Geographically corrected clustering applied to establish medical service areas

Nikita Politsinsky [1], Ilia Kuznetsov [1,2], Evgeny Panidi [1]

[1] Saint Petersburg State University, St. Petersburg, Russia - panidi@ya.ru, evgeny.panidi@spbu.ru
[2] St. Petersburg Research Institute of Phthisiopulmonology, St. Petersburg, Russia

Keywords: Medical Geospatial Data, Geospatial Data Management, Medical Service Areas, QGIS.

Abstract

In the current paper, we discover a case study of medical service areas zoning automation needed to ensure effective operation of phthisiatric service. The study was conducted for the administrative area of Saint Petersburg city (Russia). Originally, the process of phthisiatric service areas zoning bases upon outdated interpretation of medical maintenance of territories, and assumes splitting of living buildings list compiled for some administrative territory. This leads to appearing of different zoning features (service area geometry and topology, workload imbalance, etc.), which impact the quality and effectiveness of medical service. No unified and(or) open source tools for geographically corrected (in the meaning of accounting of the different spatial factors and variables) medical service areas zoning are available now. Our study is focussed onto closing of this lack basing on geospatial techniques and data management methods. To ensure the automated phthisiological medical service areas zoning we elaborated and tested a set of scripts available to be running in QGIS. Elaborated methodology was documented and considered as a possible for implementation into technological chain of Saint Petersburg phthisiatric service.

1. Introduction

The provision of primary health care to adults and children in most medical organizations is carried out in Russia (like in many countries) according to the system of serviced area division into medical lots (Grishina et al., 2019). In the case of therapeutic and tubercular medical services, this principle of the primary health care organization is based mostly upon the grouping of the serviced population according to residence. The key task in this way is to provide all citizen categories with equal access to medical care (Komarov, 2017).

Medical service areas are established traditionally with respect to observed and(or) expected (forecasted) number of patients in the serviced territory, while the service areas zoning is aimed onto balancing of attended medicals workload. In the current material, we discover a case study of service areas zoning for tuberculosis dispensaries of Saint Petersburg city (Russia). Originally, the process of phthisiatric service areas zoning bases upon outdated interpretation of medical maintenance of territories, and assumes splitting of living buildings list compiled for some administrative territory. As a rule, the precinct division of the territory appears to be carried out once during the actual commissioning of a new city district or according to a new medical clinic opening. Generally, no any additional factors and(or) restrictions are accounted when division conducting, due to the complexity of such a multifactor zoning.

Such a splitting affects the effectiveness of contraepidemic activities in its turn (Shulmin, 2013). For instance, interdependent foci of infection (or simply infection clusters, composed of a number of living buildings, where infection and reinfection cases are dependent on specific social activities of locals) can be split into several service areas, so the split infection cluster parts can be serviced without coordination of medical activities.

Service areas zoning issue is complicated also by change in actual incidence, attendance, and permanent population change year to year, which lead to change (sometimes dramatically unpredicted) of the doctors workload. Often, the schemes of the medical service areas network in large cities were formed decades ago, and subsequently adjusted basing on real attendance. Being impacted by the gradual development of new city territories and the commissioning of new residential buildings this also lead to workload imbalance.

The specific method of medical service areas division is not indicated by regulatory legal acts in Russia. To date, there are only recommended standards are presented, according to which the medical organizations allocate medical service areas independently. As the zoning update is conducted manually in many cases, it appears to be an extremely time-consuming process.

Additionally, no unified and(or) open source tools for geographically corrected (in the meaning of accounting of the different spatial factors and variables) medical service areas zoning are available now. Our study is also focussed onto closing of this lack basing on open source geospatial techniques and geospatial data operation methods.

The study is relevant due to the lack of regulation and lack of automated tools, which lead to a situation when neither the real population, nor the incidence rates, nor the convenience of a doctor responsible for the assigned territory (sometimes divided into several different parts located in different city district) are not taken into account when performing service areas zoning.

The lacks of a unified zoning methodology and appropriate software are not allow medical organizations to change the service areas network quickly, taking into account several factors. Combination of new knowledge gained in medicine and geography domains (Chistobayev, Semenova, 2013; Schweikart, Kistemann, 2013) and application of new advanced data analysis methods (including spatial data analysis) in the healthcare domain (Golovanova, 2020; Franch-Pardo et al., 2020) can ensure the lacks cover. In the context of the digital transformation of domestic healthcare in Russia, this problem acquires an interdepartmental and interdomain character and can be resolved particularly with the help of geoinformatics methods (Korovka et al., 2021).

The issue of geospatial analysis and Geographic Information Systems (GISs) implementation when establishing medical

service areas appears to be relevant also due to high social impact and dangerousness of infectious diseases (Nechaeva, 2018) and rapid spatiotemporal dynamics of their development, (Gordon, Womersley, 1997; Korovka et al., 2021). GIS-based automated analysis in this context have to ensure operative restructuring of medical service areas with respect to multiple spatial factors and variables.

2. Data and Methods

In our work we applied open source approach and built the study upon implementation of QGIS (*https://www.qgis.org*) open source software facilities (Kuznetsov et al., 2020).
The medical statistics data on registered in St. Petersburg infectious diseases cases that were used in experiments was presented in impersonalized form (according to Russian law) by St. Petersburg Research Institute of Phthisiopulmonology.
Geocoding of medical statistics data (Kuznetsov et al., 2020) was conducted using GeoMedic previously developed (as a part of described study) geocoder, that id designed specially to geocode medical service addresses. 12,257 (96%) medical service (postal) addresses were geocoded successfully and coordinated in the map, while the remaining 510 addresses were indicated as errors (containing errors that did not allow identifying the addresses even manually, since such addresses did not exist in the St. Petersburg address database). Consequently, the geocoding error was estimated as a 4%. After the automated geocoding was performed, the manual control and selective correction of the geocoded positions were performed, to enhance spatial accuracy of geocoded data and make it sufficient to ensure further research.

To observe the existing situation and form the reference dataset, the vectorization of boundaries for currently operated medical service areas was performed (in manual mode) according to publically available data. The OpenSteetMap (*https://www.openstreetmap.org*) data were used as a basic map (Fig. 1).

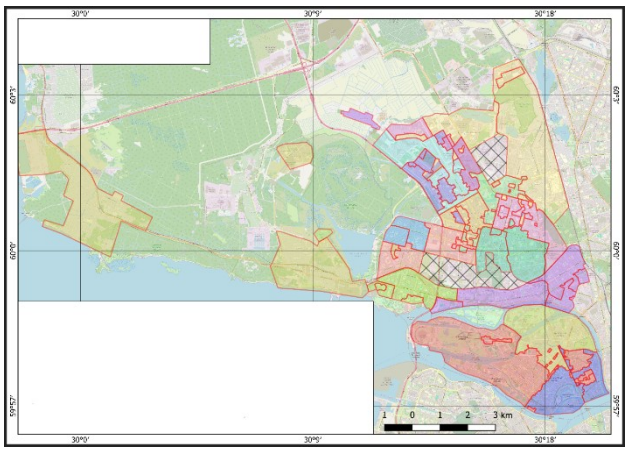

Figure 1. Contemporary network of medical service areas (different areas are shown by different colours) in Primorsky and Petrogradsky districts of St. Petersburg city (Russia). OpenStreetMap data is used as a base map.

The service area boundaries were restored along roads and rivers, buildings were not allowed to cross the boundaries. The territories where service addresses were completely absent (large parks, industrial zones, etc.) were not incorporated into service area polygons. The colour scheme in Fig. 1 reflects the affiliation of the service area or its part to the particular doctor, as currently there are cases presented when a doctor serves several areas, or several doctors serve one area.

Staff volume standards for medical service areas are fixed by orders of the Ministry of Health of the Russian Federation. Particularly the order N 932n establishes the recommendations on tuberculosis dispensary staffing (Order of the Ministry of Health of the Russian Federation dated by November 15, 2012 N 932n):

1. 1 phthisiologist per 25,000 people of the urban population
2. 3 phthisiologists per 40,000 people of the rural population

On the example of the Petrogradsky district medical service areas (Fig. 2) it is clearly seen that the existing load imbalances may appear valuable in comparison to whole population amount of a service area. In two of the three service areas, more than 20% excess of the recommended standard is observed.

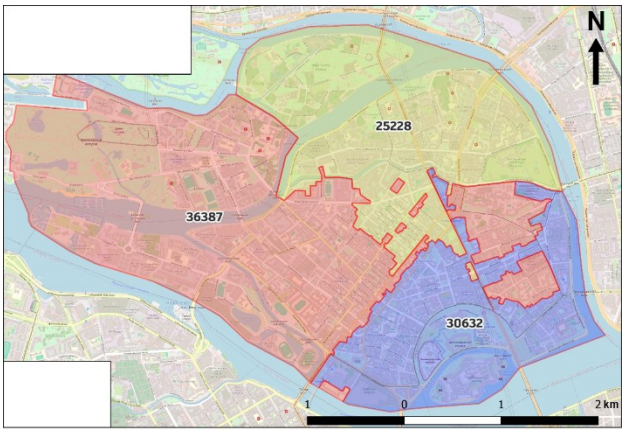

Figure 2. Population amount of the phthisiological service areas in the Petrogradsky district, labels – the population in the area. OpenStreetMap data is used as a base map.

To formalize the process the territory zoning into medical service areas, we posed the basic idea of composed of inability of the area boundaries to cross the buildings and obligation of double perception exclusion (every house have to belong to an only service area).
We decided to form service area polygons by clustering of smaller parts, which are the polygons of varying shapes covering the studied territory. Every part have to have a weight value assigned that is population amount, or number of infection cases observed, or a value of some other variable. The service area polygons are then formed by combining these parts into bigger polygons with consequent calculation and control of the summarized value of selected weight indicator. The small parts in this case can be formed using the road network through the geometrical construction of blocks surrounded by the road network segments. This approach makes it possible to combine population of multiflat houses geocoded as the point geometries and transfer them into the area geometries without any complex algoritmization.

Additional criteria of service areas zoning needed to ensure service area operation convenience were established as next:

1. Traffic accounting (the area have not cross major highways)

2. Hydrography accounting (the area have not cross rivers)
3. Shape simplicity preference (area shape have to be maximally similar to square)

Other criteria can be applied also, including complex criteria. Particularly, infection clusters can be mapped and used similarly to the traffic and hydrography data for zoning limitation.

The road graph for the studied territory was downloaded from OpenStreetMap database using the QuickOSM QGIS module (*https://plugins.qgis.org/plugins/QuickOSM/*). All types of roads were used, except for small driveways and pedestrian paths. The railways and rivers geometries were merged also with the road network geometries to ensure the service area construction restrictions set. All the layers were combined into one linear layer. City block polygons were built automatically basing on the constructed linear layer using the built-in (Lines to Polygons) QGIS tool.
However, the resulting polygon layer required additional processing. We used the Eliminate Selected Polygons tool to generalize polygons, and incorporate small and fragmented objects into large polygons valuable in the meaning of population accounting (Fig. 3).

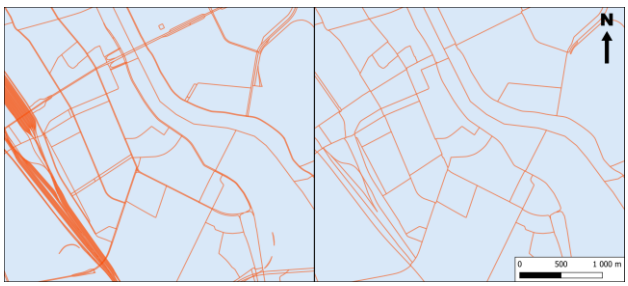

Figure 3. Generalized city block polygons (right image) compared to the blocks before generalization (left image).

Data on the observed tuberculosis infecting cases presented originally as the point geometries were merged with the generalized blocks layer using the Count Points in Polygon tool (Panidi et al., 2023). Determined number of patients in each polygon (block) was recorded in a new attribute field. Thus, the initial dataset for the formation of medical service areas was prepared, and each block was assigned by weight value (Fig. 4).

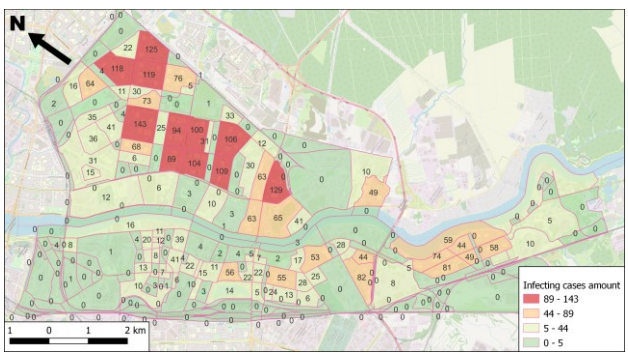

Figure 4. Network of generalized city blocks with assigned values of the infecting cases amount (numbers in the areas), Nevsky district of St. Petersburg. OpenStreetMap data is used as a base map.

At the next stage, formation of the service area polygons division was carried out in a semi-automatic mode. Zoning task for the test area (Nevsky district) assumed the district area division into 10 phthisiological service areas, since at the moment there are 10 phthisiological service areas are operated in the district. Service areas zoning was conducted according the rule of equal load for phthisiologists. According to the 3,523 of observed infecting cases in the district in total, an equal burden on each doctor is about 350 patients per service area. Using the built-in QGIS tools, the service area polygons were formed. The Statistics panel was used for tracking the total number of tuberculosis patients in one-by-one selected polygons (blocks). When a value of 350 infecting cases was reached, the selected objects were merged. Thus, a reference network of phthisiological service areas was formed (Fig. 5). Empirically, it was detected that according to the established initial parameters of the zoning, an area is formed having a strong deviation from the average number of infecting cases per service area that is 266 cases.

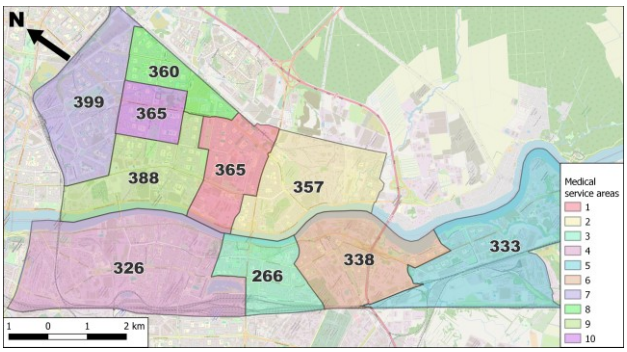

Figure 5. The result of semi-automatic zoning of the medical service areas for the Nevsky district. OpenStreetMap data is used as a base map.

At the next stage the complete automation of zoning process was elaborated. It was observer that there is no ready-made function in QGIS suitable for merging neighbouring polygons with control of summarized value of attribute. So the zoning algorithm was programed as a script in Python programming language, which is closely integrated into QGIS. The open source script performing a similar task was obtained from StackExchange (*https://gis.stackexchange.com/questions/153094/graph-network-building-and-analysis-of-linked-polygons-in-arcmap*), and was refined and adapted for the considered task. The script uses additionally the version 2.0 and higher of *NetworkX* Python library (*https://networkx.org*).

Refined algorithm was served as a geoprocessing script available in the QGIS Processing Toolbox. An additional input dataset, used at the first stage of the script running, is a layer of connections between neighbouring polygons (Fig. 6), which can be obtained by spatial combining of the original polygon layer with itself using the “touching” geometric predicate. Attribute fields of these layers are also specified when script launching: id field for the target objects and id field for the linked objects, attribute field for summation (in described case, this is the number of infecting cases field). Additionally, the number of zoned areas and tolerance parameter value are indicated. The tolerance value determines a permitted deviation from the standardized amount infecting cases for a service area. A tolerance value of 0.5 assumes that the standardized value can be exceeded by 1.5 times as much as possible, tolerance value of 1 – by 2 times.

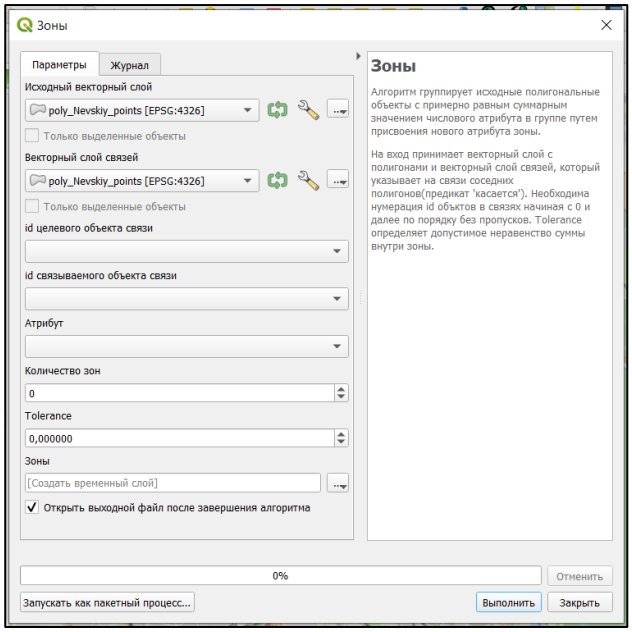

Figure 6. Graphical interface of the developed script.

The data processing is provided as follows:

1. Using the *NetworkX* library, a graph with equal weights of all edges is constructed basing on an initial polygon layer
2. All polygons are paired, cycle conducted over all pairs
3. The total number of infecting cases is calculated for other polygons, closest to the 1st and 2nd in a pair
4. When the required ratio of the infecting cases amount with respect to the established tolerance is achieved (Equation 1), a group with fewer value of infecting cases amount is discovered as a first zoned service area; otherwise, the pair search continues
5. The algorithm returns to step 2, zoning of a new area occurs using all remaining polygons that were not incorporated into the already zoned area(s)
6. The process is completed when all the areas are formed

$$\frac{S2}{S1} - (\mathrm{RATIO} - 1) \leq \mathrm{tolerance}, \qquad (1)$$

where S_1, S_2 = the numbers of infecting cases in first and second polygons in a pair
RATIO = the number of zones
tolerance = tolerance

According to the mentioned earlier additional criteria, we implemented also accounting of three follow zoning parameters:

1. Service area population amount
2. Service area intersections with rivers and major highways
3. Service area shape

The medical service areas zoning taking into account the equal load, assumes accounting not only the number of observed in the areas infecting cases. The population amount in the serviced territory is also accounted for this task. Most optimal zoning of medical service areas is achieved by simultaneous consideration of these two indicators.
Data on the number of residents were geocoded and assigned to point geometries. Using the spatial association and aggregation functions, we computed the number of residents for each city block and recorded it in the INHAB attribute field of the initial polygon layer of blocks.
To implement population amount accounting into the program logic, an additional input parameter (additional attribute field) was defined, that have to be specified by the user. Values of the infecting cases number and number of residents are read by the program and saved to the appropriate lists. It is expected that the values of these two parameters will have different dimensions, so both are normalized using standard equation (Equation 2).

$$X_{normalised} = \frac{X - X_{min}}{X_{max} - X_{min}}, \qquad (2)$$

where $X_{normalized}$ = the normalized value of a parameter (of a variable)
X_{min} = the minimal value of variable
X_{max} = the maximal value of variable
X_{min} = the value of variable to be normalized

Normalization is carried out separately for the values of the infecting cases and the population. After the normalization, all values appear to belong to the 0 to 1 range. Then two values are summed for each polygon, so the program able to consider both indicators when zoning service areas. However, deviation from the value of a complex parameter obtained basing on standards in such a case will be greater than in a case of an only indicator accounting.

An additional Linear Layer input parameter was implemented also to account the intersections with rivers and major highways. The user have to specify a linear layer selecting from available in the currently opened QGIS project.
A custom function has been developed to ensure check of the intersection of a service area with the objects in a linear layer. It takes previously set variables as input and determines the presence or absence of an intersections with a given linear layer. The function outputs True or False and messages it in the console. The function is applied after successful zoning of a service area. If the group of polygons used to build an area does not intersect the linear layer, it is reported in the console, and the cycle for pairs of polygons stops. If a group of polygons intersects a linear layer, it is reported in the console, and the cycle continues.
In result of the program running a polygon layer of medical service areas with assigned sums of indicator values is formed. In the case of intersection restriction accounting, the tolerance parameter must be set higher, as if the program cannot form a service area polygon in accordance with this tolerance and without linear layer crossing, a polygon crossing the linear layer will be formed.

Accounting of the shape restriction is also necessary for convenient logistics for doctors and patients around the service area. It is necessary to exclude strongly elongated area forms. To ensure this, Bounding Box estimation is used. The Bounding Box tool in QGIS accepts a vector layer as input. The tool outputs a polygon layer with bounding rectangular geometries. For each polygon (in our case, for a service area polygon) algorithm builds a rotated surrounding rectangle of the minimum area. It assigns also the attributes of the original objects to the constructed bounding boxes. Additionally, it calculates the width, height, area, rotation angle and perimeter values for the constructed rectangles. The values are calculated in the coordinate system of the source layer. We utilized height and width parameters to determine the elongation of service area polygons and reject unsuitable ones. If the area is strongly elongated, it will be rejected and reshaped. The elongated area polygons are those for which one side of bounding box more than 3 times longer than

another one. However, the user allowed to redefine this parameter.

3. Results and Discussion

Analysis of the currently operated network of medical service areas made it possible to identify significant features of the existing zoning. Not all service areas have a simple (from a geometric point of view) shape. A large number of enclave parts (can be observed also in Fig. 1) form a spatial complexity of service area boundaries. In some cases, streets are assigned to two service areas at once according to the public information, while (it is expected) in fact are operated as part of an only service area. In last case, it is likely that we are talking about technical errors observed when filling out and re-filling out the relevant forms and lists, performed periodically by medical specialists.
When considering the service areas of pediatric phthisiologists, we were detected no significant errors and complexities. These service areas are also geographically separated into parts, but this separation does not seem to be such a strong fragmentation as it is observed in the case of service areas of the adult phthisiologists. Most likely, it is concerned with that the pediatric service areas are rigidly linked to the service areas of pediatric clinics.

The main aspect of the service area divisioning quality assessment is its compliance with the state standards establishing the number of population in the service area. This aspect directly affects the effectiveness of doctors' work, the doctors' workload and, as a result, the quality of medical care. To estimate the population presented in currently operated tuberculosis medical service areas. We attracted publically available information on per house population, retrieved from the Housing Agencies of St. Petersburg districts.
Collected data were also geocoded as a point layer containing the data on the number of residents in multiflat houses. Actual amount of people living in every phthisiological service areas was calculated using the spatial association and aggregation tools in QGIS.

According to implemented facilities it is possible to apply developed script to medical service areas zoning with next combinations of parameters (Fig. 7):

1. Zoning according the one parameter (infecting cases value or population amount can be applied)
2. Zoning according to the one parameter (restrictions: linear layer)
3. Zoning according to the one parameter (restrictions: linear layer, shape)
4. Zoning according to the two parameters (infecting cases value and population amount)
5. Zoning according to the two parameters (restrictions: linear layer)
6. Zoning according to the two parameters (restrictions: linear layer, shape)

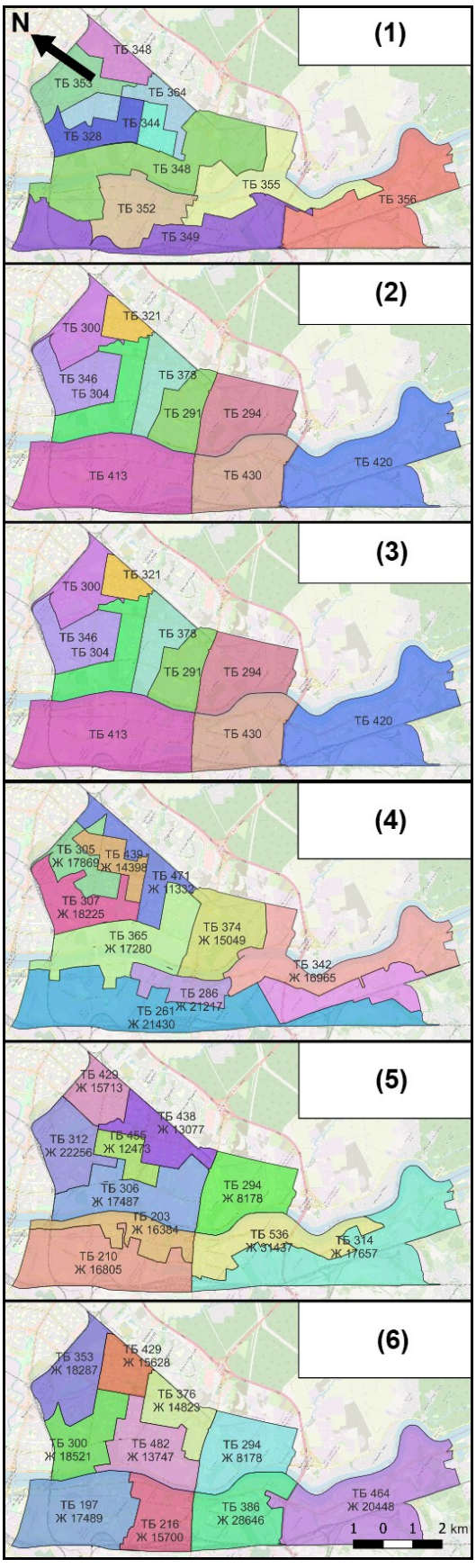

Figure 7. Examples of medical service areas zoning for the Nevsky district of St. Petersburg, generated using different combinations of zoning parameters. OpenStreetMap data is used as a base map.

4. Conclusions

To ensure the automated phthisiological medical service areas zoning we elaborated and tested a set of scripts available to be running in QGIS. The number of tuberculosis infecting cases and the population amount value are used as weight indicators. The obtained indicator value is signed to each generated service area polygon.

Gained results of test territory zoning show that the most uniform distribution of the workload in all zoned areas is achieved when zoning using one parameter without restrictions on the service area geometry. When two parameters are applied without restrictions, a less equal workload distribution is achieved.

When spatial constraints are accounted, the workload distribution across the service areas is degraded, but the compactness of the areas improved. However, all the parameters impact the effectiveness of service areas operation, so several zoning options have to be considered and the most optimal one (in the meaning of real life context) have to be chosen.

Elaborated methodology was documented and considered as a possible for implementation into technological chain of Saint Petersburg phthisiatric service.

Acknowledgements

The study presenting was funded by the Saint Petersburg State University, SPbU PureID: 118438337.

References

Chistobayev, A.I., Semenova, Z.A., 2013. Medico-geographical mapping in the former USSR and modern Russia. *Bulletin of Saint-Petersburg State University, Geology and Geography*, 4, 109-112. (in Russian)

Franch-Pardo, I., Napoletano, B.M., Rosete-Verges, F., Billa, L., 2020. Spatial analysis and GIS in the study of COVID-19. A review. *Science of the Total Environment*, 739, 140033. doi:10.1016/j.scitotenv.2020.140033

Golovanova, M.N., 2020. Improvement of anti-tuberculosis actions using a computer program for monitoring tuberculosis foci. Dissertation for the degree of Candidate of Medical Sciences, Yaroslavl, Yaroslavl State Medical University, 137 p. (in Russian)

Gordon, A., Womersley, J., 1997. The use of mapping in public health and planning health services. *Journal of Public Health*, 19(2), 139-147. doi:10.1093/oxfordjournals.pubmed.a024601

Grishina, I.F., Teplyakova, O.V., Brodovskaya, T.O., Nikolaenko, O.V. Poletaeva, N.B., 2019. *Principles of organization and structure of the district medical service. General medical examination of the population*. Ural State Medical University, Ekaterinburg. 146 p. (in Russian)

Komarov, Yu.M., 2017. *Monitoring and primary health care*. Littera, Moscow. 320 p. (in Russian)

Korovka, V.G., Galkin, V.B., Panidi, E.A., Kuznetsov, I.S., Beltyukov, M.V., Sokolovich, E.G., Panteleeva, O.V., Voronov, D.V., Kozlov, V.V., Fedorov, S.V., Yablonsky, P.K., 2021. Potential of geoinformation technologies to improve the monitoring of socially significant infections outbreaks. *Profilakticheskaya Meditsina*, 24(10), 7-13. (In Russian) doi:10.17116/profmed2021241017

Kuznetsov, I., Panidi, E., Kolesnikov, A., Kikin, P., Korovka, V., Galkin, V., 2020. GIS-based infectious disease data management on a city scale, case study of St. Petersburg, Russia. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B3-2020, 1463-1467. doi:10.5194/isprs-archives-XLIII-B3-2020-1463-2020

Nechaeva, O.B., 2018. TB situation in Russia. *Tuberculosis and Lung Diseases*, 96(8), 5-24. (in Russian) doi:10.21292/2075-1230-2018-96-8-15-24

Panidi, E., Kuznetsov, I., Panteleev, A., Yablonsky, P., 2023. Geographically corrected clustering applied to establish medical service areas. ISDE 2023 Abstract Book, 73.

Schweikart, J., Kistemann, T., 2013. Mapping health and health care [Kartographie der Gesundheit]. *Kartographische Nachrichten*, 63(1), 3-11. (in German)

Shulmin, A.V., 2013. Evaluation of main factors of the medical sector system functioning, according to healthcare managers and therapists in the districts. *Siberian Medical Review*, 1, 78-81. (in Russian)