

accuracy will also decrease. Consequently, the mapmaker must use some subjectivity in determining what magnitude of GADF is acceptable for a particular application.

5.2.2 Spatial Context Only

Another approach for simplifying map pattern is to adjust the values of enumeration units as a function of values of surrounding units, an idea developed by Waldo Tobler (1973) and promoted by Adrian Herzog (1989). This notion is based on the concept of **random error**: If we repeatedly measure a value for an enumeration unit, we will likely get a different value each time (think of trying to calculate the population for a census tract for the 2000 Census—numerous variables could affect your result). This suggests that some change in the data we have collected is permissible (Clark and Hosking 1986, 14). Moreover, the spatially autocorrelated nature of geographic data (see section 3.4.2) provides a mechanism for making adjustments: Nearby units can assist in determining appropriate values for a particular enumeration unit because similar values are likely to be located near one another.

Although this approach could be used with a wide variety of data, it is easiest to implement for proportion data (e.g., the proportion of adults who smoke cigarettes) because the statistical theory for such data is well known. To illustrate, consider the hypothetical portion of a map shown in Figure 5.9, where it is assumed we wish to change the value for the central enumeration unit as a function of the surrounding units. The simplest formula for determining the value of the central unit would be to

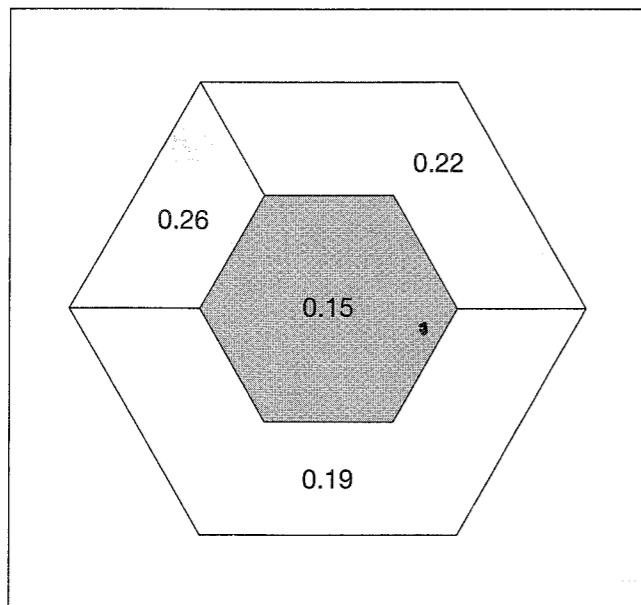


FIGURE 5.9 A hypothetical set of enumeration units for which a simplified value is desired for the central enumeration unit.

average all four values, but such a formula would not consider two factors: (1) that we have presumably taken some care in collecting the data for the central unit (and thus would like to place greater weight on it), and (2) that those units having a longer common boundary with the central unit should have greater impact. Thus, a more appropriate formula is

$$V_e = W_c V_c + W_s \left(\sum_{i=1}^n \frac{L_i}{L_T} V_i \right)$$

- where V_e = the estimated value for the central unit
- V_c = the original value for the central unit
- V_i = the original value for the i th surrounding unit
- W_c = the weight for the central unit
- W_s = the weight for the surrounding units
- L_i = the length of the boundary between the i th unit and the central unit
- L_T = the total length of the central unit boundary
- n = the number of surrounding units

If we assume W_c and W_s values of .67 and .33, respectively, then for Figure 5.9 the formula would be calculated as follows:

$$V_e = .67(.15) + .33 \left(\frac{1}{6}(.26) + \frac{2}{6}(.22) + \frac{3}{6}(.19) \right) = .17$$

Herzog indicated that a more complete algorithm should involve several other considerations. First, some limit should be placed on how much change is permitted to the central unit. For proportion data, Herzog suggested that the central value should not be moved outside of a 95 percent confidence interval around its original value (see Burt and Barber 1996, 272–274, for a discussion of confidence intervals for proportions). Second, surrounding values differing dramatically from the central value should not be used in the calculations. For proportion data, Herzog suggested a difference of proportions test (see Burt and Barber 1996, 322–324), with proportions differing significantly from the central unit not being used. Third, the process should be implemented in an iterative fashion, meaning that the entire map should be simplified several times; for example, this process could continue until all enumeration units reached the bounds of their confidence intervals or the largest change made in a unit was less than a small tolerance. Finally, the map resulting from the iterative process should be further smoothed by combining areas with nearly equal values.

Figure 5.10 illustrates the net effect of using Herzog's approach for the percentage of students in each commune of Switzerland who study at the University of Zurich

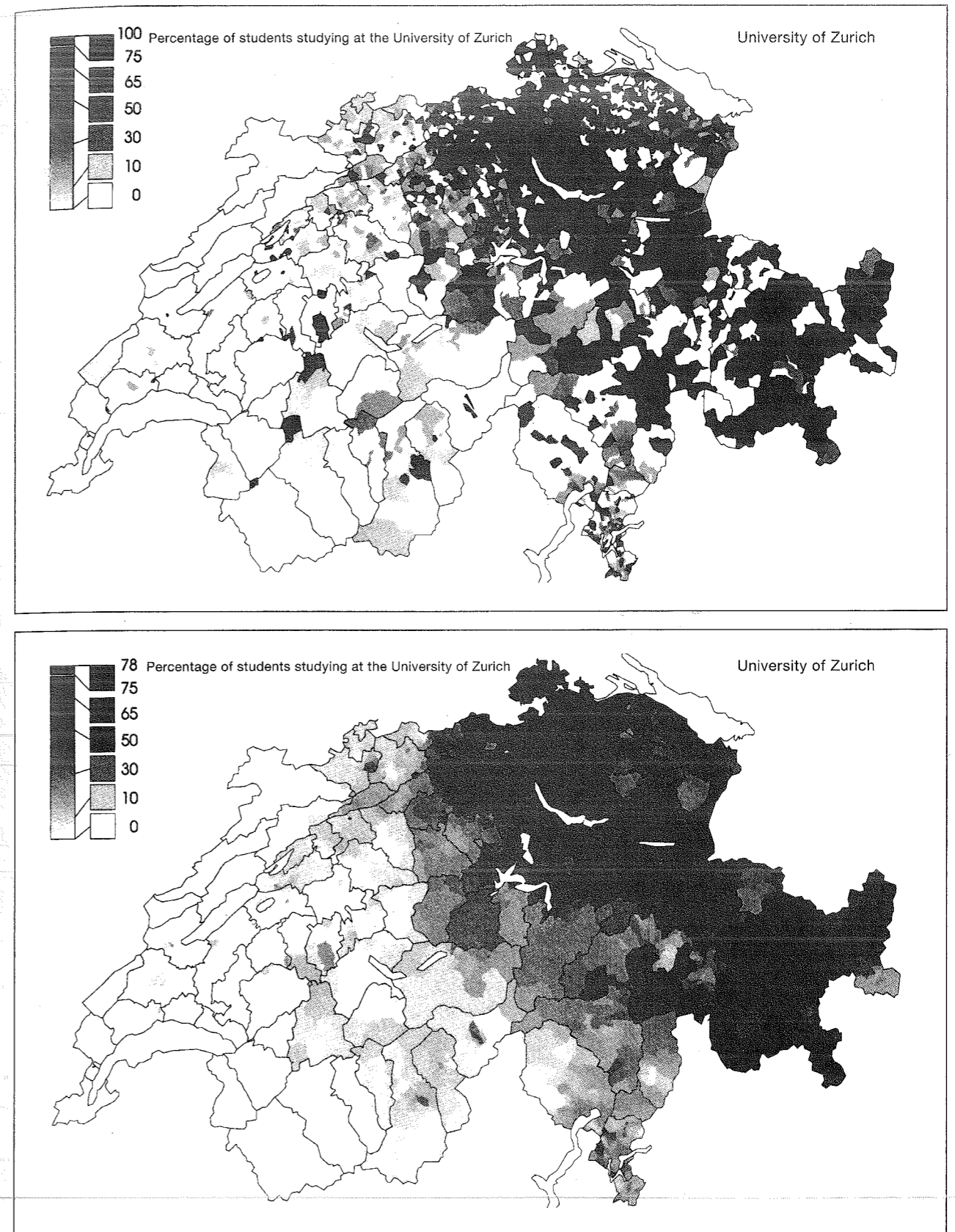


FIGURE 5.10 A comparison of an original unclassed map (top) and a simplified map (bottom) resulting from applying Herzog's method for simplifying choropleth maps. (An adaptation of Figure 1 from p. 214 of *Accuracy of Spatial Databases*, M. Goodchild and S. Gopal (eds.); courtesy of Taylor & Francis.)

Zurich: The top map is an original (unsimplified) unclassified map, and the bottom map is the simplified result. Although the bottom map is still unclassified, note that it is much simpler to interpret than the top map.

It should be noted that the methods we have examined for simplifying map appearance are relatively new and have not yet been incorporated in commercial software. Assuming that such methods do become readily available, it will be interesting to see to what extent they are used, and what impact this has on the use of more traditional classification methods that do not consider spatial context.

5.3 CLUSTER ANALYSIS

In the preceding sections of this chapter, we considered classification methods appropriate for a single numeric attribute. Here we consider classification methods appropriate for multiple numeric attributes. To illustrate, consider the data shown in Table 5.8 representing the value of various attributes for counties in New York State (excluding New York City) recorded by the U.S. Bureau of the Census for the 2000 Census. Here our interest is in whether there are certain counties that have similar scores on one or more of these attributes. For instance, you might wonder whether there are counties that have a high percentage of African Americans, a high infant mortality rate, and a decreasing population (a negative percent population change). You might tackle such questions by visually comparing separate maps of the attributes, but this would require that you mentally overlay the maps to derive various combinations of attributes for particular counties. More useful would be a method that combines the maps for you and tells you which counties are similar to one another and how they are similar—this is the purpose of the mathematical technique of **cluster analysis**, which we describe in this section.

Cluster analysis techniques can be divided into hierarchical and nonhierarchical methods. *Hierarchical* methods begin by assuming that each observation (county in the case of the New York data) is a separate cluster and then progressively combine clusters.* The process is hierarchical in the sense that once observations are combined in a cluster they remain in that cluster throughout the clustering procedure. *Nonhierarchical* methods presume a specified number of clusters and associated

* Throughout this section we assume that we are clustering enumeration units, an areal phenomenon. It is possible, however, to cluster other forms of phenomena, such as points—thus, we have used the generic term *observation*.

members, and then attempt to improve the classification by moving observations between clusters. Because hierarchical methods are a bit easier to understand and more common, we focus on them here.

5.3.1 Basic Steps in Hierarchical Cluster Analysis

To summarize the process of cluster analysis, we utilize an eight-step process that is a modification of six steps recommended by Charles Romesburg (1984).

Step 1. Collect an Appropriate Data Set

The first step is to collect a data set that you wish to cluster. Ultimately, we will cluster the data shown in Table 5.8, but for illustrative purposes we also will work with the small hypothetical data set for livestock and crop production listed in Figure 5.11A and plotted in graphical form in Figure 5.11B.

Romesburg (38) indicates that cluster analysis can be used for three purposes:

- To create a scientific question.
- To create a research hypothesis that answers a scientific question.
- To test a research hypothesis to decide whether it should be confirmed or disproved.

Because we were relatively unfamiliar with the census data for New York, we wanted to use it to create a scientific question—to explore the patterns that might arise. Those knowledgeable about the geography of New York might use the resulting patterns to look for other attributes to explain the results or consider other attributes that might be suitable for a cluster analysis.

Step 2. Standardize the Data Set

The second step in cluster analysis is to standardize the data set, if necessary. Two forms of standardization are possible. One is to account for enumeration units of varying size. Just as for choropleth maps (see section 4.4.1), this standardization is appropriate as larger enumeration units are apt to have larger values of a countable phenomenon.† For the New York State data, we standardized the raw number of African Americans living in each county by dividing by the total population of each county. For our small hypothetical agricultural data set, we will assume the data are already standardized.

A second form of standardization is to adjust for different units of measure on the attributes. For instance

† If data associated with points (as opposed to areas) are clustered, it might also be appropriate to standardize the data (see section 16.1 for appropriate methods).

TABLE 5.8 Cluster analysis

County	%
Albany	
Allegany	
Broome	
Cattaraugus	
Cayuga	
Chautauqua	
Chemung	
Chenango	
Clinton	
Columbia	
Cortland	
Delaware	
Dutchess	
Erie	
Essex	
Franklin	
Fulton	
Genesee	
Greene	
Hamilton	
Herkimer	
Jefferson	
Lewis	
Livingston	
Madison	
Monroe	
Montgomery	
Nassau	
Niagara	
Oneida	
Onondaga	
Ontario	
Orange	
Orleans	
Oswego	
Otsego	
Putnam	
Rensselaer	
Rockland	
Saratoga	
Schenectady	
Schoharie	
Schuyler	
Seneca	
Steuben	
St. Lawrence	
Suffolk	
Sullivan	
Tioga	
Tompkins	
Ulster	
Warren	
Washington	
Wayne	
Westchester	
Wyoming	
Yates	

