Swiss Virtual Campus

“Dealing with Natural Hazards“

Module 1

Topic Group: **Data presentation**

Learning Unit: **Cartographic data visualisation - Standardisation and classification of data**

Major classification methods

For thematic map presentation, the acquired and analysed thematic data values are often grouped into classes, which simplifies the reading of the map as we have learned in the previous section. If you decide to classify your data, you may wonder about what would be the best method. For this purpose we will repeat and refresh basics of your knowledge about **statistical methods** in the following.

The major methods of data classification are:

- equal intervals,
- mean-standard deviation,
- quantiles,
- maximum breaks and
- natural breaks.

To explain the procedures involved in these different methods, we use an example that deals with data of the percentage of land on which wheat was harvested in different counties (see table 1). For a detailed example and the corresponding calculations, please see the next section 'Summary of the classification method' in this learning unit. There, you can apply your achieved knowledge and understand calculations in a concrete example.

County	% Wheat	County	% Wheat
Wyandotte	0.7	Ellsworth	21.1
Greenwood	1.5	Wallace	21.2
Jefferson	2.5	Jewell	21.6
Elk	2.8	Stevens	21.8
Miami	2.9	Smith	22
Lyon	2.9	Ottawa	22.4
Wabaunsee	2.9	Cheyenne	22.7
Chase	3	Lincoln	23
Pottawatomie	3.1	Ellis	23.1
Doniphan	3.5	Decatur	23.3
Bourbon	3.6	Marion	23.3
Johnson	3.7	Rawlins	23.3
Leavenworth	3.8	Cloud	23.4
Chautauqua	3.8	Clay	23.6
Franklin	3.9	Gove	23.7
Shawnee	4.1	Seward	24
Linn	4.1	Norton	24.7
Jackson	4.1	Hodgeman	24.8
...	...	...	...

Tab.1:Part of the data of the percentage of land on which wheat was harvested in different counties of Kansas, USA. (Source: SLOCUM 1999)

Equal intervals (constant class intervals)

In this classification method, each class consists of an equal data interval along the number line shown in the figure. To determine the class interval, you divide the whole range of all your data (highest data value minus lowest data value) by the number of classes you have decided to generate.

$$\frac{range}{(number\ of\ classes)} = \frac{(high - low)}{(number\ of\ classes)} = \frac{(55.8 - 0.7)}{5} = 11.56$$

You can check this step with the formula and the given data example. After you have done that, you **add the resulting class interval to the lowest value of your data-set**, which gives you the **first** class interval. Add this interval as many times as necessary in order to reveal the number of your predefined classes.

When is it useful to choose the method of equal class intervals? It is appropriate to use equal class intervals when the data distribution has a rectangular shape in the histogram. This, however, occurs very rarely in the context of geographic phenomena. Moreover, it is useful to use this method when your classification steps are nearly equal in size.

The major **disadvantage** of this method is that class limits fail to reveal the distribution of the data along the number line. There may be classes that remain blank, which of course is not particularly meaningful on a map.

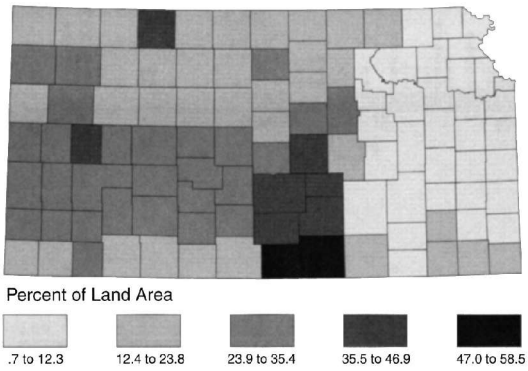


Fig.9: Choropleth map made by using the equal interval classification method. (Source: SLOCUM 1999)

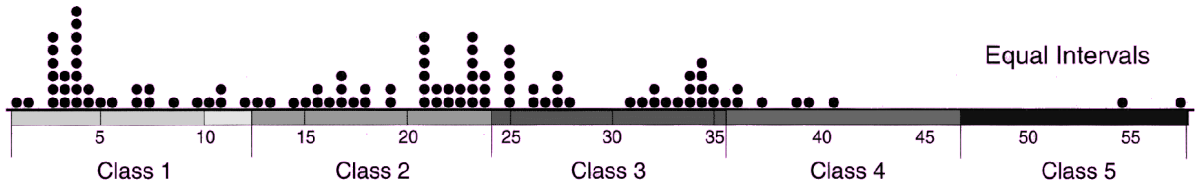


Fig.10: The number line reveals the class limits that have been created with the equal interval method. (Source: SLOCUM 1999)

Mean-standard deviation

Another method that allows us to classify our data-set is the standard deviation. This method takes into account how data are distributed along the number line (see figure). To apply this method, we repeatedly add (or subtract) the calculated standard deviation from the **statistical mean** of our data-set. The resulting classes reveal the frequency of elements in each class.

The mean-standard deviation method is particularly useful when our purpose is to show the deviation from the mean of our data array. This classification method, however, should only be used for data-sets that show an approximately "standardised normal distribution" ("Gaussian distribution"). This constraint is the major disadvantage of this method.

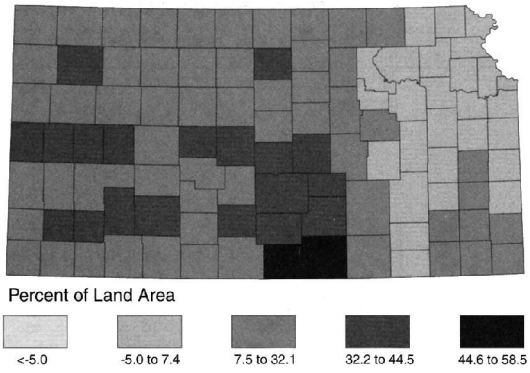


Fig.11: Choropleth map made by using the mean-standard classification method. (Source: SLOCUM 1999)

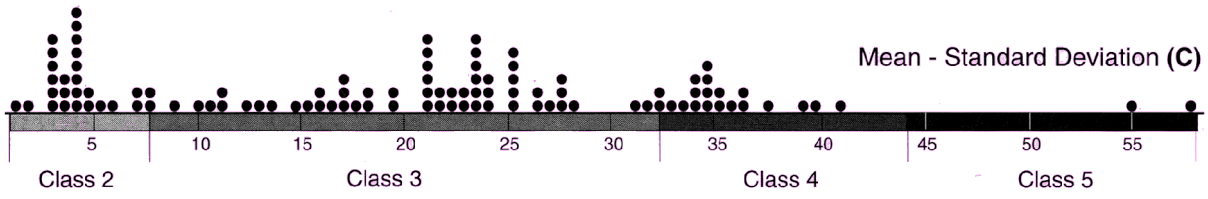


Fig.12: The number line reveals the class limits that have been created with the mean-standard method.(Source: SLOCUM 1999)

Quantiles (variable class intervals)

Another possibility to classify our data-set is to use the method of quantiles. To apply this method we have to predefine how many classes we wish to use. Then we rank and order our data classes by placing an equal number of observations into each class. The number of observations in each class is computed by the formula:

$$number\ of\ observations\ per\ class = \frac{(total\ observation)}{(number\ of\ classes)} = \frac{105}{5} = 21$$

If integer values are resulting from this calculation, we attempt to place approximately the same number of observations in each class.

An advantage of quantiles is that classes are easy to compute and that each class is approximately equally represented on the final map. Moreover, quantiles are very useful for ordinal data, since the class assignment of quantiles is based on ranked data.

The main **disadvantage** of this classification method are the gaps that may occur between the observations. These gaps sometimes lead to an over-weighting of some single detached observations at the edge of the number line. You can see such a huge value gap within the quantile class 5 in the figure below. (SLOCUM 1999)

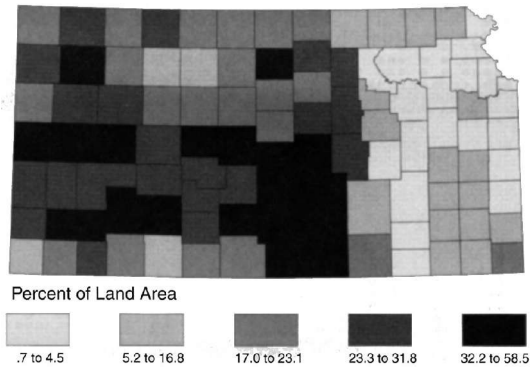


Fig.13: Choropleth map made by using the quantile classification method. (Source: SLOCUM 1999)

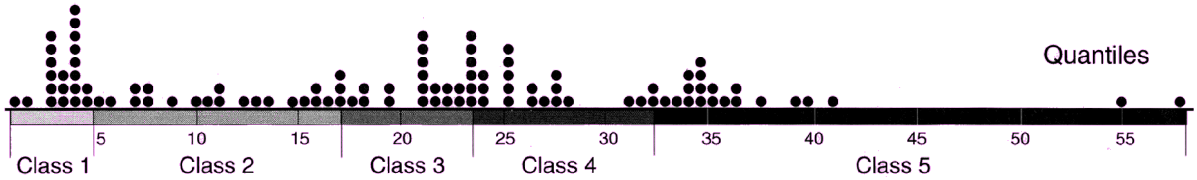


Fig.14: The number line reveals the class limits that have been created with the quantile method. (Source: SLOCUM 1999)

Maximum breaks

When we choose to use the method of maximum breaks we first order our raw data from low to high. Then we calculate the differences between each neighbouring value. Afterwards, the largest value differences will be applied as class breaks. You can also recognise the maximum breaks visually on the dispersion graph: large value differences are represented by blank spaces.

One **advantage** of working with this method is its clear consideration of data distribution along the number line. Another advantage is that maximum breaks can be calculated easily by subtracting the next lower neighbouring value from each value.

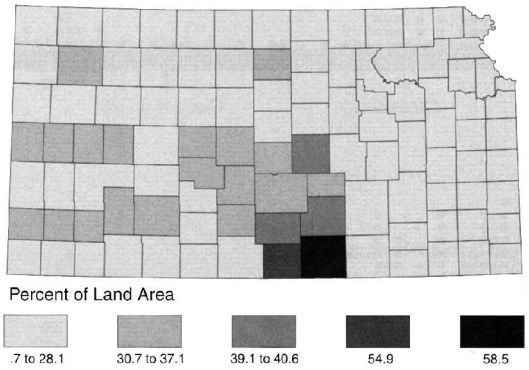


Fig.15: Choropleth map made by using the maximum break classification method. (Source: SLOCUM 1999)

A disadvantage, however, is that the systematic classification of data misses any proper attention to a visually more logical and more convenient clustering (see "Natural breaks"). This problem can be revealed when we compare the classes 4 and 5. It would probably be more useful to merge these two classes into one.

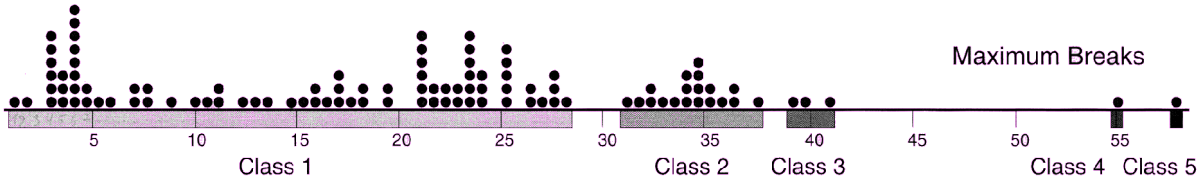


Fig.16: The number line reveals the class limits that have been created with the maximum break method. (Source: SLOCUM 1999)

Natural breaks

Applying the classification method of natural breaks we consider visually logical and subjective aspects to groupe our data-set. One important purpose of natural breaks is to minimise value differences between data within the same class. Another purpose is to emphasis the differences between the created classes.

A possible **disadvantage** of this method is that class limits may vary from one map-maker to another due to the author's subjective class definition. (SLOCUM 1999)

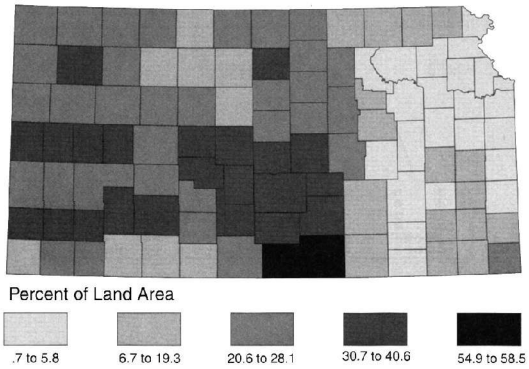


Fig.17: Choropleth map made by using the natural break classification method. (Source: SLOCUM 1999)

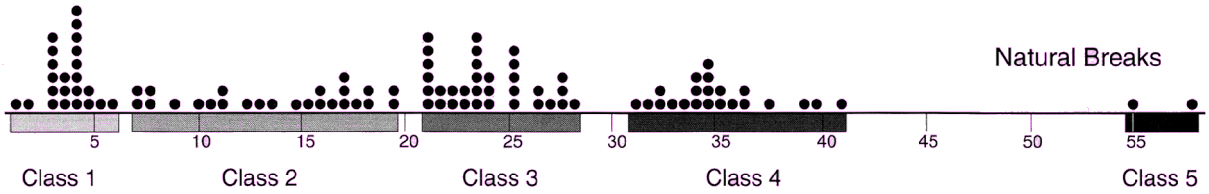


Fig.18: The number line reveals the class limits that have been created with the natural break method. (Source: SLOCUM 1999)

Summary of the classification methods

Now, that you know about the different classification methods used in thematic mapping, we will use an example that deals with **population density data in Europe** (see table 2) for a better understanding. In the following section, you will be able to comprehend each step being necessary for the decision making process for a specific classification method.

Country	Density	Country	Density
Monaco	17597	Serbia and Montenegro	104
Holy See	2023	Austria	97
Malta	1248	Slovenia	95
San Marino	441	Romania	94
Netherlands	385	Cyprus	82
Belgium	336	Macedonia	81
Turkey (european)	286	Spain	81
United Kingdom	244	Ukraine	81
Germany	233	Croatia	80
Liechtenstein	203	Greece	80
Italy	191	Bosnia and Herzegovina	77
Switzerland	176	Bulgaria	69
Luxembourg	171	Ireland	55
Andorra	144	Lithuania	55
Moldova	131	Belarus	50
Czech Republic	130	Latvia	37
Denmark	124	Estonia	31
Poland	124	Russia (european)	26
Albania	122	Sweden	20
Portugal	110	Finland	15
Slovakia	110	Norway	14
France	109	Iceland	2.7
Hungary	109		

Tab.2: Population density data Europe 2001. Data update: 2001 estimates (Source: http://www.globalgeografia.com/europe_eng/europe_sup.htm, visited Sep. 2003)

For a graphical visualisation of this data-set **see web** ('Data distribution graph'). **Tip**: It is necessary that you use both, study aid and web, to understand the example.

Equal intervals

M e t h o d

In this example you will learn how the equal intervals classification method works by using the population density data. As we learned, each class consists of an equal data interval. There are three steps in this classification:

(1)Choose the number of classes. Generally, a distribution between 4 and 8 classes are correct, 5 to 6 classes are standard.

(2)Calculation of the class interval. Each class has an equal amplitude a=A/n, where A is the total amplitude and n the number of classes.

(3)Calculation of the class limits.

Calculation and construction

(1)Choose the number of classes. In our example, we choose a number of 5 classes.

(2)Calculation of the class interval.

$$a=\frac{A}{n}$$

$$A=nMax-nMin=17597-2.7=17594.3$$

$$a=\frac{17594.3}{5}=3518.86$$

(3)Calculation of the class limits.

Class 1: nMin to (nMin+a) = 2.7 to (2.7+3518.86) =2.7 to 3521.56

Class 2: (nMin+a) to (nMin+2a) = (2.7+3518.86) to (2.7+2*3518.86) =3521.56 to 7040.42

Class 3: (nMin+2a) to (nMin+3a)= (2.7+2*3518.86) to (2.7+3*3518.86)= 7040.42 to 10559.28

Class 4: (nMin+3a) to (nMin+4a)= (2.7+3*3518.86) to (2.7+4*3518.86)= 10559.28 to 14078.14

Class 5: (nMin+5a) to (nMin+5a)= (2.7+4*3518.86) to (2.7+5*3518.86)= 14078.14 to 17597

For the dispersion graph and the map result, please **see web**.

Advantages and disadvantages

- Advantages: The equal interval classification method is simple and very easy to use, but is only satisfactory if every class is well represented (not the case for our data).
- Disadvantages: this method is not appropriate if the distribution is too asymmetrical (risk of empty classes), or if the distribution presents a few variant distribution. This method does not authorise comparisons because the range of the variable is specific of every series of data.
- Conclusion: As the data distribution is very asymmetrical this method is not suitable to the proposed data. It is confirmed by the dispersion graph (data are not represented in each class) and by the map result.

Mean-standard deviation

M e t h o d

Here, classes are formed by repeatedly adding or subtracting the standard deviation from the mean of the data. There are two steps in this classification:

(1)Choose the number of classes.

(2)Class limits are determined by using the mean and the standard deviation.

Mean: $\bar{X}=\frac{\sum fx}{\sum f}$ Standard deviation: $\sigma=\sqrt{variance}=\sqrt{\frac{(\sum X^2-N\bar{X}^2)}{N}}$

Class 1: $< \bar{X}-2\cdot\sigma$

Class 2: $\bar{X}-2\cdot\sigma$ to $\bar{X}-1\cdot\sigma$

Class 3: $\bar{X}-1\cdot\sigma$ to $\bar{X}+1\cdot\sigma$

Class 4: $\bar{X}+1\cdot\sigma$ to $\bar{X}+2\cdot\sigma$

Class 5: $> \bar{X}+2\cdot\sigma$

C a l c u l a t i o n a n d c o n s t r u c t i o n

(1)Choose the number of classes. 5 classes.

(2)Calculation of class limits.

Mean: $\bar{X}=\frac{\sum fx}{\sum f}=\frac{26073.7}{45}=579.42$

Standard deviation: $\sigma=\sqrt{variance}=\sqrt{\frac{(\sum X^2-N\bar{X}^2)}{N}}=\sqrt{\frac{301234823.9}{45}}=2587.29$

Class 1: $< \bar{X}-2\cdot\sigma < 579.42-2\cdot2587.29 < -4595.18$

Class 2: $\bar{X}-2\cdot\sigma$ to $\bar{X}-1\cdot\sigma = 579.42-2\cdot2587.29$ to $579.42-1\cdot2587.29 =$ -4595.18 to -2007.88

Class 3: $\bar{X}-1\cdot\sigma$ to $\bar{X}+1\cdot\sigma = 579.42-1\cdot2587.29$ to $579.42+1\cdot2587.29 =$ -2007.88 to 3166.71

Class 4: $\bar{X}+1\cdot\sigma$ to $\bar{X}+2\cdot\sigma = 579.42+1\cdot2587.29$ to $579.42+2\cdot2587.29 =$ 3166.71 to 5754.01

Class 5: $> \bar{X}+2\cdot\sigma > 579.42+2\cdot2587.29 > 5754.01$

For the dispersion graph and the map result, please **see web**.

A d v a n t a g e a n d d i s a d v a n t a g e

- Advantages: The mean-standard deviation method is particularly useful when the purpose is to show the deviation from the mean of the data array.
- Disadvantages: This classification method, however, should only be used for data-sets that show an approximately "standardised normal distributions" (Gaussian distribution). This constraint is the major disadvantage of this method.

- Conclusion: As the data sets do not look like a standardised normal distributions (Gaussian distribution), this method is not suitable to the proposed data. This is confirmed by the dispersion graph (data are not represented in each class) and by the map.

Quantiles

M e t h o d

Within this classification method, each class consists of exactly the same number of values. There are four steps in this classification:

(1)Choose the number of classes.

(2)Arrange all values in ascending order.

(3)Determine the number of values in each class:

$$K = \frac{(number\ of\ enumeration\ areas)}{(number\ of\ classes)}$$

(4)Beginning with the lowest value, K values are included in the first class, K values in the next class and so on. The class boundary is normally the mean value between the two adjacent values separating neighbouring classes.

C a l c u l a t i o n a n d c o n s t r u c t i o n

(1)Choose the number of classes: 5.

(2)Arrange all values in ascending order.

2.7 14 15 20 26 31 37 50 55 55 69 77 80 81 ... 233 244 286 336 385 441 1248 2023 17597

(3)Determine K. $K = \frac{(number\ of\ enumeration\ areas)}{(number\ of\ classes)} = \frac{45}{5} = 9$

(4)Arrange all values in ascending order.

Class 1:	2.7	14	15	20	26	31	37	50	55
Class 2:	55	69	77	80	80	81	81	81	82
Class 3:	94	95	97	104	109	109	110	110	122
Class 4:	124	124	130	131	144	171	176	191	203
Class 5:	233	244	286	336	385	441	1248	2023	17597

For the dispersion graph and the map result, please **see web**.

Advantage and disadvantage

- Advantages: The quantiles classification method is simple and very easy to use. Developing class boundaries of quantiles assures an equal number of values in each class and minimises the importance of the class boundaries.
- Disadvantages: Misleading if the enumeration units vary greatly in size and gaps may occur between the observations. These gaps sometimes lead to an over weighting of some single detached observations at the edge of the number line.
- Conclusion: Overall, this method is suitable for our. data even though the predominance of Monaco's density [17597 inhabitants/kmS] compared to the other European countries is not visible on the map

Maximum breaks

Method

Here, a class is determined accordingly to the distribution graph. There are three steps in this classification method:

(1)Choose the number of classes.

(2)Arrange all values in ascending order.

(3)Calculate the differences between each neighbouring value; where the largest value difference occurs, we will apply a class break. (Tip: You recognise such maximum breaks visually on the dispersion graph.)

Calculation and construction

(1)Choose the number of classes: 5.

(2)Arrange all values in ascending order.

2.7 14 15 20 26 31 37 50 55 55 69 77 80 81 ... 233 244 286 336 385 441 1248 2023 17597

(3)The difference between each neighbouring value is calculated and represented graphically. The largest value difference are applied as class breaks.

For the dispersion graph and the map result, please **see web**.

Advantage and disadvantage

- Advantages: One advantage of working with this method is its clear consideration of data distribution along the graph. Another advantage is that maximum breaks can

be calculated easily by subtracting the next lower neighbouring value from each value.

- Disadvantages: A disadvantage, however, is that the systematic classification of the data misses to pay attention to a visually more logical and more convenient clustering.
- Conclusion: Overall, this method is suitable to the proposed data even it would probably be more useful to separate the last class into more.

Natural breaks

M e t h o d

The natural break method is a graphical way of determining natural groups of similar values by searching for significant changes in frequency distribution. There are two steps in this classification:

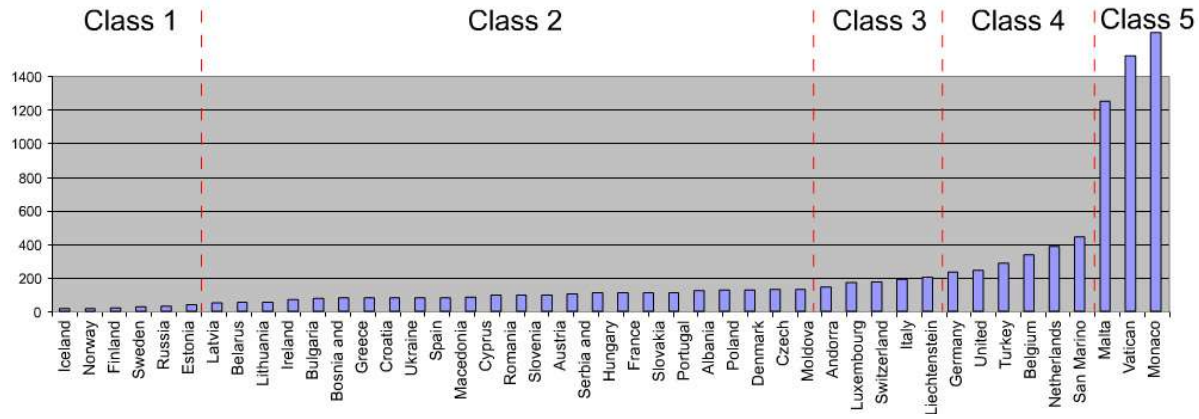
(1)Choose of the class numbers.

(2)In a visual inspection we will place class boundaries when significant changes occur (in the data distribution).

C a l c u l a t i o n a n d c o n s t r u c t i o n

(1)Choose the number of classes: 5.

(2)A visual inspection will place class boundaries in the data distribution.



For the dispersion graph and the map result, please **see web**.

Advantage and disadvantage

- Advantages: This method is simple and very easy to use. It is a graphical way of determining natural groups of similar values. Further, it can be used in conjunction with other methods.
- Disadvantages: Minor 'jumps' can be misleading and may lead to poorly defined class boundaries. As well, class limits may vary from one author to another due to its subjective point of view.
- Conclusion: This method is suitable to the proposed data.

Self test

See web.

Bibliography

- SLOCUM, T. A. (1999): Thematic Cartography and Visualization. Prentice Hall, New Jersey. ISBN 0-13-209776-1