Creation of a GIS tool for the selection of a classification method, the statistical assessment of the method's class intervals and its implementation in choroplethic mapping

M. Aza[1], M. Papadopoulou[2], D. Voutsa[3], P. Lafazani[2]

1 Msc, Surveying Engineer, AUTH

2 Assoc. Professor,,Laboratory of Cadastre and GIS, School of Rural and Surveying Engineering, AUTH

3 Phd, Surveying Engineer, AUTH, Ministry of Economics

papmar@topo.auth.gr

Abstract: This paper describes a tool for the selection and implementation of a classification method for a choroplethic depiction of data. The tool integrates all the classification methods of ArcGIS 9.3 and adds to this list another method, which belongs to the category of systematically unequal stepped class limits. At the same time, it calculates both for the new method and for each of the pre-existing methods, the index Goodness of Variance Fit (GVF). Using this index, the class intervals are evaluated for their suitability as far as an effective choroplethic depiction is concerned. Finally, it designs the choropleth map, utilizing the ArcGIS software for the selection of colour ramp and the colouration of the map units.

1. Introduction

The choroplethic depiction (see Robinson et al., 1984) is a method which is very often selected by cartographers for the thematic visualization of data. The basis of this depiction, is the classification of the data, i.e. their grouping into classes. The first step of classification is the choice of a method for the determination of class limits and the second step is the selection of the number of classes. Regarding the number of classes, it determines how detailed the data mapping will be. Ideally, the cartographer uses the maximum number of classes that can be easily perceived by the users of the map.

Many classification methods have been developed. Some of them have common characteristics as far as their "philosophy" is concerned, and they can be considered as a group. The latter does not mean that they will have the same visual results on the choropleth map.

The choice of a method has troubled and continues to trouble cartographers. The classification of a data set into distinct groups is based on the distribution of the attribute values under consideration, and in such a way, that the values inside

classes are homogeneous and the classes, in turn, are different from each other.

This paper describes a tool for the selection and implementation of a method of classification. The tool, using a statistical index, simultaneously displays the statistical evaluation of the suitability of the class limits, for a reliable representation of the data. It runs under the environment of ArcGIS 9.3, incorporates all classification methods of the software and in addition gives the user the option of another method that uses arithmetic progressions (§3). The progressions are not found in commercial software and belong to the category of methods that generate systematically unequal intervals of classes (§2). The code of the tool is written in Visual Basic 6 and utilizes the ArcObjects (§5).

2. Methods of data classification

In the literature, many methods for data classification are found. Robinson et al. (1984) divide the methods into three major groups which are the following: a) constant sequences or equal intervals of classes, b) systematically unequal stepped class limits and c) irregular intervals of class limits. Evans (1977), in a different categorization, proposed another classification of two levels. Initially, he distinguishes four different approaches: the exogenous, the arbitrary, the idiographic and the serial. Then, from these approaches sixteen class interval systems result. Slocum et al., (2005), reported a total of six methods for data classification without distinguishing groups and subgroups. Two other methods, not mentioned in the previous categorizations, but widely used are: a) the maximum breaks and b) the natural breaks (see e.g. Slocum et al., 1995). In the following paragraphs only the method of numerical progressions is described in detail, since it is new in the classification list of ArcGIS. This method belongs to the second group according to the categorization of Robinson et al. (1984).

3. Systematically Unequal Stepped Class Limits

This technique of determining the limits of classes applies only on data in ratio level or in interval level (see e.g. Papadopoulou, 2009). The method makes the limits of class intervals systematically lower towards the higher or lower end of the scale, along which the data are arranged in ascending order. The choice between the higher or lower end, depends on which part of the distribution is to be emphasized.

Two groups of sequences with unequal steps are: a) the arithmetic progressions, and b) the geometrical progressions (Robinson et al., 1984). The general form of the equation that generates the class limits for both types is:

$$L + B_1X + B_2X + \ldots + B_nX = L+\sum B_iX = H \quad (1)$$

where

L = the lowest value of the data
H = the highest value of the data
B_i = the value of the ith term in the sequence (for i from 1 to n)
n = the number of classes
X = the unknown of the equation

Such methods require only the determination of B_i. Then the equation is solved for X, for any given value of L and H, and the class limits are determined. When X is calculated, the limits of classes result as follows:

1^{st} class: L to $L+B_1X$
2^{nd} class: $(L+B_1X)$ to $(L+B_1X+B_2X)$
3^{rd} class: $(L+B_1X+B_2X)$ to $(L+B_1X+B_2X+B_3X)$
.....
n^{th} class: $(L+B_1X+B_2X+B_3X+.....+B_{n-1}X)$ to H

These methods are implemented effectively in J-shaped data distributions (see. Spiegel and Stephens, 1999) i.e., in data that show a peak at the end or at the beginning of their distribution. The methods are not included in the lists of commercial GIS software, as their logic and the way of their implementation are not simple and make them less "popular" than others, like the natural breaks, the quantiles, the equal intervals (see e.g. Robinson et al ., 1984, Slocum et al., 1995) etc.

3.1 Arithmetic progressions

In arithmetic progressions each class is separated from the next by a numerical difference. Thus, the amount of B_i is obtained from the relationship:

$$B_i = a + [(i - 1)d] \quad (2)$$

where

a = the value of the first term of the sequence
i = the number of the term calculated (the first, second, etc.)
d = the specified difference

The numerical progressions can take any of the following six types depending on d. The types of sequences with indicative values of a and d are:

1. Ascending with constant rate (a = 1, d = 1)
2. Ascending with increasing rate (a = 1, d = i–1)
3. Ascending with decreasing rate (a = 1, d = 1 / i)
4. Descending with constant rate (a = n, d = –1)
5. Descending with increasing rate (a = 10, d = – (i–1)
6. Descending with decreasing rate (a = 1, d = –1 / i)

According to Aza (2012), for the arithmetic progressions can be said that:

a) Because the data values are in ascending numerical order the ΣBi should be positive, so that the unknown X to result positive.

b) The first term (a) is chosen arbitrarily. It usually equals to 1. However, in descending sequences a number greater than 1 should be chosen so that the ΣBi will not result negative.

c) The difference (d) is a simple constant number (usually but not necessarily integer) when the sequence (ascending or descending) has a constant rate.

d) In progressions with increasing or decreasing rate, difference (d) is increased or decreased, respectively. So, in these forms the difference d can be calculated in relation to the number of classes n. In the case of increasing rate, d is > 0 and of decreasing rate d is <0.

e) From the above, the difference d can be written:

$$d = m * n^l + k \qquad (3)$$

where m, l and k are real numbers (usually integers <=> 0).

f) When $l > 0$ then the progression has an increasing rate.
When $l = 0$ then the progression has a constant rate.
When $l < 0$ then the progression has a decreasing rate.

g) For $m > 0$ the progression is ascending.
For $m < 0$ the progression is descending.
For $m = 0$ and $k \neq 0$ progression has a constant rate.

Below, two examples of calculation of X for an increasing arithmetic progression with a constant rate and for a decreasing progression with a constant rate are given. The number of classes is n = 4 in both cases.

Increasing with a constant rate:

$a = 1,\ d = 1,\ n = 4$

$B_1 = 1 + (0)1 = 1$

$B_2 = 1 + (1)1 = 2$

$B_3 = 1 + (2)1 = 3$

$B_4 = 1 + (3)1 = 4$

$L + \sum_i B_i X = H \rightarrow 1{,}66 + 10X = 98{,}78$ and $X = 9{,}712$

The limits are:

1,66 (L) to 11,37 ($L + B_1X$)

11,37 ($L + B_1X$) to 30,79 ($L + B_1X + B_2X$)

30,79 ($L + B_1X + B_2X$) to 59,92 ($L + B_1X + B_2X + B_3X$)

59,92 ($L + B_1X + B_2X + B_3X$) to 98,78 (H)

Decreasing with a constant rate

$a = 4,\ d = -1,\ n = 4$

$B_1 = 4 + (0)(-1) = 4$

$B_2 = 4 + (1)(-1) = 3$

$B_3 = 4 + (2)(-1) = 2$

$B_4 = 4 + (3)(-1) = 1$

$L + \sum_i B_i X = H \rightarrow 1{,}66 + 10X = 98{,}78$ and $X = 9{,}712$

The limits are:

1,66 to 40,51

40,51 to 69,64

69,64 to 89,07

89,07 to 98,78

4. Iterative techniques and the GVF criterion

The iterative techniques belong to the category of classification methods with irregular intervals limits of classes and are based on a rational statistical criterion. Data are grouped after iterations, so that the specified criterion is met. Jenks (1967) first introduced systems for determining the limits of classes using iterative techniques with statistical criteria, based purely on the theory of cartography.

One of the criteria is the Goodness of Variance Fit, (GVF), and is particularly useful when the cartographer wants to minimize the squared deviations associated with the mean values of classes (Aza, 2012).

The criterion to be satisfied is the maximizing of GVF where:

$$GVF = \frac{SDAM - SDCM}{SDAM} \tag{4}$$

The SDAM (Squared Deviations Array Mean) is the sum of squared deviations of each observation (x_i) from the mean value ($\overline{X}$) of the observations (array mean) and is expressed by:

$$SDAM = \sum (x_i - \overline{X})^2 \tag{5}$$

The SDCM (Squared Deviations Class Means) is the sum of squared deviations each observation x_i from the mean ($\overline{Z}_c$) of the class to which x_i belongs and is expressed by:

$$SDCM = \sum\sum (x_i - \overline{Z}_c)^2 \tag{6}$$

The difference between SDAM and SDCM is the sum of squared deviations among classes.

In the iterative technique using this criterion, the cartographer must first define an arbitrary grouping of the arranged in ascending order data, and with this grouping to calculate the GVF. The next step is the movement of observations from one class to another in an effort to minimize the sum of SDCM and thus to increase the GVF. After moving observations among classes the new values for the quantities SDCM and GVF are calculated. This process is repeated until the GVF can no longer be increased (Robinson et al.,1984).

A choropleth map in which each observation value is a separate class, is considered not to contain errors. In this case, the mean value of the class is the same as the single value of the class and thus the square of the deviation equals to 0, the quantity SDCM is also 0 and the GVF is equal to 1. So, the maximum value that the GVF can have is 1. In choroplethic mapping the data are always grouped into classes. Therefore the value of the GVF is less than 1 and the optimal solution would be the GVF to tend as much as possible to 1 (Dent, 1999).

Besides contributing to the determination of classes, the GVF index can be used to assess all other classification methods and to ascertain whether the limits of the generated classes are suitable for depicting the data. According to this logic, in the classification tool, the GVF was used as a criterion of adequacy of the class limits for all the embedded in the tool methods.

5. The Classification tool

The ArcGIS software gives the programmer the ability to create from simple applications with adjustment of the menus and toolbars to specialized applications, outside of its graphic environment. This can be achieved by using programming objects through which all the functionality of ArcGIS is exploited by object-oriented programming languages (Voutsa, 2010). These objects are called ArcObjects (see. e.g. Zeiler, 2001, Burke, 2003, Chang, 2005). The code of the tool is written in Visual Basic 6 and utilizes the capabilities of ArcObjects. The tool was created under the graphic environment of ArcGIS 9.3 and incorporated the classification methods and the choice of colours available in the ArcMap software (see e.g. Booth and Mitchell, 1999).

The basic elements in the tool are:

1. Selection of the layer on which classification will be applied.
2. Selection of the field of the layer whose values will be processed
3. Selection of the number of classes
4. Selection of the classification method

5. Selection of a colour ramp that will be applied on the spatial units of the map for all the classes.

For the implementation of the progression method a window was build, through which the user can select the first term of the sequence (a) and the difference (d) depending on the type of the arithmetic progression.

Other features offered by the tool are the maximum, the minimum and the mean value of the data set, as well as a frequency histogramme. The histogramme depicts the way the data are distributed along the number line and thereby enables the user to select a proper classification method. Apart from that, the value of GVF is calculated. If the GVF is not close to 1, the user can change the method and/or the number of classes to improve this value.

Initially, the ArcMap environment is opened. The tool is located in the toolbox "Classification". Selecting the button "Classification Tool" the main dialog window appears (Figure 1). In this window all the essential elements mentioned above are contained. With the appearance of the window the user can choose the layer on which classification will be applied and the field of the layer that contains the values to be grouped. After the selection of the field, the statistics and the frequency histogramme are displayed and the number of classes and the classification method are selected. When these actions are done, the maximum values of each class are listed in the box called "Break Values " and the value of GVF appears in the box called “GVF". Then the choice of a color ramp takes place and pressing the button “OK” the choropleth map is displayed.

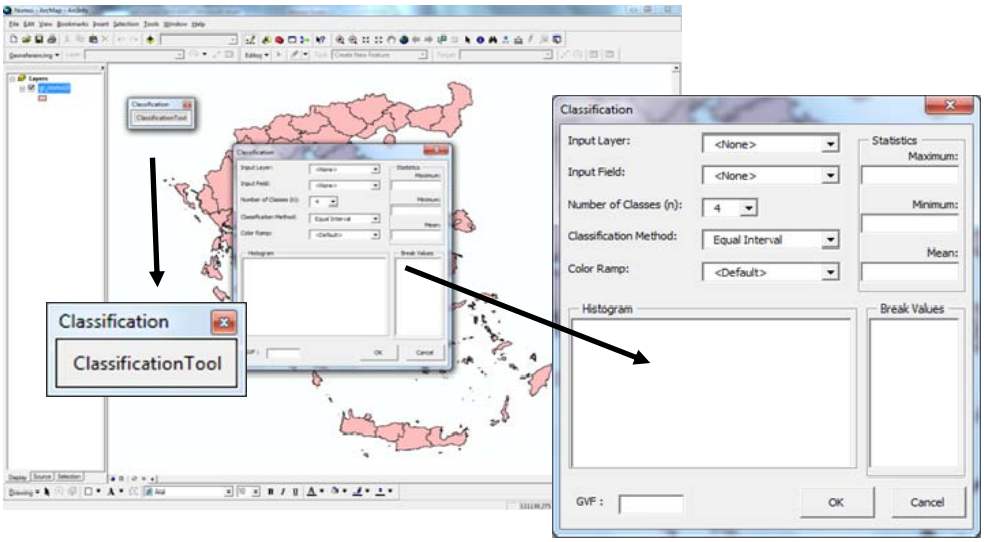

Figure 1. *The classification tool in the ArcMap environment. The main window is depicted*

The operation of the tool as well as the procedure for implementing the classification method of arithmetic progressions, are shown in detail in the following steps:

A. The layer to form the basis of the choropleth map is inserted (Figure 2a)

B. The field on which the classification will be applied is selected (Figure 2b)

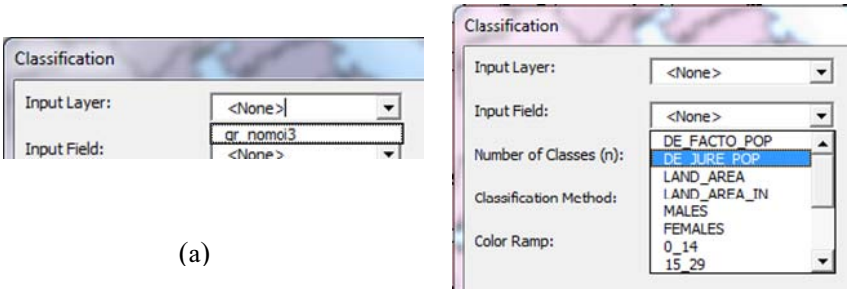

Figure 2. *(a) the layer selection and (b) the field selection*

C. After the selection of the field the maximum, the minimum and the mean value of the data as well as the frequency histogramme are displayed (Figure 3).

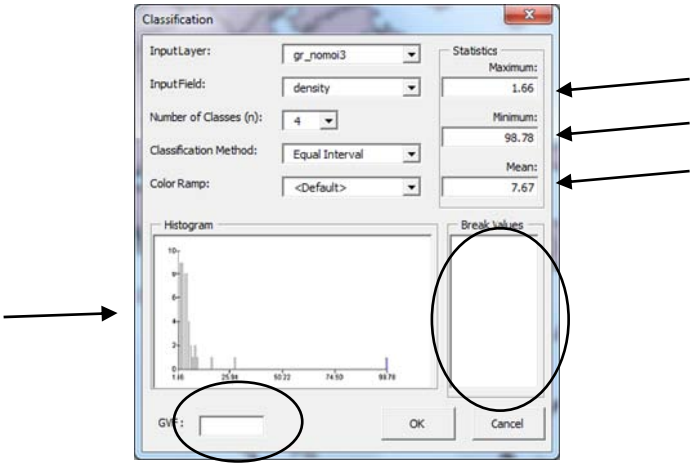

Figure 3. *Display of the maximum, the minimum, the mean and the histogramme of the data values. The boxes for the limits of classes and for GVF are still empty.*

D. The number of classes (n) is defined and after that the desirable classification method. When other methods, besides the arithmetic progression are chosen, the existing code of ArcGIS is performed (Figure 4).

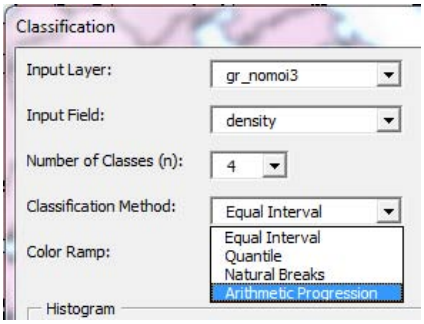

Figure 4. *Selection of number of classes and of classification method. In this example, the arithmetic progression is selected.*

E. When selecting the "arithmetic progression", a new dialogue window appears (Figure 5a). In this new window the user is asked to give values for the first term of the sequence (a) and for the difference (d) (§3.1). The window gives also the user the possibility to have assistance. By selecting the command "Help" the window of Figure 5b displays suggestions.

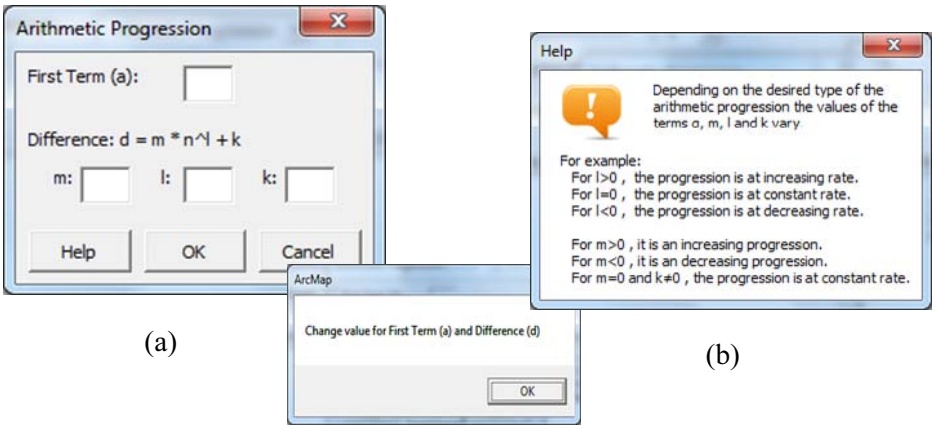

Figure 5. *(a)The window for the insertion of parameters for the arithmetic progression (b) the "help" window and (c) the error message window.*

In §3.1 it was mentioned that because the data values are in ascending numerical order the ΣB_i (Equations 1, 2) should be positive, so that the unknown X of the equation (1) is also positive. In case that the ΣB_i results to a negative value the message shown in Figure 5c is displayed.

After a successful definition of the parameters of the arithmetic progression the main dialog window appears again. The maximum limits of each class and the

value of the GVF are now displayed in the corresponding boxes (Figure 6). Next, a color ramp is selected (Figure 6) and pressing the "OK" button the code for the classification is executed and the choropleth map is displayed.

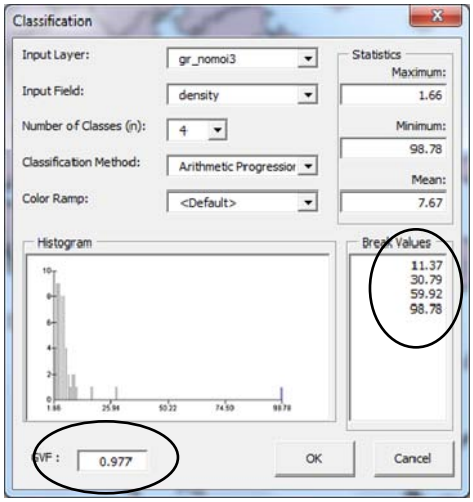

Figure 7. *Displaying of class limits and GVF value*

6. Conclusions and discussion

A common problem that arises during the compilation of a choropleth map is the choice of method to determine the class limits. Before choosing a method, however, it is necessary the cartographer to decide how he wishes to depict the data on the map. The chosen method will depend on the data, the cartographer's opinion about the quality of the data, and how easily he thinks that users can interpret the numerical values derived from the classification.

The criteria that can assist in selecting a classification method are: 1) whether the method examines how the data are distributed along the number line, 2) the easiness of understanding the classification process, 3) the simplicity of calculations needed 4) easiness in understanding the legend, 5) whether the method is acceptable for data that belong to ordinal level (see e.g. Papadopoulou, 2009) and 6) if the method can assist in selecting the appropriate number of classes (Slocum et al., 2005). In this paper, except for all the above, as a criterion for the selected classification method's suitability, the Goodness of Variance Fit (GVF) was indicated. The correct choice of a method is reflected in the value of this statistical criterion. By changing the number of classes and/or the method, the user can achieve higher values of GVF towards 1.

However, no method can be regarded as better than others even if the assessment

index GVF indicates a good classification of the data values. It is important, for the selection of a method, the cartographer to take under consideration the purpose of the map and the cognitive level of the user to whom the map is addressed. Whether the visual result of the choropleth mapping is adequate or not is left to the discretion of the chartographer.

References

Aza M., 2012. *Classification methods for choroplethic mapping and creation of an appropriate tool in a GIS environment.* Diploma Thesis, Department of Cadastre Photogrammetry and Cartography, AUTH (in Greek)

Burke R., 2003. *Getting to know ArcObjects: programming ArcGIS with VBA*. ESRI, California

Booth B., and Mitchell A., 1999. *Getting started with GIS*, ESRI, USA

Chang K., 2005, "Programming ArcObjects with VBA: a task-oriented approach". CRC Press LLC, USA

Dent B.D., 1999. *Cartography: Thematic Map Design*, 5th ed., WCB/McGraw-Hil, NYl

Evans, I.S., 1977. *The selection of class intervals*. Transactions of the Institute of British Geographers, New Series 2, Contemporary Cartography:98-124

http://www.casa.ucl.ac.uk/martin/msc_gis/evans_class_intervals.pdf (last retrieved 10-12-2014)

Jenks G.F., 1967. *The Data Model Concept in Statistical Mapping*. International Yearbook of Cartography, 7: 186-90

Papadopoulou M., 2009. *Introductory Cartography*, Lecture Notes, Department of Cadastre, Photogrammetry and Cartography, AUTH, (in Greek)
http://e-topo.web.auth.gr/index_gr.html?reload (last retrieved 2/12/14)

Robinson A.H., Sale R., Morrison J. and Muehrcke P., 1984. *Elements of Cartography*. 5th ed., John Wiley & Son, USA

Slocum T., McMaster R., Kessler F. and Howard H., 2005. *Thematic Cartography and Geographic Visualization*. Pearson Education, NJ

Spiegel M.R. and Stephens L.J., 1999. *Schaum's Outline of Theory and Problems of Statistics.* 3rd ed., McGraw-Hill Professional
http://books.google.gr/books?id=a6m_I4a2fmsC&printsec=frontcover&hl=el&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false (last retrieved 10-12-2014)

Voutsa D., 2010. *Planning and Developing a spatial decision support tool for the local authorities.* Phd Dissertation, AUTH (in Greek)

Zeiler M. (ed), 2001. *Exploring ArcObjects*, *Vol. 1 –Applications and Cartography*. ESRI, USA