A NOTE ON NEIGHBORHOOD SIZE AND THE MEASUREMENT OF SEGREGATION INDICES*

Allen C. Goodman†

ABSTRACT. This note examines neighborhood segregation measures with respect to size and validity. Conventional measures, while related to within-neighborhood homogeneity, are not necessarily related to neighborhood size. An empirical test examines racial segregation for Baltimore in 1970 and 1980 using both census tract and specially formulated neighborhood aggregates. For both years, and for all measures of segregation, the values and trends are essentially unchanged by the level of aggregation.

In the more than 40 years that social scientists have been considering segregation in housing markets or in schools, for example, two concerns that have cast doubt upon findings have involved the size and validity of the units to be measured. Size involves the problem that if racial segregation, for example, is measured at the census tract level, the amount of segregation is usually less than if measured at the block group or the block level. The implication, then, is that the smaller the neighborhood aggregate, the more segregated the measure will be, with the household typically given as the ultimate in racial segregation.[1]

Validity concerns the fact that some neighborhood aggregations are either established arbitrarily (postal zones are probably the best example), or are kept the same in the interest of comparability over time, even though the neighborhood may have changed considerably. Thus, a set of neighborhood boundaries that might have been racially homogeneous 20 years earlier could, in principle, be totally invalid during the present period. Measures of neighborhood segregation based on such aggregations would be similarly compromised.

This note makes two major points, one theoretical and the other empirical. First, it shows that two conventional measures of neighborhood segregation, while related to within-neighborhood homogeneity, are not necessarily related to neighborhood size. Second, it examines racial segregation for Baltimore in both 1970 and 1980, using both census-based neighborhood measures, and other aggregates which are presumably more valid. For both years, and for all measures of segregation, it finds both the values and the trends to be essentially unchanged by the aggregations that are used.

*I thank Noga Lewin for her research assistance. I am grateful to Jurgen Friedrichs and Baruch Kipnis for their comments, but neither is responsible for any errors that may occur.

†Research Scientist, Center for Metropolitan Planning and Research, and Associate Professor of Political Economy, The Johns Hopkins University.

[1] Clearly, this assumes no racially mixed households.

Received June, 1984; accepted February, 1985.

1. SEGREGATION MEASURES

Consider three major measures of segregation that have been found in the literature.[2] The dissimilarity index, D, measures the amount of segregation in terms of absolute deviations of neighborhood racial composition from the area mean.[3] Algebraically, D compares the sum of the absolute deviations, weighted by neighborhood size, with the theoretical maximum, to give

$$D = \left(\sum_j T_j \,|\, p_j - p^* \,|\right) \Big/ 2Tp^*(1 - p^*) \tag{1}$$

where T_j is the number of people in neighborhood j, p_j is the racial percentage in neighborhood j, p^* is the overall minority percentage in the metropolitan area, and T is the total population size in the area. Clearly, if all p_j equal p^*, no segregation exists, and D equals zero. If all neighborhoods are either all white or all minority, D is easily shown to equal one.

Although D is an intuitively appealing measure, its mathematics are problematical. Theil (1972, pp. 68–70) demonstrates how it is impossible to decompose D into components that summarize segregation within and among neighborhoods, for example. Although many researchers still use D, other measures have been adopted, whose mathematical properties are more tractable.

The segregation index, S, works from the premise that the goal of integration is to avoid deviations from mean racial composition, and that the costs of such deviations increase with the square of the deviation. Algebraically, compare the sum of these deviations from the area-wide mean, with the theoretical maximum, to get

$$S = \left[\sum_j T_j (p_j - p^*)^2\right] \Big/ Tp^*(1 - p^*) \tag{2}$$

Once again, it is easy to show that with p_j equal to p^* in all cases, S equals zero, and with p_j equal to either zero or one (total segregation), S equals one.

Consider, however, two alternative clusters of neighborhoods, j and j' such that

$$S = \sum_j T_j (p_j - p^*)^2 \Big/ Tp^*(1 - p^*) \quad \text{and} \tag{2'}$$

$$S' = \sum_{j'} T_{j'} (p_{j'} - p^*)^2 \Big/ Tp^*(1 - p^*)$$

Measure S will be greater than or less than S', depending on the sizes of the numerators, or

$$S \gtreqless S' \quad \text{as} \quad \sum_j T_j (p_j - p^*)^2 \gtreqless \sum_{j'} T_{j'} (p_{j'} - p^*)^2 \tag{3}$$

[2] Each of the three measures presented is supported by a detailed literature. For a good exposition and discussion, see Zoloth (1976).

[3] Throughout this note, racial segregation will be discussed, although all of the measures are general to other types of segregation.

Assuming for simplicity that all neighborhoods are of equal size in either set of clusters, then $T_j = \overline{T}_j$ for all j neighborhoods, and $T_{j'} = \overline{T}_{j'}$ for all j' neighborhoods. Dividing both sides of the equation by jj' yields

$$S \gtreqless S' \text{ as Var } (j) \gtreqless \text{Var } (j') \tag{4}$$

where "Var" refers to the variance. Recall that the given clusters are made up of smaller units, and assume that the smallest unit is the city block (although households could provide trivially smaller units still). For the metropolitan area, then, the total racial variance can be expressed as

$$V = \text{Var } (j) + SSW(j) = \text{Var } (j') + SSW(j') \tag{5}$$

where $SSW\ (j)$ $[SSW\ (j')]$ refers to squared deviations within neighborhoods. That is, since the squared deviations of the individual blocks from the mean are constant, irrespective of the clustering, then $SSW\ (j)$ and $SSW\ (j')$, the sums of squares within the neighborhoods under the different clusterings, vary in the opposite way to Var (j) and Var (j'). Substituting, we find that

$$S \gtreqless S' \text{ as } SSW(j) \lesseqgtr SSW(j') \tag{6}$$

This result, of course, is one that underlies any analyses that use clustering algorithms. In other words, one wishes to minimize the sum of squares within the clusters, but this is unrelated to the sizes of the clusters. The increase in S that comes from disaggregating to block groups or blocks, from tracts, comes from the explicit geographical nature of the disaggregation. That is, within a given tract, a nested disaggregation is likely to decrease the sum of squared differences within the neighborhood. A different disaggregation which only decreases neighborhood size, need not do so.

Lest it be implied that this finding is unique to the specific segregation measure used, consider the information theory index, H, defined as

$$H = 1 - \sum_j T_j E_j \Big/ TE \tag{7}$$

where

$$E_j = p_j \log (1/p_j) + (1 - p_j) \log (1/(1 - p_j)) \tag{8a}$$

and

$$E = p^* \log (1/p^*) + (1 - p^*) \log (1/(1 - p^*)) \tag{8b}$$

If logarithms are taken to the base 2, then this index is constrained to the range between 0 and 1.

As before, consider clusterings H and H'. These two can be compared such that

$$H \gtreqless H' \text{ as } 1 - \sum_j T_j E_j \Big/ TE \gtreqless 1 - \sum_{j'} T_{j'} E_{j'} \Big/ TE \tag{9}$$

As with the variance-based S, H is also related to the entropy of all of the blocks. That is, the deviation of all blocks i from the area-wide mean can be decomposed into variation among neighborhoods, and variation within them

$$(10)\quad \sum_i T_iE_i = \underbrace{\sum_j T_jE_j}_{[T_j^A]} + \underbrace{\sum_j T_j \sum_{i\in j} (T_i/T_j)(E_i - E_j)}_{[T_j^W]}$$

$$= \underbrace{\sum_{j'} T_{j'}E_{j'}}_{[T_{j'}^A]} + \underbrace{\sum_{j'} T_{j'} \sum_{i\in j'} (T_i/T_{j'})(E_i - E_{j'})}_{[T_{j'}^W]}$$

Substituting (10) into (9) shows that H and H' differ, based on the within-neighborhood variation minimized. As with S, there is no necessary relationship to neighborhood size.

2. AN EMPIRICAL EXAMPLE

This section presents an empirical example of differences between census tracts and neighborhoods whose boundaries are supposedly more valid. Differences in the resulting segregation measures are very small, suggesting that for many types of socioeconomic analysis, the much maligned census tract aggregations provide information that can not be significantly improved upon, even at significant cost.

In 1978 and 1979, a team of researchers in Baltimore City sought to redefine city neighborhoods from the block level up.[4] The 278 resulting areas (including 27 areas without a representative community organization, and several public housing areas) often differed substantially from the 202 census tracts available. They were defined through extensive consultation with planners and neighborhood residents. The only constraints were that the blocks in a neighborhood had to be contiguous, and that blocks could not be split between neighborhoods.

Given the larger number of neighborhoods, the average "real" neighborhood was approximately 30 percent smaller than the average census tract. Moreover, one would expect that if this neighborhood aggregation was significantly better than the census tract aggregation, then within-neighborhood variation should be minimized, leading to substantially higher measures of segregation in those indicators that are built up from the block level.[5] Table 1 suggests that such expectations are not validated. For both 1970 and 1980, for all three segregation measures (D is included for completeness), the neighborhood-based segregation measures are only marginally greater than those at the tract level. It might be argued that using the 1978–1979 neighborhoods in 1970 is as invalid as using the tracts, but the 1980 measure shows the same set of differences.

For example, in 1970 S equaled 0.76 (0.78) at the tract (neighborhood) level. For 1980, these measures dropped to 0.68 (tract) and 0.70 (neighborhood). H fell in

[4] For details on this process, see Taylor, Brouwer, and Drain (1979).

[5] Due to census suppression rules, many measures, such as income or education, could be built only from block group aggregations, thus requiring an allocation process assuming homogeneity. Racial measures were, in fact, built up from the block level.

TABLE 1: Effect of Level of Aggregation on Segregation Measures

	Level of Aggregation	
Time	Tract[a]	Neighborhood[b]
	Segregation Index, S	
1970	0.76	0.78
1980	0.68	0.70
	Information Theory Index, H	
1970	0.70	0.72
1980	0.62	0.65
	Dissimilarity Index, D	
1970	0.84	0.86
1980	0.78	0.80

Source: Baltimore census data.

[a] Tracts are from 1970 Third Count and 1980 STF1 files.

[b] Neighborhoods are from reformulation of census data, as noted in Goodman and Taylor (1983).

a similar manner from 0.70 (0.72) at the tract (neighborhood) level in 1970, to 0.62 (0.65) in 1980. D, although not used in the proofs above, likewise provided very similar results; D equaled 0.84 (0.86) at the tract (neighborhood) level in 1970, and 0.78 (0.80) in 1980. These results are summarized in Table 1.

Given the care with which the neighborhoods were delineated, it is useful to ask why the differences in measurements were so modest. One answer is that although an extensive reformulation of neighborhoods might better fix neighborhood centers, the boundaries are still "fuzzy." Ultimately, drawing a line between neighborhoods that are reformulated may be every bit as arbitrary as drawing a line between census tracts.

3. CONCLUSIONS

This note has examined the relationship of neighborhood size to measurement of segregation, and shows that size makes difference only if it changes within-neighborhood homogeneity. This result holds for two major measures that admit decomposition into "within" and "among" components.

It then presents the results of a comparison between the supposedly arbitrary census tract boundaries, and a very carefully redefined set of neighborhoods for Baltimore City in 1978–1979. The redefined set of neighborhoods provides no essential change in the measurement of segregation. This suggests that for many measurements of neighborhood homogeneity, census tract level aggregations are no more arbitrary than are redefined aggregations.

From a planning viewpoint this is important, since it indicates that for many purposes, time-consuming and costly redefinitions of neighborhood boundaries may not be crucial to the understanding of segregation and other spatial processes.

This is reassuring in that census tracts present well-established and comparable aggregations for analyzing patterns across areas and over time.

REFERENCES

Goodman, Allen C. and Ralph B. Taylor. *The Baltimore Neighborhood Fact Book: 1970 & 1980.* Baltimore, MD: Center for Metropolitan Planning and Research, The Johns Hopkins University, 1983.

Taylor, Ralph B., Sidney Brouwer, and Whitney Drain. *Toward a Neighborhood-Based Data File.* Baltimore, MD: Center for Metropolitan Planning and Research, The Johns Hopkins University, 1979.

Theil, Henri. *Statistical Decomposition Analysis.* Amsterdam: North-Holland, 1972.

Zoloth, Barbara S. "Alternative Measures of School Segregation," *Land Economics,* 52 (1976), 278–298.