GEODESY AND CARTOGRAPHY
ISSN 2029-6991 / eISSN 2029-7009
2015 VOLUME 41(1): 1–8
http://dx.doi.org/10.3846/20296991.2015.1024457

UDK 528.94

AUTOMATIC DATA CLASSIFICATION BASED ON THE TRIANGULAR GRAPH FOR THEMATIC MAPS

Zdena DOBESOVA

Department of Geoinformatics, Faculty of Science, Palacký University, 17. listopadu 50, Olomouc, Czech Republic,

E-mail: zdena.dobesova@upol.cz

Received 29 October 2014; accepted 9 March 2015

Abstract. A triangular point graph helps in the process of data classification for a thematic map. A triangular graph can be used for a situation that is described by three variables. The total sum of variables is 100%. The proportion of three variables is plotted in an equilateral triangular graph where each side represents a coordinate for one variable. A triangular graph displays the proportions of the three variables. The position of the point indicates the type (class) of the situation in the triangular graph. The typology of the situation can be subsequently expressed in the map.

We have created a "Triangular Graph" program which represents a helpful automatic tool for ArcGIS software. This new program classifies input data based on a triangular graph. It is realized by two python scripts located in a custom toolbox as two programs. The first program calculates X and Y coordinates in an equilateral triangular graph. The second program compares plotted points and suggested zones of a division produced by the first program. Finally, a new attribute is added to the source data. The user can create a new thematic map, based on that attribute in order to express the typology of the given situation.

The programming language Python and essential module ArcPy have been used for solving these tasks. To test the created programs several maps were made, based on the classification often used in demography. For example, the new program helped to create a sample map of age categories in districts of the Czech Republic.

The program is available to download from the Esri web pages and web pages of the Department of Geoinformatics, Palacký University Olomouc.

Keywords: cartography, thematic map, classification, triangular graph, ArcGIS, Python.

Introduction

A triangular graph is a graphic based classification method in which data are plotted into a base equilateral triangle. The triangle shows the composition of a situation that is based on three variables. This classification method has been used in geology (mineralogy and petrology), where it precisely identifies soils, rocks and minerals as well as their physical characteristics. For example, geologists determine types of bottom sediments to silt, clay, sand, silty-sand, etc. according to a triangular graph – see Shepard's Diagram (Fig. 1) (Shepard 1954).

The above classification method by a triangular graph was later used to classify data in regional geography. Several thematic typological maps based on

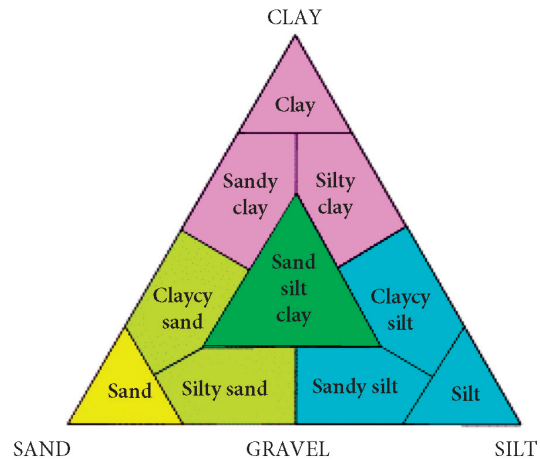

Fig. 1. Shepard's diagram of sediments classification (USGS 2000)

a triangle graph can be found in the "Population Atlas of Slovakia" (2006) (Fig. 2). Typical geographical applications include typologies of districts (towns, regions) according to:

- Typology of age groups of the population. Figure 2 shows three categories of age: 0–14, 15–60, 61+ years.
- Typology of employment sectors (primary, secondary, tertiary sector).
- Typology of level of education (basic education, secondary education and high education). It classifies the regions of the country by structure of scholarship.

Three variable data are used by the triangular graph classification method (examples mentioned above). If three variable data are not available, it is necessary to aggregate the source data to three main groups (of course, it must be logically correct) (Vozenilek *et al.* 2011). It is also possible to create two natural groups of variables and a third group which aggregates the other ones into one group (Dobesova 2014). This case is the classification to the groups according to nationalities in north-east villages in Silesia in the Czech Republic. There are two main nationalities (Czechs and Slovaks) and many other nationalities that are less numerous. Czechs and Slovaks are two natural groups. Remains nationalities are aggregated to the third group in the triangular graph classification method.

1. Triangular point graph

The term "triangular graph" is not the only one used to describe this method. The other terms include ternary graph, ternary/triangle plot, de Finetti diagram or Ossan triangle. The word "ternary" is derived from the Latin adjective "ternarius", i.e. having three parts. The name Ossan comes from the author who used this method for the first time (Kanok 1992). The use of a given term also depends on the discipline in which it is being used. For example, the term "ternary plot" often appears in geology, while the Ossan graph is used in demography.

1.1. Construction of the triangular graph

In an equilateral triangle used for classification, three sides serve as scales for three variables. The source variables are recalculated to the percentage of the total sum. The total sum of the variables is 100%. The proportion of the three variables is plotted in an equilateral triangular graph where each side has a scale for one variable from 0% to 100% (Fig. 3). The resulting point is the point of intersection of three lines in the graph. A triangular graph displays the proportions of three variables of which the sum remains the same (100%). For example, points in Figure 3 have the following partial variables:

- Point A (5, 30, 65),
- Point B (33, 25, 42),
- Point C (45, 45, 10).

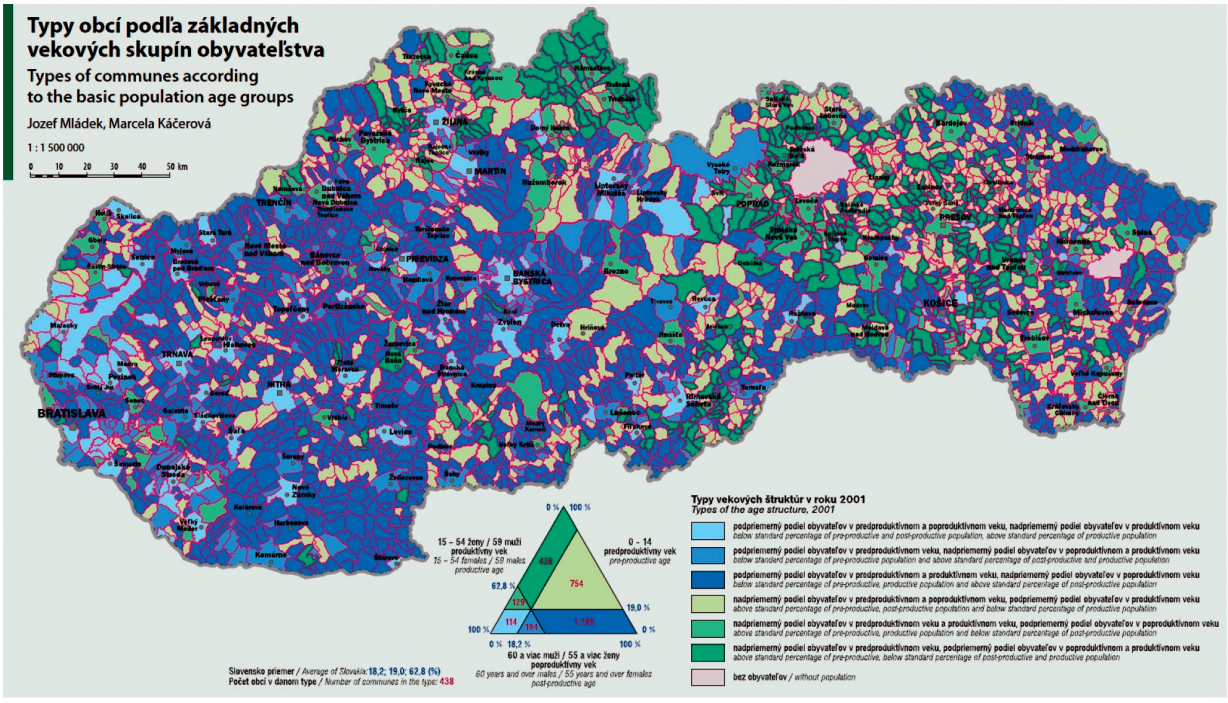

Fig. 2. Types of communes according to the basic population age groups (Population Atlas of Slovakia 2006)

When all points are plotted onto the diagram, a classification in zones can be determined. Smaller parts (triangles, rhombuses) fully fit into the base triangle. In some cases, the division is symmetrical (Fig. 1); in other cases the division is asymmetrical (legend in Fig. 2). The number of inner parts varies. The system of division and number of parts cannot be defined in advance because it depends on the character of the classified situation. The specialist must determine the proper type of division and number of inner parts. The system of division affects the final classification of the situation by grouping similar objects (e.g. regions) into respective groups.

1.2. Accessible program solutions

The geological software RockWorks contains a utility for the calculation of ternary diagrams (RockWare 2013). GIS (geographic information system) software, on the other hand, does not contain this function.

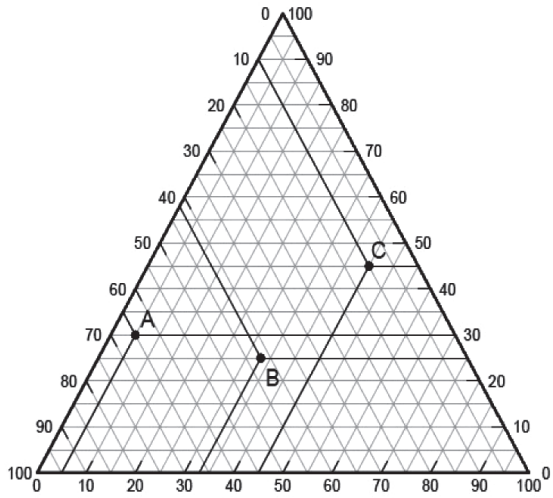

Fig. 3. Base triangle and three example points A, B, C

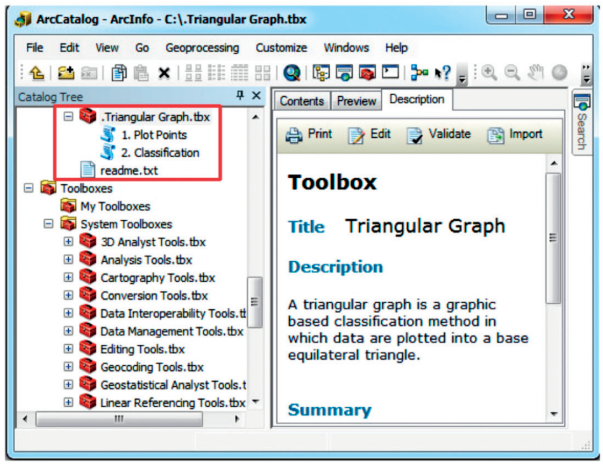

Fig. 4. Custom toolbox in Arc GIS for Desktop

When authors of thematic maps need to classify their source data based on a triangular graph, they can use Microsoft Excel sheets. Some MS Excel sheets with the predefined function for the calculation of positions in a triangular graph are freely available on the Internet. One of these sheets is the "Template for triangular diagrams" created by Aqueous Solutions Aps. (2013). Another Excel sheet can be found at the pages created by Vaughan (2011). The use of the Excel utility is a partial solution at best. When data are put in the chosen Excel sheet, only X, Y positions in the triangle are calculated and displayed. After that, the necessary classification has to be processed manually outside of MS Excel. This method of classification and creation of typological maps is not comfortable because a given Excel sheet is helpful only in the first stage of map creation.

GIS software is very often used for the creation of thematic maps. The cartographic functionality of GIS software is quite satisfactory, but some special functions for thematic mapping are missing. This deficiency was apparent from the results of research performed in the field of evaluation of cartographic functionality, in which more than ten GIS products were assessed by the CartoEvaluation method (Dobesova 2013, Evaluation of Cartography Functionality in GIS Software 2008). The result of the assessment indicates that the function for the automatic creation of triangular graphs is missing in the GIS products evaluated. Therefore, there was an opportunity to extend the software ArcGIS by a custom utility for the automatic creation of the typology based on a triangular graph. Custom utilities are a frequent solution of extending the basic functionality of GIS. The custom utility for the creation of "chart maps" in ArcGIS software is discussed in (Dobesova, Valent 2011).

2. "Triangular Graph" program

The program presented here is realised as an Esri custom toolbox under the title of "Triangular Graph". The author of toolbox is student S. Ganbaatar who created it as a part of bachelor theses (2013). This custom toolbox contains two programs for creating triangular diagrams and classification: "1. Plot Points" and "2. Classification" (Fig. 4). The utilization and functionality of both programs are explained in the following sections.

2.1. Program "1. Plot Points"

The first program processes input data. The format of input data is Esri shapefile or feature classes in

a geodatabase. The attribute table must contain three fields (columns) with input numeric data (Fig. 5). The names of columns are filled by the user into the following boxes: First part, Second part and Third part (Fig. 6). The program reads user input data and calculates a percentage value of three variable structures. After that, the program transforms percentage values to X and Y coordinates in an equilateral triangular graph. This first program also automatically generates a base triangle and lines that indicate average percentage values for further usage (Fig. 7). The geometry of the base triangle, points with X, Y coordinates and "average lines" are stored in new separate polygon, point and line shapefiles. The user can choose the names of the files. The origin of these geometries is 0, 0. The newly generated shapefiles do not have a geographical reference. A standalone second data frame is used to display these geometries in the Layout Window in ArcGIS for Desktop. A data frame with a triangular graph can be used as legend in a map layout. Data of input geometry (polygons of regions, districts) are displayed in the first base data frame. The visualization of the triangular diagram (color, size) is fully customizable as any feature based on shapefile (or feature class).

The next step is the creation of a polygon layer that contains zones of categories in the triangle. The composition of zones is the user's choice. It is a crucial step because it could hugely affect the output result. First of all, the users have to make sure that they know a lot about the classified situation (phenomenon).

There are three ways to determine these zones:

- To choose predefined symmetrical *zones from templates*. Templates are available in the template directory. There is one template with four zones and another one with 15 zones.
- To create your own *zones based on auxiliary lines*. The lines are optionally generated in the first program. Since the lines only have a snapping purpose, it is necessary to create a new extra polygon layer of zones.
- To create *own zones*. A new extra layer is drawn by the users according to their knowledge of the situation.

Each partial zone has a unique numeric code in the attribute table. The value is a code of the class in the classification. When the users create own zones (the third method), they must fill in the codes manually.

Count_Inhab1	Count_Inhab2	Count_Inhab3
1214174	1169106	1272690
91619	90625	95157
75859	75684	85500
150625	151360	159194
91637	89379	96583
75250	73628	74289
94241	94635	104511
110664	113241	124231
81247	82804	93870
101923	105541	149294
74265	82404	122759
112007	110685	113430
53545	52487	55632
173765	178506	186322
57388	59569	61945
93048	92887	92749
74614	72984	72569
71747	70300	70661
50985	51369	51313

Fig. 5. Example of data

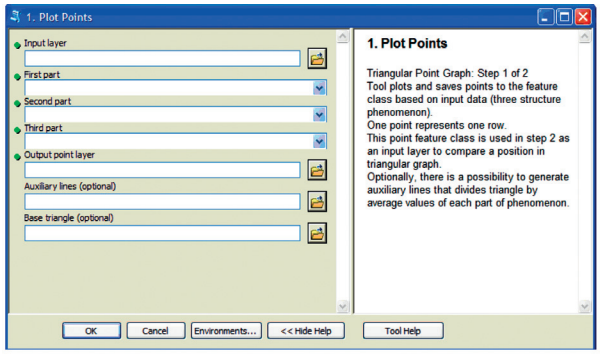

Fig. 6. User interface of the first program "1. Plot Points"

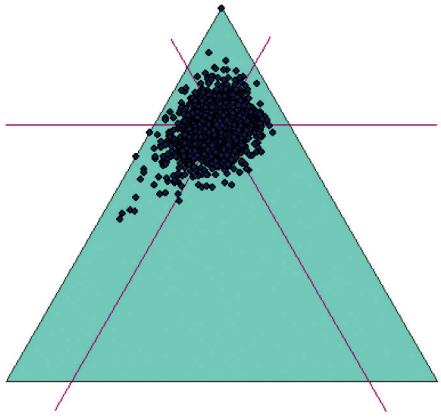

Fig. 7. Output geometries (triangular graph, "average" lines, points) from the first program

2.2. Program "2. Classification"

The second program compares plotted points from the first program and suggested zones in the triangular diagram. The result of this comparison (overlay analysis) produces a new column (attribute) added to the original input data as a classified category. The names of layers and the name of the output attribute

are the program parameters. The new attribute allows to create a thematic map based on the performed classification.

Both programs contain step by step help in the user interface on the right hand side (Figs 5, 8). Each input box is explained. The programming language Python version 2.6 and the ArcPy module were used. Programs were tested in software ArcGIS version 10 and 10.1. The error states are prevented and announced in the Result window. The custom toolbox with programs, templates for zones and Readme file are free downloadable from the Esri web page: http://www.arcgis.com/home/item.html?id=661a8e7c463a4bd2b529f01221efa8f2

3. The algorithm

The algorithm of program "Triangular Graph" is depicted in this section. In addition, some crucial ArcGIS geoprocessor methods are mentioned. The first programs "1. Plot Points" is firstly responsible for input data. The method arcpy.GetParameterAsText() is used. The important following method is arcpy.Point(), which creates a new point instance. The method arcpy.SearchCursor() maintained the reading records from attribute table. Following recalculation counts summaries for each attribute and percentage values for each attribute in rows. The coordinates X and Y of each point in the triangle are calculated by simple equations for position in the triangle (1). Coordinates of points are subsequently stored to the Python list that is created by method arcpy.Multipoint(). Points are added by method Append.

$$\begin{aligned} \text{coordX} &= \text{A1} + (\text{A2} * 0.5), \\ \text{coordY} &= \text{A2} * 86.6 / 100, \end{aligned} \tag{1}$$

where: coordX, coordY are coordinates; A1, A2 are percentage value of the attributes.

Next part creates the basic triangle. The three lines of triangle have absolute coordinates of end points:

[0, 0; 100, 0], [100, 0; 50, 86.6], [50, 86.6; 0, 0].

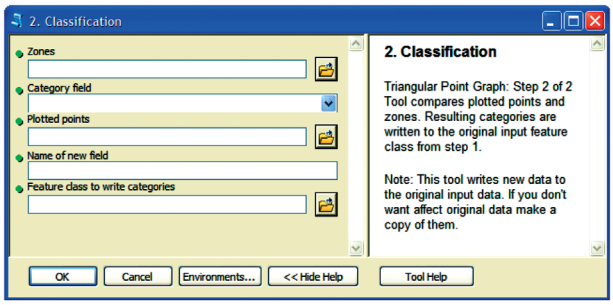

Fig. 8. User interface of the second program "2. Classification"

Next part of the program is optional. It calculates coordinates of three average lines. Firstly, the average value is calculated for each attribute (AVG1, AVG2, AVG3). The two coordinates Xn and Yn for each average line are calculated according to equation (2).

$$\begin{aligned} \text{X1} &= 0, \\ \text{Y1} &= \text{AVG2} * \sin(60), \\ \text{X2} &= \text{AVG1} * \sin(60) * \sin(60), \\ \text{Y2} &= \text{AVG1} * \sin(60) * \sin(30), \\ \text{X3} &= \text{AVG3} * \sin(60) * \sin(60), \\ \text{Y3} &= \text{AVG3} * \sin(60) * \sin(30), \end{aligned} \tag{2}$$

where: AVG1, AVG2, AVG3 are average values.

End points of average lines are:

Line 1: [X1, Y1], [100, Y1],
Line 2: [X2, X1 – Y2], [X2 + 50, X1 – Y2 + 86.6],
Line 3: [100 - X3, X1 - Y3], [100-X3-50, X1-Y3 + 86.6].

The method arcpy.Point() creates the point instance for storing coordinates for three average lines. Method arcpy.Polyline() creates lines from point list. The lines are subsequently stored in a new line feature class that is created by method arcpy.CopyFeatures_management(). To store back all data is necessary use the cursor method arcpy.UpdateCursor().

The second program "2. Classification" read the input parameters by method arcpy.GetParameterAsText(). The first is a test if the points in the triangle are within predefined zones. The method point.within() is used. New field for storing category is added to the polygon data with zones by method arcpy.AddField_management(). Categories are numbered from 0 to N (N is a count of categories). Subsequently new field is added to the source data. Cursor methods setValue() and updateRow(row) arrange the storing of new data about classified categories to the source data.

The source Python code of the program is open source (Ganbaatar 2013). It is possible download it from the web pages mentioned in section 2.2.

4. Example maps

Two maps were drawn up to verify the utilization of the "Triangular Graph" program. The first of them is the map "Types of municipalities according to the economic activity of population" (Fig. 9), describing economic activity in three basic employment sectors: primary, secondary and tertiary. This is an example in which 400 municipalities were classified into 14 categories. The division into zones is symmetrical. The main cluster of points is between industrial-service and service-industrial types in the Olomouc Region (there is only one agricultural region there). Not all the categories are

displayed in the map. An asymmetrical division of the triangle would have been better. This example map is only one of a set of maps for all 14 regions in the Czech Republic. To compare all the regions in the Czech Republic using the same classification, it is necessary to select the same symmetrical division.

The second example was the map "Types of districts in the Czech Republic according to the basic population age groups" (Fig. 10). The brown lines, which indicate average percentage values for age groups, were used for the division into zones (legend of Fig. 10).

There are six zones with different areas of partial triangles and rhombuses. This division expresses the situation better than a symmetrical division of a triangle would illustrate. The final map shows the differences in age structure for respective districts.

Summary

The article presents a useful program for data classification based on a triangular graph. The advantage is that the process is automatic for various amounts of data. It is only necessary to set input and output data. There is no need for manual calculation. Moreover, the "Triangular Graph" program creates the geometry of the equilateral triangle that can be used to display classification and to insert as legend into the map composition. The design of division zones is supported by predefined templates of a division of the triangle. The setting of the same colors for the thematic map and the triangle can be easily based on predefined (or user) color ramps in ArcGIS for Desktop. First, the colors in triangular zones are assigned. Then, the same color

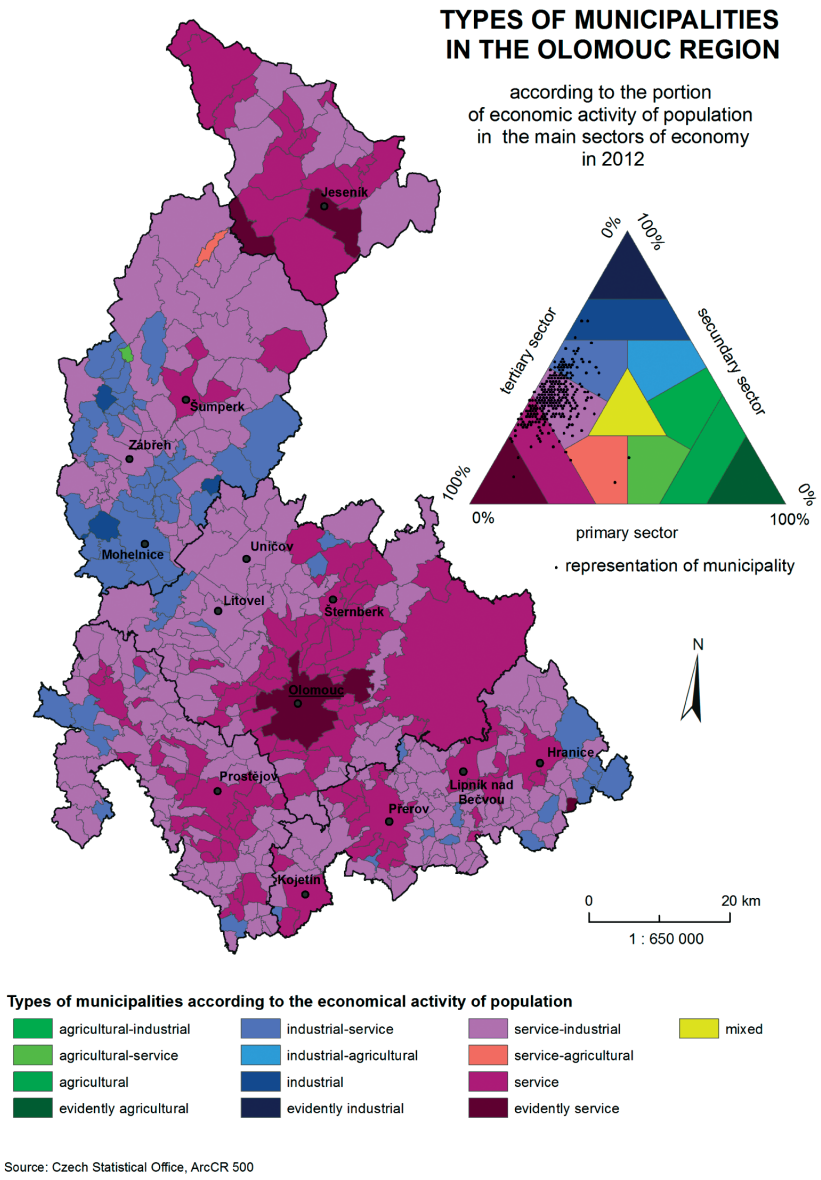

Fig. 9. Thematic map with classification based on economic activity of population (Ganbaatar 2013)

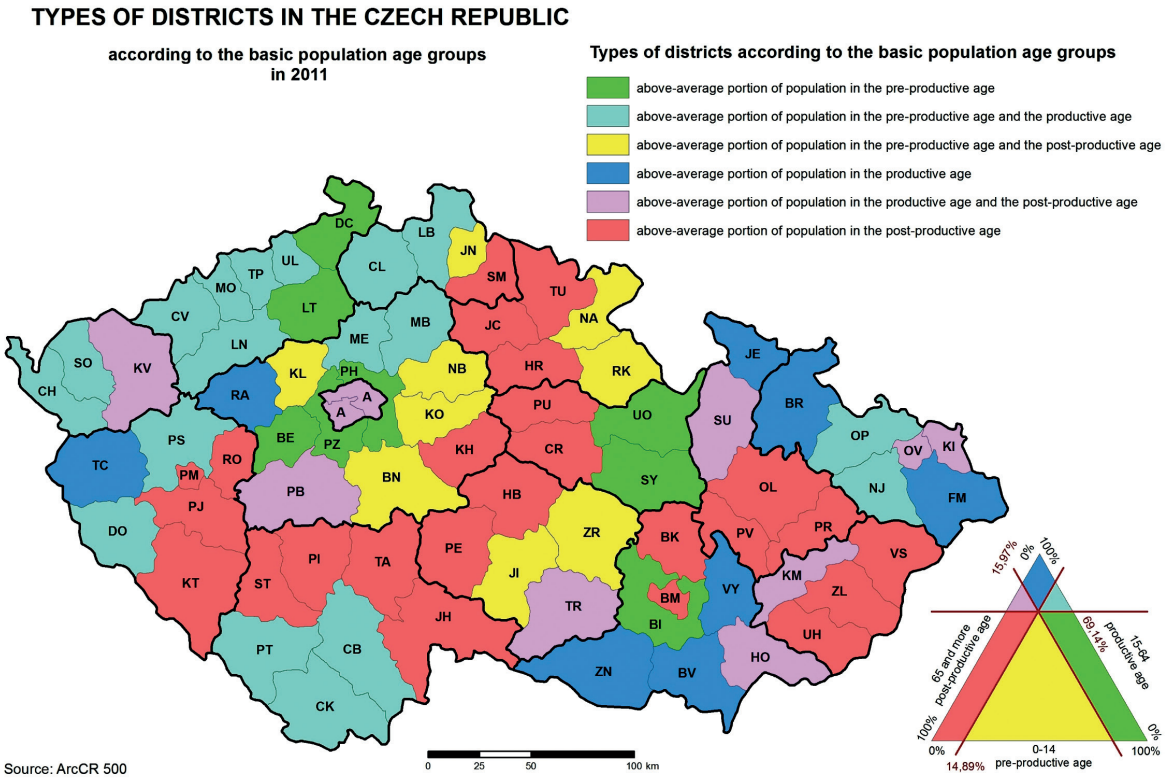

Fig. 10. Thematic map with classification based on the population age groups

scheme can be imported into the geographic layer. These steps easily guarantee that colors correspond in both the map and the legend. Two example maps document the functionality of the program.

The program code is written in the Python programming language. Both programs are adopted in the custom toolbox. The code is free to download from the Esri web pages. Users can read and modify the source program code or use it for inspiration. The code uses the base ArcObjects, and no other program extension is necessary. The advantage of the program also lies in the use of native data formats – Esri shapefile and feature class (Esri geodatabase).

The "Triangular Graph" program is expected to be of great help for cartographers. The utilization of this program should bring more thematic maps based on the triangular classification in atlases and map production. The program's ability to automatically create a classification of data provides an opportunity for more frequent use of this method by cartographers.

Acknowledgement

The author gratefully acknowledge the support by the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic).

References

Aqueous Solutions Aps. 2013. *Template for triangular diagrams in MS Excel* [online], [cited 12 December 2013]. Soborg, Denmark. Available from Internet: www.phasediagram.dk/triangular.xlsx

Dobesova, Z.; Valent, T. 2011. Program extension for diagram maps, *Geodesy and Cartography* 37(1): 22–28. Vilnius Gediminas Technical University (VGTU) Press Technika, Taylor & Francis. ISSN 2029-6991, eISSN 2029-7009. http://dx.doi.org/10.3846/13921541.2011.558330

Dobesova, Z. 2013. CartoEvaluation method for assessment of GIS software, *Geodesy and Cartography* 39(4): 164–170. Vilnius Gediminas Technical University (VGTU) Press Technika, Taylor & Francis Group. ISSN 2029-6991, eISSN 2029-7009. http://dx.doi.org/10.3846/20296991.2013.859824

Dobesova, Z. 2014. Problémy tvorby a použití typizace podle trojúhelníkoveho grafu (Problems in creation and utilization of the triangle graph in the process of classification), in *Aktivity v kartografii*. Kartografická spoločnosť Slovenskej republiky, Bratislava. 7–16. ISBN978-80-89060-23-8.

Evaluation of Cartography Functionality in GIS Software. 2008. [Online], [cited 12 December 2013]. Available from Internet: http://www.geoinformatics.upol.cz/app/visegrad

Ganbaatar, S. 2013. *Automatická typizace dat pomocí trojúhelníkového grafu* [Automatic classification of data by the triangular graph]: Bachelor theses. Palacký University, Olomouc. 44 p.

Kanok, J. 1992. Kvantitativní metody v kartografii – 1. Díl (Quantitative methods in cartography – 1st part). University of Ostrava, Ostrava. 224 p.

Population Atlas of Slovakia [Atlas obyvateľstva Slovenska]. 2006. Faculty of Science, Comenius University, Bratislava. 168 p.

RockWare. *RockWorks* [online], [cited 12 December 2013]. Available from Internet: http://www.rockware.com/.

Shepard, F. P. 1954. Nomenclature based on sand-silt-clay ratios, *Journal of Sedimentary Petrology* 24: 151–158.

USGS. 2000. Geological survey open-file report 00-358 [online], [cited 12 December 2013]. Available from Internet: http://pubs.usgs.gov/of/2000/of00-358/text/chapter1.htm

Vaughan, W. 2011. *Ternary plots* [online], [cited 12 December 2013]. Available from Internet: http://wvaughan.org/ternaryplots.html

Vozenilek, V.; Kanok, J.; Blaha, D.; Dobesova, Z.; Hudeček T.; Kozakova, M.; Nemcová Z. 2011. *Metody tematické kartografie: vizualizace prostorových jevů* [Methods of thematic cartography: visualisation of spatial phenomena]. Palacky University, Olomouc. 216 p. ISBN 978-80-244-2790-4.

Zdena DOBESOVA. Ing, PhD. She is a teacher and deputy head at the Department of Geoinformatics, Palacký University in Olomouc, Czech Republic. She holds a PhD (since 2007) degree from Technical University in Ostrava, Faculty of Mining and Geology. Research interest: GIS, spatial database, digital cartography, programming in Python for ArcGIS, visual languages, eye-tracking.