CHOROPLETH MAP ACCURACY AND THE NUMBER OF CLASS INTERVALS.

Franky A.N. Declercq
Laboratory for Cartography - Katholieke Universiteit Leuven
Redingenstraat 16, B-3000 Leuven (Belgium)
e-mail: franky.declercq@geo.kuleuven.ac.be

Abstract
For a large number of data classifications, performance of the Goodness of Variance Fit (GVF) as accuracy optimization measure was compared to two other measures. It is shown that in general a number of 7 to 8 classes is preferred to generate accurate choropleth maps. Two newly developed classification methods are suboptimal solutions of the GVF optimal classification method and are designed to minimize image fragmentation and maximize rounding of class limits. Firm rounding of the optimal class limits to less than two significant digits decreases accuracy only slightly. Suboptimal classifications that yield a less fragmented image become rather inaccurate and need 9 to 10 classes to maintain classification accuracy.

1 Aim
The Goodness of Variance Fit (GVF) is compared to the Goodness of Deviation around the Median Fit (GDMF) and Goodness of Absolute Deviation Fit (GADF) as measures to optimize classification accuracy. Two fast iterative classification methods that are developed to run on PC's are introduced. One method determines class breaks with highly rounded digits, the other seeks for class limits that result in a less fragmented map image. Both methods are designed to maintain the highest possible classification accuracy. These new classification methods are applied to a variety of data sets and compared with the Jenks optimal method and the traditional equal interval method in terms of generated classification accuracy (GVF), fragmentation index (FI) and number complexity index (NCI). The performance of each of the four classification methods is assessed for a number of classes ranging from 3 to 14. The impact of the type of classification method and the number of classes on classification accuracy, image fragmentation and class limit number complexity will consequently be judged.

2. The recommended number of classes to use in choropleth mapping
Desktop mapping software usually gives 4 to 6 classes as default and handles a varying restriction on the maximum number of classes to use (Table 1). A study of all choropleth maps published in five scientific journals over a period of six years reveals that about 50% of the maps has 5 classes and about 20% contains 6 classes. In the statistical literature an overview is given of three methods to determine the optimal number of classes for a given amount of observations [4] (Table 2). As maps easily contain 100 to 300 enumeration units, the recommended number of classes (8 to 17) is much higher than perceptual studies in the cartographic literature recommend, i.e. maximum 5 to 9 classes [7].

Definitely there is no consensus on the number of classes actually to use in choropleth mapping. As a result performance of the four classification methods will be investigated for a number of classes ranging from three to fourteen.

Number of observations n	Number of classes: $K = 1 + 3.3 \log_{10} n$	Number of classes: $K < 5 \log_{10} n$	Number of classes: $K = \sqrt{n}$
50	7	8	7
100	8	10	10
300	9	12	17
500	10	13	22

Table 1: Number of recommended classes for a given number of observations (after [4])

	histogram	quantiles	standard deviation	equal intervals	optimization	number of classes maximum	default
SPANS Map 1.3	X	X		X		N	5
MacMap 1.4		X	X	X		64	5
Atlas MapMaker 1.0		X		X		50	6
MapInfo 3.0		X	X	X	X	16	4
ArcView 2.0	X	X		X		N	5

Table 2: Classification methods and the recommended number of classes in popular desktop mapping software (N = number of enumeration units)

3. Input data used for testing

From a socio-economic database (65 variables) a sample of 30 features per municipality of Flanders (308 enumeration units) was chosen as input for the choropleth map classifications. For each data set both skewness (SK) and spatial autocorrelation (Geary Ratio) were determined (Table 3). Each of these 30 data sets was applied to classifications with 3 to 14 classes. These 360 classifications were performed using (1) the Jenks optimal accuracy method with and without rounding, (2) the equal interval method, (3) three levels of the suboptimal method minimizing image fragmentation, (4) three levels of the suboptimal method minimizing class break number complexity. This results in a total of 9*360 = 3240 different choropleth map classifications for which the GVF, FI and NCI are computed.

	normal (SK1) (SK<+0.7)	skewed (SK2) (+0.9<SK ≤ +2.0)	strongly skewed (SK3) (SK >+2.0)	
high autocorrelation (GR<0.6)	5/15	5/12	5/8	15/35
less autocorrelation (0.8<GR<1.2)	1/1	4/6	6/19	11/26
fragmented (GR>1.3)	0/0	0/0	4/4	4/4
	6/16	9/18	15/31	30/65

Table 3: Number of data sets from a global socio-economic database used as input data for the classification methods, by degree of skewness and spatial autocorrelation

4. The Goodness of Variance Fit (GVF) as a measure to optimize classification accuracy

Accuracy of choropleth maps can be measured using the GVF, i.e. Goodness of Variance Fit [2]:

GVF = 100 - 100*(SDCM/SDAM)
where SDCM = squared deviations from the class mean
where SDAM = squared deviations from the array mean)

The best possible GVF = 100, where the number of classes equals the number of observations; GVF = 0 if all observations are put into one class.

Alternative classification accuracy indices are the Goodness of Deviation around the Median Fit (GDMF) in which squared deviations from the class and array median are calculated, and the Goodness of Absolute Deviation Fit (GADF) [9 , p. 364] in which absolute deviations from the class and array median are calculated. The classification method developed by Jenks which minimizes the GVF, has been tested by Smith for five class maps and was found far more accurate than traditional classification methods [10].

The GVF-optimal method is implemented in a C-routine which yields the highest possible GVF for each of the 360 classifications tested (Figure 1). A GDMF- and GADF-optimal method was developed and applied to the same data sets. For each of the 360 classifications from each of these methods, GVF values were calculated. Grouping based on maximum GDMF values yielded almost identical classifications as with the GVF-optimal method, i.e. differences in GVF value of less than 0.6 units. GADF-optimal classifications were found more different from the GVF-optimal: differences in GVF are less than 1.0 for only 73% and less than 5.0 for 91% of the classifications, and occur especially if less than 7 classes are used for strongly skewed data sets (Figure 1).The GVF and GDMF may be consequently be considered as consistent and almost identical measures to optimize classification accuracy for all types of data and for a various number of classes, which differ from the GADF measure.

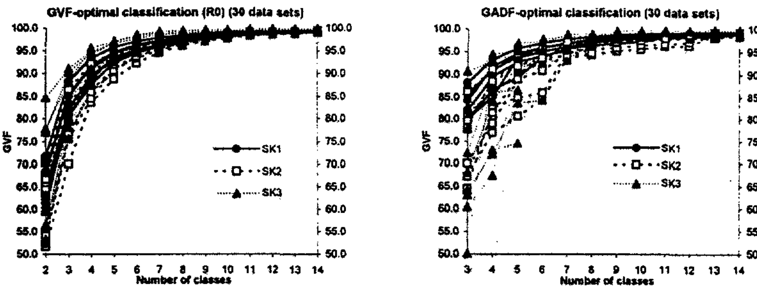

Figure 1: GVF classification accuracy of 30 data sets with a varying degree of skewness (SK) that are classed using two different accuracy optimization measures i.e. GVF and GADF

5. The suboptimal classification methods

5.1 General approach

Both newly developed methods are iterative computer algorithms that start from the class breaks of the GVF optimal accuracy method. Class breaks are shifted along a number of observations, allowing three levels of deviation from the optimal accuracy breaks, i.e. a shift of 10%, 20% or 30% of the amount of observations in a class. With each observation that switches between classes, image fragmentation or number complexity is calculated. For each level of tolerated deviation, the class intervals with minimal fragmentation or with minimal number complexity are retained.

5.1 The suboptimal method minimizing image fragmentation

Complexity of a visual image can be adequately measured using the fragmentation index FI [1, 6]:

$$FI = 100*(M-1)/(N-1)$$

where M = the total number of contiguous units of the same class and N = the total number of enumeration units. FI = 100 if each neighbouring entity belongs to another class and FI = 0 if all entities belong to the same class. This suboptimal method generates class intervals S1, S2 and S3 for each of the three possible levels of deviation from the GVF optimal classification (Figure 2).

5.2 The suboptimal method minimizing class break number complexity

A measure for the complexity of class breaks is developed based on the findings that numbers can be more easily read if they contain only one or two significant digits or if they end with a '5' [3, 5, 8]. The number complexity index (NCI) is defined as:

$$NCI = \sum_{i=1}^{n} a_i \qquad \text{where } a_i = 1 \quad \text{if } c = 1, 2, 3, 4, 6, 7, 8, 9$$

$$\text{or if } c = 0, 5 \text{ followed by numbers other than } 0$$

$$\text{where } a_i = 0.5 \quad \text{if } c = 5 \text{ followed by only 0's}$$

$$\text{where } a_i = 0 \quad \text{if } c = 0 \text{ followed by only 0's}$$

for each number consisting of n digits c.

Examples of numbers and their NCI value:

563:	3.0	85000:	1.5	4075:	3.5
500:	0.5	3550:	2.5	650087:	6.0

The rounding algorithm searches for class breaks within the range of allowed deviation from the optimal class break. Optimal class breaks in their simplest form are not rounded (R0, Figure 2) as they are the arithmetic mean of the highest and lowest value of the classes the class break divides.

The values of the input data and therefore also the R0 breaks all have maximum four significant digits. As it is advisable that class breaks should lay somewhere 'near the middle' of two neighbouring observations, an optimal class break can only be rounded as far as the rounded number (R1) still lays within one third of the range between the observations. As a consequence the suboptimal rounding algorithm thus shifts class breaks from one observation to another and determines the most rounded number that falls within the mean $\pm$ 16.7 % of the range of each two observations. From all these rounded class breaks, the most rounded within the allowed level of deviation (R2, R3, R4) is retained.

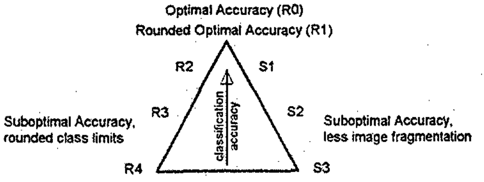

Figure 2: Classification methods and their potential classification accuracy.

6. Results.

Based on the classification accuracy of all methods (Figure 3) a GVF value of 95.0 or more may be considered as a 'satisfactorily accurate classification'. With the GVF optimal classification method (R0, R1) in general at least 6 classes are needed to attain this accuracy level. Suboptimal methods with maximum rounding (R2, R3, R4) on average need at least 7 classes and suboptimal methods with minimal fragmentation (S1, S2, S3) need at least 8 to 10 classes (cf. table 4). The traditional equal interval method (I) only achieves a GVF >=95.0 if data with a normal distribution are classed.

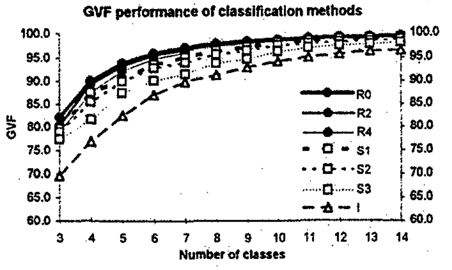

Figure 3: Accuracy of the classification methods tested

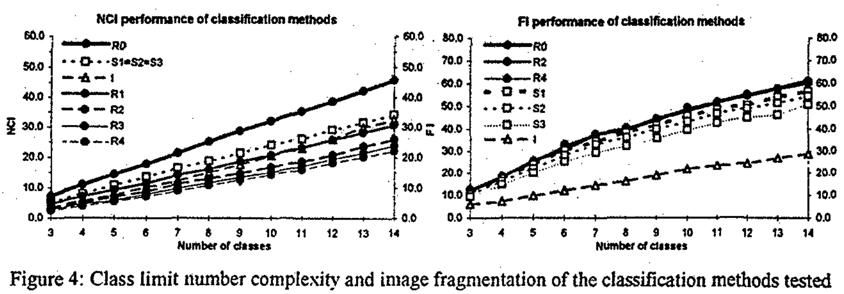

Figure 4: Class limit number complexity and image fragmentation of the classification methods tested

On average fragmentation remains constant for all suboptimal classifications with rounded class breaks (Figure 4). Optimization of fragmentation yields a decrease of FI for S1, S2 and S3 with respectively more than 5%, 10% and almost 20%. To remind the impact of FI changes on the map image: FI=20 equals about 62 visual groups in the image, FI=40 about 124 groups and FI=60 about 186 groups. The low fragmentation for the equal interval method (I) is related to the amount of data sets that are strongly skewed for which many areal units are grouped into a few classes.
The complexity of class breaks on average remains constant for all suboptimal classifications that minimize image fragmentation with NCI values somewhat higher than those of the rounded optimal method R1 (Figure 4). Not rounded optimal class breaks (R0) on average have 2.5 to 3.3 significant digits. Optimization of fragmentation yielded a decrease of NCI for compared to R1, with a mean of 1.1 to 1.8 significant digits for R2 and with 0.8 to 1.6 digits for R4 (Figure 4). Equal interval class breaks were determined by adding a rounded interval to a rounded base number, which resulted in a similar rounding as with the rule of the 1/3 range with the suboptimal methods.

7. Conclusions

A GVF value of 95.0 can be considered as normative for an accurate classification. With the use of 7 to 9 classes the suboptimal method with the most rigorous rounding (R4) gives the highest overall merits compared to the optimal accuracy method (R0) in terms of a high classification accuracy, rounded class breaks and image fragmentation (Table 4). The equal interval method on average also performs well, but this is the result of a classification with an unacceptable decrease in accuracy. The suboptimal methods obtaining lower image fragmentation (S1, S2, S3) have a loss of classification accuracy which is not compensated by the relative gain in fragmentation improvement.
The recommended number of classes in this study is determined by the nature of the geographic data tested. Complementary research on the perception of choropleths with 7 to 9 classes is necessary.

classification	with 4 to 6 classes				with 7 to 9 classes				with 10 to 14 classes			
	GVF	FI	NCI	Σ	GVF	FI	NCI	Σ	GVF	FI	NCI	Σ
R0	0	0	0		0	0	0		0	0	0	
R1	0	0	2.5	+2.5	0	0	4.0	+4.0	0	0	>4.0	+4.0
R2	0	0	4.0	+4.0	0	0	5.5	+5.5	0	0	>4.0	+4.0
R4	-0.5	0	5.5	+5.0	-0.5	0	7.0	+6.5	0	0	>4.0	+4.0
S1	-1.0	0.5	1.5	+1.0	-1.5	1.0	2.5	+2.0	0	1.0	3.0	+4.0
S2	-1.0	0.5	1.5	+1.0	-2.0	1.5	2.5	+2.0	-3.0	2.0	3.0	+2.0
S3	-1.5	1.0	1.5	+1.0	-3.0	2.5	2.5	+2.0	-4.0	3.0	3.0	+2.0
I	-2.0	7.0	2.5	+6.5	-4.0	>7.0	4.0	+7.0	-7.0	>7.0	4.0	+4.0

Table 4: Summarized comparison of the classification methods for three ranges of classes, in terms of the number of classes that a particular method needs to obtain the same accuracy (GVF), fragmentation (FI) or rounding (NCI) compared to the optimal accuracy classification method R0

References

[1] Bregt, A.K. & Wopereis, M.C.S. 1990. Comparison of complexity measures for choropleth maps. The Cartographic Journal 27: 85-91.

[2] Coulson, M.R.C. 1987. In the matter of class intervals for choropleth maps: with particular reference to the work of George F. Jenks. Cartographica 24, no. 2: 16-40.

[3] Depuydt, F. 1988. Optimization of the cartographic communication process in chororpleths by computer assistance. International Yearbook of Cartography 28: 203-211.

[4] Gardiner, V. & Gardiner, G. 1980. Analysis of frequency distributions. Concepts and techniques in modern geography nr. 19. London: Headley Ltd, 63 pp.

[5] Golde, R.A. 1966. Sharpen your number sense. Harvard Business Review 44, no. 4: 73-83.

[6] MacEachren, A.M. 1982a. Map complexity: comparison and measurement. The American Cartographer 9, no. 1: 31-46.

[7] Mersey, J.E. 1990. Color and thematic map design. Cartographica 37, no. 3: 1-157.

[8] Monmonier, M.S. 1982. Flat laxity, optimization, and rounding in the selection of class intervals. Cartographica 19, 1: 16-27.

[9] Robinson, A.H., Sale, R.D., Morrison, J.L. & Muehrcke, P.C. 1984. Elements of cartography, fifth edition. New York: John Wiley and Sons.

[10] Smith, R.M. 1986. Comparing traditional methods for selecting class intervals on choropleth maps. The Professional Geographer 38, [illegible].