ISSN 2304-3415, Russian Open Medical Journal
2023. Volume 12. Issue 3 (September). Article CID e0306
DOI: 10.15275/rusomj.2023.0306

Original article

Choosing data clustering tools for GIS-based visualization of disease incidence in the population

Roman V. Buzinov [1], Vladimir N. Fedorov [1], Aleksandr A. Kovshov [1,2], Yuliya A. Novikova [1], Nadezhda A. Tikhonova [1], Maksim S. Petrov [3], Ksenia V. Krutskaya [3]

[1] Northwest Public Health Research Center, St. Petersburg, Russia
[2] North-Western State Medical University named after I.I. Mechnikov, St. Petersburg, Russia
[3] Center for Hygiene and Epidemiology in Arkhangelsk Oblast and Nenets Autonomous Okrug, Arkhangelsk, Russia

Received 9 February 2023, Accepted 28 May 2023

Abstract: *Objective* — To substantiate the choice of optimal tools for clustering spatially referenced data on disease incidence for GIS-based analysis of their spatial distribution.
Material and Methods — We used primary data on the incidence of malignant neoplasms, chronic alcoholism, and asthma in the population of eight administrative areas in Arkhangelsk Oblast as a constituent entity of the Arctic Zone of the Russian Federation. Disease incidence was averaged over a 5-year period from 2016 to 2020. We assessed the methods for visualizing the distribution of spatially referenced indicators using the ArcMap geoinformation system tools.
Results — The study yielded differences in the outcomes of automated clustering of spatially referenced data in ArcMap, depending on the normality of the distribution in individual samples and the spread of indicator values, which was visually reflected on the resulting map. The parameter values in the samples directly affected the features of data clustering. Hence, this issue is important to consider for ensuring the correct choice of the appropriate analytical tool.
Conclusion — Our study demonstrated that when using tools for automated clustering of spatially referenced incidence data in terms of their visualization in ArcGIS, it is necessary to consider the factors that directly affect the accuracy of their presentation. We consider it most appropriate to use a clustering tool based on the geometric interval method.

Keywords: Geoportal; disease incidence; spatial analysis methods; Arkhangelsk Oblast.

Cite as Buzinov RV, Fedorov VN, Kovshov AA, Novikova YuA, Tikhonova NA, Petrov MS, Krutskaya KV. Choosing data clustering tools for GIS-based visualization of disease incidence in the population. *Russian Open Medical Journal* 2023; 12: e0306.

Correspondence to Vladimir N. Fedorov. Address: 4 Vtoraya Sovetskaya St., St. Petersburg 191036, Russia. Phone +78127179629. E-mail: v.fedorov@s-znc.ru.

Introduction

The generally accepted definition of medical geography is an interdisciplinary science at the intersection of geography (cartography) and medicine that studies the effect of environmental features on human health, as well as the laws of the spatial distribution of diseases and other pathological conditions [1]. Currently, the use of geoinformation technologies to assess the state of sanitary and epidemiological wellbeing of the population is an independent scientific field. Consequently, the clause 1.3.7, *Management of Epidemiological Risks Using GIS Technologies*, was included in the research program of the Federal Service for Surveillance on Consumer Rights Protection and Human Wellbeing (Rospotrebnadzor) for 2021-2025 [1].

Some of traditional methods of mapping events, and visualization and modeling of spatial and temporal patterns are quite complicated, such as multidimensional spatial analysis, or interpolation analysis requring a spatial analysis by deterministic mathematical methods and predicting values for points in space (e.g., modeling the presence of dust in the air).

Geostatistical analysis [2] predicts the spatial distribution of trends both in interpolation analysis and statistical analysis, which can be used to search for relationships between points, values of which reflect the dependence between different map layers [3].

Initially, geographic information systems were used mainly in epidemiology, since mapping foci of infections with geoinformation technologies allows seeing their location and concentration zones, visually encode various parameters, and investigate their dynamics over time. The multiscale imaging (transition from a larger scale to a smaller one) makes it possible to study in detail the foci of the disease, suggest the presence of latent sources of infection [4], and create models for predicting the development of the epidemic situation for each specific nosology [5].

Moreover, if there is information about the place of residence and/or the potential place of occurrence of an infectious or parasitic disease, mapping is more informative than the traditional analysis of disease incidence by district or municipality since the boundaries of disease foci and administrative boundaries may not coincide [6]. It should also be noted that the rapid assessment of the spatial distribution of cases of infectious and parasitic diseases by geographic information systems that allow creating focus density maps has great potential in planning anti-epidemic measures and can be performed in real time [7]. Hence,

geoinformation systems are gradually becoming a flexible tool for improving the system of epidemiological surveillance and expanding the possibilities in making managerial decisions [5, 8]. However, geographic information systems are widely used in analyzing the incidence of noncommunicable diseases [9], as well as in assessing the state of sanitary and epidemiological wellbeing of the entire population. In particular, the Geoportal of Sanitary and Epidemiological Welfare of the Population of the Arctic Zone of the Russian Federation, developed by the Northwest Research Center for Hygiene and Public Health, is a tool for a comprehensive assessment of the state of the living environment and health in the population of the Russian Arctic [10, 11].

At the same time, incidence rates, unlike most technogenic environmental factors, are not standardized, so it is necessary to select an adequate algorithm for their analysis. One of these algorithms includes ranking of disease incidence values based on determining the average long-term level and the root-mean-square deviation from it, and on their basis determining the boundaries of disease incidence levels characteristic for the area under assessment (e.g., high, above average, medium, below average, and low) [9]. A similar approach can also be used for a comparison of incidence rates between territories due to *Classification* tool built into software products [12].

There are other data classification tools available in ArcGIS. Among the standard tools, we can single out the following methods: equal ranges (the range of values is divided into subranges of equal size), specified range (for example, with a step of 5%), quantiles (each class contains the same number of objects), natural breaks in values sensu Jenks (for large differences between data values), geometric interval (class boundaries are set based on intervals that have a geometric sequence), and standard deviation/root-mean-square deviation (shows how much the values of feature attributes differ from the mean value). Besides, a manual method is possible as well, when the user independently sets the class boundaries. Due to the presence of a large number of tools for the spatial analysis of baseline materials, a researcher may face a difficult task of choosing the correct approach to data visualization and clustering.

The objective of our study was to substantiate the choice of optimal tools for clustering spatially referenced data on disease incidence for GIS-based analysis of their spatial distribution.

Material and Methods

We examined the methods for visualizing the distribution of spatially referenced indicators. These methods employ a standard set of data grouping tools and the graduated colors function in the ArcMap geoinformation system (ArcGIS by ESRI). The above methods allow grouping polygonal objects on the map (municipalities, towns) considering the similarity criteria for the values of the selected indicator, which simplifies the visualization of the indicator distribution and the degree of its severity on the map. These tools are present in the standard set of desktop versions of the ArcMap geoinformation system and are additionally implemented in the Geoportal mapping application built based on ArcGIS for Server Advanced Enterprise v.11 [11].

We selected five automated data clustering tools of the seven available: equal ranges, quantiles, natural breaks in values, geometric interval, and standard deviation (root-mean-square deviation). The manual data clustering tool and the specified interval clustering tool were not considered in our study.

To develop the criteria for choosing optimal tools for clustering of spatially referenced data, we used data from the Federal Information Base of Social and Hygiene Monitoring on the disease incidence in the population of the following administrative areas in the Arkhangelsk Oblast: Urban District of Arkhangelsk, Mezensky Municipal District, Urban District of Novodvinsk, Onezhsky Municipal District, Primorsky Municipal District, Urban District of Severodvinsk, Leshukonsky Municipal District and Pinezhsky Municipal District.

Due to substantial variation of year-to year disease incidence rates within the same territory, the incidence was averaged over a 5-year period from 2016 to 2020 by summing up the cases of diseases and the population numbers, and then calculating per 100,000 population. As an example for the selection of spatial analysis tools, we investigated the primary incidence of the following diseases in the population of Arkhangelsk Oblast: malignant neoplasms (for entire population), alcohol dependence syndrome (chronic alcoholism), asthma and status asthmaticus (for the population 18 years of age and older). As a part of spatial analysis tool selection, we also compared primary incidence of malignant neoplasms in children (0-14 years old) in the Arkhangelsk Oblast and chronic alcoholism in the adult population of 69 municipalities that are part of the Arctic Zone of the Russian Federation (AZRF).

Results

The various tools for automated clustering of georeferenced data in ArcGIS group datasets into classes that are visually reflected on the map. The results of clustering using five selected tools are explicitly shown in *Figures* 1-3.

In some cases, the ranges of grouped classes are similar: e.g., the geometric interval and equal ranges when grouping indicators of the chronic alcoholism in adults (18 years of age and older) (*Figure* 1); quantiles and equal ranges when grouping the indicators of the incidence of asthma and status asthmaticus in adults (18 years of age and older) (*Figure* 2); geometric interval and quantiles when grouping the incidence rates of malignant neoplasms (cumulative total for all locations) (*Figure* 3).

Some locations, with any variant of data grouping, can be classified as risk areas characterized by an increased incidence rate, compared with other studied territories of Arkhangelsk Oblast. For instance, the risk areas in terms of the primary incidence of chronic alcoholism in the adult population are Onezhsky Municipal District and the Urban District of Severodvinsk; for asthma and status asthmaticus, the risk area is Onezhsky Municipal District; for malignant neoplasms (for the entire population), these are Mezensky and Primorsky Municipal Districts.

At the same time, several study locations can be classified as risk areas only if certain methods of data clustering are employed. E.g., for the incidence of chronic alcoholism, the Leshukonsky Municipal District (method of natural breaks in values sensu Jenks); for asthma and status asthmaticus, the Leshukonsky Municipal District and Urban District of Severodvinsk (geometric interval method); for malignant neoplasms, the Leshukonsky Municipal District (method of equal ranges).

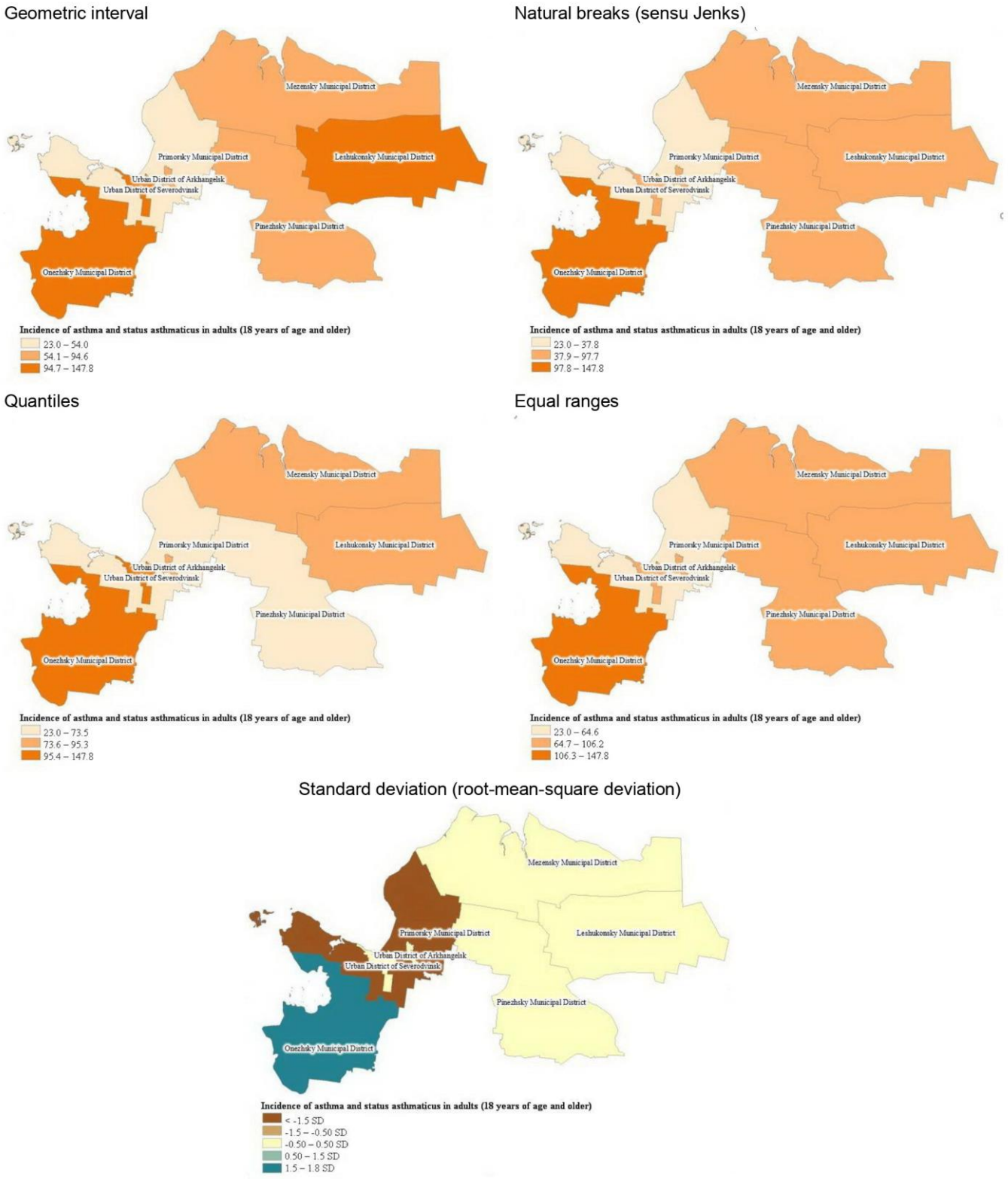

Figure 1. Clustering of data on incidence of chronic alcoholism in adults (18 years of age and older) performed by various automated techniques in ArcGIS.

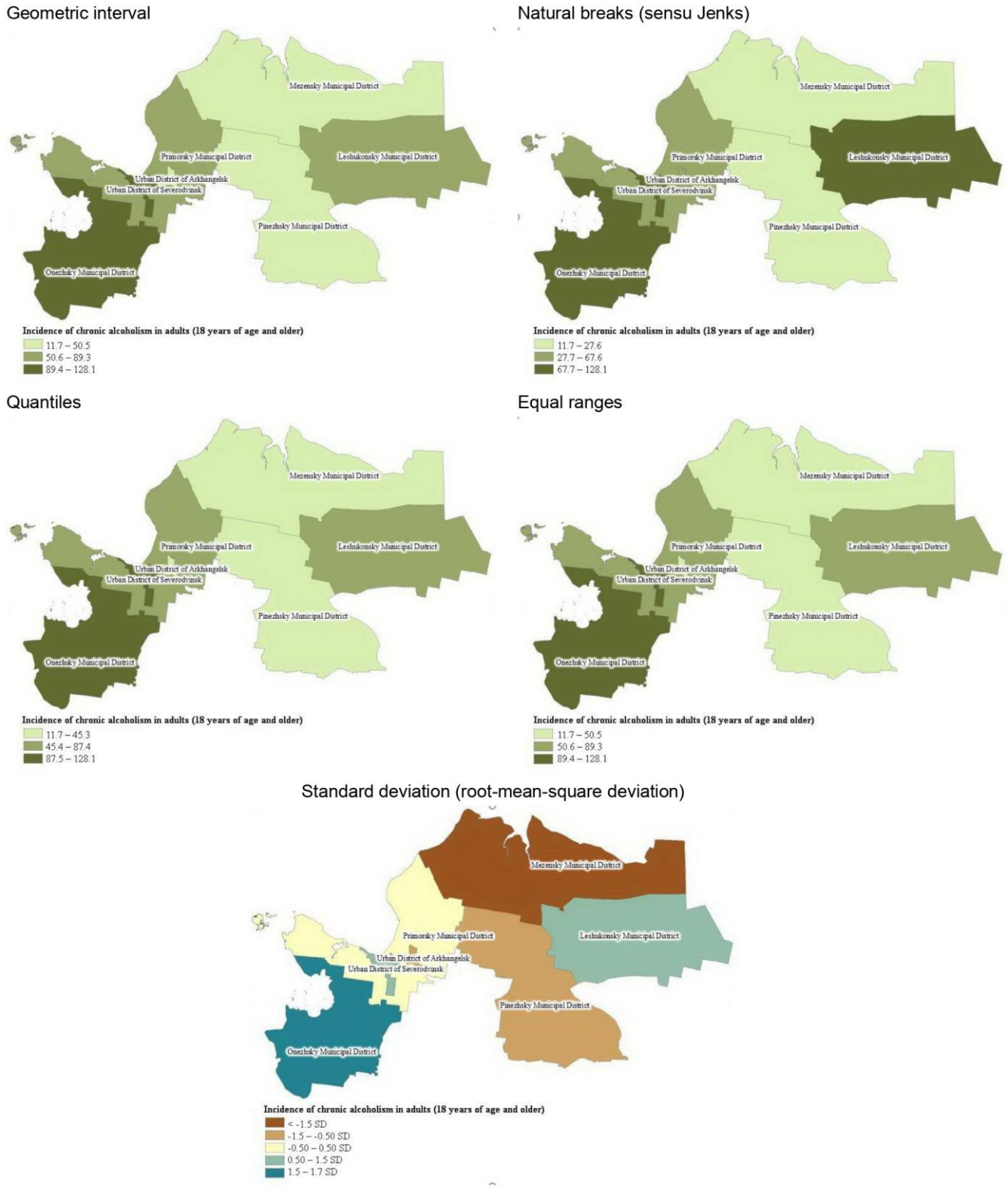

Figure 2. Clustering of data on incidence of asthma and status asthmaticus in adults (18 years of age and older) performed by various automated techniques in ArcGIS.

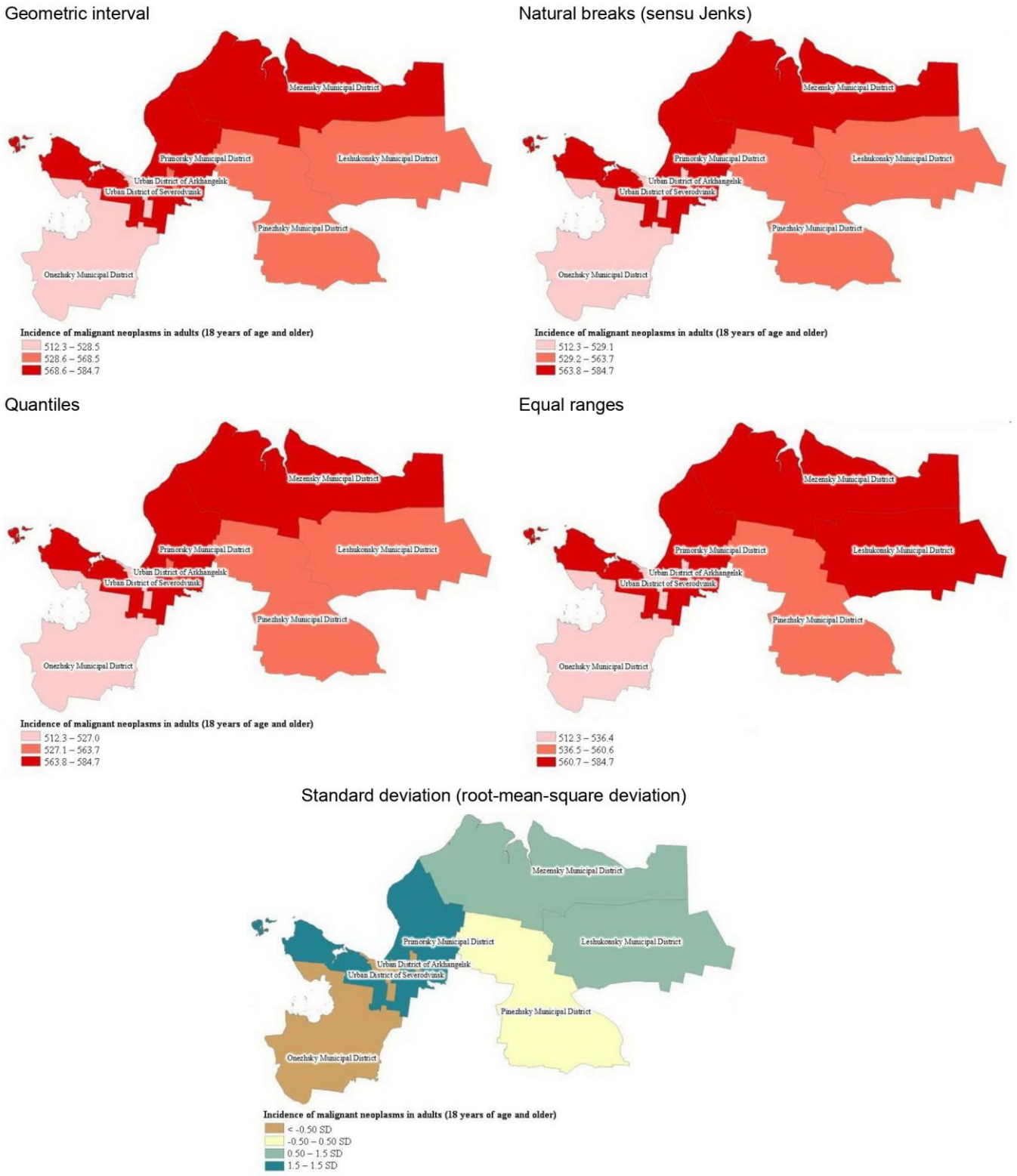

Figure 3. Clustering of data on incidence of malignant neoplasms in adults (18 years of age and older) performed by various automated techniques in ArcGIS.

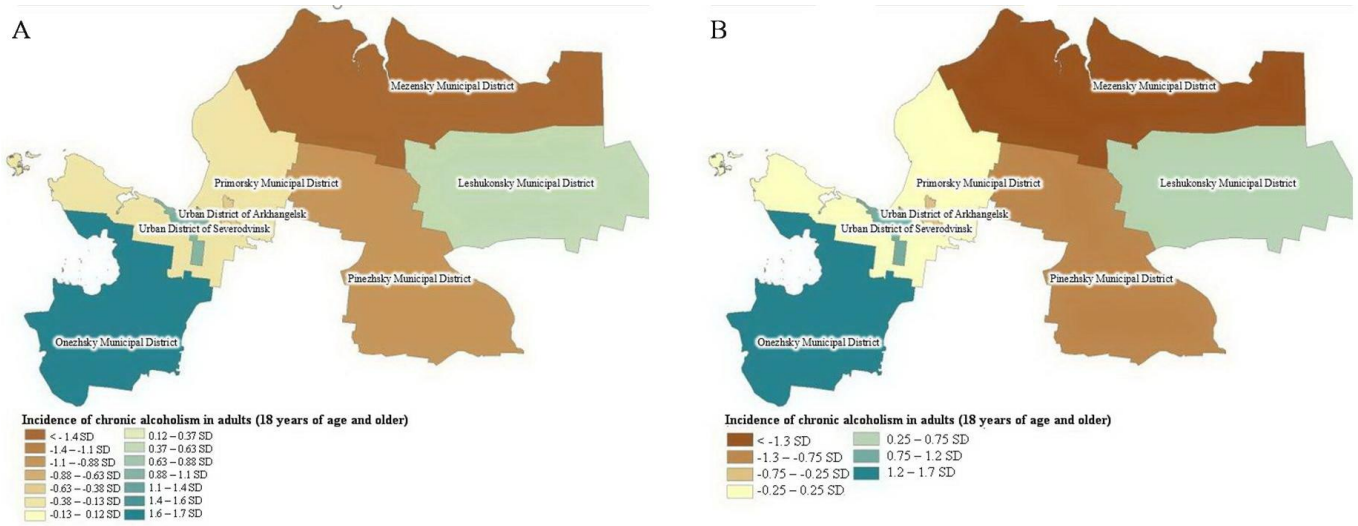

Figure 4. Clustering of data on incidence of chronic alcoholism in adults (18 years of age and older) performed using the standard deviation at an arbitrary interval step. A – Standard deviation (root-mean-square deviation) (The step is ¼ of standard deviation value); B – Standard deviation (root-mean-square deviation) (The step is ½ of standard deviation value).

Discussion

From a mathematical standpoint, the examined data clustering tools are characterized by the following criteria:

- the ability to automatically group data, which simplifies the process of their analysis.

- data grouping ranges cannot be arbitrarily changed by the user, in contrast to the manual clustering method.

The method of grouping based on the root-mean-square deviation does not allow setting the number of classes of indicator values arbitrarily: it is limited by the degree of deviation from the arithmetic mean.

The method of equal ranges is the simplest way to classify data; however, it should be used with caution in the analysis of incidence rate or medical and demographic indicators, since in this case the researcher focuses solely on the magnitude of the value of this attribute relative to other values [12]. For example, this method may show that the Mezensky, Leshukonsky and Primorsky Municipal Districts of the Arkhangelsk Oblast are included in the group of areas with the highest incidence of malignant neoplasms, but it would not be obvious whether the incidence rates in these areas differ significantly from the incidence in other studied territories of Arkhangelsk Oblast.

Classification of administrative territorial units by quantiles is possible, but it is best suited only for linearly distributed data [12]. This method groups the same number of values into each class, and since objects (administrative units) are grouped according to the principle of their equal number in each class, the resulting map may be unenlightening: areas with similar incidence rates may fall into different classes, while areas with substantially different levels of incidence may be in the same class.

In the case of a small number of study areas, this drawback, as a rule, is insignificant, whereas when analyzing many territories, the quantile method may inadequately reflect the actual patterns of incidence. E.g., when clustering all of 69 municipalities included in the AZRF into three classes (23 municipalities in each) based on the level of primary incidence of chronic alcoholism, the first and second classes of municipalities will be almost equally divided by the incidence rate (0-45.3 and 45.7-95.9, respectively). However, the third class of municipalities (risk areas) will already include the incidence range of 100.0-847.1 cases per 100,000 adults). Increasing the number of classes is possible, but manual selection of the required number of classes will complicate the researcher's task and, consequently, may make the map imperceptible due to their large number.

In the case of using the natural breaks in values method, classes of objects are defined to group similar values and maximize differences between classes. In other words, objects are clustered into classes, the boundaries of which are set where there are relatively large differences between data values [12].

Therefore, the use of spatial classification by the method of natural breaks in the analysis of disease incidence is recommended only with significant differences in values. For example, the use of this method was generally valid in the analysis of the primary incidence of chronic alcoholism in the studied territories of the Arkhangelsk Oblast, but it was not informative enough to explain the incidence of malignant neoplasms.

If the root-mean-square method is chosen, the class boundaries are set to ensure equal ranges of values proportional to the standard deviation (1, 0.5, 0.33, or 0.25 of the standard deviation), while the arithmetic mean is used as the main boundary of values. Two colors can be used for color coding: e.g., shades of red for values above average, shades of green for values below average [12].

This method stands out among all considered tools for automated clustering, since it automatically selects a certain number of ranges (classes of values). Therefore, this number cannot be set arbitrarily (unlike in case of other examined tools) (*Figures* 1-3). At the same time, the settings allow defining several options of the interval step – from one-quarter to the full value of the standard deviation in the sample under consideration, which directly affects the number of classes of values of the grouped indicator. It should be noted that in the case of manual setting of the interval step, the number of clustering ranges may exceed the number of values of the studied indicator, and the map will display an outdated legend, some of the values of which do not correlate with the visualized values of the indicator (*Figure* 4).

It is worth mentioning that this method may only be used when the data distribution is normal. Furthermore, the comparison of disease incidence rate with other intensive indicators (mortality, birth rate) in municipalities may be illegitimate due to the potentially different population sizes of municipalities. For example, the arithmetic mean of the primary incidence of malignant neoplasms in the studied territories of the Arkhangelsk Oblast is 545.7 cases per 100,000 population. However, its calculation in eight districts of the region by summing up the absolute number of disease cases, dividing by the total population of these districts and multiplying by 100,000 will yield a slightly different result, viz., 531.0 cases per 100,000 population. When analyzing the incidence of malignant neoplasms in children (0-14 years), the relative difference in incidence rates is even higher: the arithmetic mean is 11.1 cases per 100,000 child population, while the incidence value obtained by dividing the sum of absolute cases of the disease by the total population of eight districts is 19.5 cases per 100,000 child population. This is explained by the fact that the differences in the absolute numbers of the child population between the studied districts of the Arkhangelsk Oblast is almost 47-fold, while among children, malignant neoplasms are not recorded in all these districts. As a result, the use of the root-mean-square deviation method can lead to flawed grouping of districts on the map. Thus, along with the illegitimate use in the absence of a normal distribution, this method of data clustering is also not recommended as a tool for automated grouping of data on disease incidence and other intensive indicators in case of large differences in the population sizes of the compared territories. In such case, it would be more correct to compare the disease incidence in the municipal district with the incidence in the entire higher-order administrative unit (in our study, oblast), with the subsequent arbitrary establishment of classes (for example, less than 0.9, 0.9-1.1, and more than 1.1 of the incidence rates in the higher-order administrative unit of the Russian Federation).

The geometric coefficient when using the geometric interval classifier can be represented by an inverse function to optimize the class ranges. Geometric intervals are created by minimizing the sum of squares of the number of elements in each class, thereby ensuring that all classes have roughly the same number of values, while the interval sizes are also similar. This algorithm is used to process continuous data and combines the advantages of equal ranges, natural breaks, and quantile methods. It allows separating the average values from outliers, thereby yielding an outcome that is adequate from a cartographic standpoint and visually attractive. Besides, this clustering method is suitable for displaying a dataset when the attributes of most territories are equal to zero [12]. An example of the latter is the incidence of malignant neoplasms in the child population. Hence, the geometric interval method is most suitable as a tool for automated clustering of disease incidence data.

Our research limitations included restricted possibilities of using the selected research methods, limited characteristics of the research objects, the unambiguity of the context and the subject of the experiment under consideration, along with the ethical and sociocultural restraints.

Conclusion

Our results suggested that when using tools for automated clustering of spatially referenced incidence data in the context of municipalities and their visualization in ArcGIS, it is necessary to consider several factors that directly affect the accuracy of their presentation:

- Absolute and relative differences in disease incidence rates.

- Linearity of distribution of disease incidence levels.

- The nature of the distribution of disease incidence levels (normal, other than normal).

- Absolute population numbers and the number of cases of diseases.

- The number of zero values of disease incidence.

For an objective presentation of quantitative indicators of incidence on a map, it is advisable to choose a clustering tool based on the geometric interval method.

Conflict of interest

The authors declare no conflicts of interest.

References

1. Vyushkov MV, Zaitseva NN, Efimov EI, Kitaeva LS, Pobedinsky GG, Sarskov SA. Geographic Information Technologies in Epidemiology – An Up-to-Date Research Direction of Academician I.N. Blokhina Nizhny Novgorod Scientific Research Institute of Epidemiology and Microbiology. *Public Health and Life Environment – PH&LE* 2021; (4): 31-42. Russian. https://doi.org/10.35627/2219-5238/2021-337-4-31-42.
2. Krasilnikov IA, Strukov DR. The results of the first Russian conference "Geographic information systems in health care of the Russian Federation: data, analysis, decisions". *Medical doctor and information technologies* 2012; (2): 25-29. Russian. https://www.elibrary.ru/item.asp?id=17833427.
3. Strukov DR, Gorokhov VL. Geoinformation systems and multidimensional statistical methods of spatial analysis in investigating disease incidence. *Information and Control Systems* 2009; (3): 57-62. Russian. https://www.elibrary.ru/item.asp?id=12513635.
4. Korovka VG, Galkin VB, Panidi EA, Kuznetsov IS, Beltyukov MV, Sokolovich EG, et al. Potential of geoinformation technologies to improve the monitoring of socially significant infections outbreaks. *Profilakticheskaya Meditsina* 2021; 24(10): 7-13. Russian. https://doi.org/10.17116/profmed2021241017.
5. Zhukov KV, Udovichenko SK, Nikitin DN, Viktorov DV, Toporkov AV. Application of Geographic Information Systems in epidemiological surveillance for West Nile Fever and other arbovirus infections at the modern stage. Infectious Diseases: News, Opinions, Training 2021; 10(2): 16-24. Russian. https://doi.org/10.33029/2305-3496-2021-10-2-16-24.
6. Slis SS, Kovalev EV, Nenadskaya SA, Vodopyanov AS, Lyalina LV. The usage of geographic information systems for operational epidemiological analysis of influenza incidence in the territory of Rostov-on-Don including mass events. *Medical Herald of the South of Russia* 2019; 10(3): 57-61. Russian. https://doi.org/10.21886/2219-8075-2019-10-3-57-61.
7. Blokh AI, Penyevskaya NA, Rudakov NV, Mikhaylova OA, Fedorov AS, Sannikov AV, et al. Geographic information systems as a part of epidemiological surveillance for COVID-19 in urban areas. *Fundamental and Clinical Medicine* 2021; 6(2): 16-23. Russian. https://doi.org/10.23946/2500-0764-2021-6-2-16-23.
8. Asatryan MN, Gerasimuk ER, Strukov DR, Shmyr IS, Vekhov AO, Ershov IF, et al. Development of software tools based on multi-agent modeling and implemented in the new generation geographic information system for solving epidemiological problems. *Journal of microbiology, epidemiology and immunobiology* 2021; 98(4): 468-480. Russian. https://doi.org/10.36233/0372-9311-130.

9. Studenikina EM, Mamchik NP, Klepikov OV, Vinogradov PM. Geoinformation systems for assessing the incidence of mass noncommunicable diseases in urban population. In: Assessment and Geoinformation Mapping of the Medical and Environmental Situation in the City of Voronezh: Collection of research articles. Kurolap SA, Klepikov OV, Eds. Voronezh: Digital Printing Publishing House; 2019: 55-83. Russian. https://www.elibrary.ru/item.asp?id=41595128.
10. Kovshov AA, Fedorov VN, Tikhonova NA, Novikova YuA. Experience of systematizing data on the state of sanitary and epidemiological well-being of the population in the Russian Arctic. *Russian Arctic* 2020; (10): 51-60. Russian. https://doi.org/10.24411/2658-4255-2020-12105.
11. Gorbanev SA, Novikova YuA, Fedorov VN, Kovshov AA, Tikhonova NA., Rakova VV, et al. Issues of creating an information system for analysis of environmental factors in the Russian Arctic. *Hygiene and Sanitation* 2021; 100(8): 858-862. Russian. https://doi.org/10.47470/0016-9900-2021-100-8-858-862.
12. Classifying numerical fields for graduated symbology. ArcMap 10.8. ESRI ArcGIS Desktop: Classifying Data 2021. https://desktop.arcgis.com/en/arcmap/latest/map/working-with-layers/classifying-numerical-fields-for-graduated-symbols.htm.

Authors:
Roman V. Buzinov – MD, DSc, Director of the Northwest Public Health Research Center, St. Petersburg, Russia. https://orcid.org/0000-0002-8624-6452.
Vladimir N. Fedorov – MD, Senior Researcher, Head of the Department of Risk Analysis for Public Health, Division of the Social and Hygienic Analysis and Monitoring, Northwest Public Health Research Center, St. Petersburg, Russia. https://orcid.org/0000-0003-1378-1232.
Aleksandr A. Kovshov – MD, PhD, Senior Researcher, Head of the Department of Occupational Hygiene, Division of Hygiene, Northwest Public Health Research Center; Assistant Professor, Department of Radiation Hygiene and Hygiene of Education, Training and Labor, North-Western State Medical University named after I.I. Mechnikov, St. Petersburg, Russia. https://orcid.org/0000-0001-9453-8431.
Yuliya A. Novikova – Senior Researcher, Head of the Division of the Social and Hygienic Analysis and Monitoring, Northwest Public Health Research Center, St. Petersburg, Russia, https://orcid.org/0000-0003-4752-2036.
Nadezhda A. Tikhonova – MD, Junior Researcher, Department of Risk Analysis for Public Health, Division of the Social and Hygienic Analysis and Monitoring, Northwest Public Health Research Center, St. Petersburg, Russia. https://orcid.org/0000-0003-4895-4009.
Maxim S. Petrov – MD, Deputy Head Physician, Center for Hygiene and Epidemiology in Arkhangelsk Oblast and Nenets Autonomous Okrug, Arkhangelsk, Russia. https://orcid.org/0000-0001-6692-1150.
Ksenia V. Krutskaya – MD, Head of the Division of Social and Hygiene Monitoring, Center for Hygiene and Epidemiology in Arkhangelsk Oblast and Nenets Autonomous Okrug, Arkhangelsk, Russia. https://orcid.org/0000-0002-7841-767X.