ISSN: 1857-9000, EISSN: 1857-9019

CLASSIFICATION OF MULTIPLE NUMERICAL ATTRIBUTES IN ARCGIS ENVIROUMENT

Kristina KASTREVA[1] and **Penka KASTREVA** [2]

UDC: 519.85:004.6]:[528.9:616.83(497.2)

SUMMARY

The article presents the application of cluster analysis in solving practical problems in medicine. It can be used to detect models in the spatial or temporal distribution of a particular disease. Determining the location of spatial clusters is important when an epidemic of a disease occurs and can often provide guidance on what can cause it. The advantage of using cluster analysis in a GIS environment is that cluster mapping tools allow visualization of cluster locations.

In this paper, we focus on studying the clustering method that has been applied to datasets of a neurological disease. A database on the epidemiology of hereditary neuropathies in Bulgaria was used.

Key words: maps, neurological diseases, cluster analysis

1 INTRODUCTION

When mapping and analyzing statistical information, it is necessary to classify the dataset with an appropriate classification method. The most commonly used classification methods (Equal interval, Defined interval, Quantile, Natural breaks, Geometrical interval, Standard deviation) include grouping of objects / phenomena that have similar values for an individual attribute.

However, it is possible to classify objects based on multiple attributes (Slocum et all, 2005). Such process of classification (grouping) of the data is known by the term cluster analysis. It was first introduced in 1939.

Cluster analysis has long played an important role in the wide variety of fields of study such as: psychology and medicine, social sciences, biology,

[1] **Kristina KASTREVA, MD, PHD –** dr.kastreva@gmail.com, Univeristy Hospital "Alexandrovska", Department of Neurology, Medical University, Sofia, Bulgaria, Address: Bulgaria, Sofia 1431, Georgi Sofijski Str.1, Tel. +35929230462

[2] **Assoc. Prof. Penka KASTREVA, PHD**, penkakastreva@gmail.com, South-West University "Neofit Rilsky", Address: Bulgaria, Blagoevgrad 2700,66 Ivan Michailov st.

statistics, pattern recognition, information retrieval, and more. There are many applications for solving various practical problems. This analysis is useful when classifying a large body of information that is collected in the field of medicine. Cluster analysis can reduce the amount of data, making it more visible in tables and maps. In medicine, cluster analysis is used to identify patterns in the spatial and temporal distribution of a particular disease.

Clustering is a process of dividing the information into groups, known as clusters on the basis of many features (characteristics) simultaneously. The purpose of grouping is that each cluster contains similar objects (observations), and the objects from different clusters differ significantly. The similarity of the objects in the cluster is characterized by common properties (attributes). In this way, each object (observation) of the common data set will belong to a cluster with the closest average value (cluster centers).

Cluster analysis contains many numbers and different mathematical procedures. The more famous of them are: Euclidean distance, Manhattan distance, Chebishev distance and others. (Tan P. et all, 2019).

The cluster analysis methods are hierarchical and non-hierarchical. Hierarchical Cluster Analysis is used, when the number of clusters are not been previously determined. In the process of clustering, small clusters are consistently merged into larger clusters or large clusters split into smaller clusters.

Non-hierarchical cluster analysis is used for a predetermined number of clusters. The new clusters are formed by successive iterations, until a certain condition terminates the process of splitting or merging.

Three methods for non-hierarchical cluster analysis with a predetermined number of clusters are used: K-Means cluster analysis; Nearest neighbor method; Method of outermost neighbors.

K means clustering algorithm was developed to classify or group objects based on attribute properties into K number of groups, with K being an integer positive.

K – Means defines the center of gravity (centroid), which is the average of the group of objects, after which it is named. It applies to the classification of objects in a continuous n-dimensional space of attribute values. In this case, each object is perceived as a point characterized by a number of variables (attributes).

The clustering is done by minimizing the sum of squares of the distances between the data and the corresponding centroid of the cluster.

The choice of the initial centroids is a key step in the K-Means method (Tan P. et all, 2019). One approach to determine initial centroids is to select random centroids for each cluster numerous times. When the total mean squared error for all clusters has the smallest value, then we can choose the initial centroids.

The general approach is to use a hierarchical method for classifying data. In it, every object in the dataset is considered a cluster. For each cluster, the Euclidean distances between each point of the cluster and its centroid are calculated. For the smallest calculated distance between the cluster and any point, a new centroid cluster is formed, whose values are obtained as the arithmetic mean of the calculated distances of the old cluster.

2 METHODS OF STUDY

2.1 Theoretical Basis of K-Means Analysis in ArcGIS Environment

In ArcGIS, the Mapping Clusters tool contains various clustering algorithms: Cluster and Outlier Analysis, Hot Spot Analysis, andGrouping Analysis. The latter of them performs classification procedures that detect natural groups in the data. Grouping is based on attribute values of the objects and additional space or time constraints. When the "No spatial constraint" parameter is selected for the Spatial Constraints parameter, the analysis is performed with the K-Means algorithm. In this case, the objects are grouped using data only for Spatial proximity. In fact, the objects may not be in close proximity of one another in time and space in order to be part of the same group.

The successive steps in applying the K-Means grouping are (Tan P. et all, 2019):

- Choosing the number of clusters (groups), which is the parameter K;
- Repeatedly selecting randomly selected centroids for each cluster. Accepted for centroids those with minimum mean square error;
- Performing a first iteration in which the initial K number of centroids is the arithmetic mean of the values of each cluster;
- For each centroid is determined which points are closest to it by calculating the distances between them and the centroid;
- New centroids are recalculated by arithmetic mean from the calculated distances of the old cluster;
- The process continues until there is no a change in the centroids or there are no points (objects) to move from one cluster to another.

For the analysis below, we will use the terminology adopted in ArcGIS, which will help us to read correctly the report that is generated after grouping with the K-Means algorithm. For this purpose we have to implement the following remarks: The word cluster is replaced by "group", and by attribute in the attribute table we will mean "variable".

We will also make the following notations:
V - variable value

n_c - number of objects in the group;

n_v - number of variables in the group;

n_i - number of objects in the i^{th} group;
n - number of all objects in the data set;
V_{ij}^{k} - the value of the k^{th} variable of the j^{th} object in the i^{th} group (individual value of the entire data set);
$\overline{V^{k}}$ - the mean value of the k^{th} variable (centoid of the data set)
$\overline{V_i^{k}}$ - the mean value of the k^{th} variable in group i (centoid of the i^{th} group).

The Euclidean distance is used as a measure of proximity between two points. It is calculated as values between each point of the cluster and its centroid.

The centroid is the mean of the i^{th} group and it is defined by Equation 1:

$$\overline{V_i^{k}} = \frac{1}{n_i} \sum_{V_{ij}^{k} \in i^{th}\text{group}} V_{ij}^{k} \ . \qquad (1)$$

The Euclidean distance is the error or the deviation of each point from the centroid of the group (cluster), with the condition that the sum of the squares of deviations be minimum. A new group is created where the smallest distance between two points appears. The new centroid is being recalculated, etc.

The K-Means algorithm in Arc GIS software calculates the parameter R^2, which reflects how much of variation in the original data was retained after the grouping process. The greater the value of R^2 for a particular variable is, the better the clustering efficiency is (https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/how-grouping-analysis-works.htm).

The value of R^2 is calculated by the sum of the mean squared errors and these are the differences between the value of each variable and its average in the group (Euclidean distances). In ArcGIS this difference is designated as SSE (Sum of Squared Error), called global minimum.
The formula by which this parameter is calculated is:

$$R^2 = \frac{SST - SSE}{SST} , \qquad (2)$$

where SST (Total Sum of Squares) is the sum of the mean square errors of the variable V in the entire dataset, and SSE (Sum of Squares Error) only in the group.

For them, the formulas are:

$$\mathrm{SST} = \sum_{i=1}^{n_c}\sum_{j=1}^{n_i}\sum_{k=1}^{n_v}\left(V_{ij}^{k} - \overline{V^{k}}\right)^2 \tag{3}$$

$$\mathrm{SSE} = \sum_{i=1}^{n_c}\sum_{j=1}^{n_i}\sum_{k=1}^{n_v}\left(V_{ij}^{k} - \overline{V_{i}^{k}}\right)^2 \tag{4}$$

The magnitude $\left(V_{ij}^{k} - \overline{V_{i}^{k}}\right)$ in Equation 4 is the deviation of each value in i^{th} group from its centroid.

The Share factor is calculated for each variable in the group. This coefficient is a proportion calculated by the range of data in the group to the range of data of the entire data set, or

$$\mathrm{Share} = \frac{(\max - \min)^{\mathrm{group}}}{(\max - \min)^{\mathrm{data\ set}}} \tag{5}$$

This coefficient indicates what part of the values of a particular variable are contained in the group of the total range of values of the respective global variable.

2.2 Graphical Presentation of the Data

Data grouping is represented graphically as a map image, with individual groups displayed in different colors. Statistical calculations are presented in text and graphical form with a report automatically created by the software, which presents in summary the analysis of the data by groups and by variables.

For each variable, the following statistical values are calculated: Mean, Standard Deviation, Minimum, Maximum, the value of the parameter R^2, Share factor and boxplot graphics to them. These statistics are calculated for each group individually as well as for the entire data set called global values. In the report they are printed with different colors and correspond to the classification made on the map. The summary statistics are presented in black color.

Boxplot graphics in the statistical report show how the values in the group (indicated by color) are associated with the entire data set (shown in black).

Each + sign indicates the number of objects that fall outside the range of data in the group. The vertical lines of the black rectangle show, respectively, the lower boundary of the first quartile, the median, and the upper boundary of the first quartile. Outside the rectangle are the smallest and largest value for the entire data set shown with black vertical lines. The black dot indicates the location of the arithmetic mean.

Vertical color lines indicate the range of data (minimum and maximum) and the color dot is the arithmetic mean of the group.

2.3 Application of Cluster Analysis

We will look at an example of classification of objects by three variables with different statistical distributions of data. Data on the number of patients with hereditary neuropathies for the three main ethnic groups in Bulgaria, divided by administrative units, were used. Numerical data represent a radically different statistical distribution. One dataset presents numerical values with a normal distribution of Roma patients. The other two variables represent an uneven distribution of data related to the Bulgarian and Turkish ethnic groups in Bulgaria.

In this study, the K –Means function was used to analyze which districts have similar characteristics with respect to the number of affected individuals. The classification was performed for the three variables simultaneously designated as "Bulgarians", "Roma" and "Turks". The cluster analysis was performed with ArcGIS software in two, three and four groups.

The K-Means function requires specifying the number of groups, the variables by which the classification and the similarity measure are carried out (Euclidean distance.) And, as mentioned above, for the Spatial Constraints parameter “No spatial constraint” should be selected.

The output file is a new vector layer in the map. It contains all objects and analyzed variables (attributes). A field (SS_GROUP) is added to the attribute table indicating to which group belongs each object.

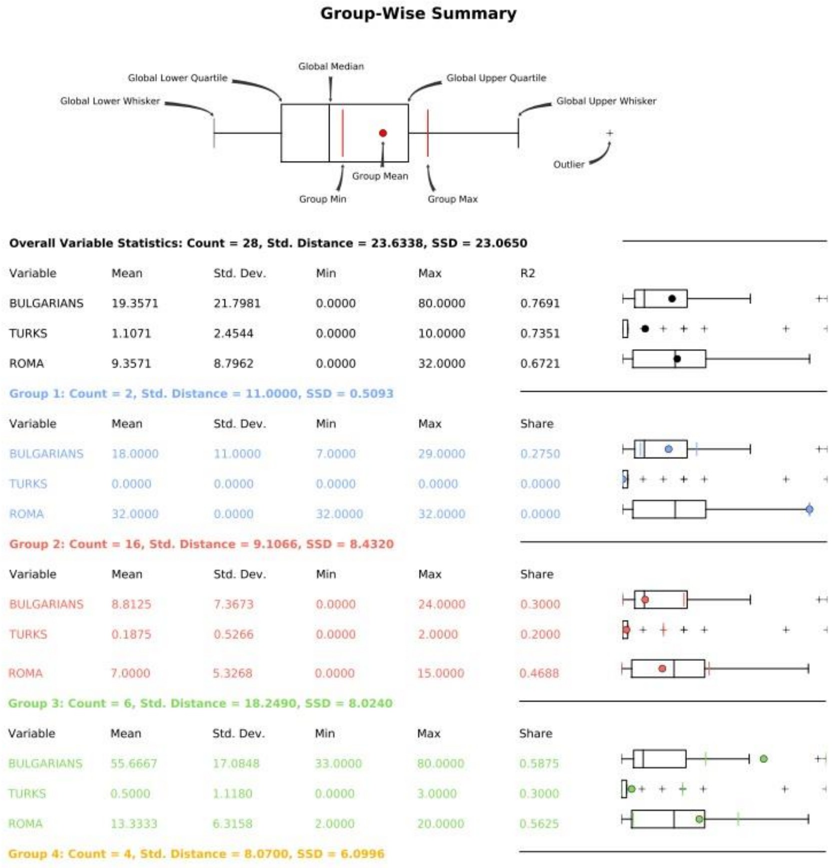
At the request of the user, an analysis report in pdf format is also created.

3 RESULTS

From the data collected in the Bulgarian Patient Registry for hereditary peripheral neuropathies, a number of studies can be made for grouping by ethnicity, age, gender and others. Research into grouping patients by district for particular genetic forms of the disease is interesting. The results of the creation of groups of district containing similar data for the three main ethnic groups in Bulgaria are presented here.

- Results of the calculated statistics for the total dataset

The performed analysis of the grouping is presented with an automatically created (by the software) report. It consists of two parts. The first part is a summary of the analysis of the data by groups presented in (Figure 1) and the second by the variables (Bulgarians, Roma, Turks) presented in (Figure. 2).



Overall Variable Statistics: Count = 28, Std. Distance = 23.6338, SSD = 23.0650

Variable	Mean	Std. Dev.	Min	Max	R2
BULGARIANS	19.3571	21.7981	0.0000	80.0000	0.7691
TURKS	1.1071	2.4544	0.0000	10.0000	0.7351
ROMA	9.3571	8.7962	0.0000	32.0000	0.6721

Group 1: Count = 2, Std. Distance = 11.0000, SSD = 0.5093

Variable	Mean	Std. Dev.	Min	Max	Share
BULGARIANS	18.0000	11.0000	7.0000	29.0000	0.2750
TURKS	0.0000	0.0000	0.0000	0.0000	0.0000
ROMA	32.0000	0.0000	32.0000	32.0000	0.0000

Group 2: Count = 16, Std. Distance = 9.1066, SSD = 8.4320

Variable	Mean	Std. Dev.	Min	Max	Share
BULGARIANS	8.8125	7.3673	0.0000	24.0000	0.3000
TURKS	0.1875	0.5266	0.0000	2.0000	0.2000
ROMA	7.0000	5.3268	0.0000	15.0000	0.4688

Group 3: Count = 6, Std. Distance = 18.2490, SSD = 8.0240

Variable	Mean	Std. Dev.	Min	Max	Share
BULGARIANS	55.6667	17.0848	33.0000	80.0000	0.5875
TURKS	0.5000	1.1180	0.0000	3.0000	0.3000
ROMA	13.3333	6.3158	2.0000	20.0000	0.5625

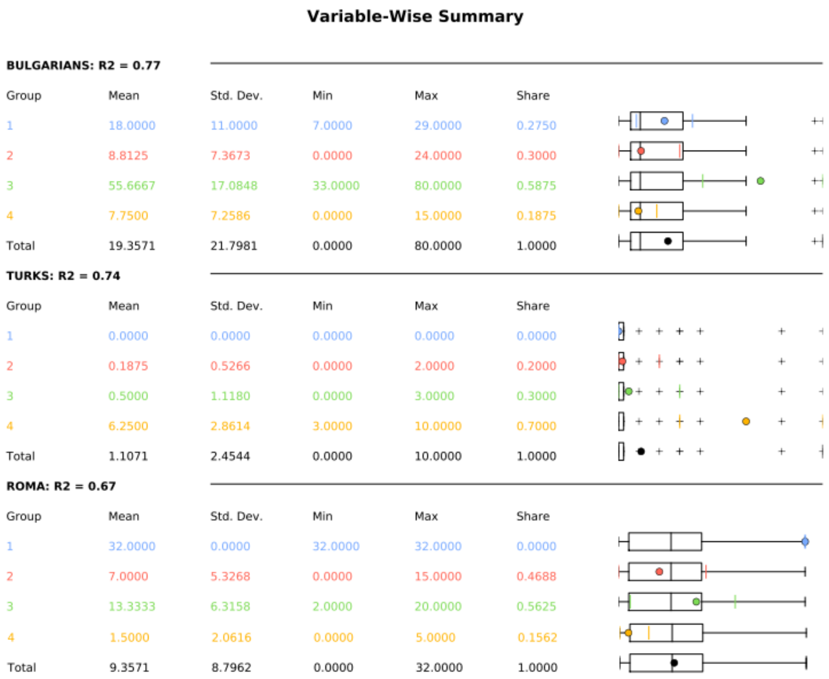
Group 4: Count = 4, Std. Distance = 8.0700, SSD = 6.0996

Figure 1. Analysis of data by groups

The general statistics Mean, Std. Dev., Min, Max, and R^2, marked in black, are calculated for the three methods of grouping of the districts for the three variables. They are used to compare with the statistics calculated for each group in order to determine in which districts the data for the three variables (number of patients from the three ethnic groups) are similar.

From the summary statistics of the whole dataset and their visualization in the Boxplot graphs, it is clear that the data for the Roma ethnic group are of a normal distribution, since the median (9.0) and the mean value (9.3571) are close, which is confirmed by the coefficient asymmetry (skewness) 1.0354. The median is indicated by a vertical black line in Boxplot and the mean value by a dot. For the other two datasets, Bulgarians and Turks, results show an uneven distribution with asymmetry coefficients of 1.4923 and 2.5138, respectively. For the same variables, the + signs indicate large differences in the values (outliers) in the data set.

The differences in the distribution of the three datasets are also judged by comparing their medians in each Boxplot. It is seen that they vary by location, but still for the variables Bulgarians and Roma, the data have a common range.



Variable-Wise Summary

BULGARIANS: R2 = 0.77

Group	Mean	Std. Dev.	Min	Max	Share
1	18.0000	11.0000	7.0000	29.0000	0.2750
2	8.8125	7.3673	0.0000	24.0000	0.3000
3	55.6667	17.0848	33.0000	80.0000	0.5875
4	7.7500	7.2586	0.0000	15.0000	0.1875
Total	19.3571	21.7981	0.0000	80.0000	1.0000

TURKS: R2 = 0.74

Group	Mean	Std. Dev.	Min	Max	Share
1	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.1875	0.5266	0.0000	2.0000	0.2000
3	0.5000	1.1180	0.0000	3.0000	0.3000
4	6.2500	2.8614	3.0000	10.0000	0.7000
Total	1.1071	2.4544	0.0000	10.0000	1.0000

ROMA: R2 = 0.67

Group	Mean	Std. Dev.	Min	Max	Share
1	32.0000	0.0000	32.0000	32.0000	0.0000
2	7.0000	5.3268	0.0000	15.0000	0.4688
3	13.3333	6.3158	2.0000	20.0000	0.5625
4	1.5000	2.0616	0.0000	5.0000	0.1562
Total	9.3571	8.7962	0.0000	32.0000	1.0000

Figure 2 Analysis of data by variables

- Results of the grouping in two groups

In the first group fall 9 of 28 districts, and in the secoand group -19. In this case there is no similarity simultaneously of all three variables. In the second (red) group, the color points of only the two variables (Bulgarians and

Roma) fall in the black rectangles (total data range). In the first (blue) group, the district with the Turks variable are the most similar.
From the calculated statistics for the general dataset we can see that the highest value of $R^2 = 0.56$ is for the ”Bulgarians” variable. This indicates that for this variable, the districts are divided into groups most effectively.

- Results of the grouping in three groups

Similarly to the above considerations, the data are analyzed when grouped into three groups. In the first a group fall 15 out of 28 districts, in the second 4 and in the third group 9. In this case, the greatest similarity of the data for all three variables simultaneously is in the first (blue) group. The analysis by groups shows that the colored dots fall into the outlines of all three Boxplot boxes. From the analysis of variables, the claim of the highest similarity in the first group is confirmed.
The mean for the three variables in the first (blue) group are closest to their respective global values. For example, for the Roma variable, the mean of the group is 9.8667, and for the whole dataset it is 9.3571. In the second and third groups, these values differ significantly. For the same variable, we have the largest value for R^2= 0.80, which is an indicator of the greatest similarity of the data in this group in terms of the variable Roma.

- Results of the grouping in four groups

When grouped into four groups, the districts with similar characteristics are allocated as follows. In the first group there are 2 districts, in the second group 16, in the third 6 and in the fourth 4. From the analysis by groups (Figure 1) it is obvious that in this distribution the districts into the second (red) group are the most similar for the three variables. The data similarity indicator with the highest value is R^2= 0.77 for the ”Bulgarians” variable.

- Comparative analysis of the three ways of grouping

In order to properly analyze the distribution of the data, we need to evaluate which of the three clustering methods is most effective. To do this, we will compare the values of R^2 for the three variables in two, three and four groups.

As it is known, the value of R^2 indicates how much of the original data is retained after the grouping process. Table 1 shows that when grouped into two and three groups, the R^2 values differ significantly for the three variables. When grouped into four groups, in addition to the high R^2 values, they are observed to be almost identical for the three variables. This allows us to interpret the distribution of the data for this variant more accurately (in four groups).

Table 1. Values of R^2

Variables	Number of groups		
	2	3	4
Bulgarians	0.56	0.29	0.77
Roma	0.42	0.80	0.67
Turks	0.05	0.29	0.74

The analyzed data relates to all forms of hereditary peripheral neuropathies. The first group includes only two districts whose data are with mean for the group (18.0) close to the global mean value (19.35). We have similarities in the data of the Bulgarian and Turkish ethnic groups.

In the second group, the most numerous, are districts with similarity of data for all three ethnic groups. They fall within the range of the average number of affected individuals.

The third group includes districts with the highest number of patients, mainly from Bulgarian and Roma ethnic groups. The fourth group is dominated by data from the Turkish and Bulgarian ethnic groups.

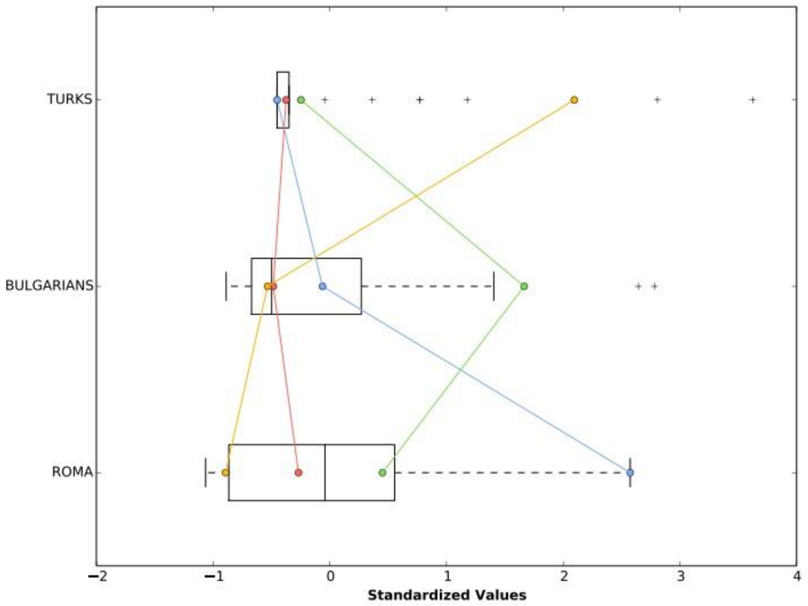


Figure 3 Parallel Box Plot

The parallel box plot graph (Figure 3) best summarizes the data, both the groups and the variables within them. The distribution of the high, medium and low values for the individual variables is more clearly distinguished here.

In the first (blue) group are the districts with the lowest number of sick patients from the Turkish ethnic group, the average values from the Bulgarian and the highest values for the Roma.

In the second (red) group, the number of diseased patients had averages for all three ethnicities, which is similar to the data for all three variables.

The third (green) group includes the districts with the highest number of affected for all three ethnic groups, but more Bulgarians.

The fourth (yellow) group reflects the districts with the highest values of the Turkish ethnic group, the average for the Bulgarians and the smallest number of affected are from the Roma ethnic group.

From the presented grouping options is obvious that the most accurate analysis is performed by grouping into four groups. This proves the importance of knowing well the data that we are analyzing in order to select the number of groups that will give us an insight of the data. In this case it is visible from the parallel boxplot chart and the compiled map that the districts colored in red have the most balanced distribution of patients in terms of their ethnicity. In the remaining districts there is a disproprotion of the affected individuals where in the yellow districts the most patients are from Turk ethnicity, in the blue from Roma ethnicity and in green from Bulgarian ethnicity.

4 CONCLUSIONS

The presented method of analyzing datasets gives us the opportunity to compile a more detailed map (Figure. 4) of the ethnic distribution of the affected individuals with hereditary peripheral neuropathies in the different districts of the country.

5 REFERENCES

1. Kastreva, K., 2018. *Development of a clinical database for phenotyping of hereditary neurological diseases*. Doctoral dissertation. Medical University Sofia. Available at:

<https://ras.nacid.bg/api/reg/FilesStorage?key=0325f27d-79cd-4d00-bf43-672541bbd512&mimeType=application/pdf&fileName=Reviewed_disertacia-04-2018.pdf&dbId=1> and <https://ras.nacid.bg/api/reg/FilesStorage?key=6d3883d6-0178-4d8b-ad66-bf11251f18a0&mimeType=application/pdf&fileName=avtoreferat_final.pdf&dbId=1 [accessed 7 May2020]

2. Kastreva, P., Kastreva, K. 2019. Classification of statistical data in GIS medium. *Geodesy, Cartography and Land Management Magazine*. Issue 5/6, 2019, pp. 20-24. ISSN 0324 – 1610

3. Slocum, T. A., McMaster, R. B., Kessler, F. C., & Howard, H. H. 2009. *Thematic cartography and geovisualization.* Pearson Education Canada, Inc., Toronto.p 518

4. Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar. 2019. *Introduction to Data Mining. Cluster Analysis: Basic Concepts and Algorithms.* 2nd edition.Publisher: Pearson. eText ISBN: 9780134080284, 0134080289 Available at: <http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf [accessed 7.05.2020]

5. ESRI. How Grouping Analysis works. Available at: <https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/how-grouping-analysis-works.htm [accessed 7.05.2020]

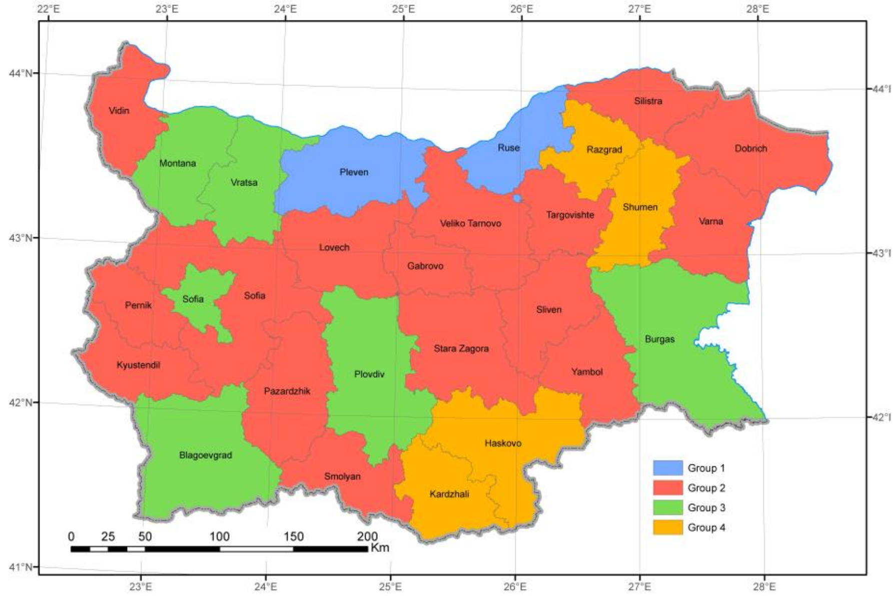


Figure4. Distribution of patients with hereditary peripheral neuropathies by group