ITM Web of Conferences **31**, 03002 (2020) https://doi.org/10.1051/itmconf/20203103002
Mathematical Modelling in Biomedicine 2019

Analyzing the spatial distribution of individuals predisposed to arterial hypertension in Saint Petersburg using synthetic populations

Vasiliy N. Leonenko[1,*] and *Sergey V.* Kovalchuk[1]

[1]ITMO University, 49 Kronverksky Pr., St. Petersburg, 197101, Russia

Abstract. Arterial hypertension (AH) is one of the most common cardiovascular diseases, and it can lead to serious complications. To optimize the delivery of patients exposed to AH to medical institutions and thus to curtail mortality in Russian cities caused by the consequences of hypertension, it is necessary to estimate the number of potential patients, along with their spatial distribution. This paper presents a method which uses synthetic population data to assess the spatial distribution of individuals potentially prone to arterial hypertension. The risk of arterial hypertension of an individual is calculated based on its demographic characteristics (age and gender). Using Saint Petersburg as a case study, we demonstrate that the mentioned approach makes it possible to perform predictions of AH cases distribution in absence of real data on hypertension status of the individuals. The results of the study will be used to assess the input flows of patients to healthcare facilities and optimize their workflow.

1 Introduction

Arterial hypertension (AH), a major cause of premature deaths worldwide, is a medical condition associated with elevated blood pressure. For 2019, hypertension affects around 1.13 billion people. The aim of reducing the prevalence of hypertension by 25% by 2025 is one of the global targets for noncommunicable diseases [1].

To prevent mortality of individuals from the consequences of hypertension in Russian cities, it is necessary to estimate the number and the spatial distribution of city dwellers exposed to hypertension. Such information will help to optimize the delivery of patients to medical institutions and their treatment [2].

This paper presents an algorithm for modeling the spatial distribution of individuals exposed to arterial hypertension. The proposed algorithm assesses the distribution of individuals prone to hypertension based on the synthetic populations technique coupled with the statistical model of the incidence of AH [3] and Monte Carlo methods. We demonstrate the output produced by the algorithm using the population of Saint Petersburg as a case study.

2 Synthetic population

A "synthetic population" is a synthesized, spatially explicit human agent database (essentially, a simulated census) representing the population of a city, region or country. By its

*e-mail: vnleonenko@yandex.ru

cumulative characteristics, this database is equivalent to the real population, but its records does not correspond to real people. Statistical and mechanistic models built on top of the synthetic populations helped tackle a variety of research problems, including those connected with public health. One of the standards of synthetic populations is proposed and developed by RTI International [4]. It was used by various research groups to create populations for 50 US states, along with another regions and countries. The authors have experience of working with the synthetic populations from RTI, using it for statistical analysis of opioid–related overdoses in Cincinnatti [5] and for the investigation of the dynamics of influenza in Russian cities [6].

According to the standard of RTI International, a synthetic population consists of several txt-files, each of them containing a table with every row being a single record corresponding to some entity – an individual, a household, a workplace, a school, etc. The principal data is stored in four files: `people.txt` (each record contains id, age, gender, household id, workplace id, school id), `households.txt` (contains id and coordinates), `workplaces.txt` (contains id, coordinates and capacity of the workplaces), `schools.txt` (contains id, coordinates, capacity). Using the RTI synthetic population protocol as a reference, we generated the corresponding files for the population of Saint Petersburg. Since the data available for Saint Petersburg was not complete, we altered or outright omitted some of the methods used in the original procedure. That resulted in a synthetic data, which might not exactly match the real population. We regard these data as an initial approximation and we plan to make our synthetic population more accurate employing additional sources of information.

The principal data source for our synthetic population is 2010 data from "Edinaya sistema ucheta naseleniya Sankt Peterburga" ("Unified population accounting system of Saint Petersburg") [7]. The data is represented in a form of Excel spreadsheets containing records with house addresses and the corresponding number of dwellers of certain age and gender. To match the household addresses with the geographical coordinates and assess the plausibility of the obtained geographical data, a computational algorithm was developed and implemented using Python programming language. The named algorithm employed a geocoder to detect locations of the addresses and filtered out errors in the results with multiple addresses matching only one pair of coordinates. In the same fashion the coordinates of schools were assigned, based on the school list from the official web–site of the Government of Saint Petersburg [8].

The distribution of working places for adults and their coordinates were derived from the data obtained with the help of Yandex.Auditorii API [9]. Initially the data was available in a form of a file in GeoJSON format (it is an open standard format designed for representing simple geographical features, along with their non-spatial attributes). The file consisted of relative workplace size assessments for each of the cells in a hexagonal grid. This data was normalized using the official cumulative employment numbers [10]. Synthetic workplace records were created by assigning the calculated number of employees in each hexagonal cell to imaginary geographical location coinciding with the center of this cell.

We assumed that young people aged 7 to 17 attend schools, and the adults of working age (18 to 55 for females and 18 to 60 for males) might be working. Iterating through the list of records for the city dwellers generated earlier, we were assigning each person to a closest school or working place, until they are filled to capacity or there is no more people to be assigned. More details on the population used and the algorithms for its generation employed could be found in [6].

3 Assessing AH risk and individual AH status

When the synthetic population is created, we assess the health conditions of individuals associated with arterial hypertension. There are two types of corresponding data that we generate and add to the individual records of the synthetic population:

- The AH risk (the probability of having arterial hypertension). Based on [3], we assumed that the mentioned probability depends on age and gender of an individual. The corresponding cumulative distribution function is shown in Fig. 1. The example of individuals distribution on the map with their AH risk divided into three categories (low, average and high) is shown in Fig. 2.
- The actual AH status (positive or negative). The corresponding value (0 or 1) is generated by the Monte Carlo algorithm according to the AH risk calculated in the previous step. The AH status might be used in simulation models which include demographic processes and population-wide simulation of the onset and development of AH.

At the moment, none of the parameters of the synthetic population were used for the hypertension status assessment aside from age and gender, but we consider adding more factors to the statistical model, where the workplace distribution information might also be utilized. For example, adding workplace type into the synthetic population data will allow us to assume the amount of daily physical activity of individuals working there, which is one of the factors correlated (inversely) with the AH probability [1]. Also, the geographical locations of the individuals' workplaces may be used in simulation models to calculate the distribution of cases of acute conditions ocurring in working hours and thus to facilitate balancing of the incoming flows to healthcare facilities.

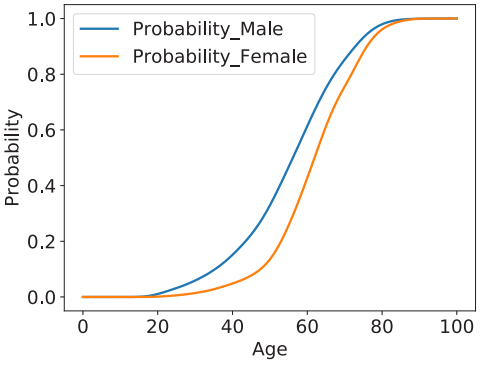

Figure 1. The cumulative distribution function used to define the AH status of an individual, based on data from [3].

4 Results

To visualize the spatial distribution of the individuals prone to AH (shortly, AH+ individuals), we convert the coordinates of an individuals' household locations from degrees to meters using Mercator projection. After that we form a grid with cell size 250m x 250m, defining its bounds with maximum and minimum coordinates of the regarded synthetic individuals. Then we calculate the overall number of dwellers and the number of individuals exposed to AH for each cell of the grid. The algorithm was implemented as a collection of scripts written

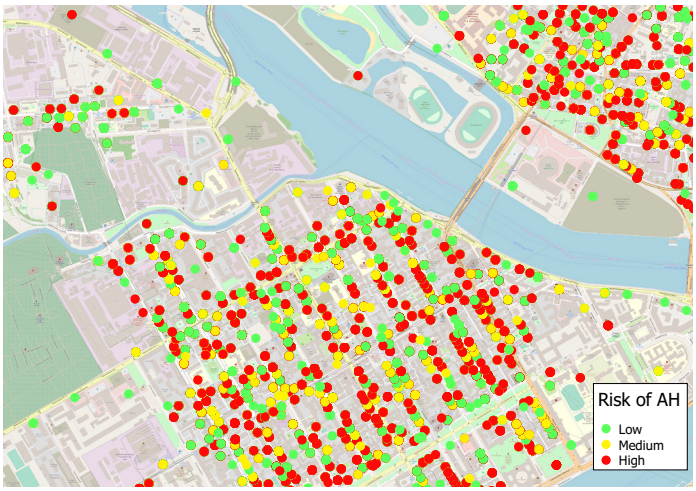

Figure 2. The distribution of individuals of different risk categories to arterial hypertension

in Python programming language with the libraries `numpy`, `matplotlib`, and `pandas`. The output of the algorithm is a txt-file with the coordinates of the cells and the cell statistics (overall number of individuals, number of AH+ individuals). As a final step, we use QGIS open software to draw cells on the map of Saint Petersburg (OpenStreetMap background layer). The result is demonstrated in Fig. 3.

Figure 3. Density of the individuals exposed to arterial hypertension in the cells with size 250m x 250m

As it is seen from the map, the discrete representation of the density of individuals with arterial hypertension sometimes leads to sharp borders between the cells, which is not plausible. Also, the selection of the cell size strongly affects the visualization outcome. To resolve these issues, we used an interpolation function in QGIS to smoothen the transition between the regions with different densities. In Fig. 4, we demonstrated a part of the city map cor-

responding to Vasilyevsky island, one of the central districts of Saint Petersburg, with the interpolated data on the density of AH+ individuals.

The proportion of synthesized AH+ individuals in the population obtained with the help of our algorithm is not uniformly distributed in space. Particularly, it can be seen in Fig. 4 that there are certain spots of high density of the people with AH in the south and north–west, and there is a spot of low density of AH in the south–west of the island. Since the statistical model utilized to assess the AH risk considers only age and gender as AH development factors, the heterogeneity of AH+ clusters basically reflects the heterogeneity of demographic parameters distribution of the population. To confirm or deny the hypothesis that those factors are essential for AH assessment and that our method reproduces AH+ distribution with satisfactory accuracy, additional spatially explicit datasets should be used. We plan to use the hospital data on emergency calls in Saint Petersburg in 2015 which were related to acute coronary syndrome (ACS) [11]. Since ACS is one of the most widespread consequences of AH [1], our idea is to compare different neighborhoods by the value of the ratio of total number of ACS cases to the assessed number of AH+ individuals. The same technique was earlier developed and applied by the authors to the ratio between the number of calls to emergency services related to opioid overdoses and the assessed number of opioid drug users in Cincinnatti [5].

Figure 4. Assessment of the density of AH+ individuals in Vasilyevsky island, St Petersburg, using the interpolated data

5 Discussion

In this study we presented an approach based on the synthetic population concept that makes it possible to perform analysis of the spatial distribution of individuals predisposed to arterial hypertension. The algorithms for assessing and visualizing the density of city dwellers prone to AH are implemented using Python and QGIS open software. It is worth noting that the approach could be used to analyze the distribution of any chronic condition if the corresponding population data is available. The next planned improvement is to implement hotspot analysis based on Getis–Ord Gi* statistics and to mark statistically significant anomalies on the map (unusually high and unusually low density of AH cases compared to the city average). Also,

a separate algorithm will be developed to assess the prospected incoming flow of patients with AH to a hospital in a fixed map location.

There is a number of drawbacks of the study which should also be addressed, namely:

- At the moment we use the synthetic population based on the 2010 census data. This population does not take into account the newly constructed housing and the increased number of dwellers in Saint Petersburg. We plan to employ the 2018 data to bring the population up to date.
- The current visualization approach makes it somewhat difficult to distinguish the areas with low density of AH cases and scarcely populated areas (for instance, industrial zones). The way to overcome this issue might be in regarding a special indicator instead of AH density.
- The statistical model could be modified to include the dependency of AH probability on additional parameters, apart from age and gender, for instance, alcohol and tobacco consumption and body mass index of a person [1].

We assume that these modifications will further enhance the accuracy of the algorithm and will make it a useful tool for the healthcare specialists.

6 Acknowledgements

This research is financially supported by The Russian Science Foundation, Agreement #19-11-00326. The authors thank Dr. Anna Semakova for providing the data for the cumulative distribution function which was used to assign the arterial hypertension status.

References

[1] WHO, *Hypertension. Fact sheet*, [online], https://www.who.int/news-room/fact-sheets/detail/hypertension

[2] S.V. Kovalchuk, A.A. Funkner, O.G. Metsker, A.N. Yakovlev, Journal of Biomedical Informatics **82**, 128 (2018)

[3] A. Semakova, N. Zvartau, Procedia Computer Science **136**, 433 (2018)

[4] W.D. Wheaton, J.C. Cajka, B.M. Chasteen, D.K. Wagener, P.C. Cooley, L. Ganapathi, D.J. Roberts, J.L. Allpress, Methods report (RTI Press) **2009**, 905 (2009)

[5] S. Bates, V. Leonenko, J. Rineer, G. Bobashev, Computational and Mathematical Organization Theory **25**, 36 (2019)

[6] V. Leonenko, A. Lobachev, G. Bobashev, *Spatial Modeling of Influenza Outbreaks in Saint Petersburg Using Synthetic Populations*, in *International Conference on Computational Science* (Springer, 2019), pp. 492–505

[7] Government of Saint Petersburg, *Labor and employment committee. information on economical and social progress*, [online], http://rspb.ru/analiticheskaya-informaciya/razvitie-ekonomiki-i-socialnoj-sfery-sankt-peterburga/ (In Russian)

[8] Government of Saint Petersburg, *Official web-site*, [online], https://www.gov.spb.ru/

[9] Yandex, *Auditorii*, [online], https://audience.yandex.ru/

[10] Government of Saint Petersburg, *Edinaya sistema ucheta naseleniya Sankt Peterburga (Unified population accounting system of Saint Petersburg)*, [online], https://reestr-gis.spb.ru (In Russian)

[11] S.V. Kovalchuk, M.A. Moskalenko, A.N. Yakovlev, *Towards model-based policy elaboration on city scale using game theory: application to ambulance dispatching*, in *International Conference on Computational Science* (Springer, 2018), pp. 404–417