Доклады Международной конференции “Математическая биология и биоинформатика”. Под ред. В.Д. Лахно. Том 8. Пущино: ИМПБ РАН, 2020. Статья № e33. doi: 10.17537/icmbb20.29

A demographic microsimulation model for the long-term evolution of synthetic populations in Saint-Petersburg

Arzamastsev S.A., Leonenko V.N.

ITMO University, Saint-Petersburg, Russian Federation

vnleonenko@itmo.ru

In this paper, we present a microsimulation model (MSM) that allows to simulate demographic processes in agent populations. As its input, the model uses a synthetic population, which consists of independent agents with defined dwellings and workplaces. The model allows not only to monitor the changes in demographics of the population over time but also to consider the geospatial distribution of individuals and their activities in a regarded urban setting. Using open-access 2010–2018 demographic data for St. Petersburg, we have assessed the changes in the states of the agents associated with aging, birth, emigration, immigration, and formation of new households. The modelling algorithm is implemented in Python programming language. To demonstrate the capabilities of the model, we derived a synthetic population of Saint-Petersburg for 2018 from the synthetic population of 2010. The results of the modelling are aligned with the available aggregated statistical data. The synthetic populations created with the help of the model allow fine-scale simulations of the outbreaks of influenza and COVID-19 in Russian cities.

Key words: Microsimulation Modeling, Population Projections, Demographic Processes, Artificial Societies, Python.

Демографическая микросимуляционная модель для долгосрочной эволюции синтетических популяций в Санкт-Петербурге

Арзамасцев С.А., Леоненко В.Н.

Университет ИТМО, г. Санкт-Петербург, Российская Федерация

В данной работе представлена микросимуляционная модель (МСМ), которая позволяет моделировать демографические процессы в человеческих популяциях. В качестве входных данных модель использует синтетическую популяцию, которая состоит из независимых индивидов с определенными местами жительства и рабочими местами. Модель позволяет не только отслеживать изменения в демографии населения с течением времени, но и учитывать пространственное распределение людей и их активность в рассматриваемой городской среде. Используя демографические данные открытого доступа за 2010–2018 гг. для Санкт-Петербурга, мы оценили вероятности изменений состояний индивидов, связанных со старением, рождением, миграцией и образованием новых домашних хозяйств. Алгоритм моделирования реализован на языке программирования Python. Для демонстрации возможностей модели мы построили синтетическую популяцию Санкт-Петербурга 2018 года на основе известной синтетической популяции 2010 года. Корректность результатов моделирования подтверждена их сравнением с имеющимися агрегированными статистическими данными. Синтетические популяции, созданные с помощью модели, позволяют моделировать вспышки гриппа и COVID-19 в городах России на уровне отдельных индивидов.

Ключевые слова: *микромоделирование, демографические прогнозы, демографические процессы, искусственные сообщества, Python.*

1. Introduction

While there is a growing popularity of agent-based modelling as a tool of tackling the challenges of epidemiology and public health, a bottleneck in the approach which hinders its successful application is the necessity in highly detailed individual-level demographic data. Most of the time, the corresponding temporal data is at best fragmentary and lacking details. Sometimes the parts of the time series are outright absent which makes it impossible to model the processes in the population. Since the changes in number and composition of population are the key factors for influenza outbreak determination [1–4], crime rate prediction [5], for evaluation of geospatial patterns of opioid drug users [6] and people with chronic diseases [7], obtaining plausible demographic information to operate the models is a must. This data is also necessary for the ongoing research in the area of COVID-19 epidemic modeling and forecasting [8].

A method which allows to synthesize missing data and thus to obtain the input for agent-based models consists in the utilization of individual-level demographic models [9], [10], also known as microsimulation models [11], or MSMs. Demographic model are used to reconstruct the changes in initial synthetic population by modelling the events which happen with each agent in the population over the years. As a result, one can get an "evolved" synthetic population without detailed demographic data (in this case the justification of the obtained populations is possible using the aggregated data) and even project it into the future. In the current study, we develop such a model using the idea from RTI International team [12]. We use the same standard of the synthetic populations [13]. However, since the Russian statistical data is incompatible with the similar US data (for instance, the public microsurveys, wildly used for demographic modeling in the US, are, to authors' knowledge, not available in Russia), we had to build our demographic model from scratch. The main aim of the presented model is to integrate the impact of demographic indicators on population structure and density in medium- and long-term period using Saint Petersburg as a case study. To use the model with the data that includes other groups of population or cities, it is necessary to update characteristics related to mortality, birth, migration, as well as the actual list of the populated households. The current modular implementation of the model allows considering additional demographic processes in synthetic populations by adding new methods to existing ones.

2. Methods

An algorithm that is described in this article is a microsimulation model (MSM) with discrete time, one step being equal to one year. The algorithm connects the demographic events and changes in household structures. The implemented methods reflect such processes as aging, death, birth, and migration [14]. Every event is expected to happen only once in the regarded period of time (a year).

2.1. Data

To calibrate the algorithm, we use the data for Saint-Petersburg in 2011–2018 that has been taken from open sources [15]. Namely, the following information was extracted from the database:

- Yearly number of the deceased,
- Mortality coefficient of men and women per 1000 of the population,
- Age composition of population,
- Number of fertile women per 1000 women according to age categories,
- Yearly number of newborn boys / girls,
- Yearly number of the migrated,
- Emigrants/immigrants fraction for both genders in the each given age group.

Due to the fact that migration data from 2011, 2012, 2015 and 2016 is absent, we have used a Lagrange polynomial to complete the gaps in data. In particular, to determine the portion of emigrants and immigrants according to the gender we have used migration data of 2010 and 2019 in order to increase the accuracy of prediction.

2.2. Filling the gaps

At each step we gradually determine the number of agents in every age group (the period is 5 years). After that we need to determine the number of agents in each age (with the step of 1 year) to collect the data set. Value of series divided by 5 is assigned to the node $n + 2$ from age category $n..n + 4$. Thereafter the updated data is interpolated by Lagrange polynomial in order to get the data for the age groups with the age step equal to one year (15-year-olds, 16-year-olds, etc.). At the last stage, we consider the influence of the demographic processes on the new age distribution. For example, if there is an emigration, we should remove the corresponding agent data from population.

To get data according to the age split encountered in statistical data, we use the Lagrange polynomial [16]. This is a polynomial $P(x)$ of the degree $\leq n-1$ that passes through n points: $(x_1, f(x_1))$, $(x_2, f(x_2))$,… This polynomial is determined in the following way:

$$P(x)=\sum_{i=1}^{n} p_i(x),$$

$$p_i(x)=y_i\prod_{\substack{k=1\\ k\neq j}}^{n}\frac{x-x_k}{x_j-x_k},$$

where x is the unknown point's distance to the first point,
x_i are all the roots of the polynomial, except x_k,
j is the sequence number of point j,
k is the sequence number of point k, $k \neq j$,
n is the degree of the polynomial.

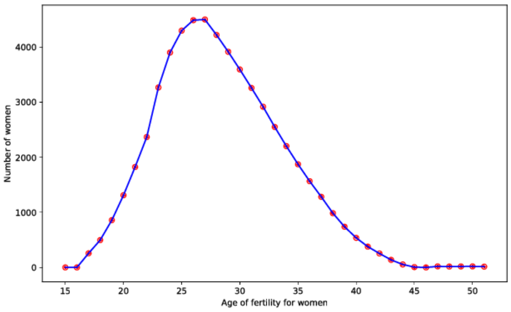

Fig. 1. Lagrange interpolation algorithm application results for the number of fertile women.

We use the curves with $n = 3$ because the biases increase with the growth of n. In Figure 1, we demonstrate the implementation of Lagrange method that determines the number of fertile women in each age.

2.2. Model structure

The model consists of 5 main modules:

- Mortality. In each age group we determine a number of agents to be removed from the population using the coefficients of mortality. Then we apply the Lagrange interpolation to distribute the removed among the age groups. In the end, the agents of relevant age are removed.
- Fertility. In every age group of fertile women (those with the ages from 16 to 52, according to [13]), we determine the number of agents which gave birth, along with the gender of their child. By the application of Lagrange interpolation method, we assign child's genders using data of the number of newborn boys and girls. The newborns are added to the household of a mother. We also consider the possibility for a mother to have more than one child.
- Emigration. In each age group we determine a number of men and women that will be removed from the population. Modeling is done in the similar way as the modeling of mortality but in this case the used data is the age group size and the fraction of emigrants that leave the city, taking into account their gender.
- Immigration. In each age group we determine a number of agents of each gender that will be added in population according to the known fraction of immigrated men and women in age groups and total number of immigrated people. Having implemented Lagrange method on two data sets, we only need to distribute immigrants into households. This is determined according to the population density in each household. We consider number of residents living in every household as the indicator of frequency of the household where agents live.
- Aging. At this stage we add to each agent one year to his/her current age.

While modeling demographic processes in Saint-Petersburg from 2011 to 2018 we take into account the aspect of building the new households. We had information about the list of objects that were built during this period and the households' capacity but we did not have information about the particular years of entry into service for each of the corresponding building. Every household was given a generated year of entry into service to match the total annual number of new households equal to the corresponding aggregated data available for Saint-Petersburg.

3. Results

In Figure 2, we present how the population fractions of every age category presented in Petrostat demographic data (0–17, 18–30, 31–50, 51–62, 63+) is changing over the years. There is a trend of increasing number of minors that is related to increased birth rate. There is also a decreased portion of young population (18–30) because of relatively low number of minors in the beginning of simulation.

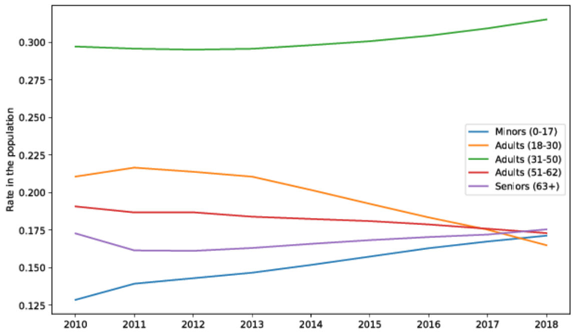

Fig. 2. Dynamics of age distribution in the population of St. Petersburg over time.

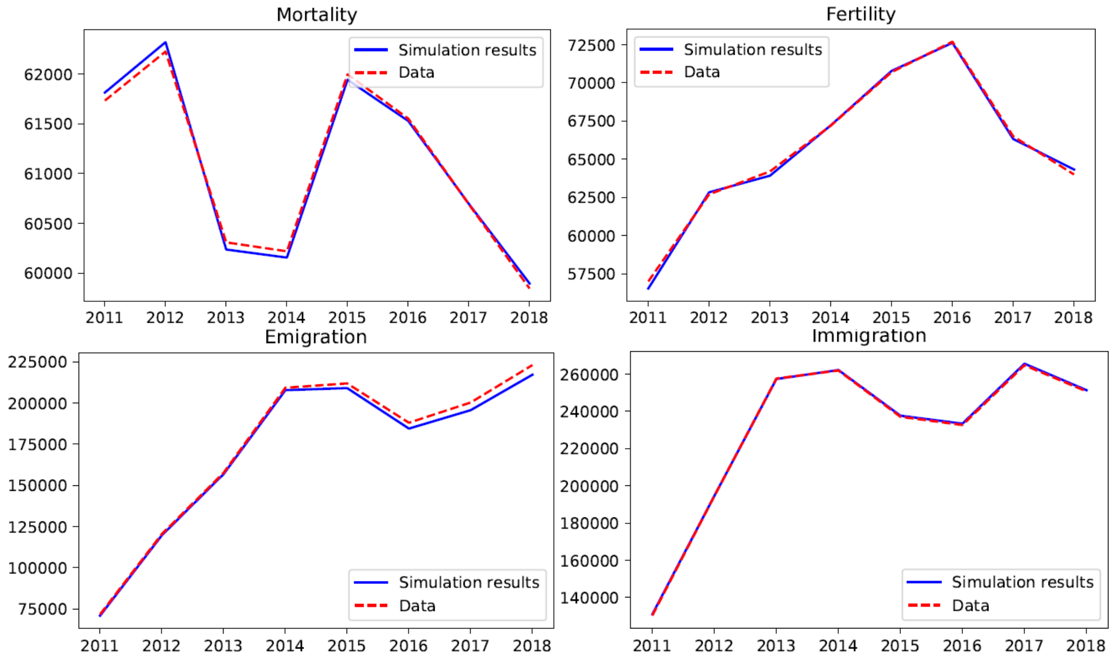

Fig. 3. The comparison of simulated demographic data against the statistics.

In Table 1 and Figure 3, we compare the obtained numbers of agents participating in demographic processes with the real data. Values of MAPE (Mean Absolute Percentage Error) for computational results of life processes in MSM are presented in table 1.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|Y_{\text{fact}} - Y_{\text{perd}}\right|}{Y_{\text{fact}}}$$

Table 1. MAPE for the simulated demographic processes

	Mortality	Fertility	Emigration	Immigration
MAPE	0.09 %	0.31 %	1.39 %	0.23 %

In Figure 4, we present the distribution of households in Saint-Petersburg on the city map. We depict households built before 2011 and used in 2010 synthetic population [17] as green points and households reconstructed by the demographic model according to data [15] as blue ones.

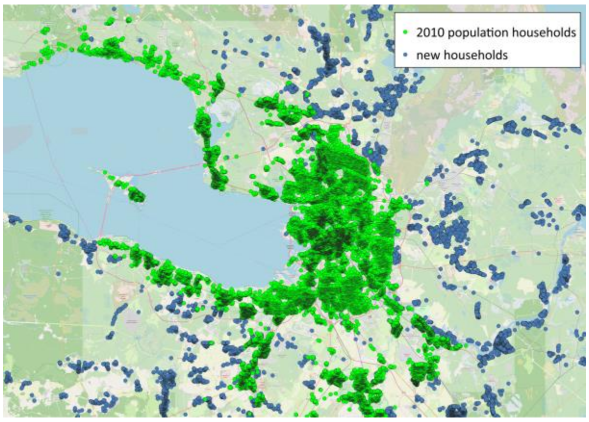

Fig. 4. The distribution of households in St Petersburg.

4. Discussion and future work

In this article, we presented a microsimulation model that portrays the demographic processes of Saint-Petersburg population at the fine scale. The aim of the model is to simulate as accurately as possible the processes in the population over time based on available initial data (2010 synthetic population dataset in our case). We demonstrate that this approach is an effective method to obtain a set of synthetic populations when they cannot be generated directly from the detailed data. The model can be extended by adding other processes which are currently not considered, for example, internal migrations, marriages, and divorces. The model is also planned to be used to simulate demographic processes in other Russian cities.

The current MSM was created within the research aimed at getting more accurate and relevant populations of Saint-Petersburg. The next step is to accurately assign the agents to their activity places such as schools or workplaces. This task requires further research related to employing additional demographic parameters. The resulting populations will have all the necessary attributes to simulate interactions between individuals, which makes it possible to study infection propagation through this population, including COVID-19 outbreak simulations.

5. Acknowledgments

This research is financially supported by The Russian Science Foundation, Agreement #20-71-00142.

6. References

1. Leonenko V., Bobashev G. Analyzing influenza outbreaks in Russia using an age-structured dynamic transmission model. *Epidemics*. 2019. V. 29. doi: 10.1016/j.epidem.2019.100358.
2. Cooley P.C., Bartsch S.M., Brown S.T., Wheaton W.D., Wagener D.K., Lee B.Y. Weekends as social distancing and their effect on the spread of influenza. *Computational and Mathematical Organization Theory*. 2016. V. 22. № 1. P. 71–87.
3. Cooley P., Brown S., Cajka J., Chasteen B., Ganapathi L., Grefenstette J., Hollingsworth C.R., Lee B.Y., Levine B., Wheaton W.D. et al. The role of subway travel in an influenza epidemic: a New York City simulation. *Journal of Urban Health*. 2011. V. 88. № 5. P. 982.
4. Grefenstette J.J., Brown S.T., Rosenfeld R., DePasse J., Stone N.T., Cooley P.C., Wheaton W.D., Fyshe A., Galloway D.D., Sriram A. et al. FRED (a framework for reconstructing epidemic dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC Public Health*. 2013. V. 13. № 1. P. 940.
5. Ingilevich V., Ivanov S., Crime rate prediction in the urban environment using social factors. *Procedia Computer Science*. 2018. V. 136. P. 472–478. doi: 10.1016/j.procs.2018.08.261.
6. Bates S., Leonenko V., Rineer J., Bobashev G., Using synthetic populations to understand geospatial patterns in opioid related overdose and predicted opioid misuse. *Computational and Mathematical Organization Theory*. 2019. V. 25. P. 36–47. doi: 10.1007/s10588-018-09281-2.
7. Leonenko V.N. Analyzing the Spatial Distribution of Acute Coronary Syndrome Cases Using Synthesized Data on Arterial Hypertension Prevalence. In: *Computational Science – ICCS 2020*. Springer International Publishing, 2020. P. 483–494. doi: 10.1007/978-3-030-50423-6_36.
8. Hoertel N., Blachier M., Blanco C., Olfson M., Massetti M., Limosin F., Leleu H. Facing the COVID-19 epidemic in NYC: a stochastic agent-based model of various intervention strategies. *medRxiv*. 2020. doi: 10.1101/2020.04.23.20076885.

9. Spielaue M., Dupriez O. A portable dynamic microsimulation model for population, education and health applications indeveloping countries. *International Journal of Microsimulation*. 2019. V. 12. № 3. P. 6–27. doi: 10.34196/ijm.00205.
10. Fatmia M.R., Habib M.A. Microsimulation of life-stage transitions and residential location transitions within a life-oriented integrated urban modeling system. *Computers, Environment and Urban Systems*. 2018. V. 69. P. 87–103. doi: 10.1016/j.compenvurbsys.2018.01.003.
11. Bae J.W., Paik E., Kim K., Singh K., Sajjad M. Combining microsimulation and agent-based model for micro-level population dynamics. *Procedia Computer Science*. 2016. V. 80. P. 507–517. doi: 10.1016/j.procs.2016.05.331.
12. Rogers S., Rineer J., Scruggs M., Wheaton W., Cooley P., Roberts D., Wagener D. A geospatial dynamic microsimulation model for household population projections. *International Journal of Microsimulation*. 2014. V. 7. № 2. P. 119–146.
13. 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. V. 2009. № 10. P. 905. doi: 10.3768/rtipress.2009.mr.0010.0905.
14. Symonds P., Hutchinson E., Ibbetson A., Taylor J., Milner J., Chalabi Z., Davies M., Wilkinson P. Microenv: A microsimulation model for quantifying the impacts of environmental policies on population health and health inequalities. *Science of the Total Environment*. 2019. V. 697. P. 1–10. doi: 10.1016/j.scitotenv.2019.134105.
15. *Federal State Statistic Service open data*. URL: https://petrostat.gks.ru/folder/32168 (accessed 14.09.2020).
16. Zheng Y., Li X., Li M., Li X., Tang D. Modeling road surface and network from a 3d perspective. In: *2nd International Conference on Computer Engineering and Technology*. 2010. P. 186–190. doi: 10.1109/ICCET.2010.5485244.
17. 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*. 2020. V. 44. doi: 10.1016/j.jocs.2020.101156.