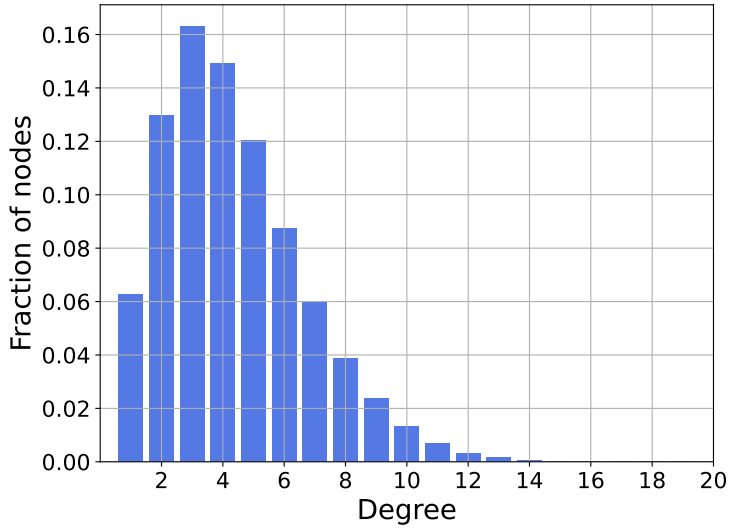

Fig. 4. Degree distribution of the contact network in a sampled synthetic population of Chelyabinsk

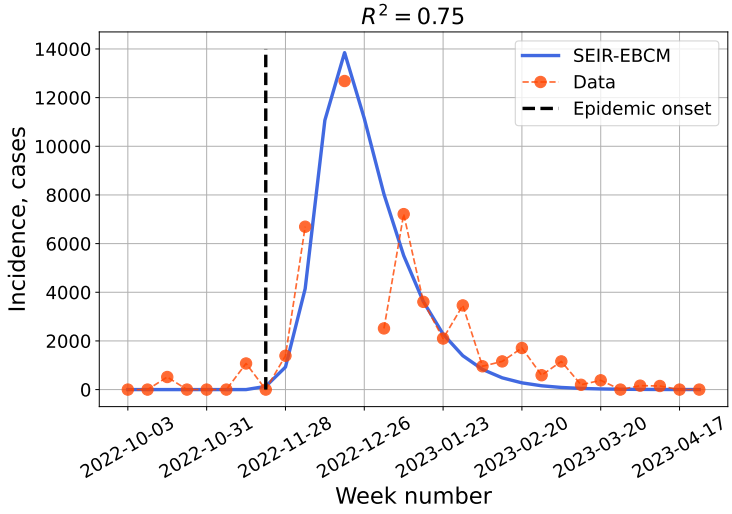

Fig. 5. SEIR EBCM calibration to influenza incidence data of 2022-2023 epidemic season in Chelyabinsk

demonstrate the comparison of simulation curves for two different types of input contact networks. While the incidence curves for the EBCM and network models are well approximated when using a Barabási-Albert contact network, these incidence curves may have significant discrepancies when modeled on a

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_13

Table 5. Parameter values for SEIR EBCM obtained by calibration

Parameter	Value
β	0.0038
α	0.1
γ	0.14
ρ	0.0005

Table 6. Comparison of simulation time of different models.

Model	Avg. simulation time, sec	Complexity
SEIR ODE	$\approx 10^{-3}$	Low
SEIR EBCM	$\approx 10^{-1}$	Low
SEIR network model	$\approx 10^{2}$	Medium
Multiagent model	$\approx 10^{4}$	High

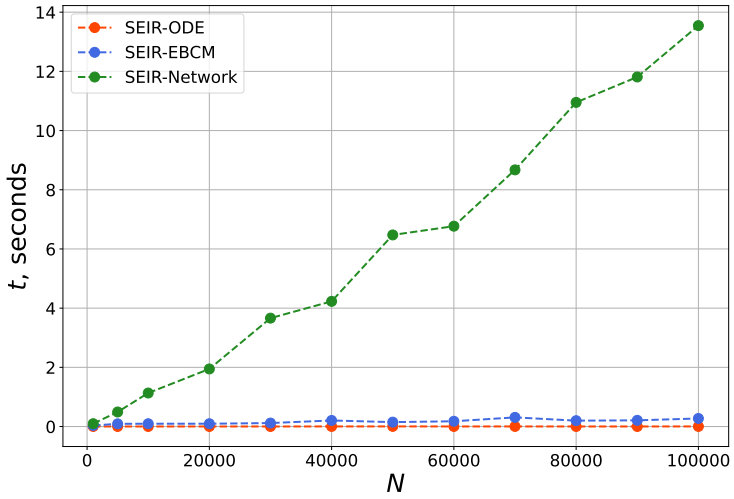

Fig. 6. Simulation time for different models depending on the number of nodes of the contact network graph. Experiments were conducted using Barabási-Albert network with $m = 5$

contact network constructed from the synthetic population data. Particularly in this case, EBCM failed to reproduce the bimodality of the incidence curve. To our knowledge, this effect was not reported in the studies we relied upon. Consequentially, the usage of EBCM seems to be limited at least to unimodal curves, otherwise it changes the disease dynamics. Further studies are planned to quantify this limitation, exploring dependency of curves approximation quality on sampling ratio r and synthetic population structure of different cities.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_13

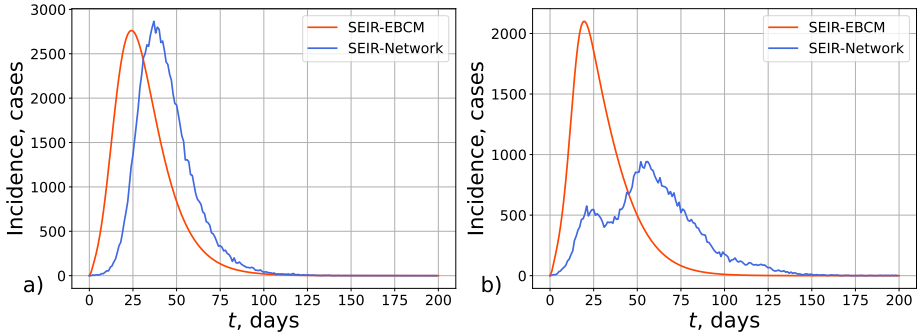

Fig. 7. Simulated daily incidence with different contact network graphs: a) simulation on Barabási-Albert network; b) simulation on a contact network graph based on synthetic population data.

4 Discussion

In this article, we proposed a method for fast and detailed modeling infection propagation which is based on combining synthetic population transformation and SEIR EBCM usage. To assess the applicability of the method, we compare different aspects of its usage with other modeling methods. Traditional methods, such as the SEIR ODE approach, offer fast simulation time and relatively straightforward calibration, making them appealing for rapid assessments of epidemic indicators and incidence prediction. However, these models fail to account for heterogeneous effects within populations. In contrast, multiagent modelling with synthetic populations provides a detailed view of epidemic spread at the individual level, capturing complex interactions and behavior patterns of individuals. Despite its advantages, multiagent modeling is hindered by high simulation time, which greatly complicates model calibration and data assimilation. Additionally, creating synthetic populations for these models is challenging and time-consuming, and these populations quickly become outdated due to demographic changes. In our method, using SEIR EBCM for modeling offers low simulation time and makes it possible to capture effects connected with population of concrete city using a transformed synthetic population. SEIR EBCM approach makes it an attractive choice for handling the complexities of epidemic modeling effectively. By integrating the benefits of different modeling techniques, SEIR-EBCM can provide more accurate predictions and better support public health strategies. However, to capture the effects connected with topology of each city the construction of synthetic populations is also needed.

In future studies, we plan to enhance our approach by upgrading the sampling techniques to preserve the distribution of apartment and work office sizes more accurately. Additionally, we aim to assess the feasibility of approximating contact networks based on synthetic populations using typical network models such as Barabási-Albert or Erdős-Rényi. Furthermore, we intend to employ this

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_13

model for epidemic surveillance as part of a modeling framework that will allow ensemble forecasting and accurate assessing of epidemic indicators, such as R_t and fractions of immune population across different age groups.

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

References

1. Open Street Map, https://www.openstreetmap.org/, April 23, 2025
2. Yandex Audience, https://audience.yandex.ru/, April 23, 2025
3. Alota, C.P., Pilar-Arceo, C.P., de los Reyes V, A.A.: An edge-based model of seir epidemics on static random networks. Bulletin of Mathematical Biology **82**(7), 96 (2020)
4. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. science **286**(5439), 509–512 (1999)
5. Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. The journal of physical chemistry **81**(25), 2340–2361 (1977)
6. Kiss, I.Z., Miller, J.C., Simon, P.L., et al.: Mathematics of epidemics on networks. Cham: Springer **598**(2017), 31 (2017)
7. Korzin, A.I., Kaparulin, T.I., Leonenko, V.N.: Assessing the applicability of the multiagent modeling approach to the epidemic surveillance of COVID-19 in Russian cities. In: 2024 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). pp. 237–242. IEEE (2024)
8. Leonenko, V., Arzamastsev, S., Bobashev, G.: Contact patterns and influenza outbreaks in Russian cities: A proof-of-concept study via agent-based modeling. Journal of Computational Science **44**, 101156 (2020)
9. Leonenko, V.N.: Herd immunity levels and multi-strain influenza epidemics in Russia: a modelling study. Russian Journal of Numerical Analysis and Mathematical Modelling **36**(5), 279–291 (2021)
10. Miller, J.C., Slim, A.C., Volz, E.M.: Edge-based compartmental modelling for infectious disease spread. Journal of the Royal Society Interface **9**(70), 890–906 (2012)
11. Miller, J.C., Ting, T.: EoN (epidemics on networks): a fast, flexible Python package for simulation, analytic approximation, and analysis of epidemics on networks. arXiv preprint arXiv:2001.02436 (2020)
12. Rineer, J., Kruskamp, N., Kery, C., Jones, K., Hilscher, R., Bobashev, G.: A national synthetic populations dataset for the United States. Scientific Data **12**(1), 144 (2025)
13. Shang, Y.: SEIR epidemic dynamics in random networks. International Scholarly Research Notices **2013**(1), 345618 (2013)
14. Wang, Y., Cao, J., Alsaedi, A., Ahmad, B.: Edge-based SEIR dynamics with or without infectious force in latent period on random networks. Communications in Nonlinear Science and Numerical Simulation **45**, 35–54 (2017)
15. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. nature **393**(6684), 440–442 (1998)
16. Wheaton, W.D., Cajka, J.C., Chasteen, B.M., Wagener, D.K., Cooley, P.C., Ganapathi, L., Roberts, D.J., Allpress, J.L.: Synthesized population databases: A US geospatial database for agent-based models. Methods report (RTI Press) **2009**(10), 905 (2009)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_13