Appendices to:

"Distinguishing the Geographic Levels and Social Dimensions of U.S. Metropolitan Segregation, 1960-2000."

by Claude S. Fischer, Gretchen Stockmayer, Jon Stiles, Michael Hout

***Demography*, February 2004**

Main Appendix: *Data Sources, Definitions, and Matching Issues*

Data for these analyses were drawn from summary files produced by the Bureau of the Census.[1] The requirements of our analyses demanded two things of the data. First, it required the existence of cell counts which were defined for a) common characteristics which captured the elements of race and ethnicity, class, and life cycle, over b) the entire period from 1960 through 2000,[2] or c) equivalent universes of individuals or households. The second principal requirement

[1] For 1960, we used Census Tract-Level Data obtained from DUALabs, Inc. and redistributed by ICPSR. For 1970, tract-level data was drawn from the *Fourth Count A* tallies for sample data (for all states except New Jersey, which was generously provided to us by Anne Gray from the Princeton Population Lab). For 1980 and 1990, we used the Summary Tape File 3A data, provided on the internet by CIESIN. Finally, for 2000, we are using the SF1 counts for the data items which can be drawn from the short form, which have been supplemented with SF3 for items drawn from the long-form. Some researchers have found large discrepancies between segregation measures based on sample data (e.g., SF3) and those drawn from full count files (e.g., SF1) for individual metropolitan areas with small (1%-2%) minority populations. We compared our results for 1980, 1990, and 2000 for measures available from both full count and sample tabulations. The resulting segregation measures were virtually identical for the two sources

[2] Originally, we addressed a longer span of censuses by employing the Elizabeth Mullen Bogue files for 1940, 1950, and 1960. Use of these files permit, after creating a map between metropolitan area codes, the tracing of segregation levels for selected MAs much further back. Unfortunately, because the extent of tracting within an area changed substantially over time and the coverage of MAs was incomplete, the years prior to 1960 were unsuitable for the types of analyses we used. (In 1940, for example, only tracted cities are included, while in 1950, only 59 of the 114 tracted areas are represented in the files).

was that these counts were available for units of geography and levels of geography that were consistent over the same period.

Each of these requirements appears simpler on the surface than in practice. Our first concern is with the availability and consistency of our measures. The most obvious way in which one of our measures will be inconsistent is if the cells of the table from which it was collected are chosen differently in different periods. Alternatively, the same cells may be collected, but the universe they are collected for may differ in some way. A third issue which can give rise to inconsistencies occurs when both the cells collected and the universe remain the same, but the underlying distribution changes in a way that the "same" cells now represent an analytically distinct group.

By and large, we selected measures for which the first problem either does not arise or can be handled by collapsing the data. In some cases, we wind up having to collapse into categories we might not prefer, but the final categories are consistent if not ideal. (One notable exception to this is the occupational categories, which were extensively revised between 1970 and 1980, and which, despite their centrality to social class, had to therefore drop from analyses.)

The second case arises for marital status variables. In both cases, the universe in earlier periods includes persons at younger ages than those in later periods. The exclusion of 14 year olds from the numerator and denominator for whom marital status is determined in later years would increase the percent never-married, but more-or-less universally across all areas. Hence, segregation measures which are not affected by prevalence (like entropy or dissimilarity) should not be unduly affected by this problem.

The final issue may arise for items like income. Clinging to specific dollar ranges for income measures may ignore or misstate the effects of inflation and will miss the effects of across-the-board increases in income which place similar dollar amounts at different locations in the income distribution over time. We chose to emphasize distributional issues, and examined the 20% at the top and bottom of the distribution.

Educational attainment, while collected in categories which can be made fairly consistent by collapsing more detailed categories, will similarly reflect a changing distribution over the period. For income, we chose to emphasize distributional issues, and examined the 20% at the top and bottom of the distribution, while for education we focused on the specific credentials. (Because of the difficulty in separating age and education effects, trends in educational segregation are not included in this paper.)

The second broad requirement of our analyses is consistently available geographies – regions, metro areas, central cities, suburban places, and tracts. The tract issues are possibly the most complex and are discussed separately in Appendix 3 below; inconsistencies in place availability required an estimate-based solution, which is documented in Appendix 2 below. Here we discuss only the decisions we made regarding definitions of comparable geographic units and the attribution of geographic identifiers.

Metropolitan areas are, Census Bureau definition, intended to be composed of a "core area containing a large population nucleus, together with adjacent communities that have a high degree of economic and social integration with that core." Metropolitan areas (as their first incarnations as SMAs) were first defined for the 1950 census. The number of MAs grew from

169 in that year to 331 in 2000. During the same period, the proportion of the U.S. population that lived in such areas rose from 56% to 80%. By and large, as the table below illustrates, growth in the percent of population covered is drawn from the creation of new areas rather than population growth in the previously defined areas. For each of the contemporary definitions of metropolitan areas for 1950 through 1980, population shares in the year of definition and 1990 are virtually identical. Nor, if we look at those areas prior to their definition as an MA, are population shares greatly lower.

Percent of US Population Living in:	**1960**	**1970**	**1980**	**1990**	**1998**
MAs as defined in 1950	59.3	60.1	56.9	56.2	54.9
MAs as defined in 1960	63.0	64.4	62.1	62.4	61.6
MAs as defined in 1970		68.6	66.9	67.5	67.1
MAs as defined in 1980			74.8	75.9	76.0
MAs as defined in 1990				77.5	77.7
MAs as defined in 1999					80.1

Source: Statistical Abstract of the United States, 2000. Population counts based on the territory bounded by each decennial census years' defined metropolitan boundaries; these counts divorce territory changes from shifts of population to or from metropolitan areas. This source does not list the percent of the population living in 1950 MA's, but Bogue (1959) indicates that that the figure is 56.8%, virtually the same as for those areas in 1998.

These tables suggest that changes in segregation of the metropolitan population may reflect both changes in the population in particular areas, but also differences between established MAs and newly defined MAs. As discussed in the text, we examined the robustness of our results with respect to the differences by age of metropolitan area. Changes in segregation of the metropolitan population may reflect both changes in the population in particular areas, but also differences between established MAs and newly defined MAs. As discussed in the text, we

examined the robustness of our results with respect to the differences by age of metropolitan area and size of metropolitan area.

In addition to newly created MA's, existing MAs may add population either through the addition of population within previously defined boundaries, or through the incorporation of new counties or areas into the MA. We chose to regard the contemporary boundaries of metropolitan areas as definitive of the analytically appropriate boundaries: we are examining the segregation of persons, not of territory. As stated in the text, analyzing the data with constant boundaries yields similar findings.

Measures:

Race and Ethnicity. We explore three measures – race, Hispanic origin, and nativity – of segregation. For the period from 1960 through 2000, we use three mutually exclusive race categories – white, black, and other. We draw additional distinctions within and across racial categories by Hispanic origin for the period from 1970 through 2000. Finally, we take population counts for native- and foreign-born persons from 1960 on. Idiosyncrasies in measures for selected years are discussed below.

In 1960, the three base race categories were directly reported. We formed the same three categories for 1980 and 1990 by simply collapsing more detailed categories. In 1970, a substantial proportion of race counts were suppressed at the tract level; for those tracts, we imputed counts by applying the proportional shares by race identified prior to the census-performed allocations and substitutions to the total population counts. In 2000, when the questionnaire permitted the identification of multiple races, we used the single race counts for

calculations, and verified trends using racial identifications in "alone-or-in-combination." The identification of Hispanic-origin persons in 1960, based on surname and parentage, was incomparable with the self-identification available in later years, and we excluded it from our analyses. The 1970 instrument offered multiple modes for identification of the Hispanic population – parentage, surname, self-identification, and mother tongue; we used the counts generated from the self-identification item on the 5% long-form questionnaire.

Class. We use two measures – family income and homeownership – as our indicators of segregation by class. (We also considered two additional measures of class – educational attainment and occupation – but set them aside, the former because its close correlation with age makes it difficult to separate from the effects of life cycle and the latter because the way that occupational categories are constituted changed substantially over the period.)

Income categories for our analyses are tabulated at a family level. For 1960, family income is provided in 13 categories; in 1970, 1980, 1990, and 2000, family income is reported, respectively, in 15 categories, 17 categories, 25 categories, and 16 categories. In order to provide a more consistent measure over time of counts of the well-off and the poor, we estimated the dollar figures which establish the 20th and 80th percentiles of family income for families in each year's sample, assuming an even distribution of families within each income category. This linear interpolation which we used to estimate the top and bottom quintiles we also applied at the tract level to apportion families to one of the three income strata formed. (Thus, for example, a set proportion of counts for a category which straddles the 20th percentile would be allocated to the bottom 20th percentile, and the remainder allocated to the middle 60%). In each year, the

primary distinction between housing units which are owned, regardless of mortgage status, and those which are rented, with or without cash rent, is the basis for our measure of homeownership.

Life Cycle. We captured life cycle effects by counts of population according to age and marital status. Individual cell counts, reported in categories ranging from single years to 10 year age groups, were collapsed into counts of those 14 and younger, 18 to 29 year-olds, and persons aged 65 and older (as well as the complementary groups, e.g., persons 15 and older, or those 64 and younger). We identified marital status for men and women aged 14 and older before 1980, and for those 15 and older after that point.

Detailed Appendices

Detailed Appendix 1. *The Significance of Differences in the Theil Index.*

To interpret our measure of segregation – *H* – we need to know the effect of some reasonable changes in population on *H*. James and Tauber (1985, eq. 9) provide an equation that contains the answer, but it is so complex that some explication might be useful. First, consider these definitions:

T_s = size of population in tract *s*,

T = total population,[3]

[3] Our national statistics refer to the metropolitan population. These results also apply to specific geographical units, for example, the segregation in one region or metropolitan area.

p_{is} = proportion of the population in tract *s* that is in the segregated group *i* (blacks, Latinos, poor people, etc.),

p_i = proportion of the total population that is in the segregated group *i*,

$p_{\sim i,s}$ = proportion of the population in tract *s* that is NOT in the segregated group *i* (nonblacks, non-Latinos, non-poor, etc.),

$p_{\sim i}$ = proportion of the total population that is NOT in the segregated group,

$E_i = p_i \ln(1/p_i) + p_{\sim i} \ln(1/p_{\sim i})$,

$E_{is} = p_{is} \ln(1/p_{is}) + p_{\sim i,s} \ln(1/p_{\sim i,s})$, and

x = the population exchange – the movement of x people of the segregated group from one tract (i) to another (j) and an equal number of people of the other group from j to i. *H* is defined in equation [2] of the paper.

James and Tauber work out the derivative of *H* with respect to *x*. Remember that the derivative is the generalization of the slope – it gives the amount of change in *H* for a small change in *x*. In a population of millions, we can think of it as the amount of change in *H* for a one-person exchange between tracts *s* and ~*s*. Their equation is:

$$\frac{dH}{dx} = \frac{1}{T} \ln\left(\frac{p_{i,\sim s} / p_{\sim i,\sim s}}{p_{is} / p_{\sim i,s}} \right) / E_i \quad . \qquad [A.1]$$

Since we are mainly interested in the consequences of dx for H, we move dx over to the right-hand side by multiplying both sides of equation [A.1] by dx, i.e.,

$$dH = \frac{1}{E_i} \ln\left(\frac{p_{i,\sim s} / p_{\sim i,\sim s}}{p_{is} / p_{\sim i,s}} \right) \frac{dx}{T} \quad . \qquad [A.2]$$

We have rearranged terms slightly for convenience. As *T* is total population, the term *dx/T* re-expresses the exchange of people as a proportion of the total population instead of as a raw number of people.[4] E_i is a constant that depends on the relative sizes of the segregated group

[4] It will prove important later that the number of people moving is x from one group and x from another or 2x altogether.

and the rest of the population. E_i reaches its maximum – and $1/E_i$ reaches its minimum – when the segregated group is exactly one-half of the population. If the segregated group is either very small or very large, then E_i will be small, too. Thus, all else being equal, an exchange has a bigger impact on H if the number of people being exchanged is a non-trivial proportion of either the segregated group or the rest of the population than if the segregated group is about as big as the rest of the population.

To interpret the middle term – what we call the "exchange accelerator" – of equation [A.2], think of three kinds of neighborhoods: "ghettos," "proportionally mixed," and "isolated majority" neighborhoods. To quantify these distinctions, let's use the numbers that stem from the segregation of African Americans in the 1990s. At that point in American history, African Americans were about 12 percent of the U.S. population, so $p_i = .12$ and $p_{\sim i} = .88$ in a "proportionally mixed" neighborhood. We can think of a "ghetto" neighborhood as one in which $p_i = .95$ and $p_{\sim i} = .05$. Finally let us say that a neighborhood is "isolated majority" if $p_i = .05$ and $p_{\sim i} = .95$. With these kinds of neighborhoods in mind, we can calculate the exchange accelerator for exchanges between each possible pair involving different kinds neighborhoods; Table A.1 shows the results. All three kinds of exchange involve integration, so they all reduce segregation (i.e., have a negative sign with respect to dH). An exchange that integrates both neighborhoods (an exchange between a ghetto and an isolated majority neighborhood) reduces segregation more than exchanges between segregated and proportionally mixed neighborhoods. Because African Americans are in the minority, an exchange between a ghetto neighborhood and a proportionally mixed neighborhood represents more integration than an exchange between a proportionally mixed neighborhood and an isolated majority one. We only have to consider changes between different types of neighborhoods because an exchange between a pair of neighborhoods with identical distributions results in an accelerator of zero and no change in H.[5]

[5] In the formula for the exchange accelerator – $\ln((p_{i,\sim s}/p_{\sim i,\sim s}) / (p_{is}/p_{\sim i,s}))$ – exchanges between neighborhoods of the same type mean that $p_{is} = p_{i,\sim s}$ and $p_{\sim i,s} = p_{\sim i,\sim s}$ so the numerator equals the denominator and the

So now we are in a position to say what we expect from an exchange involving two percent of the total population – a group of people from the segregated population equal to one-percent of the total population moving to a new neighborhood and an equal number of people from the other group moving the other way. As we are talking about racial segregation, p_i is the proportion of the total population that is African American (.12) which implies that E_i = .367. Exchanges between proportionally mixed and ghetto neighborhoods would lower *H* by (–4.937/.367) x .01 = –.13. Exchanges between ghettos and isolated majority neighborhoods would reduce segregation even further: *dH* = (–5.889/.367) x .01 = –.16. African Americans from proportionally mixed neighborhoods who exchange with nonblacks in isolated majority neighborhoods further reduce segregation (though not nearly as much as African Americans who move out of isolated minority neighborhoods) *dH* = (–.952/.367) x .01 = –.03.

Two considerations suggest that very modest changes in *H* probably have substantive significance. First, a one-percent-point exchange refers to one percent of the total population; a group of African Americans equal to one percent of the total U.S. population is about 8.3 percent of the African American population. That is a very large population redistribution. The consequences of such a large population shift can be as great as .16 or as small as zero, depending on the concentration of African Americans at the source and at the destination. Second, moves from ghettos to isolated majority neighborhoods are probably pretty rare. Exchanges between more similar neighborhoods get less acceleration, and have less impact on overall segregation. Together, these two considerations suggest that observed changes in *H* as small as .02 or .03 represent substantively significant reductions in segregation. Third, while we are talking about a rather small fraction of moves that nonblacks make every year, we are supposing that they make different choices than the typical black mover makes. About 20 percent of Americans move in a given year. We are supposing in these calculations that about 10 percent of the nonblacks who are moving choose to live in either a proportionally mixed

---

ratio equals 1; ln(1) = 0. Thus exchanges between neighborhoods of the same type produce exchange accelerators of 0 and that results in *dH* = 0.

neighborhood or a ghetto. We are also supposing that over 40 percent of the African Americans who move leave ghetto neighborhoods for proportionally mixed or isolated majority neighborhoods. In other words, the scenarios behind the forgoing calculations assume some atypical behaviors.

The other segregated groups we work with involve minority populations with shares of the total population that resemble the African American population's share. Those living below the poverty line are 12 to 14 percent in most years, the richest and poorest fifths are 20 percent by definition, children have been about 22 percent of the population in recent years, and seniors are about 8 percent of recent years' populations. So changes of *H* that range from .05 to .10 must be thought of as large shifts in the distribution of the population. Changes in *H* as small as .02 and .03 are probably changes worth discussing.

Table A.1

Values of the Exchange Accelerator for Exchanges Between Pairs of Neighborhoods, By Neighborhood Types

Minority originates in:	Minority moves to:	Formula	Value
Ghetto	Proportionally mixed	ln((.12/.88) / (.95/.05))	-4.937
Ghetto	Isolated majority	ln((.05/.95) / (.95/.05))	-5.889
Proportionally mixed	Isolated majority	ln((.05/.95) / (.12/.88))	-.952

Detailed Appendix 2. *Statistical Estimation of Small Place Segregation in 1960*

The analyses in this paper draw on decennial census data from 1960 to 2000. From 1970 through 2000, the census linked tracts to the places in which they were located when those places had populations of at least 2,500. But in 1960, the census identified places for tracts only when the places were 25,000 or larger in population. As a result, the percent of individuals in our metropolitan universe who live in identifiable places jumps from 69% in 1960 to 79% in 1970. Since we are seeking to apportion segregation between different geographic levels over the period since 1960, this place identification change presents two problems. First, segregation can only be apportioned to place to the extent that places are identifiable in our data. Thus, the smaller proportion of places uniquely defined in 1960 artificially decreases segregation apportioned to place (versus tracts within places) in 1960 compared to the other decades. Second, the data from 1970 onwards indicates that most of the between-place segregation occurs in places with fewer than 25,000 persons. Thus we fail to pick up an important source of place-level segregation in 1960 that is observed in the other decades.

For these reasons, we implement a procedure to estimate place-level segregation in 1960 so that it is consistent with the other decades. The time trend in the proportion of between-place segregation attributable to small places (those with a population of less than 2,500) is stable or approximately linear for the characteristics we analyze. As a result, we can estimate the proportion of segregation attributable to small places for 1960 using the trends derived from the years in which more detailed place identifiers were available. The dependent measure we use is the *difference* between the proportion of segregation attributable to all places with a population of 2,500 or more and that attributable only to large places with a population of 25,000 or more. Observed values are then fit using a linear time trend and dummies for size of the metropolitan area, and extrapolated to 1960.

The purpose of the metropolitan area size dummies is to increase accuracy. Three groups are defined by metropolitan area population size: over 4 million persons, 1 to 4 million, and less than 1 million. The rationale for this is that metropolitan areas of different sizes tend to have

consistent differences in the apportionment of segregation by geographical level. Thus, performing the procedure in three separate groups and then recombining those results increases the overall accuracy. (This procedure assumes that there is no major discontinuity from 1960 to 1970 in the trend of segregation in small places versus large places.) The entire procedure is recapitulated in symbolic form below.

Indices:

i = 1,2. Index of place definition. (1: places of >2,500, 2: places of >25,000)

j = 1,2,3. Index of group by population of metropolitan area. (1: metro areas of >4 million, 2: metro areas between 1 and 4 million, 3: metro areas of <1 million)

k = 1960, 1970, 1980, 1990, 2000. Index of decade.

Known

H_{jk}^{T+P} =Total segregation due to tract and place for group *j* in decade *k* (for all *k*).

H_{jk}^{T+P} =Proportion of due to place, with place definition *i* for group *j* in decade *k* (for *k* ≠ 1960 when *i* = 1; for all *k* when *i* = 2).

$P_{1jk} - P_{2jk}$ =(for k ≠ 1960).

Estimates:

$\underline{D_{jk} = b_0 + kb_1}$ = Estimate from OLS regression line of model .

$\hat{D}_{j,1960} + P_{2,j,1960}$ = .

$H_{1,j,1960}^{T+P} \times \hat{P}_{1,j,1960}$ = .

$\hat{H}_{1,1,1960}^{P}, \hat{H}_{1,2,1960}^{P}, \hat{H}_{1,3,1960}^{P}$ = Population-weighted average of

$H_{1,1960}^{T+P} - \hat{H}_{1,1960}^{P}$

Detailed Appendix 3. *Handling Changes in Tracts*

For 1960, tract data are available from 172 Metropolitan Areas (MA's), but only 133 were completely tracted. Nonetheless, tracted populations in that year include 92% of the total population of MA's and about 59% of the national population. For the remaining years, coverage of metropolitan areas is complete, but there is incomplete coverage outside metropolitan areas.

Tracts are intended to represent roughly equal chunks of the population – with somewhere between 2,500 and 8,000 inhabitants – which make them a suitable unit for identifying a "neighborhood," and the historical coverage of the tracted population permits reasonable estimates of changes in neighborhood characteristics for metropolitan areas since 1960. However, maintaining consistent geographical size conflicts with maintaining roughly comparable population sizes. As population and density increase, the census alters tract boundaries. The principal alteration is splitting existing tracts into multiple tracts, although sometimes the census merges tracts or forms new ones by a complex mix of splitting and merging. Depending on their needs, researchers interested in segregation have either chosen to hold historic boundaries constant and re-combine separated tracts for later years, or alternatively, to keep tracts at a roughly constant population size, using more finely divided geographies over time.

Following the logic of the decisions we made about metropolitan boundaries, we use tract boundaries as they were contemporaneously defined. These boundaries more appropriately define a local neighborhood than do out-of-date physical boundaries when there has been much growth. However, by using the tract equivalency files for 1970-1980, 1980-1990, and 1990-2000, we also examined trends in levels of segregation using constant boundaries.[6] Those

[6] Areas not tracted in previous censuses were treated as residuals of the metro-state-county unit. Areas previously tracted were assigned the tract identification number from the prior census in which most of its population resided. We assigned earlier tract identifications by chaining these year-to-year equivalencies. The tract equivalency files were available from ICPSR: Census of Population and Housing, 1980 (United States): 1970-Pre-1980 Tract Relationships (ICPSR 7913); *idem.*, 1990: Tiger/Census Tract Comparability File (ICPSR 9810); *idem.*, Census Tract Relationship Files (CTRF) (ICPSR 13287).

analyses suggest that our findings are robust to changes in the definition of our tract neighborhoods.

The decomposition of segregation across geographic levels also requires that smaller geographies "nest" within larger geographies. However, while tracts nest within counties and metropolitan areas, they can be split by place boundaries. Use of split tracts rather than full tracts will result in higher estimates of segregation, since the base geographic unit will be finer. To maintain consistency with the full tract data available in 1960 and 1970, and to retain more comparable tract sizes and definitions, we aggregated split tracts counts to the full tract level, and attributed the tract counts to the place, if any, which contained the largest proportion of the total population of that full tract.

Methodological Appendix to
"Distinguishing the Geographic Levels and Social Dimensions
of U.S. Metropolitan Segregation, 1960-2000

December 2003

Calculation of Theil's H at Nested Geographic Levels

The basic calculation of *H* is described in the text, and several articles noted in the text review different calculations by which *H* can be decomposed into additive contributions from different groups within the population, or different levels of geography at which segregation might occur (esp., Reardon and Firebaugh 2002; Reardon and Yun 2001). This appendix reviews the calculations for the particular type of analysis of greatest interest in our paper: additive decomposition of *H* into the contributions of nested geographic levels when measuring segregation among *K* groups. In the text of the paper, all results shown are for segregation between only two groups, a special case of the formulas that follow.

Notation. In the paper, we work with five nested levels of geography: metropolitan United States (*U*), regions within the United States (*R*), metropolitan areas within regions (*M*), the center city/suburb divide within metropolitan areas (*C*), places within central cities or suburbs (*P*), and census tracts within places (*T*). For the geographic areas, the capital letters above will refer to the total number of such areas in the study or to the level in general, while the lower case of that letter will refer to a particular area, often used as an index in a sum of some quantity over all area units. That is:

Geography:

G = $\{U, R, M, C, P, T\}$, number of areas at a geographic level, or the level in general

g = $\{u, r, m, c, p, t\}$, indexes specific units at a geographic level

$G_2 \subset G_1$ = indicates nesting: level G_2 is nested within level G_1.

Our other notational conventions are:

Population:

k = indexes K groups (sub-populations) within the population

N_g = total population in geographic area g

$p_{g,k}$ = geographic area g's proportion in group k, $\sum_{k=1}^{K} p_{g,k} = 1$.

Calculation of H. The calculation is described in the paper, but is reviewed here as well. *H* compares the proportion of each of the *K* groups in some total area to the proportions found in its smaller constituent areas. In order to compare proportions, we need a summary measure of the overall distribution

across the *K* groups. Entropy (*E*) measures the diversity of groups in an area. It is calculated for a unit of geography *g* as follows:

$$E_g = \sum_{k=1}^{K} p_{g,k} \ln\left(\frac{1}{p_{g,k}}\right)^{1}$$

Say that you are considering a total area G_1, which has G_2 sub-areas nested within it. Theil's *H* for G_2 nested within G_1 (or $H_{G_2 \subset G_1}$ in the notation used here) is a weighted average of the differences between the diversity in the large area (E_{G_1}) and each of the G_2 smaller areas (E_{g_2}), with the weights supplied by the smaller area's share of the larger areas's population:

$$H_{G_2 \subset G_1} = \sum_{g_2=1}^{G_2} \left(\frac{N_{g_2}}{N_{G_1}}\right)\left(\frac{E_{G_1} - E_{g_2}}{E_{G_1}}\right).$$

Substituting some of the geography used in the paper, $H_{T \subset U}$, for example, measures segregation of tracts within metro U.S. by comparing tract-level entropies (E_t) with the entropy of metro U.S. (E_U). $H_{P \subset U}$ is a similar calculation to measure segregation at the place level within metro U.S. $H_{t \subset p}$ indicates tract-level segregation occurring within a particular place p and would only be calculated using those tracts nested within place *p*.

Decomposition of *H* into Contributions of Two Levels of Geography. We now want to extend the *H* calculation above to two nested geographic levels, specifically tracts nested within each of the four major census regions, which are nested within metro U.S. The equation is as follows:

$$H_{T \subset U} = H_{R \subset U} + \sum_{r=1}^{R} \left(\frac{N_r}{N_U}\right)\left(\frac{E_r}{E_U}\right) H_{t \subset r}$$

---

[1] Note that for any $p_{g,k} = 0$, let $p_{g,k} \ln\left(\frac{1}{p_{g,k}}\right) = 0$, by definition.

The first term on the right-hand side of the equation above is the portion of total tract-within-metro U.S. segregation that is due to segregation at the level of region. The second term is the remainder that is due to segregation of tracts-within-regions. While the equation above does show how to calculate the remainder directly (as a weighted average of the $H_{t\subset r}$'s calculated for each of the *R* regions within metro U.S.), in practice it is often easier to calculate $H_{T\subset U}$ and $H_{R\subset U}$ and subtract for the remainder term.

Decomposition of H into Multiple Levels of Geography. The previous equation showed how to decompose total tract-within-metro U.S. segregation into that accounted for at the regional level versus the remainder occuring at all lower geographic levels. Following the logic of that decomposition, we could also decompose each of the four regional $H_{t\subset r}$'s into the portion of segregation due to the segregation of metropolitan areas within the region ($H_{m\subset r}$) versus the remainder segregation among tracts within each of the region's metropolitan areas (a weighted average of the $H_{t\subset m}$'s for each metropolitan area within each region). The sums telescope into an equation that can be expressed in terms of the *H* indices calculated for each level of geography nested in the total area (metro U.S.). Each term represents the portion of segregation due to a particular level of geography, after all higher levels of geography have been taken into account:

$$
\begin{aligned}
H_{T\subset U} = \quad & H_{R\subset U} && \leftarrow \text{ Remainder Portion} \\
+ \quad & \left(H_{M\subset U} - H_{R\subset U}\right) && \leftarrow \text{ Metropolitan Area Portion} \\
+ \quad & \left(H_{C\subset U} - H_{M\subset U}\right) && \leftarrow \text{ Center City-Suburban Divide Portion} \\
+ \quad & \left(H_{P\subset U} - H_{C\subset U}\right) && \leftarrow \text{ Place Portion} \\
+ \quad & \text{Remainder} && \leftarrow \text{ Tract Portion}
\end{aligned}
$$

Alternately, the total and each term of the sum above can be calculated directly in terms of the original entropies and populations:

$$
\begin{aligned}
H_{T\subset U} &= \frac{1}{N_U E_U}\sum_{t=1}^{T} N_t (E_U - E_t) \\
H_{R\subset U} &= \frac{1}{N_U E_U}\sum_{r=1}^{R} N_r (E_U - E_r) \\
\left(H_{M\subset U} - H_{R\subset U}\right) &= \frac{1}{N_U E_U}\sum_{r=1}^{R}\sum_{m=1}^{M_r} N_m (E_r - E_m) \\
\left(H_{C\subset U} - H_{M\subset U}\right) &= \frac{1}{N_U E_U}\sum_{r=1}^{R}\sum_{m=1}^{M_r}\sum_{c=1}^{C_m} N_c (E_m - E_c)
\end{aligned}
$$

$$\left(H_{P\subset U} - H_{C\subset U}\right) \quad = \quad \frac{1}{N_U E_U} \sum_{r=1}^{R} \sum_{m=1}^{M_r} \sum_{c=1}^{C_m} \sum_{p=1}^{P_c} N_p (E_c - E_p)$$

$$\text{Remainder} \quad = \quad \frac{1}{N_U E_U} \sum_{r=1}^{R} \sum_{m=1}^{M_r} \sum_{c=1}^{C_m} \sum_{p=1}^{P_c} \sum_{t=1}^{T_p} N_t (E_p - E_t)$$