DELINEATION OF DEMOGRAPHIC AREAS AND THE CONTIGUITY RATIO

By: James M. Beshers, Purdue University

In this paper several problems in defining and delineating demographic areas will be discussed, and the potential value of the contiguity ratio in solving these problems will be indicated.[1] Let us regard a demographic area as a bounded area with a set of distinctive demographic characteristics, as compared with neighboring areas. Thus a demographic area is defined in terms of the characteristics of sub-areas which lie within and outside any given area. Therefore every demographic area should be delineated in terms of (1) the distribution of demographic characteristics within a larger area, of which the particular area is a part, and (2) the distribution of demographic characteristics within sub-areas of this area.

Note that hierarchies of areas can exist—the area at one level of discussion may become the sub-area at the next higher level. Census tracts and their combination in urbanized areas may each in turn be examples of demographic areas. However, the implications of this paper are not limited to these areas.

We shall consider the usefulness of a particular delineation of areas for research purposes. Criteria should be developed for the suitability of a given delineation for research purposes. Such criteria would have two functions: first, they would serve as a guide to the delineation of new areas, and second, they would enable us to evaluate existing areas for their appropriateness in research. We may evaluate existing areas without reference to the original purposes for which they were delineated—whether for administrative or for research purposes.

But we must see immediately that there is no general solution to this problem. No single criterion can be posed for a given set of areas. As the purposes of given research projects differ, so must the criteria which determine the usefulness of a set of areas for these purposes. In particular, the variables of importance will differ from project to project, as well as the statistics computed from these variables. The Shevky-Bell[2] literature asserts that there are three types of demographic variables with different spatial distributions. Nevertheless, we can consider the properties of criteria for particular variables and estimates.

Any criterion which we propose will be relative to the number and size of areas into which a larger area is to be subdivided. Assuming equal sized areas (either geographic size or population size), we must consider the specific number of areas, for if the number of areas is increased, then better delineations may become possible. There are two problems—first, determining the best delineation relative to a specific number of areas; second, determining the adequacy of this delineation with respect to the purposes of a specific project. If this delineation is not adequate, then the number of areas may be insufficient.

In order to illustrate these statements, we may consider two of the purposes for which a research project may use observations on areas. The area data may be used to obtain estimates on sub-units, either sub-areas or individuals, or the area data may be related to structural, or aggregate, concepts. In the former case, the appropriate criterion must be a function of the ratio of variation between areas to variation within areas, i.e., a function of the correlation ratio.[3] In the latter case a criterion is not so readily apparent. However, if areas meet a correlation ratio criterion, it seems likely that they meet most criteria appropriate for the latter case as well.

Although a correlation ratio criterion seems appropriate to the discussion of single variables, two problems remain. First, this criterion will vary from project to project as the need for greater or less accuracy differs. Translation of a correlation ratio criterion into a statement specifying the accuracy of conclusions based on certain areas might be a solution. Confidence intervals within areas could be constructed. Second, when two or more variables are considered, and the purpose of research is to estimate the relationship between these variables, then the covariance should be maximized by the delineation of areas.

A correlation ratio criterion may enable us to determine the best delineation of areas relative to a given number of areas. It may also permit us to evaluate this relative optimum in terms of the accuracy needed. Further work must be done on the latter possibility. The contiguity ratio, however, may shed some light upon the former problem (as well as many other demographic problems.)

The contiguity ratio, developed by R. C. Geary,[4] is an appropriate measure of the spatial clustering of characteristics of areas. The contiguity ratio is a two-dimensional generalization of the Von Neuman ratio used in time series analysis. The contiguity ratio compares the sum of squared deviations between the value for each area with the values for its contiguous areas summed over all areas in the numerator with the variance in the denominator. Constants are chosen so that the expected value for a random distribution is unity.

In Geary's notation, let the number of areas be n, the measure of the T-th area Z_T, with number of connections K_T. The contiguity ratio C is given by

$$C = \frac{n-1}{2k} \cdot \frac{\sum_T \sum_{T \neq T'}{}' (Z_T - Z_{T'})^2}{\sum_T (Z_T - \bar{Z})^2}$$

where

$K = \sum' K_T$ — T′ designate the areas contiguous to area T,

$\sum$ = sum over all counties

$\sum'$ = sum over contiguous counties — $Z_{T'}$ is the measure of the contiguous area,

Suppose we have three areas lying end-to-end, with values 5, 4, and 3 for each area. Then we have

$$\frac{2}{4} \cdot \frac{4}{2} = 1$$

revealing no effect of contiguity.

Let us consider the uses of the contiguity ratio. By itself, the ratio tells us whether a single variable has a significant clustering effect. (Significance may be determined either by randomization or by classical normal theory.) Further, for a given area, the degree of clustering between several variables may be compared.

How does this measure of clustering relate to the delineation of demographic areas? Recall our definition of a demographic area as a bounded area with a set of distinctive demographic characteristics, as compared with neighboring areas. The measurement of clustering effects of characteristics therefore has a two-fold significance for the delineation problem. Distribution of characteristics by sub-areas has significance outside and inside a particular area. The fact of clustering itself, as may be demonstrated by the contiguity ratio, must be evidenced before an area may be delineated.

But the contiguity ratio may guide delineation more specifically when used in conjunction with regression analysis. After the existence of a clustering effect has been demonstrated, we may seek an explanation for this effect. Regarding the clustered variable as a dependent variable, we may select independent variables and compute a regression equation. The effects of the independent variables may be removed, and the residuals tested for a clustering effect. If the residuals are not clustered, then the independent variables have "explained" the clustering effect. Subject matter theory must supply the meaning of this "explanation."

If the independent variables are distance measures, then the regression is equivalent to fitting a surface to the original variables with the contiguity ratio employed as a criterion of "goodness of fit." These distance measures may be represented in rectangular co-ordinates or polar co-ordinates. The distance measures in the regression may be supplemented by classifications of areas, which may be introduced into the analysis by covariance methods.

By these means the gradient hypothesis and various zonal hypotheses of urban ecology may be tested. The amount of variation attributable to each effect may be determined. The contiguity ratio may be used to determine whether the clustering effect has been accounted for by these hypotheses.

Thus the contiguity ratio aids us directly in studying the distribution of demographic characteristics, and therefore in the delineation of demographic areas. But these methods may be turned to the evaluation of a given delineation as well.

Recall that we wish to determine the best delineation of areas relative to a given number of areas. The surface representing the distribution of a variable must be considered. If the variation can be represented by a continuous smooth surface, then almost any delineation of areas will be as good as the best delineation. But if the variation is characterized by sharp fluctuations— canyons, deep gorges, and isolated peaks, so to speak— then the delineation must be tailored carefully to these configurations. The contiguity ratio used with a regression surface provides a partial answer to this question. The "goodness of fit" of smooth surfaces can be evaluated by these techniques.

If our variables have smooth surfaces, then we may neglect detailed delineation problems, and concentrate our attention on insuring that a sufficiently large number of areas are used. If our variables have surfaces which are "almost smooth", then we may be able to smooth them out by increasing the number of areas. The choice of smooth surfaces is by no means an easy task.

In conclusion, we need criteria for the usefulness of areas. These criteria should derive from the consequences of using the areas, from the risk involved. These criteria should determine the best delineation relative to a given number of areas, and they should evaluate this "best" delineation. A correlation ratio criterion might be used for both of these problems. An alternative approach using the contiguity ratio calls attention to the smoothness of the surface of distribution of a variable. If the surface is smooth, then more attention should be paid to providing a large enough number of areas, and less attention should be paid to the particular delineation of boundaries for areas.

---

1 This discussion is drawn in part from the author's unpublished Ph.D. thesis, "Census Tract Data and Social Structure: A Methodological Analysis," University of North Carolina, 1957. Daniel O. Price, Rupert B. Vance and James A. Norton have been of assistance in formulating these ideas. This paper was presented at the Annual Meeting of the American Statistical Association, Dec. 1958, under the title, "The Definition of Population Clusters and the Contiguity Ratio,"

2 In particular, Social Area Analysis by Eshref Shevky and Wendell Bell, (Stanford: Stanford University Press, 1955).

3 The coefficient of intraclass correlation, rho, may be the best statistic for this purpose. See Leslie Kish, "Differentiation in Metropolitan Areas", American Sociological Review, 19 (August, 1954).

4 "The Contiguity Ratio and Statistical Mapping," by R. C. Geary, The Incorporated Statistician, Vol. 5, No. 3.