# HOW MANY REGIONS?
# TOWARD A DEFINITION OF REGIONALIZATION EFFICIENCY

Ferko Csillag° and Sandor Kabos°°
° Department of Geography, University of Toronto, Erindale College, Mississauga, ONT, L5L 1C6, Canada I <fcs@eratos.erin.utoronto.ca>
°° Mathematical Statistical Group, Institute of Sociology, Eotvos Lorand University, Pollack tér 10., Budapest-1088, Hungary I <h56kab@ella.hu>

## *ABSTRACT*

This paper revisits a more than twenty-five-year old idea of G. F. Jenks and W. R. Tobler about the relationship between accuracy, information and map complexity of choropleth maps. The problem of regionalization (*sensu* aggregation) is treated within a spatial statistical context. For a map with N regions to be aggregated into G groups, nonspatial hierarchical classification schemes disregard spatial pattern and are prone to lead to non-contiguous classes (i.e., each group may consist of a large number of patches). Restricting merges during clustering according to neighborhood-topological relationships rewards contiguous patches of classes, but may impose too strict, potentially misleading, constraints. To obtain more efficient, less complex aggregate representations (e.g., maps) we propose to evaluate efficiency by a modified version of Akaike's information criterion: AIC' = (-2 x loglikelihood + 2 x number of patches). It follows from the general principle of model selection, by minimizing the sum of fitting error and some measure of model complexity, Socio-economic, environmental and simulated data are used to highlight the characteristics of this approach, which appears particularly useful when no additional information is available to select the number of groups.

## *INTRODUCTION*

The art and science of creating beautiful and meaningful maps based on some two-dimensional distributions has attracted people for several hundred years. In particular, major efforts have been focused on creating "the" spatial/cartographic analogy of classification; i.e., to put the N elements (data representation units, DRUs) of a two-dimensional lattice into G<<N "spatial groups". Such tasks often emerge in studies of socio-economic variables (e.g., defining wealthy/poor neighborhoods), in environmental studies (e.g., finding locations of suitable habitats) and in many other geographically-oriented fields.

Considering the frequent occurrence and diversity of applications of such tasks, it is not surprising that several detailed studies and overviews have focused on the series of "map-making" decisions and their optimization. Classical *cartographic* treatises, typically under the "error and classification of choropleth maps" keywords, can be found in Jenks and Caspall (1971), Monmonier (1973), Stegena and Csillag

(1986) and, in textbook format, in Robinson et al. (1995, p.517 ). More analytical approaches to similar problems are dealt with in *spatial statistics* (with widespread applications in econometrics, epidemiology, soil science) generally under the "aggregation and the modifiable areal unit problem" headings, for example, in Unwin (1981), Haining (1990) and Cressie (1993). A somewhat closely related array of techniques have emerged in *image processing* usually referred to as "image segmentation" (see, e.g., Schowengerdt, 1983, Kertesz et al., 1996). Several reports recognized the relationships, and interactions, among these procedures and some attempted to define a more general framework for "spatial data representation" (e.g., Maguire et al., 1991). Within the context of geographical information systems (GIS), often linked with statistical software packages, "spatial grouping" is also a frequently occurring common task, even if it is performed with diverse goals ranging from illustration, detection and verification of spatial patterns, optimization of visual and/or functional representation.

The real impetus for this paper, however, is an intriguing idea illustrated on the last, an apparently neglected, figure (see Figure 1) from the seminal paper of Jenks and Caspall (1971).

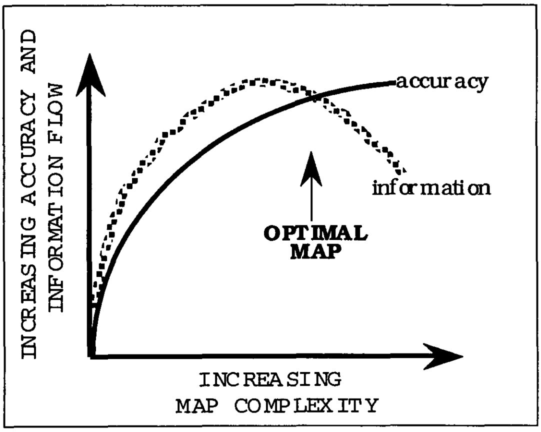


*FIGURE 1.*
Relationship between accuracy and information flow on maps; redrawn from to Jenks and Caspall (1971). The accuracy vs. complexity relationship was supported by empirical data; the information vs. complexity relationship was based on Waldo Tobler's personal communication. Interestingly, the choice of an "optimal map" does not correspond to either maximum information flow, or to maximum accuracy.

This figure seems to suggest more challenges than conclusions:

- How does one measure information flow as a function of map complexity?
- What evidence supports the shape of the "information vs. complexity" curve? What determines its position along the complexity axis?
- What evidence supports the existence of a unique intersection of the "information vs. complexity" and the "accuracy vs. complexity" curves? What does its position depend on?
- What algorithm is suitable for finding the optimal map?
- Assuming there is a unique intersection, what justifies the selection of the marked "optimal map" *instead* of maximum information (or maximum accuracy)?

## (RE)DEFINING THE PROBLEM: MEASURING INFORMATION LOSS

Let us define the problem as follows. The object set (T) consists of N units (e.g., polygons in a coverage, pixels in an image). The data set (Y) is a K-dimensional variable observed at each element of T. $P^{(G)}$, a partition of T, consists of G collectively exhaustive disjoint classes. Each class covers a region that may consist of one or more patches. The special regionalization where all classes are spatially contiguous (i.e., the number of classes (G) equals the number of patches (R)) is called segmentation. Note that the finest partition, the only N-partition is $P^{(N)}$, and the coarsest partition, the only 1-partition is $P^{(1)}$, and the definition of fine/coarse (and finer/coarser) is the usual. Let D(P) denote the *discrepancy* between a selected regionalization and the observed phenomenon. In light of general statistical model selection (Linhart and Zucchini, 1986), our model of choice should be parsimonious, i.e., it should not have more parameters than the ones which can be reliably estimated, for example, to avoid "overfitting". Thus, discrepancy consists of two parts: one due to *approximation*, and another due to *estimation*. The first component, $D_A(P)$, practically measures model complexity, and is often completely neglected. The second component, $D_E(P)$, measures the goodness of fit between the sample and the chosen (approximating) model. In the above outlined classification example it is the “loss of information due to grouping”, and it is most commonly measured by the expectation of the negative loglikelihood. This leads us to Akaike’s information criterion as a measure of discrepancy (Akaike, 1973):

AIC = (2 x number of parameters -2 x loglikelihood).

Our task is to minimize the discrepancy over the set of all possible partitions P*. Note that P* consists of potentially very large number of elements ($2^N$), thus there is no real chance to find the exact solution. Clustering procedures, therefore, are typically confined to some subset of P* while minimizing D(P). An acceptable way to avoid the problem of comparing, and thus choosing from, models of different complexity by computing $D(P)=D_A(P)+D_E(P)$, is to set G, i.e., to reduce the problem to finding $P^{(G)}$, the G-class map, with the smallest $D_E(P)$. Cromley (1996) provides an extensive recent review of comparing different "estimation discrepancies" with given number of classes.

In this paper we will consider the problem when the number of classes (G) is not known. Hierarchical clustering algorithms, for example, are suitable to scan a subset of P*, the monotone aggregating (coarsening) sequence of $P^{(N)}$, $P^{(N-1)}$, ..., $P^{(1)}$, i.e., they start from the finest partition (all elements form a separate class) and the number of classes decreases by one in each step (by merging two classes) until the coarsest partition, $P^{(1)}$, is reached. The algorithms differ from each other in the way they decide which two classes to merge. The most common choice for $D_E(P)$ is the ratio of within-groups variance/total variance. The value of $D_E(P)$ can be regarded as a measure of separation of partition P. The Ward-method of clustering (Ward, 1963) in each step selects the pair of clusters to merge by minimizing the increase in the above defined "estimation discrepancy" leading to the monotone increasing sequence of $D_E(P)^{(N)}$, ..., $D_E(P)^{(1)}$. Note that there is no guarantee that any member of this sequence is close to the minimum of $D_E(P^*)$.The likelihood in the case of a simple product MVN($\mu$,$\sigma$) is:

$$l(\mu,\sigma,Y) = \text{const} \times \exp\{-(1/2) \times (1/\sigma^2) \times \Sigma_n[y_n-\mu_n]^2]\}$$

for which the (-2 x loglikelihood) reduces to the within-groups sum-of-squares if $\sigma^2=1$. As Jenks and Caspall (1971) also note, accounting for $D_E(P)$ only during aggregation, one would always choose the map with each DRU being a separate class, because $D_E(P^{(N)})$ provides the "best" separation by *value*. Following from AIC, the discrepancy due to approximation, $D_A(P)$, should equal the number of classes (G).

In geographical applications, when judging whether a partition is "good" or not, one is frequently concerned with pattern, the separation by *location* as well. Assuming that we are looking at "dirty pictures", i.e., realizations, where some "crisp" regions are blurred by noise, it is essential to use methods which are robust in "finding" the regions, thus accounting for $D_A(P)$ as well. One approach in this direction is the restriction of the subset from which a clustering algorithm chooses classes to merge according to neighborhood-topological information. Such "patch"-versions can be implemented for any hierarchical clustering, similarly to several region-growing algorithms developed in image analysis (Landgrebe, 1980). If we restrict merges to neighbors, the number of classes (G) equals the number of patches (R) in each step.

## *AN EXAMPLE*

Let us illustrate how these measures of discrepancy work with a simple example. Figure 2a. shows a simple map of 64 DRUs with three classes (0, 8 and 9 represent values), which form three "crisp" regions, or patches. Figure 2b. and 2c. are "standard" cartographic representations with three equal-count and three equal -interval choropleth maps, respectively. Aggregating this map with Ward-clustering and its "patch"-version, we can plot the within-group sum-of-squares (SSQw), the number of classes and the number of patches for each iteration (Figure 3.).

To generate a measure of information (see Figure 1.), both clustering procedures can be characterized by AIC (in this case SSQw+2 x number-of-classes); cAIC denotes the case of Ward clustering and pAIC denotes the case of Ward_patch clustering (Figure 4.). Note that cAIC practically serves as a stopping rule, but it "stops" a little bit "early". Therefore, we propose to investigate a modified version, cAIC'= SSQw+2 x number-of-patches for the Ward-clustering, because it retains essential information about the pattern, while it is not prone to the restrictions of Ward_patch. The minimum of the AIC-plots corresponds to the "minimum information loss" due to the model, and such measures are particularly useful in comparing the nested series of models generated by hierarchical clustering.

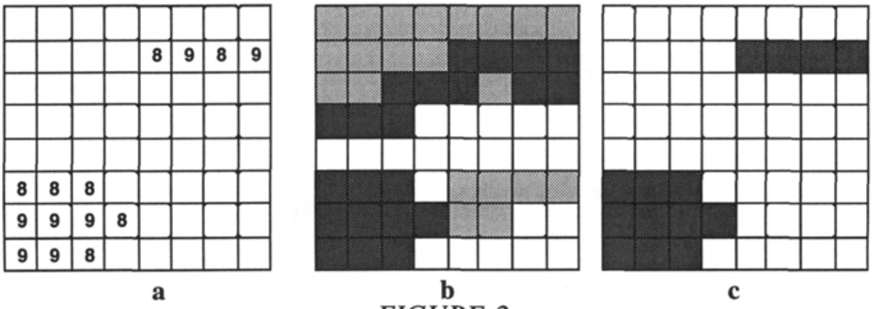


*FIGURE 2.*

Sample map (a) with three classes forming three "crisp" regions (0, 8 and 9 are values). Its three-class equal-count (b) choropleth map represents seven patches and its three-class equal-step (c) choropleth map represents three patches.

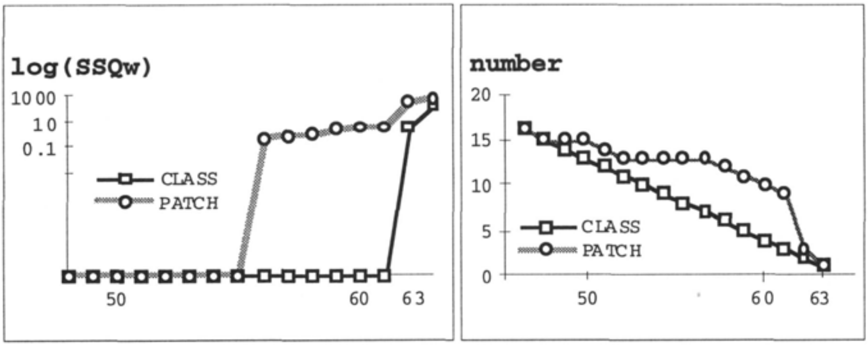


*FIGURE 3.*

SSQw (left) and the number-of-classes/patches (right) for the last 16 iterations of Ward and Ward_patch clustering of Figure 2a. SSQw is the discrepancy due to estimation, and the number-of-classes or the number-of-patches is the discrepancy due to approximation.

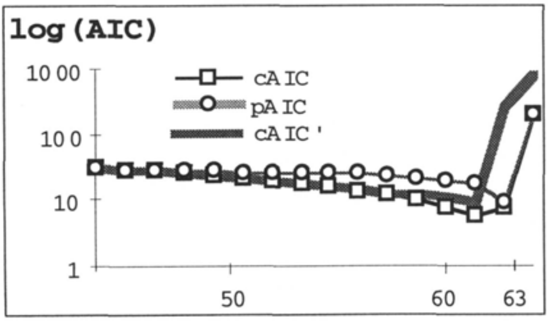


*FIGURE 4.* Plot of measures of "information loss" for the last 16 iterations of Ward (cAIC, cAIC') and Ward_patch (pAIC). The minimum of these functions corresponds to minimum discrepancy between the model and the observation.

Both cAIC and pAIC have minima at 3-classes/9-patches (i.e., each value in Figure 2a. forms a separate class), while the minimum of cAIC' coincides with 2-classes/3-patches. In other words, both cAIC and pAIC seem "to overfit" resulting in a tendency to reduce SSQw at a cost of greater number of patches. One can apply cAIC' as an alternative stopping rule for agglomerative hierarchical clustering of geographical phenomena.

## *A SIMULATION EXPERIMENT*

To investigate the behavior of these measures of discrepancy we have conducted a simulation experiment. Because we are interested in deriving some information about spatial pattern (c.f. regionalization), noise with various levels of spatial autocorrelation was added to Figure 2a, and these realizations were aggregated using both clustering algorithms (Ward and Ward_patch). The spatial structure of noise was controlled by the conditional autoregressive (CAR) parameter ρ (Cressie, 1993):

$$\mathbf{Y} \approx \mathrm{MVN}[\mu,\sigma^2(\mathrm{I}-\rho\mathbf{W})^{-1}]$$

where $w_{ij}$=1 for neighbors (0 otherwise), and we set $\mu_{ij}$=0 and $\sigma^2$=1. Fifty realizations were analyzed for ρ set to 0.0, 0.1, 0.2, and 0.245, respectively. Table 1. summarizes the results for the two extreme values of ρ, and Figure 5. shows example outputs.

*TABLE 1.*
Summary of fifty simulations for extreme values of spatial autocorrelation. Rows contain mean values and standard deviations for various measures of aggregation quality. Columns refer to different merging and stopping rules for Ward clustering.

| ρ=0 | cAIC | | pAIC | | cAIC' | |
|---|---|---|---|---|---|---|
| minimum | 21.37 | 1.58 | 48.45 | 5.45 | 54.09 | 6.87 |
| class | 7.57 | 0.79 | 14.43 | 3.87 | 5.00 | 1.53 |
| patch | 35.14 | 5.05 | 47.86 | 6.52 | 15.29 | 6.50 |
| iteration | 56.43 | 0.79 | 49.43 | 3.69 | 59.00 | 1.53 |
| SSQw | 6.22 | 1.82 | 19.31 | 4.73 | 23.52 | 7.90 |

| ρ=0.245 | cAIC | | pAICp | | cAIC' | |
|---|---|---|---|---|---|---|
| minimum | 22.90 | 2.03 | 46.26 | 4.79 | 55.04 | 4.92 |
| class | 8.14 | 1.35 | 15.14 | 3.44 | 5.00 | 1.29 |
| patch | 33.29 | 5.47 | 46.86 | 5.90 | 16.29 | 3.35 |
| iteration | 55.86 | 1.35 | 48.86 | 3.44 | 59.00 | 1.29 |
| SSQw | 7.72 | 3.20 | 18.64 | 8.47 | 26.21 | 3.80 |

One would expect that as the spatial autocorrelation of noise increases the higher the chance to mislead the clustering algorihtm by forming "artificial" patches.

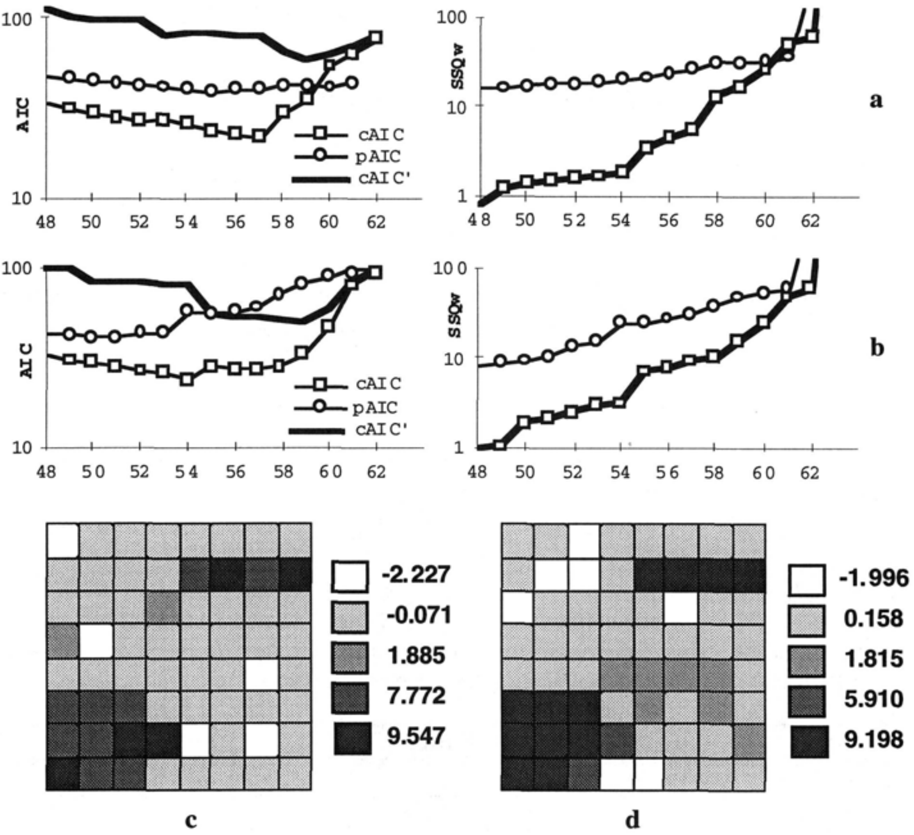


*FIGURE 5.*

Sample outputs from the simulation-aggregation study. Values of SSQw and AIC for a "typical" run for the sum of Figure 2a. and spatially not correlated noise (a) and spatially highly correlated noise (b). The corresponding 5-class maps corresponding to the minimum of cAIC' are shown on (c) and (d), respectively.

Clearly, the cAIC' stopping rule is the most resistant to the increasing ρ; its average choice for the number of classes remains the same (5.0), while both other measures tend to choose more classes and even more patches ($D_A(P)$), at each ρ. Of course, it comes at a cost of greater $D_E(P)$; the values of SSQw are significantly higher than for the other two measures. It is important to note that in real applications, typically, there is no information about the relationship between the amount of noise and the nature of boundaries, therefore, there is always a chance to overfit to "islands" (when using cAIC), or to "awkward patches" (when using pAIC).

## *TOWARD APPLICATIONS*

The implementation of using any of the above described measures in geographical analysis is relatively simple in commercially available GIS software.

Below, two very different mapping problems are used to illustrate the applicability of the findings where regionalization is the task. We have intentionally selected examples, where no *a priori* information can be easily used.

Case-1: Regions of high acid deposition in the northeast US are intensively studied to understand and predict its impact on the soil-water-plant systems. Since long-term acid deposition measurements are only sporadically available, elevation has been used as a surrogate for the amount of wet acid input into lake ecosystems (for example, for defining sampling strata).

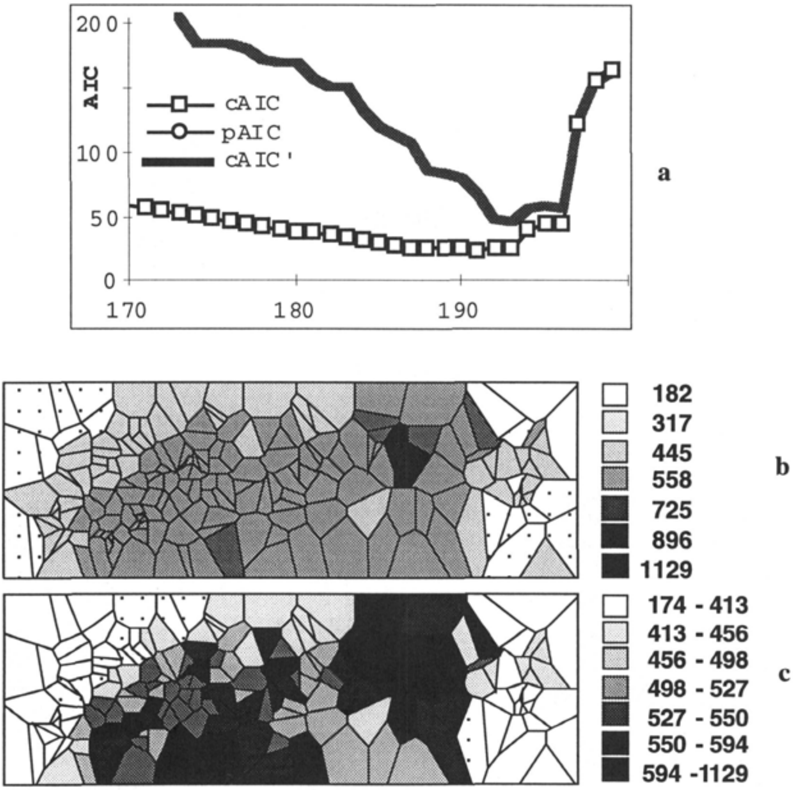


*FIGURE 6.*

An east-west cross section of the Adirondack Mountains, NY, with the Voronoi-polygons for 200 lakes from the Adirondack Lake Survey. A relatively smooth variable, elevation (m) is recorded as a surrogate for acid deposition to delineate variously impacted areas. Discrepancy values (last 30 iterations shown, (a)) for the different clustering procedures coincide at 7 classes (b). The 7-class equal-count choropleth map (c) gives a vastly exaggerated impression.

A subset of 200 lakes from the Adirondack Lake survey along the major elevation gradient is used in this test. Figure 6. summarizes the results, which indicate that even in case of relatively smooth variation, traditional choropleth mapping (simple histogram-partitioning) can lead to quite misleading results.

Case-2: In urban socio-economic research, the delineating regions (e.g., for market, services, voting behavior) often aims to identify "areas of action" or "areas of influence". Below, we show an example using percentage of unemployment based on 121 enumeration areas in the Greater Toronto Area. The three clustering procedures result in significantly different regionalizations (Figure 7.). Because of the small, intensively segmented southwestern section, according to pAIC, cAIC and cAIC' one would select 13 classes (77 patches), 8 classes (72 patches) and 3 classes (19 patches), respectively. The "closest" equal-step choropleth map is shown for comparison.

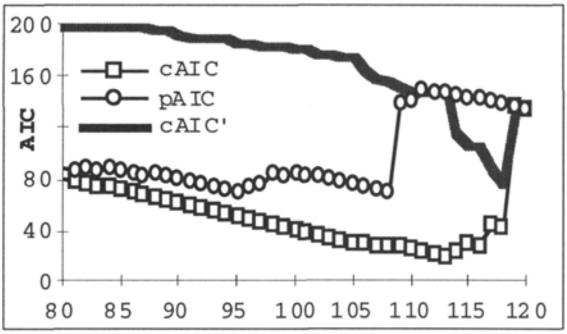
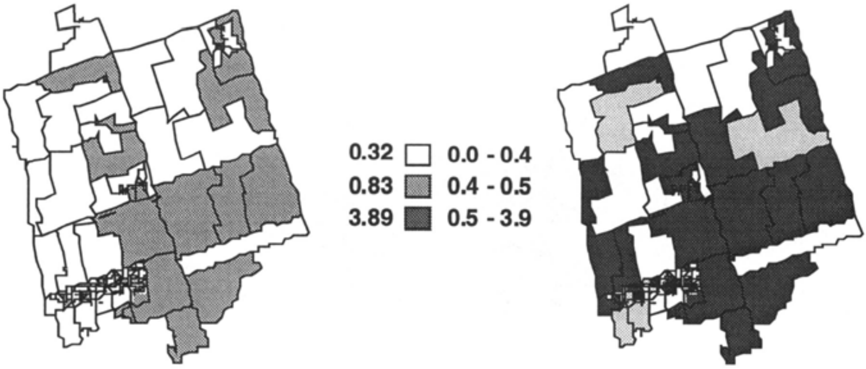


*FIGURE 7.*
Regionalization of %unemployment for 121 enumeration areas. The plot of AIC (left) for the clustering indicates vastly different patterns. The map corresponding to cAIC' and the "closest" equal-count choropleth map (below).



## CONCLUDING REMARKS

There are many conceptual models of geographical regions. When landscapes are represented by some variables attached to some data representation units, spatial statistical tools can be applied to "finding" homogeneous regions, especially when no other ancillary information (constraint, requirement) is available. Within this context we revisited the proposition of optimizing "information flow" (Jenks and Caspall, 1971) and compared three different measures of it using hierarchical (Ward) clustering.

In a simulation study, accounting for the spatial autocorrelation of noise, we found the modified, topologically sensitive Akaike information criterion a robust measure to avoid "overfitting" and moderately "reward" contiguous patches. The immediate next step should be to implement the CAR-based likelihood in AIC. The proposed type os measure is relatively simple to implement in commercially available software, at least to be used as guidelines in creating choropleth maps. It is also an advantage, that the computation is straightforward to extend to the multivariate case (i.e., regionalization based on more than one variable).

## *ACKNOWLEDGMENT*

We thank N. Fogarasi, an undergraduate at the University of Toronto, for programming. Parts of this research was conducted while S. Kabos visited Toronto for which the financial support of NSERC is gratefully acknowledged.

## *REFERENCES*

Akaike, H., 1973., Information theory and an extension of maximum likelihood principle. Second Int. Symp. Information Theory (eds. B. N. Petrov and F. Csaki), pp. 267-281. Akademiai Kiado, Budapest.

Cressie, N. A., 1993. Statistics for spatial data., J. Wiley, New York.

Cromley, R. G., 1996., A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data. Int.J. Geog. Info. Sys. 10:405-424.

Haining, R. P., 1990. Spatial data analysis in the social and environmental sciences. Cambridge University Press, Cambridge.

Jenks, G. F. and Caspall, F. C., 1971., Error on choroplethic maps: definition, measurement reduction. Annals of the Assoc. Amer. Geog. 61:217-244.

Kertesz, M., Csillag, F. and Kummert, A., 1996. Optimal tiling of heterogeneous images. Int. J. Remote Sensing 10

Linhart, H. and Zucchini, W., 1986., Model Selection. J. Wiley, New York.

Maguire, D., Goodchild, M. and Rhind, D., 1991. Geographical information systems: Principles and applications. J. Wiley, New York.

Monmonier, M., 1973., Analogs between class-interval selection and location-allocation models. The Canadian Cartographer. 10:123-131.

Robinson, A.H., Morrison, J. L., Muehrcke, P.C., Kimerling, A.J. and Guptill, S.C., 1995. Elements of cartography. J. Wiley, New York.

Schowengerdt, R. A., 1983. Techniques for image processing and classification in remote sensing. Academic Press, New York.

Stegena, L. and Csillag, F., 1986., Statistical determination of class intervals for maps. The Cartographic Journal 24:142-146.

Unwin, D., 1981., Introductory spatial analysis. Methuen, London.

Ward, J. H., 1963., Hierarchical groupings to optimize an objective function. J. Amer. Stat. Assoc. 58:236-244.