Web Appendix.
"Inequality and the Measurement of Residential Segregation by Income in American Neighborhoods"
Tara Watson

A. Data Description

Tract-level census data for 1980, 1990, and 2000 are taken from Census CDs and for 1970 they are taken from the Urban Institute Underclass Database. I also use county-level information from the Census CDs and the City and County Data books. I use the IPUMS to estimate national trends in industrial mix and job centralization.

The income data for families in the tract-level data sets is the number of families in different income bins in the year previous to the census year. There are 15 income bins in the 1970 data, 17 in the 1980 data, 25 in the 1990 data, and 16 in the 2000 data. For data on family income by race, there are 9 bins the years 1970-1990 and 16 bins in 2000.

Gainesville, FL is excluded from the analysis due to missing data. There are 216 remaining metropolitan areas in the analysis.

B. The Centile Gap Index

The Centile Gap Index (CGI) estimates how far the average family income within a tract deviates in percentile terms from the median tract family income, compared to how far it would deviate under perfect integration. A panel data set with the income segregation indices is associated with this web appendix and is available on-line. Please cite this working paper or the future published version when using the indices. Contact the author (Tara.Watson@williams.edu) with questions or corrections.

As noted in the text, the formula for the Centile Gap Index (CGI) is:

$$CGI_m = (0.25 - (1/J_m) \Sigma_j |P_j - P_{medtj}|) / 0.25,$$

where CGI_m is the Centile Gap Index in metropolitan area m, J_m is the number of families in metropolitan area m, P_j is the estimated percentile in the metropolitan area m income distribution of family j, and P_{medtj} is the estimated income percentile of median family in the tract of family j. That is, the term $|P_j - P_{medtj}|$ represents the estimated income percentile distance of a given family from the median family in their tract. Note that, although I refer to income percentiles for clarity, I do not divide families into 100 discrete groups. Rather, the ranking is continuous on a scale from 0 to 1. In principle, the Centile Gap Index goes from 0 to 1, with 1 meaning perfect segregation. In practice, as I will discuss below, perfect segregation could never be observed with a small number of income bins.

As a benchmark, consider what I will call the "true" Centile Gap, CGI*, which hypothetically could be computed if the full distribution of income in each census tract were reported without error by the census. The CGI* is defined by the formula:

$$CGI_m^* = (0.25 - (1/J_m) \Sigma_j |P_j^* - P_{medtj}^*|) / 0.25.$$

where the CGI_m^* is the "true" CGI of metropolitan area m, the term $|P_j^* - P_{medtj}^*|$ represents the true percentile distance between family j and the median family in family j's tract, and J_m is the number of families in metropolitan area m.

Because the full income distribution in a census tract is unavailable in practice, P_j^* and P_{medtj}^* must be estimated using the information available in 15, 17, 25, or 16 family income bins (for the years 1970, 1980, 1990, and 2000 respectively). Some assumptions are necessary in order to estimate P_j^* and P_{medtj}^*. As is noted in the text, I assume that income percentiles are distributed uniformly within an income bin within a census tract. A simple example illustrates the calculation. Suppose I know that in a given metropolitan area, income bin x represents exactly 4 percent of the population, from the bottom of the 60^{th} through the top of the 63^{rd} percentile. A given tract has 60 families in bin x. I assume these families are uniformly spread among the four percentiles. In other words, 15 of them are uniformly spread in the 60^{th} percentile, 15 are uniformly spread in the 61^{st}, and so on. Further, suppose I know that there are 1000 families in the tract, 450 of which are in bins below bin x and 490 of which are in bins above bin x. I therefore know that the median income percentile in the tract is in bin x. Using the uniformity assumption, I estimate the median family in the tract as having income rank 0.633.

The CGI computed using the uniformity assumption systematically understates the CGI* to the extent that income sorting is an important phenomenon. This can be seen be considering two extreme cases. If there is no income segregation, the CGI correctly estimates income segregation to be zero. On the other hand, if there is perfect income segregation, the CGI estimates segregation to be less than one.

The degree of deviation between CGI and CGI* also depends on the number of income bins and where they fall in the income distribution. In particular, the deviation of the "typical" family from the boundaries of its income bin is important. If more families are in wide income bins (i.e., bins that represent a many income percentiles), income segregation is understated to a greater degree by the CGI. This implies that the CGI computed using fewer bins will tend to be lower than the CGI computed using more bins. One concern that I will put to rest below is that the rise in income segregation over time is a spurious result due to a change in the number of income bins.

I will now argue that the bias in the Centile Gap Index induced by the uniformity assumption is small relative to changes over time and differences across metropolitan areas. With the publicly available data, one cannot test the assumption directly. Rather, I will make use of the observed patterns across income bins to infer that the uniform assumption within bins is likely to be "approximately" correct. It is important to note that the inference can be made because income segregation levels are low in an absolute sense. That is, most census tracts contain at least some families from most income bins.

If it were common that only one or two out of 25 income bins were represented in a census tract, the uniformity assumption would be likely induce a substantial bias which would swamp differences across time or metropolitan areas. Suppose, for example, there were perfect segregation in a metropolitan area. The number of income bins would clearly affect the amount of segregation that could be observed in the data.

The simulation exercise I perform evaluates the importance of the number of bins and bin boundaries to the estimated CGI. For each metropolitan area in 1970 and 1990 (the years with the smallest and greatest number of bins), I combine the 15 or 25 actually observed bins into fewer bins, each representing a greater part of the income distribution. I start by assuming only one income bin is observable. Therefore, no information about segregation is available and the uniformity assumption implies that the full income distribution is represented within each tract. The estimated CGI is 0 for every metropolitan area. Next, I assume two large bins are observable. In 1990, for example, the bottom large bin contains the bottom 12 actual bins and the

top large bin contains the top 13 actual bins. I continue to increase the number of large bins by one and estimate the CGI using the uniformity assumption.

If the actual distribution of family income percentiles within the constructed large bins deviates substantially from uniformity, the estimated CGI should be very sensitive to the number of bins. For example, as shown in Figure A1, when one moves from 1 bin to 2 bins or 2 bins to 3 bins, the estimated CGI changes substantially, suggesting that that the uniformity assumption is bad. On the other hand, as one increases the number of bins, there is convergence towards an estimated CGI. Based on the simulation, I infer that the additional information that would be provided by additional bins if they were available would not substantively change the estimated CGI. In other words, the CGI estimated with 15 or more bins using the uniformity assumption is "close enough" to the true CGI* so that one can feel comfortable making comparisons within metropolitan areas and over time.

I have not addressed two types of measurement error which might also be a problem in estimating income segregation, as discussed in Davidoff (2003). Families misreport income and annual income is a noisy indicator of permanent income. Davidoff (2003) suggests that correcting for measurement error might lead to double the estimated fraction of variance that is attributable to sorting between neighborhoods. Note that, in my case, the measurement problem would arise from noise in a family's rank in the distribution rather than income per se. Assuming changes in measurement error are uncorrelated with inequality, measurement error would tend to attenuate my results. I do not address this issue in the analysis.

C. Other Measures of Income Segregation

The Herfindahl Index is given by the formula

$$HI_m = (1/J_m) * \Sigma_j \Sigma_q (Ptjq)^2 \text{, where}$$

J_m = number of families in metro area m, and
Ptjq = the fraction of quintile q residents in the tract of family j.

The NSI, or Neighborhood Sorting Index, is the square root of the ratio of between tract variance to total variance of income:

$$NSI_m = \sigma_{Nm} / \sigma_{Jm,} \text{ where}$$

σ_{Nm} = standard deviation of mean tract income in metro area, and
σ_{Jm} = standard deviation of family income in metro area.
See Jargowsky (1995, 1996) for more details.

The formula for an exposure index of quintile x to quintile y in metropolitan area m is

$$Exposure_{xym} = \Sigma_t (X_t/X_m) * (Y_t/J_t), \text{ where}$$

X_t = number of quintile X families in tract t,
X_m = number of quintile X families in metro area m,
Y_t = number of quintile Y families in tract t, and
J_t = number of families in tract t.
The exposure of quintile x to quintile y can be interpreted as the average fraction of quintile y families in the typical quintile x family's census tract.

Also see Cutler, Glaeser, and Vigdor (1999). In the present analysis, no adjustment is necessary for group size since all family income quintiles are the same size. Exposure indices were developed to study segregation between discrete racial groups and are not ideal for analyzing

segregation along a continuous dimension such as income. Nevertheless, they are useful as a supplement to overall income segregation measures.

Note that none of the income segregation measures I use here are explicity spatial. That is, I do not use geographic information about proximity of neighborhoods to one another.

A panel data set with the income segregation indices is associated with this web appendix and is available on-line. Please cite this working paper or the future published version when using the indices. Contact the author (Tara.Watson@williams.edu) with questions or corrections.

D. Metropolitan area definitions.

The boundaries of metropolitan areas change over time and a researcher must make a decision about how to deal with this fact in the analysis. One values consistency, but at the same time wants to capture genuine changes in the area of the residential market. I pursue an intermediate approach. I use the consolidated metropolitan areas defined by the Census Bureau as of 2003, including all of the counties that were tracted in a particular year. Therefore, the boundaries of about half of the metropolitan areas in the sample change over time. I throw out metropolitan areas that had no tracted counties in 1970. For New England, I use the county-based metropolitan area definitions developed by the Census (NECMAs) rather than the standard town-based definitions. Metropolitan areas outside of New England are always based on counties or county-equivalents.

E. Families versus Households.

I use data for families because they are available for all four Census years. Families, which are households in which at least two residents are related by blood or marriage, make up a large fraction (ranging from 68 percent in 2000 to 75 percent in 1980) of households. Comparing In 1980-2000, families have higher segregation levels than all households. If I run the baseline regression using household segregation and household income inequality for 1980-2000, the results are very similar to the family-based analysis and highly significant.

F. Industrial Mix Variables

Ten initial metropolitan area industry shares are interacted with national industry changes over time to predict the level of total employment relative to 1970 total employment, the level of employment of less-skilled men relative to 1970 total employment, and the fraction of metropolitan employment in the central city. Less-skilled workers are defined as those with a high school degree or less. The national trends for centralization and skill level are computed using IPUMS data on 18-65 year olds in metropolitan areas who worked at least 15 hours in the previous week. Predicted variables are used rather than direct measures of employment growth, demand for less-skilled men, and job centralization because these characteristics may be endogenous to segregation. Therefore, the industrial mix variables in some sense "under-control" because they do not capture the effect of idiosyncratic changes in industrial composition.

The formula for Predicted Employment is

$$PredEmp_{mt} = \Sigma_i \, (Emp_{im70}/Emp_{m70}) * (NatEmp_{it}/NatEmp_{i70}),$$

where $PredEmp_{mt}$ is the predicted employment level in metropolitan area m at time t, Emp_{im70} is the employment in metropolitan area m in 1970 in industry i (from aggregated county-level data), Emp_{m70} is the total employment in metropolitan area m in 1970 (from aggregated county-level data), $NatEmp_{it}$ is the total employment in all metropolitan areas in industry i at time t (from aggregated county-level data), and $NatEmp_{i70}$ is the total employment in all metropolitan areas in industry i in 1970 (from aggregated county-level data). It is clear from the formula that all metropolitan areas have a predicted employment of 1 in 1970.

The formula for Predicted Employment of Less Skilled Men is

$$PredEmpLSM_{mt} = \Sigma_i \, (Emp_{im70}/Emp_{m70}) * (Nat2LSM_{it}/Nat2Emp_{it}) * (NatEmp_{it}/NatEmp_{i70}),$$

where $PredEmpLSM_{mt}$ is the predicted employment of less-skilled men in metropolitan area m at time t, $Nat2LSM_{it}$ is the employment of less-skilled men in all metropolitan areas at time t in industry i (from the PUMS), $Nat2Emp_{it}$ is the total employment of less-skilled men in all metropolitan areas at time t (from the PUMS), and other variables are as above. The variable is a prediction of employment of less-skilled men in year t relative to total metropolitan area employment in 1970.

The formula for Predicted Job Centralization is

$$PredCent_{mt} = \Sigma_i \, (PredFrac_{imt}) * (Nat2CC_{it}/Nat2Emp_{it}),$$

where $PredCent_{mt}$ is the predicted fraction of employment in the central city, $Nat2CC_{it}$ is the employment in industry i at time t in all central cities (from the PUMS), $Nat2Emp_{it}$ is the employment in industry i at time t in all metropolitan areas (from the PUMS), and $PredFrac_{imt}$ is the predicted fraction of employment in industry i in metropolitan area m at time t and is defined by:

$$PredFrac_{imt} = Emp_{im70} * (Nat2Emp_{it}/Nat2Emp_{i70}) \; / \Sigma_i \, [Emp_{im70} * (Nat2Emp_{it}/Nat2Emp_{i70})].$$

In the formula, Emp_{im70} is the employment in industry i in metropolitan area m in 1970 (from aggregated county-level data), $Nat2Emp_{i70}$ is the employment in industry i in 1970 in all metropolitan areas (from the IPUMS), and other variables are defined as above.

The formula for Predicted Manufacturing Employment Share is

$$PredManEmp_{mt} = (ManEmp_{m70} * (NatManEmp_t/NatManEmp_{70})) / (Emp_{m70} * (NatEmp_t/NatEmp_{70}))$$

where $PredManEmp_{mt}$ is the predicted manufacturing employment share in metropolitan area m at time t, $ManEmp_{m70}$ is the manufacturing employment in metropolitan area m in 1970 (from aggregated county-level data), $NatManEmp_t$ is the manufacturing employment in all metropolitan areas in year t (from aggregated county-level data), $NatManEmp_{m70}$ is the manufacturing

employment in all metropolitan areas in 1970 (from aggregated county-level data), Emp_{m70} is the total employment in metropolitan area m in 1970 (from aggregated county-level data), $NatEmp_t$ is the total employment in all metropolitan areas at time t (from aggregated county-level data), and $NatEmp_{70}$ is the total employment in all metropolitan areas in 1970 (from aggregated county-level data).

G. Data on Family Income by Race

The data for investigating family income by race is less than ideal for several reasons.

Suppression of data is an important problem for the analysis of segregation within racial groups. Income data is suppressed when the number of members of a group is so low that the Census Bureau determines that including the information would violate the privacy of the residents. This is known as primary suppression. In addition, the 1980 census data includes complimentary suppression, in which another group besides the primarily suppressed group is also suppressed in order to prevent inferences using aggregate data. I have not adjusted the analysis to take account of data suppression. However, in some parts of the paper, the sample is limited to metropolitan areas with at least 5,000 black residents or to those with at least 10,000 black families to minimize the problem.

The 1990 and 2000 censuses include information about the income distribution for black households rather than black families at the census tract level. To correct for this fact, I have assumed that the relationship between segregation of black families and segregation of black households in 1990 and 2000 is the same as that for all families and all households. The qualitative results are similar with or without the adjustment.

In 1970-1990, there are nine bins in the data representing the income distribution within racial groups, and there are 16 bins in 2000. The properties of the overall CGI suggest that an analysis using only nine bins is likely to understate the "true" degree of segregation in a metropolitan area. The results should be considered suggestive rather than conclusive.

H. Racial Segregation Measures

Two measure of racial segregation are used in the analysis. One is the exposure of black families to white families, which is the typical fraction of white families in a typical black family's neighborhood. This measure is analogous to the exposure indices used to measure income segregation.

The second measure of racial segregation, which is used in the regressions, is a person-based measure (rather than family-based) similar to that used by Cutler, Glaeser, and Vigdor (1999). It is an exposure index of black residents to other black residents, adjusted for the exposure that would be expected given the number of black residents and white residents in the metropolitan area. For details, see p.459 of Cutler, Glaeser and Vigdor (1999). Note, however, that they consider isolation of black residents from all other residents while I consider the isolation of black residents from white residents.

Hispanic families may be of any race. Some of the rise in exposure of black families to white families may be due to an increase in exposure to white Hispanic families.

I. Measurement of Metropolitan Area Income Inequality

The primary measure of income inequality used throughout the paper is the 80-20 percentile ratio of family income within the metropolitan area. Other income ratios, such as the 90-10, show a similar relationship to income inequality. I prefer the 80-20 ratio because it does not require making any assumptions about the income distribution of the highest (top-coded) income bin, and therefore is less subject to measurement error.

The construction of metropolitan area income inequality measures is based on a methodology described and tested in Jargowsky (1995). In particular, metropolitan area income is assumed to be distributed with a linear distribution below the mean and a pareto distribution above the mean.

J. Central City and Suburbs

Central cities are those places identified by the Census Bureau as such in 2003 based on metropolitan area residential and commuting patterns, and represent a consistent geographic area over time. There may be more than one central city in a metropolitan area; these are combined for the purpose of the analysis. The suburbs include all remaining tracted portions of the metropolitan area in a given year. Suburbs are also combined for the purpose of the analysis. Census tracts in 1980-2000 are matched to places, which in turn are matched to central cities. For cases in which a census tract includes both central city places and suburban places, it is considered part of the central city if at least half of the tract area is within the central city. In 1970, neither central cities nor places are identified in the data.

Appendix Figure 1.
Simulated Measurement of Average CGI by Number of Income Bins, 1970 and 1990