

MULTIVARIATE REGIONALIZATION: AN APPROACH USING INTERACTIVE STATISTICAL VISUALIZATION

Jonathan R. Hancock
Department of Geography
Syracuse University
343 H.B. Crouse Hall
Syracuse NY, 13244
e-mail: jrhanoc@rodan.syr.edu

ABSTRACT

This paper discusses a new way to use computers in determining geographic regionalizations for the purposes of creating descriptive maps. A program for **computer aided regionalization** is described. Rather than relying on automatic algorithms to delineate regions, this method uses interactive statistical graphics to help the user gain an understanding of data and allows the user to group areas into regions in a trial-and-error fashion. An adaptation of the frame-rectangle symbol, called a "multiple-frame" symbol, is used to communicate the attribute values of areas and the statistical properties of regions. The chief advantages of this approach are that it is flexible, it increases the user's awareness of the data, and it assists in the labeling of regions.

REGIONALIZATION AND DESCRIPTION

Geographic regionalization¹ may be defined as grouping contiguous areas, or more precisely basic spatial units (BSU's), into regions such that *something can be said* about each region. This implies that areas within a region have something in common, such as similar attribute characteristics. In general, regionalization is based on the statistical interpretation of one or more attribute variables from such sources as census, electoral, or property tax data. The method described in this paper, for example, uses the statistical measures of mean and variance and allows the user to consider as many as seven attribute variables.

Regions are delineated for a variety of purposes, and often these purposes require optimizing or standardizing the regions according to well-defined criteria. In election districting, for example, the primary goal is to create regions that are as equal in population and as compact as possible. Regionalizations such as this can be called *functional* because the derived maps have practical importance -- in this case, they determine where people may vote. In contrast, *descriptive* regionalizations are used in making maps that communicate geographical ideas or illustrate spatial

¹Regionalization differs from classification in that the members of each region must be contiguous.

patterns. Descriptive maps of labeled regions are common in atlases, newspapers, journals, textbooks, and on television; examples are maps of ethnic neighborhoods, political regions, and socio-economic zones. Appropriately, functional regionalizations must be generated according to strict rules; descriptive regionalizations, on the other hand, tend to be more subjective and are not always made in a systematic or scientific fashion. Nonetheless, descriptive regionalizations can be influential and should express geographic facts as accurately as possible. The method introduced in this paper provides a new way of creating such descriptive maps of regions.

Of course, descriptive maps do not need to contain regions -- there are other techniques for conveying spatial information through maps. Examples are classed maps such as choropleth or isarithmic maps, flow maps, and maps of point phenomena. In many instances, however, especially when several attribute variables are considered, a map of labeled regions is the simplest and most effective way to communicate geographic ideas. Maps of labeled regions do not require any knowledge of cartographic techniques and do not require the reader to decipher any cartographic symbols. Also, a map of labeled regions tends to be more memorable and emphatic than other types of maps because it forces the reader to associate shapes (regions) with words (labels).² Thus, while maps of labeled regions are less precise and objective than other types of maps, they are easier for the layman to understand.

THE DIFFICULTIES OF REGIONALIZATION

Over the last few decades, geographers have developed a variety of methods for generating regionalizations. The difficulty is that these methods are based on different theories and they yield varied results. Most researchers do not have the opportunity to try different approaches and tend to stick with one method. Often regionalizations are defended on the basis of the method used to create them rather than on the statistical qualities of the derived regions, and researchers seem to have too much faith in the methods they have chosen. Even if a researcher did try different methods, he or she would find it difficult to compare the results and ascertain which method yields the "best" solution, because there is no universally accepted way of evaluating or comparing regionalizations.

Many of regionalization methods use automatic algorithms to try to maximize the homogeneity of regions.³ In statistical terms, this is the

²For more on the importance of associating verbal descriptions with regions, refer to Rodoman.

³There are fundamentally different ways of creating regionalizations including maximizing the difference between regions, maximizing the contrast at borders between regions, and maximizing the spatial autocorrelation of a regionalization.

same as minimizing the variance within regions.⁴ The method introduced in this paper also has the goal of maximizing regional homogeneity, but it does not use a solution-finding algorithm. The problem with the algorithmic approaches is that they might miss subtle but important patterns in the data or might emphasize the wrong criteria in judging regions. Also, there may be additional, perhaps even non-quantifiable, factors that the regionalizer wishes to consider in setting up his or her regionalization.⁵ An automatic algorithm is not equipped to consider factors that are not clearly defined and numerically expressed. In short, an automatic algorithm may be too inflexible.

A NEW APPROACH TO REGIONALIZATION

I propose a non-algorithmic approach to regionalization which takes advantage of the computer to speed up and facilitate the delineation of regions but does not actually determine regions. I call this approach **computer aided regionalization**, and it is the basis of a program I am developing named *Look 'n Link*.⁶ It is important to emphasize that computer aided regionalization is *not* automatic; it is an environment that enables the user to delineate regions in a trial-and-error fashion by providing interactive feedback concerning the quality (i.e., homogeneity) of regions as they are being created.

This regionalization method is open ended -- the process is complete when the user is satisfied with the quality of the regionalization and the degree of aggregation. The number of regions, the degree of homogeneity, and the amount of time spent deriving regions are entirely up to the user's discretion. As a result, computer aided regionalization cannot guarantee that the regions will be homogeneous or compact,⁷ and it could be used to create bad regionalizations. Regionalizations made in this manner cannot be defended on the basis of how they were generated; they can only be defended on the basis of the statistical properties of the results.

With *Look 'n Link*, the user builds regions by repeatedly joining adjacent areas or regions. The program starts off by displaying a base map showing the areal unit boundaries. Centered over each area is a multiple-variable graphic symbol, called a **multi-frame** (described below), that indicates the values of several variables for each area. After viewing these

⁴Variance expresses how loosely the values of constituent areal units are dispersed about a region's mean (average) value, and hence describes the degree of heterogeneity within a region; it is defined as "the average squared difference between an observed value and the arithmetic mean." (Griffith and Amrhein, 1991)

⁵Examples of these non-quantifiable factors are personal impressions of regional patterns, traditionally accepted affiliations among areas, and major physical barriers.

⁶I call the program *Look 'n Link* because it allows the user to *look* at the map to gain an understanding of the data and then *link* areas together to create regions. The program is written in Think Pascal and runs on the Macintosh.

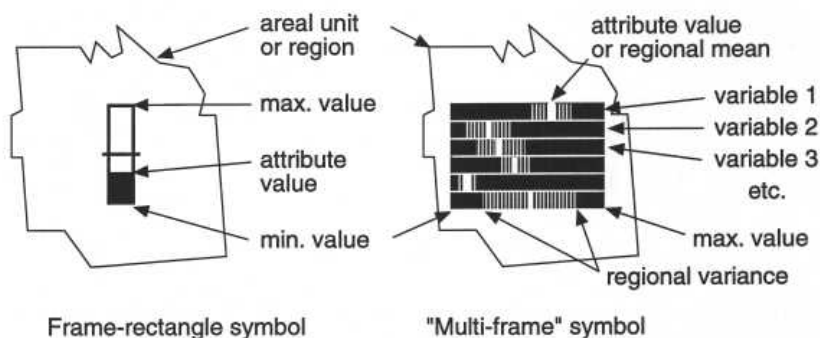
⁷It *does* guarantee that the regions will be contiguous.

symbols, the user selects a pair of adjacent areas to join and clicks on the boundary between these areas to "link" them. The two areas' multi-frames are replaced by a single multi-frame representing the combined region and indicating the values and variances for that region. Each time the user adds an area to a region he can see how that addition effects the region's homogeneity. The user can also remove an area from a region, so he or she is free to try out different combinations of areas.

THE MULTI-FRAME SYMBOL

I had to invent a way of displaying multivariate data so that it is easy to compare neighboring areas based on a set of variables. I rejected the idea of using multiple maps because this would force the user to switch back and forth between different views; I therefore needed some kind of multiple-variable point symbol. I considered using Chernoff's faces,⁸ but I felt that the way these symbols present variables (i.e. by the shapes of different physiognomic features) could distort the results because some people might notice noses more than eyes. I tried star symbols (glyphs),⁹ but when combined with a base map these created an overload of directional information. The best solution seemed to be an adaptation of the frame rectangle symbol.¹⁰

The operative metaphor for a frame rectangle symbol is that of a thermometer: the height of the "fluid" corresponds to the associated attribute value. Place several frame rectangles together and turn them on their side, and you have the **multi-frame symbol**.



Each of the horizontal bars in the multi-frame symbol corresponds to a particular attribute variable. The location of the vertical white stripe on a horizontal bar indicates the attribute value of an area (BSU) or the mean attribute value of a region (group of BSU's). If the symbol represents a

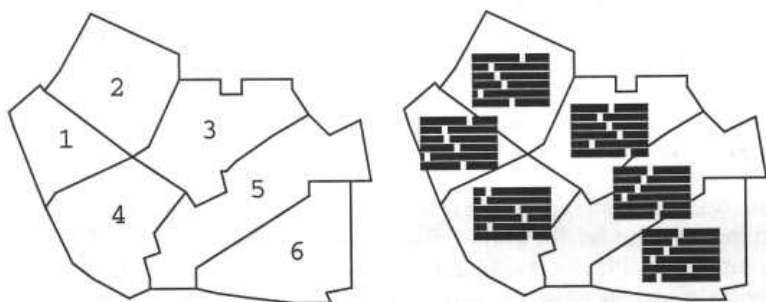
⁸See Feinberg, p. 173.

⁹See Dunn and Walker, p. 1400.

¹⁰The frame-rectangle symbol has been used by Mark Monmonier.

region, there may also be a shaded area around the white stripe, the size of which indicates the variance of the region with respect to that variable. This shaded zone can be thought of as a representation of the "fuzziness" of a region.

The different horizontal bars of the multi-frame are color coded to help the user associate them with their respective variables. The user can change the order and number of the variables represented, but all multi-frames will display the same set of variables. In comparing multi-frame symbols, the user need not think about individual attribute values. Instead, he or she can focus on the collective configuration of white stripes, which I refer to as the **profile** of a region or area. The profile visually summarizes a region's attribute characteristics. To select areas for joining, the user scans the map for neighboring areas with similar profiles.

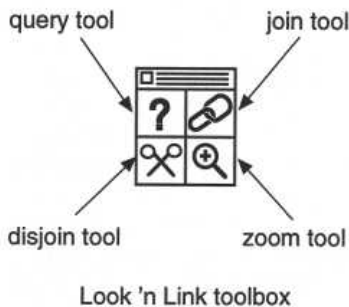


In the hypothetical example above, the user might start by joining areas 1 and 2 or areas 5 and 6 because these pairs of regions have the most similar profiles.¹¹ Areas 1 and 2 seem to be the closest match. Areas 5 and 6 have significantly different values for the last variable, but are strikingly similar in terms of the other variables. The multi-frames for areas 3 and 4 have some similarities, especially with regard to the last four variables.

CREATING REGIONS

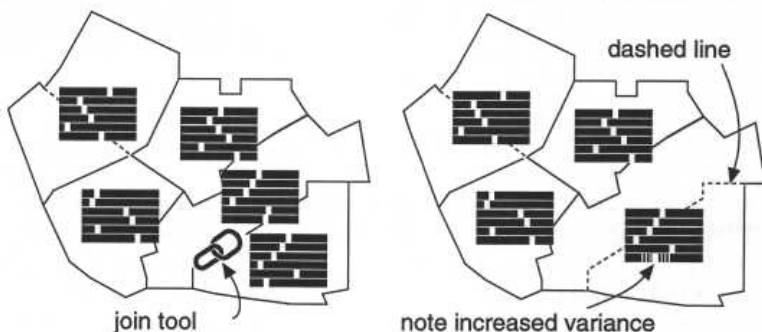
Like other Macintosh applications, *Look 'n Link* features a graphic menu or "toolbox" that allows the user to select one of several modes of operation or "tools". When a tool is selected, the appearance of the cursor changes accordingly. The chosen tool is positioned with the mouse and activated by pressing the mouse button. The use of **zoom tool** (magnifying glass) is obvious -- it adjusts the scale of the displayed map so that the multi-frame symbols (which remain constant in size) do not obscure each other or the areal boundaries.

¹¹These multi-frame symbols do not indicate any variance because they represent BSU's, not regions.



The **query tool** (question mark) is used to access information about the map. When a user clicks on an area or region, a balloon appears containing information such as the names and populations of the areal unit(s). By clicking within a multi-frame symbol, the user can see a display of a variable's name, numerical expressions of the regional mean and variance, and the name and value of the weighting variable (if

appropriate). For more detailed information on regions, the user may print out statistical summaries that list the regional means, variances, minimums, and maximums for all variables.









Regions are created with the join and disjoin tools. The **join tool** (chain links) is for connecting two adjacent areas to create a region, which the user does by clicking on the boundary between the areas. The boundary changes from a solid line to a dashed line and the two areas' multi-frames are replaced by a single multi-frame for the composite region. The figure above shows what happens when areas 5 and 6 are joined; note that the multi-frame indicates an increase in variance for the last variable. Only adjacent (contiguous) areas may be linked; this limitation is necessary to guarantee that regions remain contiguous.¹² One can undo a join between two areas by clicking the **disjoin tool** (scissors) on a dashed boundary.

¹²While most thematic mapping packages store boundary information in a non-topological ("spaghetti") format, *Look 'n Link* requires the more complicated topological (point-line-polygon) data model because the program uses chain processing to determine the areas or regions along a boundary, to alter the appearance of boundary segments, and to check the contiguity of areas.

LABELING THE REGIONS

While the program described here could be used merely as a data exploration tool, it is intended to help people create presentation maps, although it will not actually print them out¹³. The user is discouraged from using the multi-frame symbol on a static map; it would be too hard for most audiences to decipher. Instead, the map maker should translate these symbols into concise and understandable *labels*. Writing labels is far from trivial, and may be the most crucial step in creating a good map of regions. To determine these labels, the map maker must study the multi-frames and the statistical summaries for each region.

Var. ID	Variable Name	Regional Mean	Regional Variance		
1	%Mondale84	28.2	8.3	1	
2	%Reagan84	62.1	9.7	2	
3	%Dukakis88	46.5	25.1	3	
4	%Bush88	53.2	25.0	4	
5	%Clinton92	55.5	6.7	5	
6	%Regan92	31.2	11.1	6	

In this hypothetical example, a researcher is interested in political regions, or more specifically, regionalization based on voting in the last three presidential elections. Illustrated here are the statistical summary and multi-frame for one particular region. This region was strongly pro-Reagan in 1984 and was pro-Clinton in 1992. The variances for the 1984 and 1992 election results are relatively low, implying that this region is fairly homogeneous in terms of the voting patterns for these years. The situation for 1988 is different. Not only is Reagan's margin of victory slimmer, but the variances are much higher. This means that the region is not as consistent with regard to 1988 voting patterns. In describing this region, the user should focus on those years for which the variances are low, ignoring the year with high variances. Thus, he or she might label the region "REAGAN IN '84, CLINTON IN '92." Other regions on the same map might be labeled, "CONSISTENTLY REPUBLICAN," "REAGAN REPUBLICAN," or "DUKAKIS IN '88." To justify his or her selection of regions and regional labels, the user could append to the map detailed summaries of the regional statistics .

COMMENTS

Some might criticize computer aided regionalization on the basis that it is too "subjective." R.J. Johnston demonstrated, however, that the so-called "objective" approaches to classification are actually subjective, mainly because the results of regionalization depend largely on the choice of the

¹³The program will allow users to export regional boundary data to an illustration package, such as FreeHand, wherein labels, legends, titles, etc. may be added.

method used (Johnston, 1968). Many of these objective methods themselves have subjective aspects, such as the choice of indices or factors, the choice of a cut-off level in hierarchical grouping, or the choice of significance levels in methods involving inferential statistics. Most regionalization algorithms fail to adequately deal with these unintended subjective influences. In contrast, computer aided regionalization welcomes subjective influences. The way the user creates regions (i.e., based on perception) is subjective; on the other hand, the way the computer creates the multi-frame symbols (i.e., using statistical measures) is objective. Computer aided regionalization is thus a compromise between an objective and a subjective approach.

A legitimate criticism of computer aided regionalization, as well as other approaches to regionalization, concerns the aggregation problem.¹⁴ Several geographers, most notably, S. Openshaw, have written about the aggregation problem, also known as the modifiable areal unit problem (MAUP).¹⁵ In a 1991 article, A.S. Fotheringham and D.W.S. Wong wrote that,

The modifiable areal unit problem is shown to be essentially unpredictable in its intensity and effects in multivariate statistical analysis and is therefore a much greater problem than in univariate or bivariate analysis. The results of this analysis are rather depressing in that they provide strong evidence of the unreliability of any multivariate analysis undertaken with data from areal units. (Fotheringham and Wong, 1991, p. 1025)

My attitude about the aggregation problem is that it *is* rather depressing, but, like a lot of things in life, you just have to accept it and go on. The only consolation I can offer is that at least computer aided regionalization does not presume to be entirely objective, so it cannot be said that MAUP spoils an otherwise "valid" result. On the other hand, it is essential that the user be aware of the aggregation problem and that he or she qualify his interpretation and acknowledge the limitations that are due to the given set of areal units.¹⁶

Some might say that computer aided regionalization, because of its inherent flexibility, could make it easier for ill-intentioned individuals to create gerrymandered or otherwise deceptive maps. My response is that, if people want to lie or cheat with maps, they'll find a way to do it whatever software they're using. A computer program cannot be a policeman! *Look 'n Link* does indicate the reasonableness of a regionalization, so at least one could not make a bad map without being aware of it. In contrast,

¹⁴For most applications, data is available only for a predefined set of enumeration districts (such as census tracts or counties) because the census bureau or other agency, for reasons of confidentiality, does not release individual level data. Therefore, while it is possible to know the total or mean values for each BSU, it is impossible to know how those values are distributed, spatially or otherwise, within that BSU.

¹⁵See Openshaw, 1981.

¹⁶I thank Mark Monmonier for this insight.

many of the automatic approaches provide no mechanism for verifying how good the results are. Computer aided regionalization encourages a healthy awareness of the quality of the results, and such awareness should engender honesty. I might also respond that automatic regionalization algorithms may be unethical because they defer interpretation to a machine. Are we to let computers define our world and describe our society? Computers can be wonderful tools, but they should not determine how we understand ourselves.

Whatever method is used -- automatic or non-automatic -- a regionalization necessarily involves some loss of information. It is the responsibility of the map maker to see that this loss of information is not harmful. Nonetheless, any regionalization, whether automatically derived or not, should be critically evaluated and viewed with an appropriate measure of skepticism.

REFERENCES

- Dunn, R., and R. Walker. 1989. District-level variations in the configuration of service provision in England: A graphical approach to classification. *Environment and Planning, A* 21:1397-1411.
- Feinberg S.E. 1979. Graphical methods in statistics. *The American Statistician* 33:165-178.
- Fotheringham A.S., and D.W.S. Wong. 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning, A* 23:1025-1044.
- Gilbert, E.W. 1960. The idea of the region. *Geography* 45:157-75.
- Griffith, Daniel A., and C.G. Amrhein. 1991. *Statistical Analysis for Geographers*. Englewood Cliffs, NJ: Prentice Hall.
- Harley, J. B. 1991. Can there be a cartographic ethics? *Cartographic Perspectives* 10:9-16.
- Johnston, R.J. 1968. Choice in classification: The subjectivity of objective methods. *Annals of the Association of American Geographers* 58:575-589.
- MacEachren, A.M. 1991. The role of maps in spatial knowledge acquisition. *The Cartographic Journal* 28:152-162.
- MacEachren, A.M., and M. Monmonier. 1992. Introduction. *Cartography and Geographic Information Systems* 19(4):197-200.

- Monmonier, Mark 1991. *How to Lie with Maps*. Chicago: University of Chicago Press.
- Openshaw, S. 1977. A geographical solution to the scale and aggregation problems in region building, partitioning and spatial modeling. *Transactions, Institute of British Geographers* n.s. 2:459-72.
- Openshaw, S. 1984. Ecological fallacies and the analysis of areal census data. *Environment and Planning, A* 16:17-31.
- Openshaw, S., and P.J. Taylor. 1981. The modifiable areal unit problem. In *Quantitative Geography: A British View*, N. Wrigley and R.J.Bennet, eds., Boston: Routledge & Kegan Paul.
- Rodoman, B.B. 1967. Mathematical aspects of the formalization of regional geographic characteristics. *Soviet Geography* 8:687-708.