The Multigroup Entropy Index
(Also Known as Theil's H or the Information Theory Index)

John Iceland[1]
University of Maryland
December 2004

[1] These indexes were prepared under contract to the U.S. Census Bureau.

Table of Contents

Summary

This website contains three sets of residential-pattern indicators for 1980, 1990, and 2000.

1. The highlighted measure is the "multigroup entropy index," which is also known as the multigroup version of Theil's H or the multigroup information theory index. This is a measure of "evenness."
2. "Diversity" scores are also available; these are used in the calculation of the multigroup entropy index. A diversity score measures the extent to which several groups are present in a metropolitan area, regardless of their distribution across census tracts.
3. Dual-group entropy indexes are included here, where the reference group consists of all people not of the main group in question.[2] Two-group entropy indexes are computed for Non-Hispanic Whites, Non-Hispanic African Americans, Non-Hispanic Asians and Pacific Islanders, Non-Hispanic American Indians and Alaska Natives, Non-Hispanics of other races, and Hispanics.

Data Source

These indexes are based on data from the 1980, 1990, and 2000 decennial censuses (the 100 percent data). The main data issues involved in calculating racial and ethnic residential patterns revolve around the definition of racial and ethnic categories, geographic boundaries, and residential-pattern measures.

[2] Dual-group entropy indexes were also included in the 2002 report by Iceland, Weinberg, and Steinmetz. In that report, the entropy index indicated the segregation of each of several groups (Blacks, Hispanics, Asians and Pacific Islanders, and American Indians and Alaska Natives) from non-Hispanic Whites.

Race and Ethnicity

In 1977, the Office of Management and Budget (OMB) issued its Statistical Policy Directive 15, which provided the framework for federal data collection on race and ethnicity to federal agencies, including the Census Bureau for the 1980 decennial census. The OMB directed agencies to focus on data collection for four racial groups – White, Negro or Black, American Indian, Eskimo, or Aleut; and Asian or Pacific Islander – and one ethnicity – Hispanic, Latino, or Spanish origin. The questions on the 1980 and 1990 censuses asked individuals to self-identify with one of these four racial groups and whether they were Hispanic or not.[3]

After much research and public comment in the 1990s, the OMB revised the Nation's racial classification to include *five* categories – White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or other Pacific Islander. An additional major change was to permit the self-identification of individuals as "one or more races." While a small fraction of the population had already been doing so on previous census forms, this new directive made this practice permissible in data collection activities.

This change naturally challenges researchers to determine the best way to present historically-compatible data. To facilitate comparisons across time, minority race/ethnicity definitions that could be rather closely reproduced in the three different decades were used, and which closely approximate 1990 census categories. Six mutually exclusive and exhaustive categories were constructed: Non-Hispanic Whites, Non-Hispanic African Americans, Non-Hispanic Asians and Pacific Islanders, Non-Hispanic American Indians and Alaska Natives, Non-Hispanics of other races, and Hispanics. Having mutually exclusive and exhaustive

[3]The Population Censuses have a special dispensation from OMB to allow individuals to designate "Some Other Race" rather than one of those specifically listed. Because of Congressional directives, the decennial census questions also ask about specific Asian and Pacific Islander races (e.g., Chinese).

categories is essential for constructing a single multiracial index. For Census 2000, this involved combining the Asian and Native Hawaiian or other Pacific Islander groups. In addition, non-Hispanic people who identified themselves as being of two or more races in 2000 were also categorized as "Other" since people could not mark more than one race in 1980 or 1990. Census 2000 figures indicate that 4.6 million, or 1.6 percent of the population, designated themselves as multiracial (and non-Hispanic). Because of the relatively small number of multiracial people, the impact of the creation of this category in Census 2000 on segregation is small.[4] People who reported being Hispanic were categorized as such, regardless of their response to the race question.

Geographic Areas

Residential pattern indexes often measure the distribution of different groups across units within larger areas. Thus, to measure residential patterns, one has to define both the appropriate larger area and its component parts. The larger areas here are represented by metropolitan areas, as these are reasonable approximations of housing markets. These are operationalized by using independent and primary metropolitan statistical areas, referred to hereafter as metropolitan areas, or MAs. To facilitate comparisons over time, the definition of MA boundaries in effect during Census 2000 (issued by the Office of Management and Budget on June 30, 1999) were

[4] As a way of testing the sensitivity of the information theory index calculated here to differences in race categories, an alternative race classification scheme with the Census 2000 data was tested: instead of the six categories described above, eight were constructed. The two extra were created by splitting the Asian and Pacific Islander category into two (Asians, and Native Hawaiians and Other Pacific Islanders), and splitting the non-Hispanic Other category into non-Hispanic "Other," and non-Hispanics who marked two or more races. The mean entropy index for all 331 metropolitan in 2000 was 0.181 using six categories, and 0.180 using the eight categories, indicating the very small effect of using these two alternatives. The correlation between the two is over 0.99.

used. Minor Civil Division-based MAs were used in New England. To address the second geographic consideration, this analysis uses census tracts. These units are designed with the intent of representing neighborhoods, are delineated with substantial local input, and thereby a reasonable choice from a heuristic perspective.

In 2000, there were 331 MAs in the U.S. For this analysis, six MAs were omitted (Barnstable-Yarmouth, MA, Flagstaff, AZ-UT, Greenville, NC, Jonesboro, AR, Myrtle Beach, SC, and Punta Gorda, FL) because they had fewer than 9 census tracts and populations of less than 41,000 in 1980. All other MAs used had populations of at least 50,000 in 1980, which is typically one of the criteria for defining an area an MA.

Residential Pattern Measures

Residential pattern measures, usually referred to as "residential segregation" measures in the social scientific literature, have been the subject of extensive research for many years, and a number of different measures have been developed over time (e.g., see Massey and Denton, 1988; Iceland, Weinberg, and Steinmetz, 2002). Reardon and Firebaugh (2002) note that all major reviews of such indexes limit their discussion to dichotomous measures (e.g. Duncan and Duncan, 1955; James and Taeuber, 1985; Massey and Denton; 1988; White, 1986; Zoloth, 1976; Massey, White, and Phua, 1996). The earliest of the multigroup indexes is the information theory index (H) (sometimes referred to as the entropy index), which was defined by Theil (Theil, 1972; Theil and Finezza, 1971).

The entropy index is a measure of "evenness"—the extent to which groups are evenly distributed among organizational units (Massey and Denton 1988). More specifically, Theil described entropy index as a measure of the average difference between a unit's group

proportions and that of the system as a whole (Theil 1972). H can also be interpreted as the difference between the diversity (entropy) of the system and the weighted average diversity of individual units, expressed as a fraction of the total diversity of the system (Reardon and Firebaugh 2002).

The entropy score, which is a measure of diversity, and the entropy index, which measures the distribution of groups across neighborhoods, are discussed below. A measure of the first is used in the calculation of the latter. The entropy score is defined by the following formulas, from Massey and Denton (1988). First, a metropolitan area's entropy score is calculated as:

$$E = \sum_{r=1}^{r} (\Pi_r) \ln[1/\Pi_r]$$

where Π_r refers to a particular racial/ethnic group's proportion of the whole metropolitan area population. All logarithmic calculations use the natural log.[5]

Unlike the entropy index defined below, this partial formula describes the *diversity* in a metropolitan area. The higher the number, the more diverse an area. The maximum level of entropy is given by the natural log of the number of groups used in the calculations. With six racial/ethnic groups, the maximum entropy is log 6 or 1.792. The maximum score occurs when all groups have equal representation in the geographic area, such that with six groups each would comprise about 17 percent of the area's population. This is typically not referred to as a measure of "segregation" because it does not measure the distribution of these groups across a metropolitan area. A metropolitan area, for example, can be very diverse if all minority groups

[5] When the proportion of a particular group in a given census tract (Π_r) is 0, then the log is set to 0. This is the preferred procedure here, as the absence of a group (or multiple groups) should result in a 0 increase in the diversity score (where a higher score indicates more diversity).

are present, but also very highly “segregated” if all groups live exclusively in their own neighborhoods.

A unit within the metropolitan area, such as a census tract, would analogously have its entropy score, or diversity, defined as:

$$E_i = \sum_{r=1}^{r} (\Pi_{ri}) \ln[1/\Pi_{ri}]$$

where π_{ri} refers to a particular racial/ethnic group’s proportion of the population in tract i.

The entropy index is the weighted average deviation of each unit’s entropy from the metropolitan-wide entropy, expressed as a fraction of the metropolitan area’s total entropy:

$$H = \sum_{i=1}^{n} \left[\frac{t_i (E - E_i)}{ET} \right]$$

where t_i refers to the total population of tract i, T is the is the metropolitan area population, n is the number of tracts, and E_i and E represent tract i's diversity (entropy) and metropolitan area diversity, respectively. The entropy index varies between 0, when all areas have the same composition as the entire metropolitan area (i.e., maximum integration), to a high of 1, when all areas contain one group only (maximum segregation). While the diversity score is influenced by the relative size of the various groups in a metropolitan area, the entropy index, being a measure of evenness, is not. Rather, it measures how evenly groups are distributed across metropolitan area neighborhoods, regardless of the size of each of the groups.

Other multigroup segregation indexes exist, such as a generalized dissimilarity index and an index of relative diversity. In a detailed review of 6 multigroup indexes (dissimilarity, gini, entropy, squared CV (coefficient of variation), relative diversity, normalized exposure), Reardon

and Firebaugh (2002) conclude that the entropy index is clearly the superior measure. They note, for example, that entropy is the only index that obeys the "principle of transfers," (the index declines when an individual of group m moves from unit i to unit j, where the proportion of persons of group m is higher in unit i than in unit j). The entropy index can also be decomposed into its component parts. For these reasons, the entropy index was calculated here.

Dual-Group Entropy Indexes

In addition to the multigroup entropy index, indexes for particular groups are also available here. These employ a two-group entropy index (H) calculation, which uses the same formulas specified above, where the distribution of each of six groups in question (Non-Hispanic Whites, Non-Hispanic African Americans, Non-Hispanic Asians and Pacific Islanders, Non-Hispanic American Indians and Alaska Natives, Non-Hispanics of other races, and Hispanics) is compared to the distribution of all other groups combined. In other words, the reference group for these calculations consists of those who are not of the racial/ethnic group being considered. Additional discussion and analyses of these indexes is contained in Iceland (2004).

References

Duncan, Otis Dudley and Beverly Duncan. 1955. "A methodological analysis of segregation indexes." *American Sociological Review* 20: 210-17.

Iceland, John. 2004. "Beyond Black and White: Residential Segregation in Multiethnic America." *Social Science Research* 33, 2 (June): 248-271.

Iceland, John, Daniel H. Weinberg, and Erika Steinmetz. 2002. *Racial and Ethnic Residential Segregation in the United States: 1980-2000*. U.S. Census Bureau, Census Special Report, CENSR-3, Washington, DC: U.S. Government Printing Office.

James, David R. and Karl E. Taeuber. 1985. “Measures of segregation.” *Sociological Methodology* 14: 1-32.

Massey, Douglas S. and Nancy A. Denton. 1988. "The Dimensions of Residential Segregation." *Social Forces* 67:281-315.

Massey, Douglas S., White, Michael J., and Voon Chin Phua. 1996. "The Dimensions of Segregation Revisited." *Sociological Methods and Research* 25, 2 (November): 172-206.

Reardon, Sean F., and Glenn Firebaugh. 2002. “Measures of MultiGroup Segregation.” *Sociological Methodology* 32, 1 (January): 33-67.

Theil, Henri. 1972. *Statistical decomposition analysis*. Amsterdam: North-Holland Publishing Company.

Thiel, Henri and Anthony J. Finezza. 1971. “A note on the measurement of racial integration of schools by means of informational concepts.” *Journal of Mathematical Sociology* 1: 187-94.

White, Michael J. 1986. “Segregation and diversity measures in population distribution.” *Population Index* 52: 198-221.

Zoloth, Barbara S. 1976. “Alternative measures of school segregation.” *Land Economics* 52: 278-298.