Analyzing the spatial distribution of acute coronary syndrome cases using synthesized data on arterial hypertension prevalence *

Vasiliy N. Leonenko[1,2][0000−0001−7070−6584]

[1] ITMO University, 49 Kronverksky Pr., St. Petersburg, Russia 197101
[2] Almazov National Medical Research Centre, 2 Akkuratova st., St. Petersburg, Russia 197341 vnleonenko@yandex.ru

Abstract. In the current study, the authors demonstrate the method aimed at analyzing the distribution of acute coronary syndrome (ACS) cases in Saint Petersburg. The employed approach utilizes a synthetic population of Saint Petersburg and a statistical model for arterial hypertension prevalence. The number of ACS–related emergency services calls in an area is matched with the population density and the prospected number of individuals with arterial hypertension, which makes it possible to find locations with excessive ACS incidence. Three categories of locations, depending on the joint distribution of the above-mentioned indicators, are proposed as a result of data analysis. The method is implemented in Python programming language, the visualization is made using QGIS open software. The proposed method can be used to assess the prevalence of certain health conditions in the population and to match them with the corresponding severe health outcomes.

Keywords: Acute coronary syndrome · Arterial hypertension · Synthetic populations · Statistical modeling · Python.

1 Introduction

Acute coronary syndrome (ACS) is a range of health conditions associated with a sudden reduced blood flow to the heart. This condition is treatable if diagnosed quickly, but since the fast diagnostics is not always possible, the death toll of ACS in the world population is dramatic [6]. The modeling approach for forecasting the distribution of ACS cases would allow the healthcare specialists to be better prepared for the ACS cases, both in emergency services and in stationary healthcare facilities [3]. One of the simple forecasting methods is related to the application of statistical analysis to the retrospective EMS calls data associated with acute heart conditions. However, if the corresponding time series data set is not long, the accurate prediction is impossible without using additional data

* This research is financially supported by The Russian Science Foundation, Agreement #19-11-00326.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_36

related to the possible prerequisites for acute coronary syndrome calls, such as health conditions that increase the risk of ACS.

One of the factors in the population which might raise the probability of acute coronary syndrome is arterial hypertension (or, shortly, AH) — a medical condition associated with elevated blood pressure [14]. Arterial hypertension is one of the main factors leading to atherogenesis and the development of vulnerable plaques, which in turn might be responsible for the development of acute coronary syndromes [10]. Thus, we might assume that the urban area populated predominantly by individuals with AH might demonstrate higher rates of ACS. Based on that assumption, it might be possible to use spatially explicit AH data as an additional predictor of prospective ACS cases. Unfortunately, the data on AH prevalence with the geographical matching are rarely found, and for Russian settings, they are virtually non–existent. Nevertheless, they could be generated synthetically, which adds uncertainty to the analysis but, on the other hand, makes possible the analysis itself.

In this paper, we describe methods and algorithms to analyze the distribution of ACS–associated emergency medical service calls (shortly, EMS calls) by matching them with synthesized data on arterial hypertension prevalence. Using Saint Petersburg as a case study, we address the following question: may the synthesized AH data combined with EMS calls dataset provide additional information connected with ACS distribution in the population, compared to absolute data and relative data on EMS calls alone?

2 Data

2.1 EMS calls

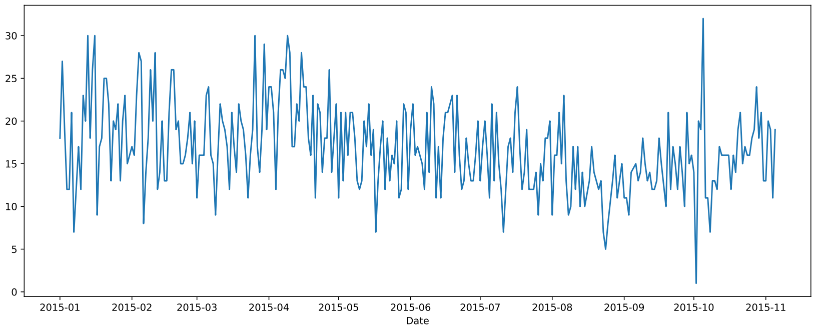


Fig. 1. The daily dynamics of emergency service calls connected with acute coronary syndrome (Jan – Nov, 2015)

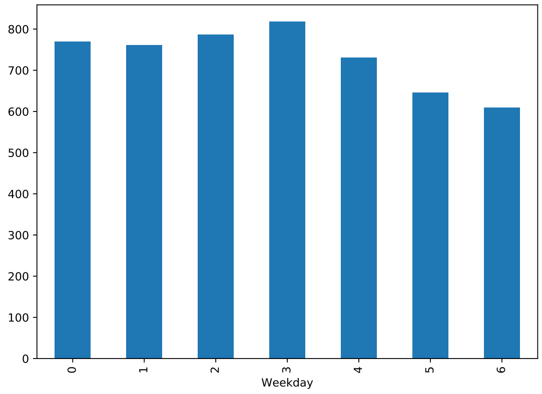
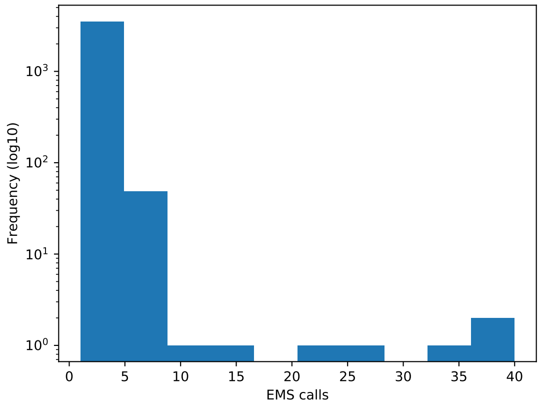


Fig. 2. The cumulative number of ACS emergency service calls in different weekdays.



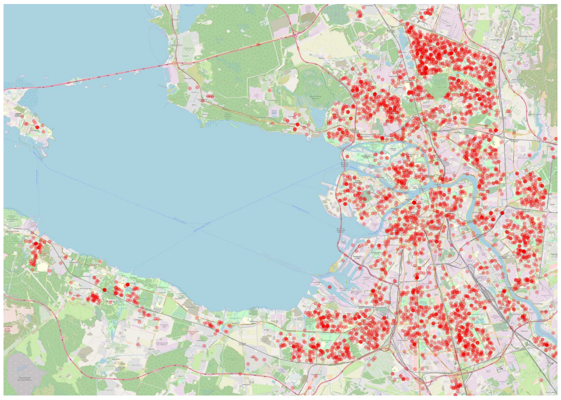


Fig. 3. The spatial distribution of ACS emergency service calls in Saint Petersburg

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_36

The EMS data we used in the research contain 5125 ACS–related EMS calls registered in Saint Petersburg from January to November 2015 [7]. The back–of–the–envelope analysis of the time series corresponding to daily number of calls (Fig. 1) and the weekly EMS calls distribution (Fig. 2) did not reveal any statistically significant patterns connected with distribution of calls over time, although it is clear that the number of EMS calls has a decline in the weekends. Thus, there is no straightforward prediction method to forecast fluctuations of the cumulative number of daily EMS calls connected with ACS.

The spatial distribution of calls for the whole time period based on the addresses from the database is shown in Fig. 3. The histogram for cumulative distribution was built by calculating the total number of EMS calls in a given spatial cell with the size 250 m x 250 m, with empty cells (0 EMS calls) excluded from the distribution. It was established that the form of the histogram does not change significantly if the cell sizes vary (up to 2 km x 2 km). It can be seen that the predominant majority of the spatial cells had 1 to 5 EMS calls, and only for single cells this number exceeds 8. Based on general knowledge, we assumed that the increased concentration of the EMS calls within particular cells may be caused by one of the following reasons:

- The cell has higher population density compared to the other cells;
- The cell has higher concentration of people with arterial hypertension, which might cause higher ACS probability;
- The cell includes people who are more prone to acute coronary syndrome due to unknown reasons.

To distinguish these cases and thus to be able to perform a more meaningful analysis of EMS calls distribution, we assess the spatial distribution of city dwellers and people with high blood pressure using the synthetic population approach.

2.2 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 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. In this study, we have used a synthetic population generated according to the standard of RTI International [13].

According to the standard of RTI International, the principal data for any given synthetic population 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), and `schools.txt` (contains id, coordinates, capacity). Our synthetic population is based on 2010 data from "Edinaya sistema ucheta naseleniya Sankt Peterburga" ("Unified population accounting system of Saint Petersburg") [4], which was checked for errors and complemented by the coordinates of the given locations. The schools records were based on the school list from the official web–site of the Government of Saint Petersburg [5]. The distribution of working places for adults and their coordinates were derived from the data obtained with the help of Yandex.Auditorii API [15]. The detailed description of the population generation can be found in [8].

2.3 Assessing AH risk and individual AH status

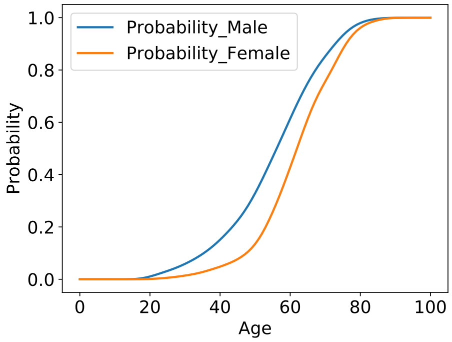


Fig. 4. The cumulative distribution function used to define the AH status of an individual, based on data from [12].

Further on we assess the probability for an individual in the synthetic population to have arterial hypertension. Each individual receives two additional characteristics [9]:

- The AH risk (the probability of having arterial hypertension). Based on [12], we assumed that the mentioned probability depends on age and gender of an individual. The corresponding cumulative distribution function was found using the data of 4521 patients during 2010–2015 and is shown in Fig. 4.
- 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.

The proportion of the synthetic population affected by arterial hypertension is found to be 26.6 % which roughly correlates with the AH prevalence data in the USA according to American Heart Association Statistical Fact Sheet 2013 Update (1 out of every 3) [1] and is lower than the estimate for the urban

DOI: 10.1007/978-3-030-50423-6_36

population in Russia (47.5%) [11]. The cumulative and spatial distributions of AH+ individuals in Saint Petersburg are shown in Fig. 5. It can be seen that the non–uniformity in ages and genders of the citizens potentially causes an uneven distribution of individuals exposed to arterial hypertension.

Further in the paper we match the number of AH+ dwellers of every cell with the number of EMS calls within this same cell and propose an indicator to analyze the relation between them.

2.4 Calculating the indicators related to EMS calls.

We convert the coordinates of EMS calls location from degrees to meters using Mercator projection. After this, we form a grid with a fixed cell size (250m × 250m) which covers the urban territory under consideration. Finally, using the EMS calls dataset, we calculate the overall number of EMS calls which was made within each cell of the grid. In the same way, we calculate the overall number of dwellers and AH+ individuals for the cells. This algorithm was implemented as a collection of scripts written in Python 3.7 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, overall number of EMS calls).

In order to understand the relationship between the numbers of AH+ users and the number of EMS calls, we follow our earlier research [2], where the ratio r_1 between the overdose–related EMS calls and the assessed number of opioid drug users was studied. In this paper, we compare r_1 with the alternative indicator r_2 which depends on the cumulative number of people in the cell under study instead of the assessed quantity of AH+ individuals. The formulas to calculate the mentioned ratios are the following:

$$r_1 = \frac{n_{ems} + 1}{n_{ah} + 1} \quad \text{and} \quad r_2 = \frac{n_{ems} + 1}{n_p + 1}$$

where n_{ems} is the number of registered EMS calls in a cell, n_{ah} is the assessed number of AH+ users in a cell, and n_p is the number of dwellers in a cell based on the synthetic population data. These quantities represent the number of calls per AH+ individual and calls per dweller, respectively. By adding 1 to the numerator and denominator we are able to avoid a divide by zero error, and although it provides a small skew in the data, its consistent application across all cells leaves the results and their interpretations unhindered. We use the ratio r_1 to understand which cells have large differences in the orders of magnitude compared to other cells. The ratio r_2 is introduced to compare its distribution with r_1 and thus decide whether the statistical model for AH+ probability assessment helps more accurately detect the anomalies connected with EMS calls distribution.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_36

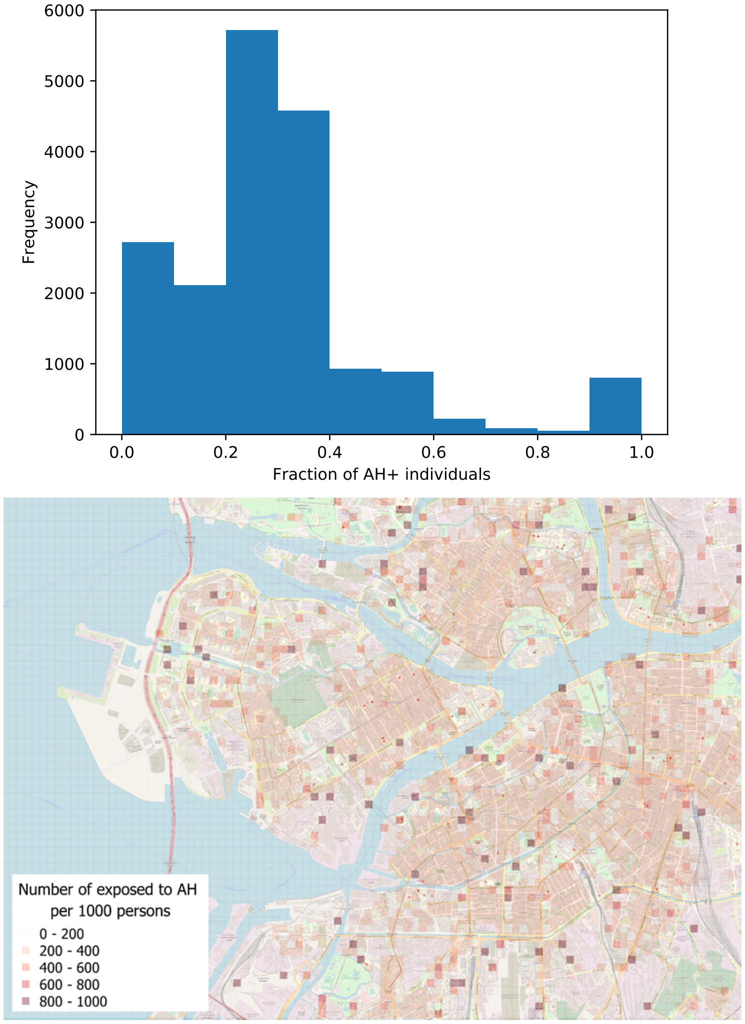


Fig. 5. The aggregated and geospatial distributions of AH+ individuals in Saint Petersburg

DOI: 10.1007/978-3-030-50423-6_36

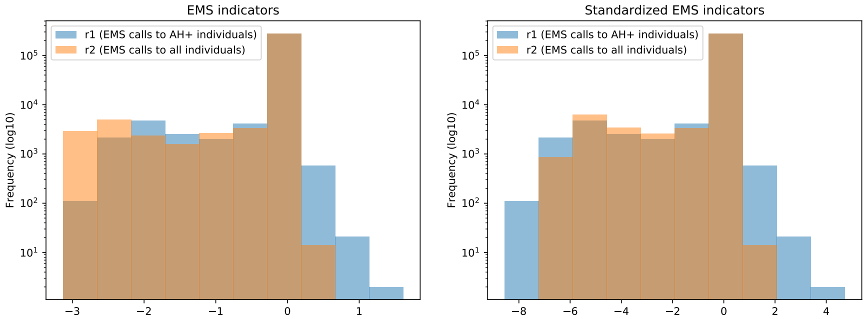


Fig. 6. The distributions of r_1 and r_2 (original and standardized).

3 Results

3.1 Cumulative distribution

In Fig. 6, the aggregated distributions of the r_1 and r_2 values for our data are shown. On the left graph, the distributions are given in their original form, and in the right one the standardized distributions are demonstrated, i.e. with means equal to 0 and standard deviations equal to 1. Although the shape of the histograms is similar, the difference between the corresponding distributions is statistically significant, which is supported by the results of Chi–square test performed for the standardized samples. The crucial difference is in the histogram tails, i.e. in the extreme values of the indicators, which, as it will be shown further in the paper, is also accompanied by their different spatial distribution.

3.2 Spatial distribution

In Fig. 7, a distribution of 20 cells with the highest values of r_1 and r_2 is shown (shades of blue and shades of green correspondingly). The lighter shades corresponds to the bigger cell side lengths (250, 500, 1000 and 2000 meters).

The results demonstrate that the locations of high r_1 values change less with the change of cell side length, compared to r_2 (it is demonstrated on the map by several points with different shades of blue situated one near another). Also it is notable that the high r_2 values were found in lined up adjacent cells (see left and right edges of the map). This peculiarity of r_2 distribution requires further investigation, because it hampers the meaningful usage of the indicator.

The locations marked with three blue points represent concentration of high EMS calls in the isolated neighborhood with few assessed number of AH+ individuals. Most of these locations happen to be near the places connected with tourism and entertainment (1 – Gazprom Arena football stadium, 2 – Peterhof historical park) or industrial facilities (3 – bus park, trolleybus park, train depot; 4 – Izhora factory, Kolpino bus park). Location 5 corresponds to Pulkovo airport, a major transport hub (it is marked by only two blue points though).

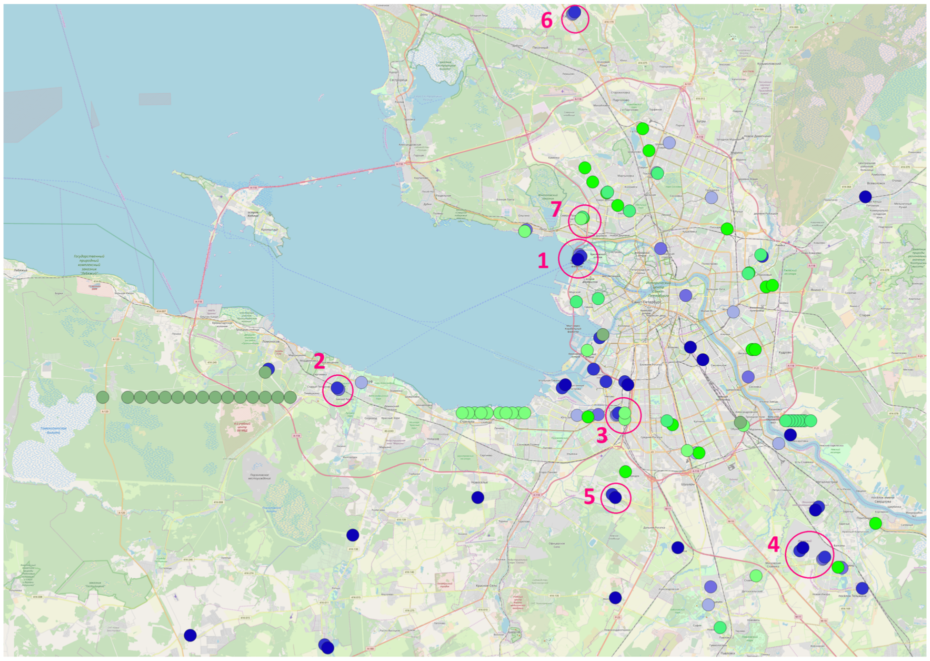


Fig. 7. Points of high r_1 and r_2

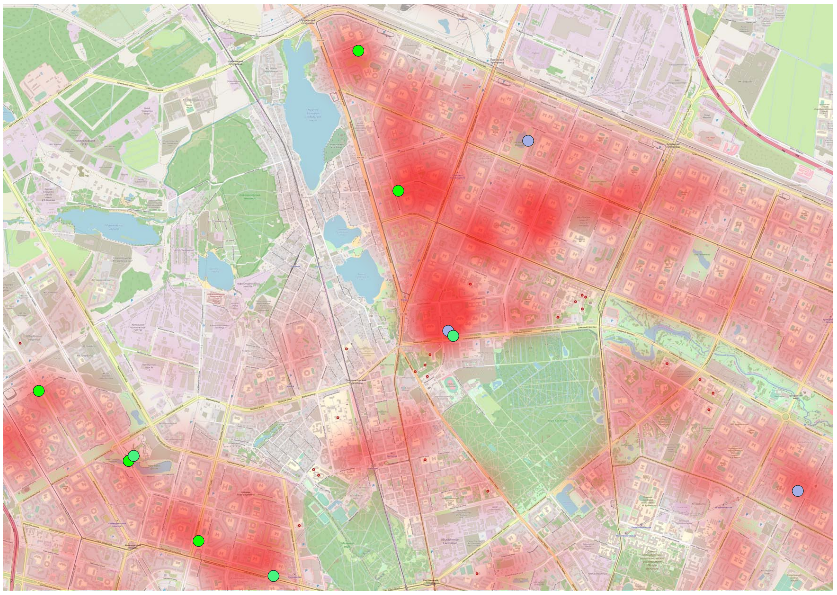


Fig. 8. Heatmap of EMS calls matched against high r_1 and r_2 locations

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_36

Location 6 is the one which cannot be easily connected with excessive EMS calls — it is situated in a small suburb with plenty of housing. The possible interpretation of why it demonstrates high r_1 is the discrepancy between the actual number of dwellers for 2015 (a year for EMS calls data) compared to the 2010 information (a year for populational data). This zone was a rapidly developing construction site and subsequently witnessed a fast increase in the number of dwellers. Location 7 is also an expectational one – it is the only one which is marked by three green points (high r_2). Additionally, this zone was not marked by high r_1, although it is easily interpreted as yet another industrial district (Lenpoligraphmash printing factory). Increasing the number of points in a distribution to 100 does not change significantly the results: isolated areas with meaningful interpretation are mostly marked by the blue points, except Lenpoligraphmash at location 7.

Whereas the exceptional values of r_1 indicate isolated non–residential areas (industrial objects and places of mass concentration of people) which might be connected with the increased risk of ACS and thus require attention from healthcare services, the extreme values of r_2 indicator might be useful when we need to assess the excess of EMS calls in the densely populated residential areas. In Fig. 8, where r_1 and r_2 values are plotted against a heatmap of EMS call numbers, we see that there are two types of peak concentrations of EMS calls (bright red color). Ones are not marked with green dots (the r_2 values are not high) and thus might be explained by high concentration of dwellers in general. Others, marked with green dots, show the locations with high number of EMS calls relative to population. In case the locations does not demonstrate high r_1 values (no blue dots in the same place), they might correspond to the category of neighborhoods with ACS risk factors not associated with arterial hypertension (to be more precise, not associated with the old age of dwellers, since it is the main parameter of the statistical model for AH prevalence used in this study).

4 Discussion

In this paper, we demonstrated a statistical approach which uses synthetic populations and statistical models of arterial hypertension prevalence to distinguish several cases of ACS–associated EMS call concentration in the urban areas:

- High r_1 values for any corresponding number of EMS calls (Fig. 7) might indicate locations where acute coronary syndrome cases happen despite the low AH+ population density (for instance, particular industrial zones).
- Average to low r_2 values for high number of EMS calls (Fig. 8, red spots without green points) correspond to areas with high population density.
- High r_2 values and low r_1 values for high number of EMS calls (Fig. 8, red spots with green points) might indicate areas where the excessive number of ACS cases cannot be explained neither by the high population density, nor by AH prevalence, thus they might indicate neighborhoods with unknown negative factors.

It is worth noting that due to the properties of our EMS dataset (see Section 2.1 and Fig. 3) most of the locations with extremely high r_1 and r_2 correspond to the number of EMS calls in a grid cell equal to 1. Ascribing EMS calls to one or another property of the area based on such a small number of observations is definitely premature, and thus our interpretations given earlier in the text should be continuously tested using the new data on EMS calls. Despite the fact that we cannot draw any definite and final conclusions, in the author's opinion, the study successfully introduces the application of the concept of using synthesized data for health conditions of unknown prevalence (arterial hypertension) to categorize spatial distribution of their acute repercussions (acute coronary syndrome). As it was demonstrated by the authors before [2], the same approach can be successfully used in case of opioid drug usage, and we expect to broaden the scope of its application by applying it in other domains.

As to the current research, we plan the following directions of its further development:

- Currently, the time periods of the EMS calls information and synthetic population data do not match, which might cause the bias in the estimated values of the indicators. We plan to reproduce the results of this study using the actualized data sets.
- The enhanced statistical model for AH is considered to make the calculation of the number of AH+ individuals more accurate.
- The values of r_1 are almost the same for the cases of (a) 1 EMS call in presence of 0 AH+ individuals, and (b) $2n$ calls in presence of n AH+ individuals, so those cases cannot be distinguished by using indicators such as r_1, although they are essentially different. We want to explore the possibility of using a yet another indicator which will take into account the absolute number of dwellers in the neighborhood and will have a meaningful interpretation.
- We have access to a number of health records of the people hospitalized with ACS in a human–readable format, which contains information about their AH status. Using natural language processing tools, we plan to obtain a digital version of this data set and consequently to assess numerically the connection between AH and ACS cases in Saint Petersburg. This result will help reduce uncertainty in the results of the current study connected with analyzing the distribution of r_1.

References

1. AHA: American heart association statistical fact sheet 2013 update. [online], ”https://www.heart.org/idc/groups/heart-public/@wcm/@sop/@smd/documents/downloadable/ucm_319587.pdf”. Last visited: 04/10/2020.
2. 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 **25**(1), 36–47 (2019)

3. Derevitskiy, I., Krotov, E., Voloshin, D., Yakovlev, A., Kovalchuk, S.V., Karbovskii, V.: Simulation of emergency care for patients with ACS in Saint Petersburg for ambulance decision making. Procedia Computer Science **108**, 2210–2219 (2017)
4. 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) Last visited: 04/13/2020.
5. Government of Saint Petersburg: Official web-site. [online], https://www.gov.spb.ru/ Last visited: 04/13/2020
6. Jan, S., Lee, S.W., Sawhney, J.P., Ong, T.K., Chin, C.T., Kim, H.S., Krittayaphong, R., Nhan, V.T., Itoh, Y., Huo, Y.: Catastrophic health expenditure on acute coronary events in asia: a prospective study. Bulletin of the World Health Organization **94**(3), 193 (2016)
7. Kovalchuk, S.V., Moskalenko, M.A., Yakovlev, A.N.: Towards model-based policy elaboration on city scale using game theory: application to ambulance dispatching. In: International Conference on Computational Science. pp. 404–417. Springer (2018)
8. Leonenko, V., Lobachev, A., Bobashev, G.: Spatial modeling of influenza outbreaks in Saint Petersburg using synthetic populations. In: International Conference on Computational Science. pp. 492–505. Springer (2019)
9. Leonenko, V.N., Kovalchuk, S.V.: Analyzing the spatial distribution of individuals predisposed to arterial hypertension in Saint Petersburg using synthetic populations. ITM Web of Conferences **31**, 03002 (2020)
10. Picariello, C., Lazzeri, C., Attana, P., Chiostri, M., Gensini, G.F., Valente, S.: The impact of hypertension on patients with acute coronary syndromes. International journal of hypertension **2011** (2011)
11. S.A., B., et al.: Arterial hypertension among individuals of 2564 years old: Prevalence, awareness, treatment and control. by the data from eccd. Cardiovascular Therapy and Prevention **13**(4), 4–14 (2014), in Russian.
12. Semakova, A., Zvartau, N.: Data-driven identification of hypertensive patient profiles for patient population simulation. Procedia Computer Science **136**, 433–442 (2018)
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**(10), 905 (2009)
14. WHO: Hypertension. Fact sheet. [online], https://www.who.int/news-room/fact-sheets/detail/hypertension Last visited: 04/13/2020
15. Yandex: Auditorii. [online], https://audience.yandex.ru/ Last visited: 04/13/2020.