Automating and utilizing equal-distribution data classification

Gennady Andrienko[1,2], Natalia Andrienko[1,2], Ibad Kureshi[3], Kieran Lee[4], Ian Smith[4], and Toni Staykova[5]

[1] Fraunhofer Institute IAIS, Sankt Augustin, Germany

[2] City, University of London, UK

[3] Inlecom Systems, Brussels, Belgium

[4] Royal Papworth Hospital, Cambridge, UK

[5] Cambridge Medical Academy, Cambridge, UK

Abstract. Data classification, i.e., organising data items in groups (classes), is a general technique widely used in data visualisation and cartography, in particular, for creation of choropleth maps. Conventionally, data are classified by dividing the data range into intervals and assigning the same symbol or colour to all data falling within an interval. For instance, the intervals may be of the same length or may include the same number of data items. We propose a method for defining intervals so that some quantity represented by values of another attribute is equally distributed among the classes. An example is dividing a set of geographic regions into classes according to the values of the attribute "Birth rate" so that the classes have approximately equal total values of the attribute "Population" or "Arable land area". This kind of classification supports exploratory analysis of relationships between the attribute used for the classification and the distribution of the phenomenon whose quantity is represented by the additional attribute. The approach may be especially useful when the distribution of the phenomenon is very unequal, with many data items having zero or low quantities and quite a few items having larger quantities. With such a distribution, standard statistical analysis of the relationships may be problematic. We demonstrate the potential of the approach by analysing data referring to a set of spatially distributed people (patients) in relationship to characteristics of the areas in which the people live.

Introduction

Data classification (Slocum et al. 2013, Chapter 4; Kraak and Ormeling 2021, Section 7.3) involves organising data items into groups, called classes, according to some principle, for example, in the basis of values of one or more attributes, or according to geographic location, or time reference. The most frequently used type of classification is based on dividing the range of values of a numeric attribute into intervals, so that each of the data classes corresponds to one interval and consists of data items with the values of the attribute lying in this interval. The intervals defining the classes are called *class intervals*. There are many methods of defining class intervals. The most common methods, according to Slocum et al. (2013), include equal intervals, quantiles, mean-standard deviation, natural breaks, nested means, and optimal. The latter, given a certain measure of diversity, minimises the diversity within the classes and maximises it between the classes. Kraak and Ormeling (2021) additionally consider arithmetic series, geometric series, and harmonic series. The choice of an appropriate method of classification depends on the properties of the data (particularly, the distribution of the attribute values) and the purpose of the classification.

One of the common purposes of classification is to simplify visual representation of data, as is done in choropleth maps. Another purpose, which is in the focus of our paper, is to support exploratory data analysis. Organising data items in groups is a generic analytical technique that not only helps analysts

to deal with large data amounts but also facilitates abstractive perception and cognition by hiding excessive details and enabling an overall view of the data.

In this paper, we consider the use of data classification for the task of exploring relationships between values of thematic attributes. Let us illustrate it by an example. We have data describing the Lower Layer Super Output Areas (LLSOA) of London, which are small units of territory division used for collecting census and other statistical data. The data contain values of a numeric attribute "% BAME", which represent the percentages of non-white (i.e., Black, Asian, and minority ethnic) residents in the population of the areas (data source: UK Census 2011). Besides, we have data specifying the geographic locations of the London pubs, from which we count the number of pubs in each area (data source: OpenStreetMap). We want to know how the geographic distribution of the pubs is related to the ethnic structure of the population, in particular, to the proportions of non-white residents. As can be seen in Fig.1, a scatterplot where the districts are represented by dots positioned according to the values of the attributes "% BAME" and "Number of pubs" does not show us the character of the relationship clearly enough. It is because most of the districts have no or very few pubs. It is also hard to understand the relationship when using a map that shows the variation of the proportions of non-white people over the territory and the distribution of the pubs (Fig.2).

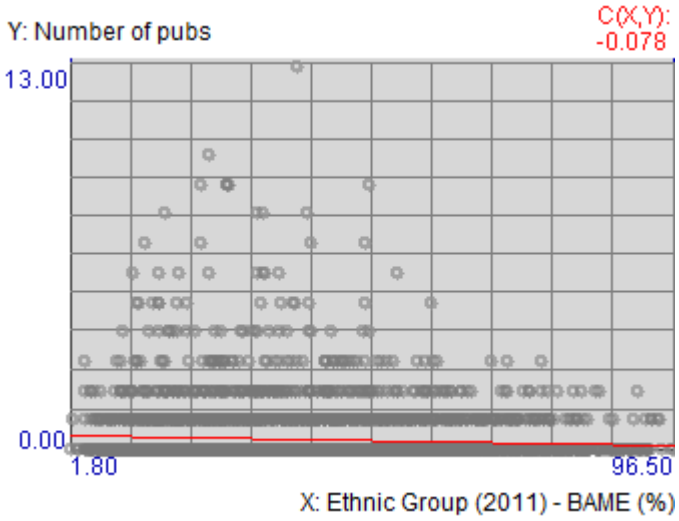

Figure 1. A scatterplot may not show the relationship between two attributes well enough when there are many data items with very low values of one of the attributes[1].

[1] All illustrations in this paper are screenshots made with the software system V-Analytics, which is publicly available at URL http://geoanalytics.net/V-Analytics/.

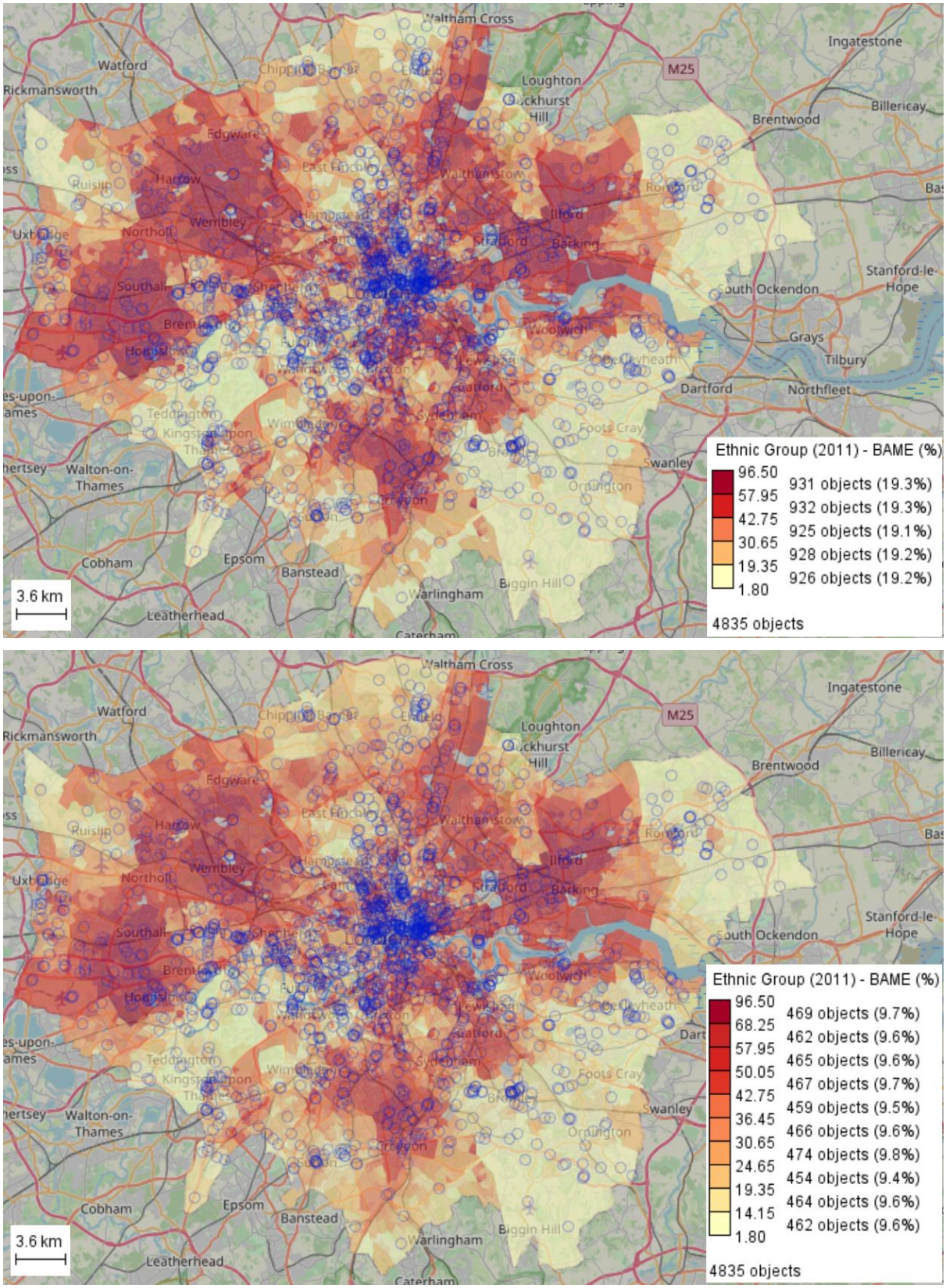

Figure 2. Classified choropleth maps represent the percentages of non-white population by statistical districts of London. In the upper image, the data are divided into quintiles, in the lower image – into deciles. Blue semi-transparent circles show the locations of the pubs in London.

Classification of the districts according to the values of the attribute "% BAME" helps us to approach the problem. We divide the data into quantiles (i.e., classes with approximately equal number of members) based on the values of this attribute and look how the pubs are distributed among the quantiles. In the upper map in Fig.2, the districts are divided into quintiles (i.e., each class includes about one fifth of all districts), and in the lower map into deciles (i.e., each class includes approximately one tenth of the districts). The distributions of the pubs among the classes are shown in the tables in Fig.3. In each table, we can see the number of the districts in each class, the total count of pubs per class, and the mean number of pubs per district of each class. The pub counts and the means tell us that there are notably fewer pubs in the districts with high percentages of non-white residents than in the remaining districts. More precisely, we see a non-linear relationship between the percentage of BAME and the number of pubs in a district. The number of pubs in the districts with the lowest percentages of non-white residents is somewhat smaller than in the groups of districts starting from the second quintile (Fig. 3, left) and the second decile (Fig. 3, right). Except for these lowest classes, the classes of districts with low and medium proportions of non-whites have relatively high numbers of pubs, and there is a notable decreasing trend in the number of pubs as the proportions of non-whites exceed 50% and further increase. By varying the number of data classes, we check and refine our observations. We see that the decreasing trend for the proportions of BAME exceeding 50% preserves and becomes even more prominent when we divide the data into more than 10 classes. This increases our confidence concerning the relationship observed. As the next step of the analysis, it is appropriate to confirm our observations by means of statistical analysis of the difference between the districts with the high percentages of BAME and the rest.

	N members	Number of pubs, sum	Number of pubs, mean
[01.80..9.35)	926	357	0.386
[19.35..30.65)	928	381	0.411
[30.65..42.75)	925	399	0.431
[42.75..57.95)	932	304	0.326
[57.95..96.50]	931	192	0.206

	N members	Number of pubs, sum	Number of pubs, mean
[01.80..14.15)	462	158	0.342
[14.15..19.35)	464	199	0.429
[19.35..24.65)	454	176	0.388
[24.65..30.65)	474	205	0.432
[30.65..36.45)	466	215	0.461
[36.45..42.75)	459	184	0.401
[42.75..50.05)	467	180	0.385
[50.05..57.95)	465	124	0.267
[57.95..68.25)	462	100	0.216
[68.25..96.50)	469	92	0.196

Figure 3. Table displays with coloured bars in the cells show how the pubs are distributed among the quintiles (left) and deciles (right) of the districts with respect to the proportions on non-white residents.

This example demonstrates that data classification can be a useful instrument for exploration of relationships between attributes. A similar example of analysis was described in more detail by Andrienko et al. (2020).

Apart from dividing data into quantiles, it may be meaningful to divide the data so that a certain set of objects or some quantitatively measured phenomenon is approximately equally distributed among the classes. For example, geographic districts can be classified so that the classes have approximately equal areas or amounts of population. Such a division may be desirable when the task is to explore the relationship between the attribute used as the basis for the classification and attributes that may be related to the objects or phenomenon distributed over the classes. For example, division of the London districts into equal-area classes may be useful for exploring the relationship between the proportions of non-white population and the amounts of green area, and division into equal-population classes is meaningful when the second attribute is the income per capita.

Earlier, Andrienko and Andrienko (2004) proposed a visualisation-based interactive approach to defining classes with approximately equal distribution of some quantity. In this paper, we propose an algorithm for obtaining such divisions in an automated way, which may be beneficial when the analysis process involves repeated data classifications with varying the number of classes or the attribute used as the basis for the classification. For such an analysis, interactive manual creation of data classes may be too laborious and time-consuming. Besides, the previous authors mostly focused on describing the method to define data classes but not so much on the use of the resulting classes in further data analysis.

The contribution of our paper is two-fold. First, we introduce the method for automated division of data into classes based on values of a numeric attribute A_C so that a given quantity Q is approximately equally distributed among the classes. Second, we show by example how the classes obtained in this way can be used for studying the relationship between the attribute A_C and another attribute A_X.

After a brief discussion of the related work concerning data classification, we present our approach in general terms and then demonstrate utilisation of it in a case study with data describing geographically distributed patients. The goal of the analysis is to investigate the relationships of characteristics of the patients to socio-demographic characteristics of the underlying territory.

Related work

Starting from Jenks' (1977) pioneering work on defining class intervals for optimizing homogeneity of the data classes, a number of approaches have been developed by cartographers and geovisualisation researchers. Slocum et al. (2013) present and discuss the multitude of common methods for data classification based on values of a numeric attribute and the problem of choosing an appropriate method. Egbert and Slocum (1992) were the first to implement an interactive tool for data classification to support exploratory analysis of spatial data with the use of choropleth maps. Further implementations combined interactive definition of class breaks with aggregation of attributes of interest for the classes (Andrienko and Andrienko 1999) and computation of indicators of the statistical optimality (Andrienko et al 2001). It was proposed to use a cumulative frequency curve for informed interactive selection of class intervals, and the idea was generalised to representing not only the numbers of the geographical objects being classified but also the corresponding sums of values of a quantitative attribute, such as population or area (Andrienko and Andrienko 2004); see an example in Fig.4, left. The authors mentioned that generalised cumulative curves can be used for interactive definition of classes of geographic units with approximately equal total population or covered area, as shown in Fig.4, right. However, it was not discussed how such classes can be used in further data analysis, in particular, for exploration of relationships between attributes.

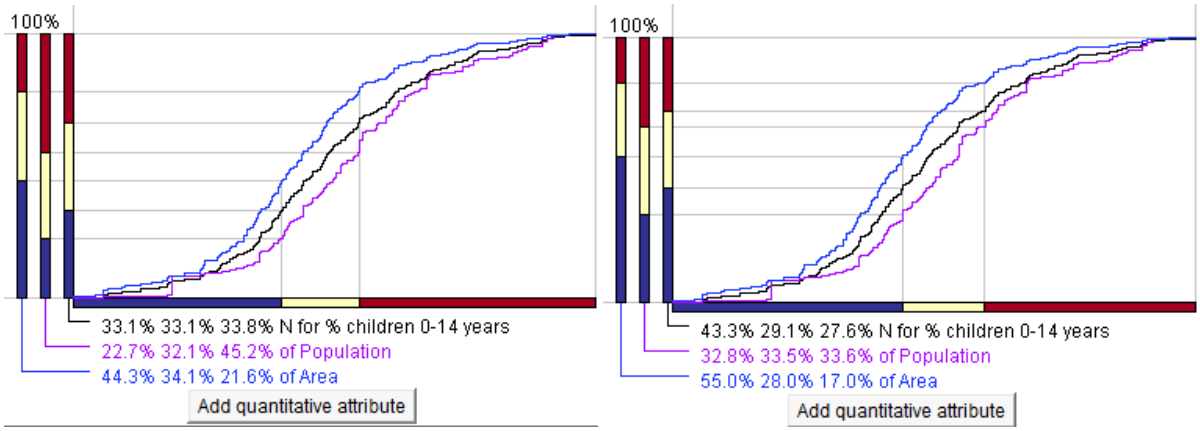

Figure 4. A cumulative curve display for supporting interactive data classification. The horizontal axis represents the value range of a numeric attribute (% of children 0-14 years old) used for the classification. The value range is divided into 3 intervals. In black is the cumulative frequency curve, in

purple the cumulative population curve, and in blue the cumulative area curve. On the left, the class breaks make classes with approximately equal counts of objects; on the right, the classes have approximately equal total population.

In the process of data analysis, it may be necessary to perform data classification multiple times by varying the number of classes or using different attributes as the classification basis. Manual work on defining classes by means of interactive tools may require much time of the analyst. It is generally desirable to save the valuable time of human analysts by supporting data analysis wherever possible by computational techniques. Particularly, automation of the data classification process saves the analyst's time that would be required for interactive classification.

We build on the work by Andrienko and Andrienko (2004), and we extend it by, first, proposing an algorithm for automated classification and, second, showing how resulting data classes can be used for exploration of relationships between attributes.

The classification method

Let ***B*** (base of the distribution) be a set of objects characterised by a numeric attribute $\boldsymbol{A_C}$, which is used for the classification of the objects, and let ***Q*** be another numeric attribute representing some quantities associated with the objects of ***B***. The goal is to divide the range of the attribute $\boldsymbol{A_C}$ into class intervals so that the corresponding object classes have approximately equal total quantities of ***Q***, i.e., sums of the values of ***Q***. The key idea of the algorithm is to order the data records describing the objects according to the values of $\boldsymbol{A_C}$ and then compute cumulative sums of the values of attribute ***Q*** along the ordering. The class breaks are chosen so that the sums of the values of ***Q*** between the breaks do not exceed the total sum of the values of ***Q*** divided by the desired number of classes. The details are provided in the following pseudo-code.

Algorithm: Define class intervals for attribute $\mathbf{A_C}$ equalizing the distribution of the quantity **Q**.
Inputs: data $<v^C_i, v^Q_i>$, i=1..N, where v^C_i is a numeric value of attribute $\mathbf{A_C}$, v^Q_i is a positive numeric value of attribute **Q**; desired number of classes K
Output: *breaks*[] – array of K-1 class interval breaks

```
begin
    for j := 1 to K-1 breaks[j] := null
    Sort the data so that v^C_{i+1} >= v^C_i, i=1..N-1
    totalQ := sum(v^Q_i), i=1..N
    currentSum := 0
    k := 1
    for i := 1 to N while k < K-1
        currentSum := currentSum + v^Q_i
        if (currentSum > totalQ /K)
            breaks[k] := (v^C_{i-1} + v^C_i) / 2
            k := k+1
            currentSum:= v^Q_i
    return breaks
end
```

This pseudo-code can be explained as follows. Initially, all class breaks are set to null. First, the data records are sorted in the order of increasing values of the attribute $\boldsymbol{A_C}$. Next, the sum of the values of the attribute ***Q*** from all records is computed and assigned to the variable *totalQ*. Then, the algorithm goes along the ordered sequence of records one by one. In each step, it adds the value of the attribute ***Q*** from the current record to the variable *currentSum*. When the value of the variable *currentSum* exceeds the value of *totalQ* divided by the desired number of classes K, it is the signal that there must

be a class break between the current record and the previous record, so that the previous record goes to the lower class and the current record to the next class. Hence, a new class break is generated by taking the middle value between the values of the attribute $\boldsymbol{A_C}$ in the previous record and in the current record. After that, the variable *currentSum* gets the value of the attribute $\boldsymbol{Q}$ from the current record and thereby starts accumulating the values of $\boldsymbol{Q}$ for the next class. The algorithm then proceeds to the following record in the sequence. The algorithm terminates when all K-1 class breaks have been defined.

It is usually not possible to achieve an absolutely equal division of the total quantity *totalQ* among K classes, i.e., such that the sum of the values of $\boldsymbol{Q}$ in each class exactly equals *totalQ*/K. If the division is just slightly unequal, it may not be too problematic for subsequent data analysis. When the classes are strongly unequal, it may be reasonable to try divisions with a larger or smaller numbers of classes.

In the following, we shall show by an example how data classes defined using the proposed method can be used for analysing relationships between the classification attribute $\boldsymbol{A_C}$ and other thematic attributes.

Case study

The data we analyse describe the appointments of patients to clinics for testing the existence and severity of symptoms of the obstructive sleep apnoea condition (OSA), which is a potentially serious health disorder. Tests are done using special devices, called oximeters, given to patients for taking measurements at home. A patient that needs to be tested gets an appointment to one of multiple sites (clinics) distributing oximeters. A test requires two journeys to a site, first for picking up an oximeter and then for bringing it back. It happens quite often that patients do not attend their appointments. It may happen even several times with the same patient. Health care researchers hypothesised that patients may be more likely to skip their appointments when they live far away from the clinic they need to travel to. However, a previous study ascertained that the no-shows were not related to the patients' distances from the sites where they had appointments (Lee et al. 2019).

It was also checked whether the distance to a clinic might affect the patients' inclination to undergo a test. It was conjectured that people living far away from the distribution sites may be unwilling to spend their time and money travelling when they do not feel really serious problems. However, this hypothesis was also not confirmed: no correlation was found between the distance to the clinic and the severity of the patient's condition according to both the objective result of the OSA diagnosis and the patient's subjective estimation.

The study described in this paper aimed at checking whether the patients' behaviours could be related to the living conditions, such as the level of poverty and unemployment, in their communities. We used official statistics reporting the indices of population deprivation by the statistical districts called lower layer super output areas (LLSOA). This is the lowest level of statistical data aggregation in the UK; hence, the corresponding deprivation data have the highest spatial resolution and accuracy from all openly accessible official statistics for the UK. In the subset of the data covering our study area (see Fig. 5), the population of the LLSOA varies from 937 to 5,438 inhabitants, with the mean 1,734, median 1,619, lower and upper quartiles 1,441 and 1,906, respectively, and 90th percentile 2,313. These numbers demonstrate that the districts are quite small in terms of population amounts.

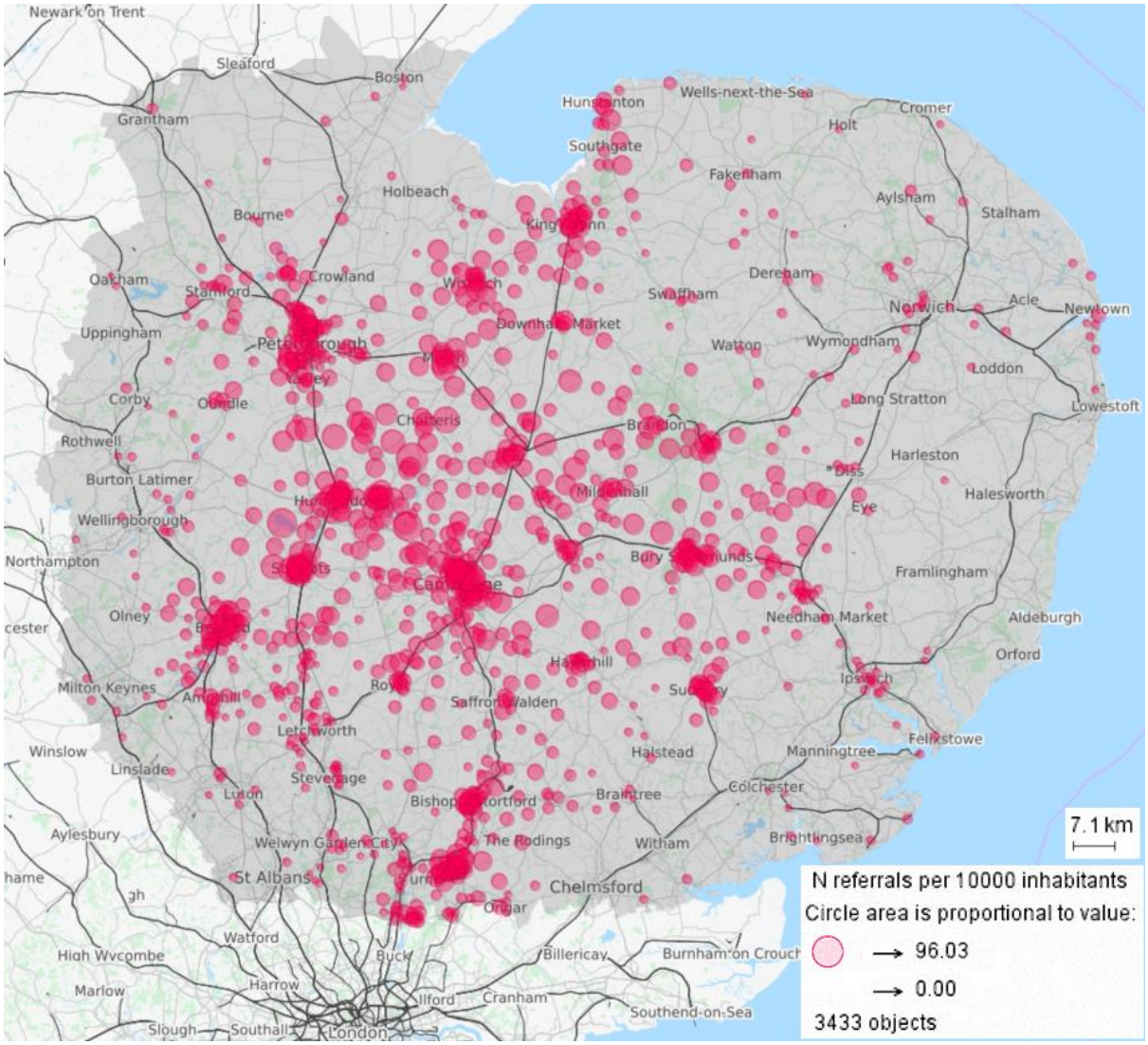

Figure 5. The study area with the spatial distribution of the patients represented by the counts of the patients per 10,000 residents of the districts (LLSOA). The counts are visually encoded by proportional sizes of the circle symbols coloured in red.

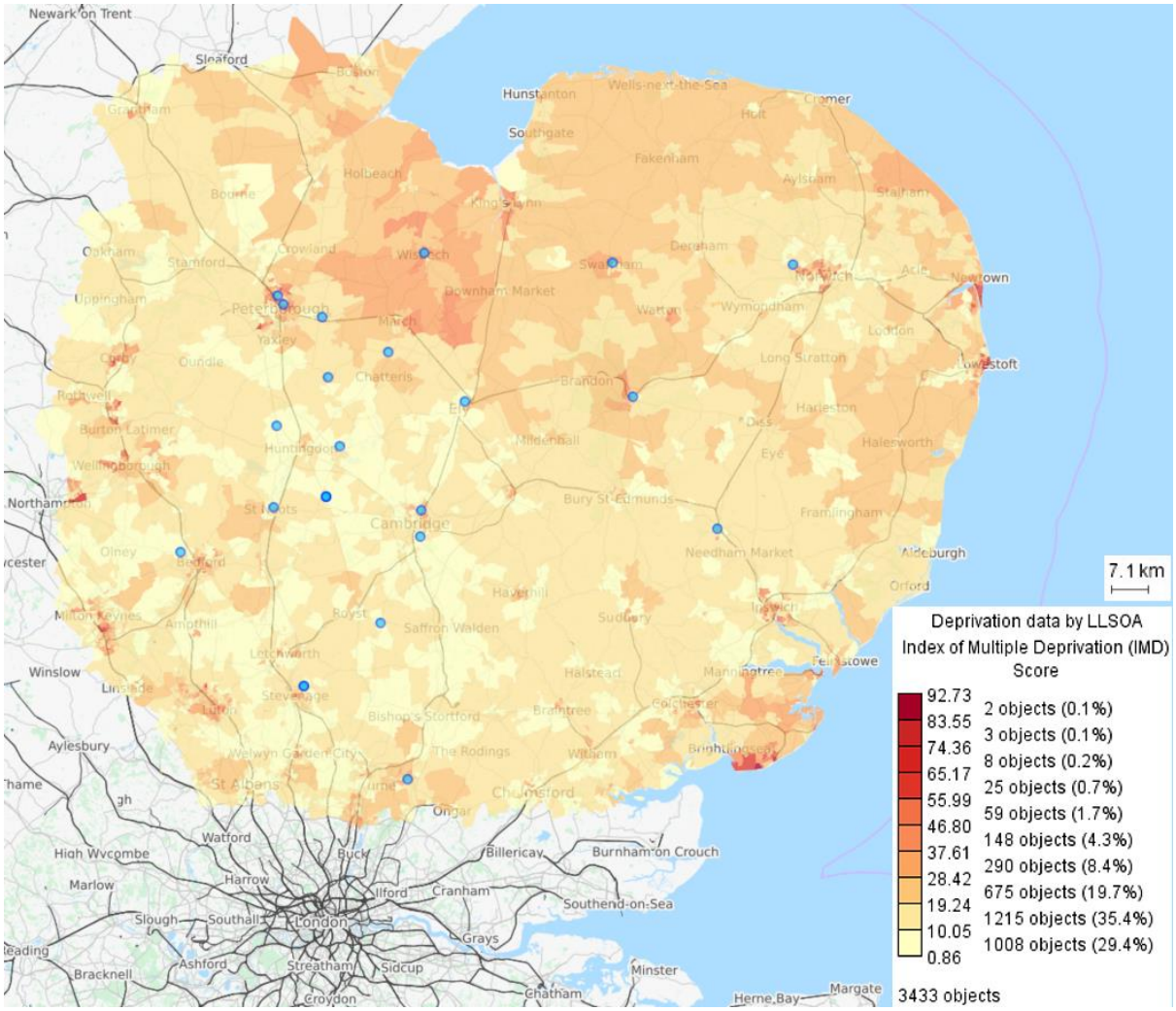

Figure 6. The districts of the study area are painted according to the values of the index of multiple deprivation (IMD). The value range is divided into 10 equal-length intervals, which are visually encoded by colours from light yellow to dark red. The dots in light blue show the locations of the clinics.

Figures 5 and 6 show the study area, which is located in the east of England, north of London. The map in Fig. 5 represents the spatial distribution of the patients' homes. In order to protect the privacy of the individual location data, they have been aggregated by the LLSOA. The map shows the counts of the patients, including those who did not attend the test (referred to as no-shows), per 10,000 residents of the districts. It can be seen that the distribution of the patients is very uneven. The patients are densely concentrated in relatively few areas and sparsely distributed over the remaining areas. The map in Fig. 6 shows the variation of the index of multiple deprivation (IMD), which is a composite attribute taking into account the income, employment, education, health, and other aspects of people's life. The LLSOA data also include the deprivation values for each of these aspects.

While it is possible to attach the deprivation scores of the areas to the records of the patients, it is not reasonable to look for direct correlations between the patients' individual data and the data of the areas in which they live. There can be no correlations at the individual level due to the highly uneven distribution of the patients, such that a few areas include a large number of patients with very diverse health characteristics.

The approach we use for revealing possible relationships between the area deprivation scores and the patients' health conditions and behaviours is based on the use of data classification. The idea is to divide the set of areas into classes according to the deprivation scores so that the patients are approximately equally distributed among the classes. Since the classes have almost equal numbers of

patients, they are comparable in terms of the characteristics of the patients. We can thus look whether and how the subset of patients living in poorer areas differs from the subset of patients living in wealthier areas.

We used the patients' data to compute for each output area the total number of patients, the number of no-shows, the counts of the patients with hypertension, diabetes, and those who smoke, and the counts of the patients by the quintiles of the attributes Age, BMI (body mass index), and ESS (Epworth Sleepiness Scale Score) (Johns 1991). The latter is the self-estimated degree of sleepiness expressed as a number from 0 (no sleepiness) to 24 (extremely severe). The main test result is an objective measure of the severity of OSA, that is the value of the oxygen desaturation index (ODI) measured by the overnight oximetry at home. There are four categories for ODI, normal, mild, moderate, and severe. So, for each area, we also obtained the counts of the patients by these four ODI categories.

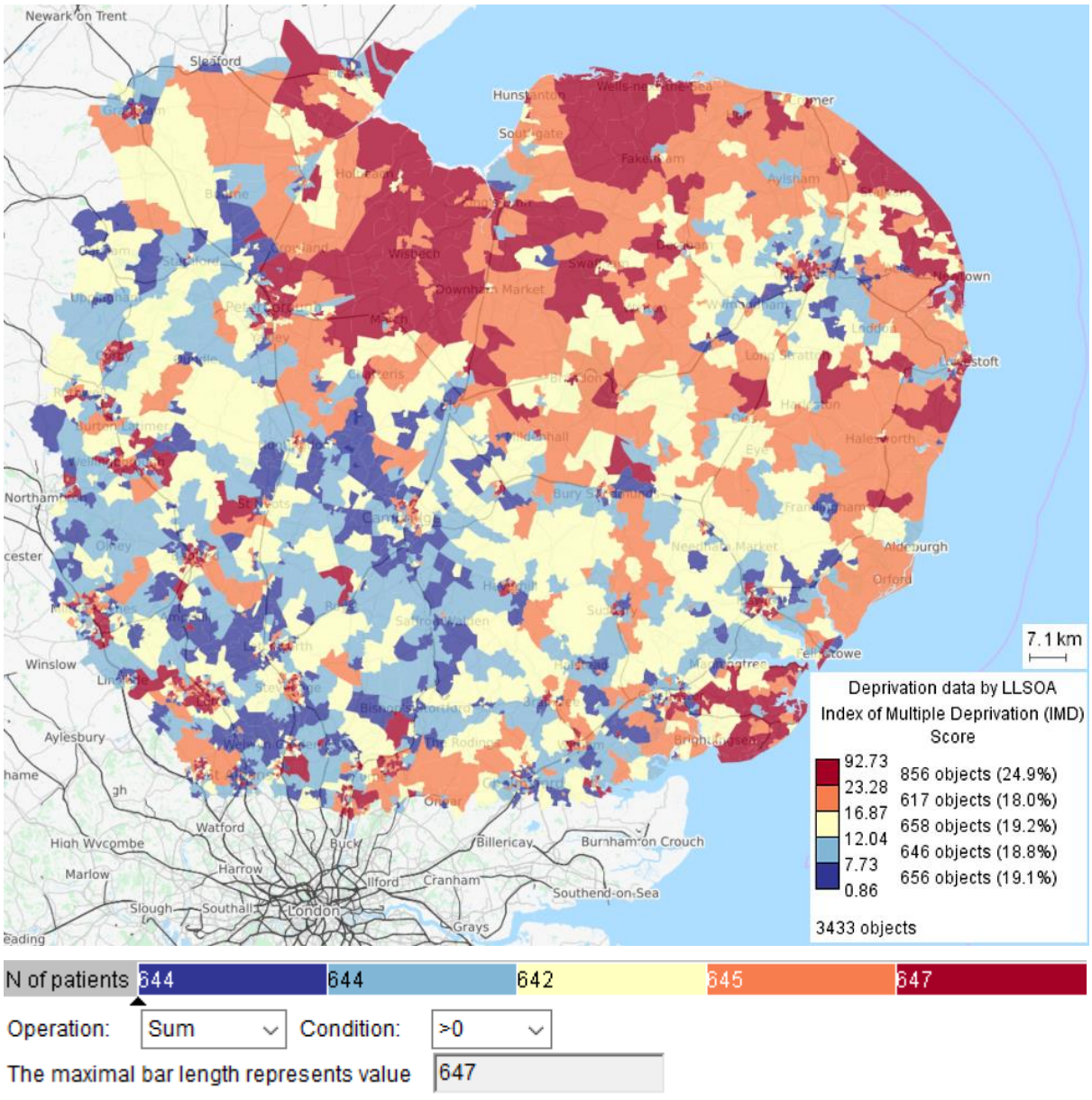

Figure 7. The output areas are classified according to the IMD (index of multiple deprivation) scores so that the patients are approximately equally distributed among the classes. The bar chart below the map shows the number of the patients in each class of the areas.

The map in Fig.7 shows the output areas classified according to the values of the attribute IMD (index of multiple deprivation) so that the patients are approximately equally distributed among the five

classes. In terms of our notation, ***B*** is the set of output areas, A_C is the IMD score, and ***Q*** is the number of the patients. The area classes are represented by the colours from dark blue for the lowest interval of the IMD values to dark red for the highest value interval. We use a diverging colour scale (Harrower and Brewer 2003) to distinguish between below-average and above-average values (encoded in blue and red, respectively). The bar charts in figures 8 and 9 represent the summary statistics of the patients and no-shows by the area classes. The bars are painted in the same colours as the areas in the map. The bar lengths in each row are proportional to the total counts of the patients from a particular category by the area classes. The first row in Fig. 8 presents the counts of the no-shows.

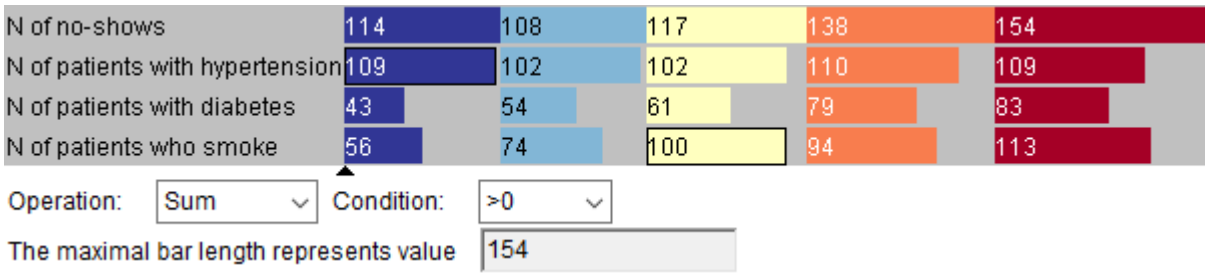

Figure 8. Counts of no-shows and special categories of patients by the classes of the output areas.

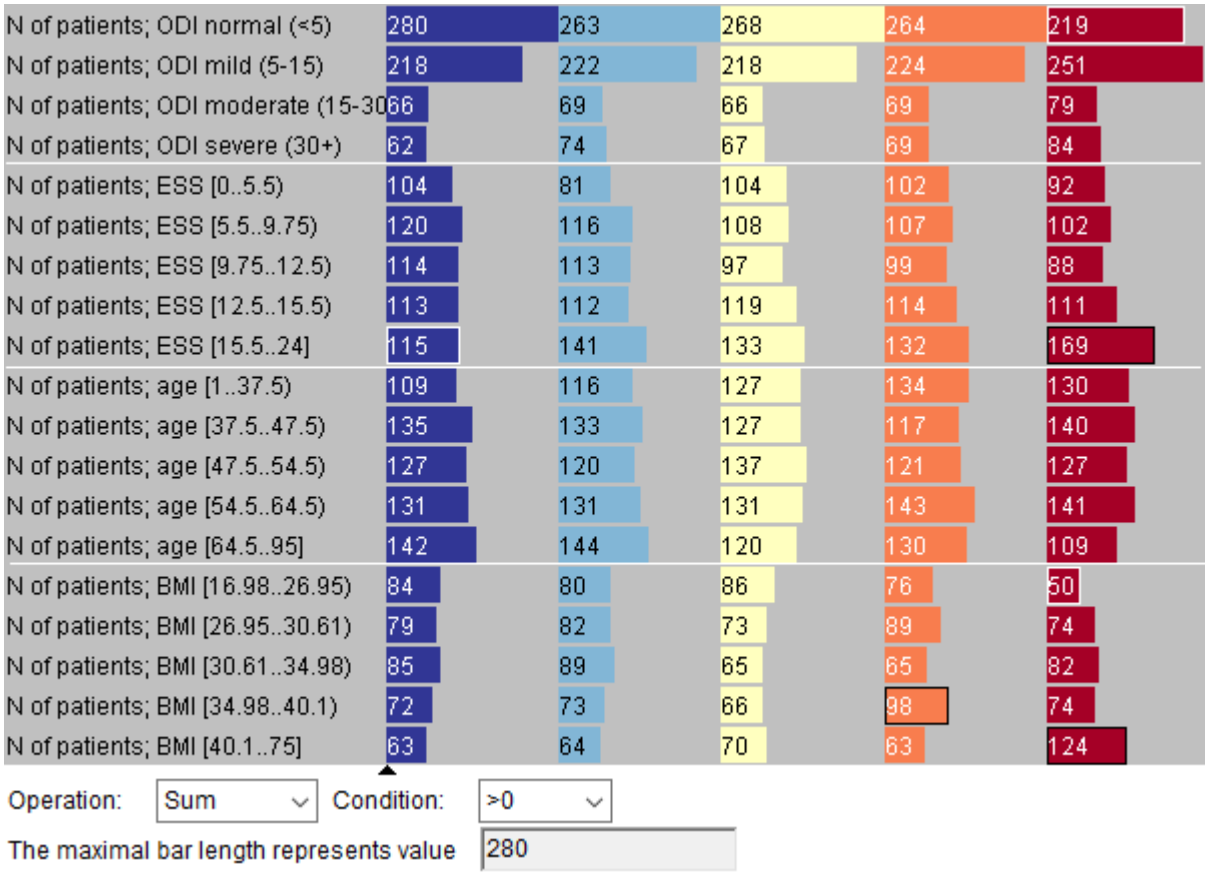

Figure 9. Counts of the patients by the area classes and value intervals of the attributes ODI (oxygen desaturation index), ESS (self-estimated sleepiness), age, and BMI (body mass index).

The statistics presented in figures 8 and 9 indicates the existence of notable differences between the classes of the areas in terms of the no-shows and the characteristics of the patients. Thus, the number of no-shows is higher in the areas with the larger IMD scores, i.e., in the poorer areas. In the poorest area class, the number of the patients with the normal level of ODI is much smaller and the number of the patients with the severe ODI level is higher than in the wealthier area classes. The statistics of the self-estimated sleepiness (ESS) supports the hypothesis that the patients living in poor areas are

less inclined to go for a test until they feel that they have a serious problem. Specifically, for the class of the districts with the highest deprivation levels (dark red), the counts of the patients with low to average values of ESS are notably smaller than the counts of patients with high values of ESS. In addition, when we compare the numbers of the patients with the highest values of ESS across the classes of the districts, we see that the top class has prominently the highest number of such patients.

Apart for the statistics of the no-shows, ODI, and ESS, differences exist also in terms of the attributes "diabetes", "smoking", "age", and "BMI". In the poorer areas, there are more patients with diabetes, overweight (high BMI), and those who smoke, whereas the number of elderly people who take the test is smaller than in the other area classes. Similar observations have been done regarding specific deprivation indexes, such as income, employment, education, health, and other.

Having detected these differences, we checked their statistical significance. Specifically, we compared the class of the most deprived districts against the class of the least deprived districts. For interval data comparisons (i.e., ODI, ESS and BMI), we used two-tailed independent sample (homoscedastic) t-tests. For ratio data comparisons (i.e., no-shows, smoking, diabetes, and referrals per 10,000 inhabitants), we used z-scores test and calculated p values from the resulting z scores. According to the tests, the differences between the lowest and highest classes are statistically significant for the percentages of the no-shows ($p=0.001$), ODI ($p=0.011$), ESS ($p=0.001$), BMI ($p<0.00001$), the number of smoking patients ($p<0.00001$), and the number of patients with diabetes ($p=0.0002$). Besides, a significant difference exists in the number of referrals (i.e., people who wished to be tested) per 10,000 residents of the areas ($p=0.005$); namely, poorer areas have fewer referrals. Since almost all of the comparisons were statistically significant, it was not necessary to apply formal methods for correction of p values.

Hence, by means of the equal-distribution data classification, we were able to reveal the relationships between the behaviours of the patients and the degree of deprivation in the areas where they live. Most of the detected relationships have been statistically confirmed. We would like to make a special note that these relationships do not manifest as correlations at the level of individual patients' data due to the extremely uneven spatial distribution of the patients. Given the particular properties of the data under analysis, the use of the equal-distribution data classification for the exploratory analysis proved to be very helpful.

Discussion and conclusion

The principal possibility of the equal-distribution data classification was mentioned long ago (Andrienko and Andrienko 2004). However, the authors did not discuss in detail how the resulting classification can be used in further data analysis, in particular, for exploration of relationships between attributes. Our paper complements the previous work by demonstrating an example of the use of equal-distribution data classes for exploration of the relationships between the attribute that served as classification bases and other attributes available in the data. The essence of the approach is to summarize the data by the classes and detect significant differences between the classes in terms of the attributes whose values have been summarised.

Our work presented in this paper was motivated by a practical need to understand relationships between two or more space-based phenomena. Our study, which was performed within an EU-funded research project Track&Know, has shown that equal-distribution classification may be especially helpful when there is a phenomenon with a highly uneven distribution over the (spatial) base. In such cases, other approaches to detecting relationships may not work.

While equal-distribution data classification can be done interactively using a generalised cumulative curve display, as was proposed in the earlier work, it is quite time-consuming when the classification needs to be done multiple times for different classification attributes and/or different numbers of classes. Therefore, we have devised an algorithm enabling automated classification.

We believe that our work, including the algorithm of the equal-distribution classification and demonstration of the way of utilising it in data analysis, can be useful for researchers and practitioners in spatial data analysis. We also think that the method may have wider use, as it is applicable not only to spatially distributed phenomena but also to phenomena distributed over any kind of common base that can be represented as a set of objects with attributes reflecting the distributions of the phenomena.

On the other hand, we deem it necessary to note that we do not propose the use of equal-distribution classification as a brand-new approach superior to the state-of-the-art methods for geographic data analysis. We would like to stress that this is a merely exploratory technique intended to support visual detection of possible presence of relationships between attributes. Once being detected, relationships need to be tested using statistical methods, as we did in our case study. Explorations similar to the presented one can be done using existing software tools, such as GeoDa[2] (Anselin and Rey, 2014), while explorers can benefit from the proposed way to automate the generation of equal-distribution data classes.

It is also appropriate to note the general limitations of data classification as a method to represent data visually. Such a representation may seriously distort the perception of the data distribution. Thus, classification masks the differences between the extreme values and the rest. It also exaggerates small differences between data items that happened to be put in different classes but can mask much larger differences between data items put in the same class. These kinds of distortions can be noticed by comparing the maps in Fig. 7 and Fig. 6. However, the limitations of classification do not deny its usefulness when it is utilised cautiously and purposefully. The way of obtaining classes, the class boundaries, and the sizes of the classes are very important pieces of information that need to be taken into account in interpreting data visualisations and in data analysis.

Acknowledgements

This research was supported by Fraunhofer Center for Machine Learning within the Fraunhofer Cluster for Cognitive Internet Technologies, by DFG within Priority Programme 1894 (SPP VGI), and by EU in projects Track&Know and SoBigData++.

References

Andrienko, G. and Andrienko, N. (1999) Interactive Maps for Visual Data Exploration. International Journal Geographical Information Science, 1999, 13(4), pp.355-374

Andrienko, G., Andrienko, N., and Savinov, A. (2001) Choropleth Maps: Classification revisited. In Proceedings ICA 2001, Beijing, China, vol.2, pp.1209-1219

Andrienko, N., and Andrienko, G. (2004) Cumulative Curves for Exploration of Demographic Data: a Case Study of Northwest England. Computational Statistics, 2004, v.19 (1), pp. 9-28

Andrienko, N., Andrienko, G., Fuchs, G., Slingsby, A., Turkay, C., and Wrobel, S. (2020) Visual Analytics for Data Scientists. Springer

Anselin, L., and Rey, S. (2014) Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace, and PySAL. GeoDa Press LLC, Chicago, IL

Egbert, S.L. and Slocum, T.A. (1992) EXPLOREMAP: an exploration system for choropleth maps. Annals of the Association of American Geographers, 82, pp.275-288.

Harrower, M., and Brewer, C. (2003). ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. The Cartographic Journal. 40. 27-37. 10.1179/000870403235002042.

[2] https://en.wikipedia.org/wiki/GeoDa

Jenks, G.F. (1977) Optimal data classification for choropleth maps: Occasional Paper No. 2, Dept. Geography, Univ. Kansas, 24 p.

Johns, M.W. (1991) A New Method for Measuring Daytime Sleepiness: The Epworth Sleepiness Scale, Sleep, Volume 14, Issue 6, November 1991, Pages 540–545, https://doi.org/10.1093/sleep/14.6.540

Kraak, M.-J. and Ormeling, F. (2021) Cartography. Visualization of Geospatial Data. Fourth Edition. CRC Press

Lee, K., Andrienko, N., Andrienko, G., Kureshi, I., Staykova, T., & Smith, I. (2019). P045 Aggregated patient journeys and no-show rates of oximetry outreach network in East Anglia. BMJ Open Respiratory Research 2019; 6 :doi: 10.1136/bmjresp-2019-bssconf.45

OpenStreetMap. URL: https://www.openstreetmap.org/

Slocum, T. A., MacMaster, R.B., Kessler, F.C., and Howard, H.H. (2013): Thematic Cartography and Geovisualization, Pearson New International Edition. Pearson Higher Education.

UK 2011 Census - Office for National Statistics. URL: https://www.ons.gov.uk/census/2011census