



National Poverty Center Working Paper Series

#05-2

March 2005

The GINI Coefficient and Segregation on a Continuous Variable

Jeongdai Kim, University of Texas at Dallas

Paul A. Jargowsky, University of Texas at Dallas

This paper is available online at the National Poverty Center Working Paper Series index at:
http://www.npc.umich.edu/publications/working_papers/

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the National Poverty Center or any sponsoring agency.

**The GINI Coefficient and Segregation
on a Continuous Variable***

Jeongdai Kim
University of Texas at Dallas
(jdkim@utdallas.edu)

Paul A. Jargowsky
University of Texas at Dallas
(jargo@utdallas.edu)

Abstract. The Gini Index is a widely used measure of income inequality. It has also been used as a segregation measure, but only in the case of binary variables, e.g. race or gender. We develop a general version of the Gini Index that can accommodate either continuous or binary variables, and discuss its relationship to existing measures. Using the Public Use Microdata Sample from the 2000 Census, we illustrate the calculation of the Index and show that it is highly correlated with an existing measure of economic segregation.

Approximate Word Count: 1,800

Classification Code: C43, D31, D63

Key words: Gini, Segregation, Inequality, Economic Segregation

* The authors wish to thank Brian Berry, Marie Chevrier, and Rebecca Yang for helpful comments and suggestions. We also thank the *Centre de Sciences Humaines* in New Delhi, India, where Paul Jargowsky was a Visiting Scholar when this paper was written.

The GINI Coefficient and Segregation on a Continuous Variable

1. BACKGROUND

Segregation is not only a social issue but also an economic issue that is closely related to the economic inequality and development. However, the issue has not received much attention in the literature. One of the reasons, we believe, is that there are few measures of segregation for economic variables such as income, since segregation measures are mainly developed for dichotomous variables such as the gender gap in occupations (Deutsch et al. 1994; Fluckiger and Silber 1990) and Black/White segregation (Carrington and Troske 1998; Duncan and Duncan 1955). A typical approach for a continuous variable has been to divide the continuous variable into two or more discrete categories, and then employ the traditional group-based measures (Massey and Eggers 1991). This approach has been criticized for discarding information and for potentially confounding changes in the underlying parameters of the continuous variable's distribution with changes in segregation (Jargowsky 1996). The Neighborhood Sorting Index (NSI), a version of eta squared, has been used to measure income segregation (Jargowsky 1996). However, without comparable measures, the performance of NSI has not been examined. This note aims to provide a valid measure of segregation for continuous variables based on the Gini coefficient.

From a measurement perspective, segregation is by definition a way of grouping units, whereas inequality is not. Segregation concerns the allocation of primary units with different levels of a characteristic, e.g. persons of different races or households of different incomes, among the subgroups of a larger group, e.g. neighborhoods within a

metropolitan area. Inequality, in contrast, examines the distribution of a characteristic of the primary units within a given group, with no regard for the membership of those units to subgroups.

The GINI coefficient (G) is a common measure of income inequality (Gini 1912, 1921). It has also been employed as a measure of segregation, although in that context, its use has generally been limited to dichotomous groups (Silber 1989; James and Taeuber 1985; Massey and Denton 1988).¹ We show how a modified version of the Gini Coefficient can be used as a segregation measure for continuous variables, a task for which there are few alternative measures (White 1986). To measure segregation along a continuous dimension, we modify G based on the idea of Jahn, Schmid and Schrag (1947) and compare it to NSI. In this note, we use household income as an example, but the argument is applicable to any continuous variable.

2. THE GEOMETRY OF THE INDEX

As shown in Figure 1, the Lorenz curve relates the cumulative percentage of total household income, sorted from lowest to highest income, as a function of the cumulative percent of total households (Lorenz 1905). The Gini Coefficient of Income Inequality is defined as the area between the Lorenz curve and the line of equality, divided by the total area under the line of the equality. In Figure 1, assume that L_i is a Lorenz curve based

¹ These measures have been extended to handle multiple racial and ethnic categories (James 1986; Morgan 1975; Reardon and Firebaugh 2002; Sakoda 1981; Silber 1989).

on data sorted by individual household income, and L_n , sorted by neighborhood means of household incomes.² The inequality index of individual households (G^i) is defined by:

$$G^i = (A_1 + A_2) / (A_1 + A_2 + B). \quad (1)$$

The Gini Coefficient of Neighborhood Inequality (G^n) with L_n is calculated in an analogous manner to G^i :

$$G^n = A_1 / (A_1 + A_2 + B) . \quad (2)$$

G^n is conceptually inappropriate as a measure of segregation, because it simply measures the inequality of groups without conveying any information on how the distribution changes when individual units are aggregated into groups while segregation is the degree of group separation *given* the distribution of individual units.³ Nor does G^n provide any sense of the degree of segregation because its upper bound varies depending on the distribution of individual units. When all neighborhoods have the same mean income, G^n takes a value of zero. However, when there is maximal segregation – every household living in neighborhoods with other households of an income – the Lorenz curve for neighborhoods (L_n) converges to the Lorenz curve for individuals (L_i) and the upper bound of G^n is determined by G^i .

Jahn, Schmid and Schrag (1947) introduced the ratio of G^n to the value of G which would obtain if there were complete segregation as a measure of Black-White segregation. Although it was proposed for a dichotomous case, extending their idea by

² It is obvious that L_n is always above or on L_i .

³ Hutchens (2004) is an example of a work that assumes G^n is a measure of segregation.

scaling G^n with G^i solves the conceptual and measurement problems. The Gini Coefficient of Segregation (G^s) is defined as:

$$G^s = \frac{G^n}{G^i} = \frac{A_1 / (A_1 + A_2 + B)}{(A_1 + A_2) / (A_1 + A_2 + B)} = \frac{A_1}{A_1 + A_2}. \quad (3)$$

Maximum segregation, given the existing income distribution, is indicated by G^s being equal to one, and minimum segregation, by zero. G^s is a general form of the GINI segregation index that includes binary and continuous variable cases as shown in the next section.

3. FORMULATION

Literature on the decomposition of Gini provides various computation methods for G^i and G^n (Mussard 2004; Pyatt 1976; Silber 1989; Yao 1999). Following the geometrical definition in Figure 1, and the area calculation method shown in Figure 2,

$$\begin{aligned} G^i &= (A_1 + A_2) / (A_1 + A_2 + B) \\ &= (0.5 - B) / 0.5 \\ &= 1 - 2B \end{aligned} \quad (4)$$

$$2B = \sum_{i=1}^N (Y_{i-1} + Y_i)(H_i - H_{i-1}) \quad (5)$$

where Y_i and H_i are, respectively, cumulative percentages of income and number of households at the i^{th} household ranked in ascending order by household income from one to N . G^n is calculated in an analogous manner, ranking the households in ascending order by mean neighborhood income. Then, given $H_i - H_{i-1} = 1/N$,

$$G^s = \frac{G^n}{G^i} = \frac{1 - \sum_{i=1}^N (Y_{ni-1} + Y_{ni})(H_i - H_{i-1})}{1 - \sum_{i=1}^N (Y_{i-1} + Y_i)(H_i - H_{i-1})} = \frac{N - \sum_{i=1}^N (Y_{ni-1} + Y_{ni})}{N - \sum_{i=1}^N (Y_{i-1} + Y_i)} \quad (6)$$

In (6), Y_{ni} is the cumulative percentage of average household income of the neighborhood in which the i^{th} household resides. Formula, (6) is a general form applicable to both binary and continuous variable cases. Extending Anand (1983), (6) can be also expressed in the form of mean difference Gini as follows.

$$G^s = \frac{G^n}{G^i} = \frac{\frac{1}{2N^2m} \sum_{i=1}^N \sum_{j=1}^N |y_{ni} - y_{nj}|}{\frac{1}{2N^2m} \sum_{i=1}^N \sum_{j=1}^N |y_i - y_j|} \quad (7)$$

where y_i is the income of the i^{th} households, y_{ni} is the average household income of the neighborhood in which the y_i household resides, and m is the overall mean income.

To see the applicability of (6) and (7) to binary cases, assume that there are B Blacks and W Whites in T population ($T = B + W$), assign a value of one for each Black and zero for each White, define $P = (B/T)$, and consider Lorenz curves in two dimensional space of the percentage population (horizontal axis) and the percentage Black (vertical axis). Then, in (7), m is replaced with P , y_i equals one for Black and zero for White, Y_i will be the cumulative percentage of Black to the total population, y_{ni} is the proportion Black in the i^{th} household's neighborhood, and its cumulative percentage is Y_{ni} . Since G^i , the denominator of G^s in (7), becomes $1 - P$,

$$G^s = \frac{\sum_{i=1}^N \sum_{j=1}^N |P_{ni} - P_{nj}|}{2N^2 P(1 - P)} \quad (8)$$

where the proportion of Black in the i^{th} household's neighborhood, y_{ni} , is replaced with P_{ni} for the consistency of terminology. (8) is the formula frequently used in the literature for binary segregation cases. Thus, it is clear that (6) and (7) are general forms of the Gini segregation index that include binary cases.

We will contrast G^s with the Neighborhood Sorting Index (NSI), which is based on the idea that there is a distribution of individual household incomes and a distribution of neighborhood mean incomes. To the extent that individuals of different incomes reside in neighborhoods together, the dispersion of neighborhood income will be less than the dispersion of household income (Jargowsky 1996). Thus, the NSI may be expressed as:

$$NSI = \sqrt{\frac{\text{Variance Between Neighborhoods}}{\text{Total Variance}}} = \sqrt{\frac{\sum_{i=1}^N (y_{ni} - m)^2}{\sum_{i=1}^N (y_i - m)^2}} \quad (9)$$

where y_{ni} is the average income of the neighborhood to which the i^{th} household belongs. NSI attains a minimum of 0 when all neighborhood mean incomes are equal to area-wide mean income, so there is no variance between neighborhoods. It attains a theoretical maximum of 1 when all persons live only with others who have identical incomes, so that all variance in household income is between neighborhoods. The square of this measure is equivalent to eta squared from a one-way analysis of variance in which the categorical variable is neighborhood.

4. EMPIRICAL APPLICATION

For the purpose of illustration, we use the 5-percent Public Use Microdata Sample (PUMS) from the 2000 U.S. Census (Ruggles 2004). The data is a 5 percent sample of all U.S. households, a subset of the households that completed the 2000 Census “long form” survey. Two levels of geography are employed: 1) the 25 largest U.S. metropolitan areas, and 2) Public Use Microdata Areas (PUMAs), which function as neighborhoods in this analysis.⁴ We exclude households in non-metropolitan areas. Table 1 shows the number of PUMAs, number of sample households, weighted number of total households, and mean household income for the 25 largest metropolitan areas.

Table 2 shows the individual, neighborhood, and income segregation Gini coefficients, as well as the standard deviations of the household and neighborhood income distributions and the Neighborhood Sorting Index. Metropolitan areas vary little from each other in terms of income inequality, but have very different levels of sorting across neighborhoods. G^s and NSI rank metropolitan areas in a very similar fashion.

Figure 3 shows the near perfect correspondence between the NSI and the G^s in the sample of 25 metropolitan areas. The Pearsonian correlation coefficient is 0.9771, and it is highly significant. Clearly, these measures are tapping into a common underlying phenomenon. However, the NSI, as a function of squared deviations, is known to give greater weight to higher incomes than GINI, which does not overweight any segment of

⁴ PUMAs are bigger than census tracts, which more frequently served as neighborhood proxies (White 1987). The 25 largest metropolitan areas contain 693 PUMAs. A few PUMAs on the fringes of metropolitan areas including territory in neighboring metropolitan areas or non-metropolitan areas, and thus did not have a metropolitan area code. These PUMAs and the households they contain were excluded from the analysis. The average PUMA size was 50,000 households (weighted), based on an average sample size of 2,250.

the distribution (James and Taueber 1985; White 1986). This may be considered an advantage in certain applications, particularly those involving smaller, more heterogeneous neighborhood units. However, the optimal weighting depends on the researcher's implicit social welfare function (Atkinson 1970; Dalton 1920).

In the absence of individual-level data, both measures present a challenge to researchers. In the case of NSI, the total household variance must be estimated from the aggregated data (Jargowsky 1995: Appendix A). In the case of G^s , a distribution of the individual household incomes must be assumed. In a highly skewed distribution, as is typical for income, the variance may be more difficult to estimate accurately than the mean, because of the higher penalty for misestimating the incomes and deviations in the open-ended top bracket. This is an advantage for the Gini approach.

5. CONCLUSION

The Gini Index of Neighborhood Inequality, G^n , the application of the Gini Index to neighborhood-level income, is not a measure of economic segregation. Rather it is a mixture of information about the inequality of the income distribution and the ameliorative effects of neighborhood heterogeneity. Normalizing the neighborhood-level Gini by the individual-level Gini provides a measure of segregation, and provides an alternative measure of economic segregation that may be easier to estimate in many cases than the Neighborhood Sorting Index. In our application to household income in U.S. metropolitan areas, the two measures were highly correlated across 25 metropolitan areas. In general, the Gini Index of Segregation may be used with either continuous or dichotomous variables.

References

- Anand, S. (1983) *Inequality and Poverty in Malaysia: Measurement and Decomposition*. Oxford: Oxford University Press.
- Atkinson, A. B. (1970) On the measurement of inequality. *American Sociological Review*, 43, 865-80.
- Carrington, W.J. & Troske, K. R. (1998) Interfirm segregation and the black/white wage gap. *Journal of Labor Economics*, 16, 231-260.
- Dalton, H. (1920) Measurement of the inequality of income. *The Economic Journal*, 30, 348-61.
- Deutsch, J., Fluckiger, Y., & Silber, J. (1994) Measuring occupational segregation. *Journal of Econometrics*, 61, 133-146.
- Duncan, O. D. & Duncan, B. (1955) A methodological analysis of segregation measures. *American Sociological Review*, 20, 210-217.
- Fluckiger, Y. & Silber, J. (1990) Career choice and the measurement of segregation by sex in employment: the case of students in Switzerland. Pp. 97-108 in *Modeling the Labor Market*, Proceedings of the Meeting of the Applied Econometrics Association, Stasbourg.
- Gini, C. (1912) Variabilita e mutabilita. Reprinted in E. Pizetti & T. Salvemini, eds., *Memorie di Metodologia Statistica*. Rome: Libreria Erendi Virgilio Veschi, 1955.
- _____. (1921) Measurement of Inequality of Incomes. *The Economic Journal*, 31, 124-126.
- Hutchens, R. (2004) One Measure of Segregation. *International Economic Review*, 45, 555-578.
- Jahn, J., Schmid, C. F. & Schrag, C. (1947) The measurement of ecological segregation. *American Sociological Review*, 3, 293-303.
- James, D. R. & Taeuber, K. E. (1985) Measures of Segregation. *Sociological Methodology*, 15, 1-32.
- James, F. J. (1986) A new generalized 'exposure-based' segregation index. *Sociological Methods and Research*, 14, 301-316.
- Jargowsky, P. A. (1995) *Take the money and run: economic segregation in U.S. metropolitan areas*. Discussion Paper 1056-95 (January 1995). Madison, Wisconsin: Institute for Research on Poverty.
- _____. (1996) Take the money and run: economic segregation in U.S. metropolitan areas. *American Sociological Review*, 61, 984-998.
- Lorenz, M. O. (1905) Methods for measuring the concentration of wealth. *American Statistical Association*, 9, 209-219.

- Massey, D. S. & Denton, N. A. (1988) The dimensions of racial segregation. *Social Forces*, 67, 281-305.
- Massey, D. S. & Eggers, M. L. (1991) The ecology of inequality: minorities and the concentration of poverty, 1970-1980. *American Journal of Sociology*, 95, 1153-1188.
- Morgan, B. S. (1975) The segregation of socioeconomic groups in urban areas: a comparative analysis. *Urban Studies*, 12, 47-60.
- Mussard, S. (2004) The bidimensional decomposition of the Gini ratio. A case study: Italy. *Applied Economics Letters*, 11, 503-505.
- Pyatt, G. (1976) On the interpretation and disaggregation of Gini coefficients. *Economics Journal*, 86, 243-255.
- Reardon, S. F. & Firebaugh, G. (2002) Measures of multigroup segregation. *Sociological Methodology*, 32, 33-67.
- Ruggles, S., Sobek, M., Alexander, T., Fitch, C. A., Goeken, R., Hall, P. K., King, M. & Ronnander, C. (2004) Integrated Public Use Microdata Series: Version 3.0 [Machine-readable database]. Minneapolis, MN: Minnesota Population Center. Accessed at <http://www.ipums.org> on November 11, 2004.
- Sakoda, J. (1981) A generalized index of dissimilarity. *Demography*, 18, 245-250.
- Silber, J. (1989) Factor components, population subgroups and the computation of the Gini index of inequality. *The Review of Economics and Statistics*, 71, 107-115.
- White, M. J. (1986) Segregation and diversity measures in population distribution. *Population Index*, 52, 198-221.
- _____. (1987) *American Neighborhoods and Residential Differentiation*. New York: Russell Sage Foundation.
- Yao, S. (1999) On the decomposition of Gini coefficients by population class and income source: a spreadsheet approach and application. *Applied Economics*, 31, 1249-1264.

Figure 1: Lorenz Curves for Individual
Households and Neighborhoods

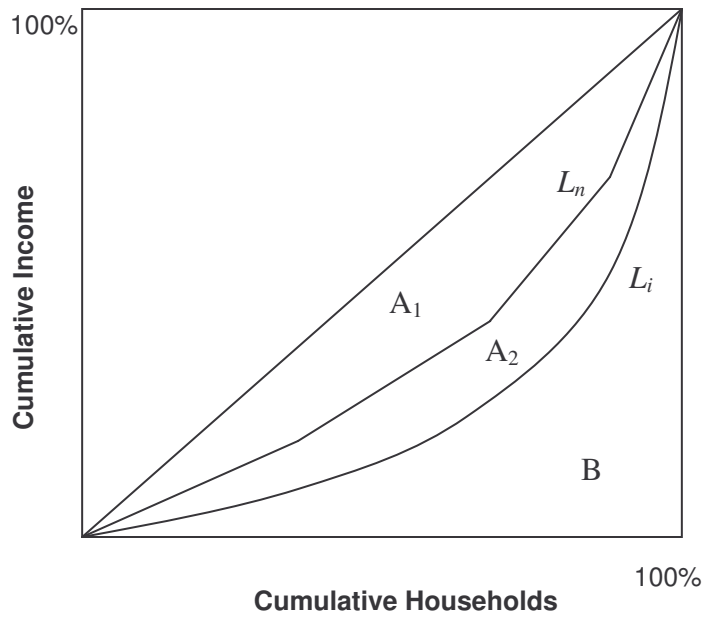


Figure 2: Area Under the Lorenz Curve

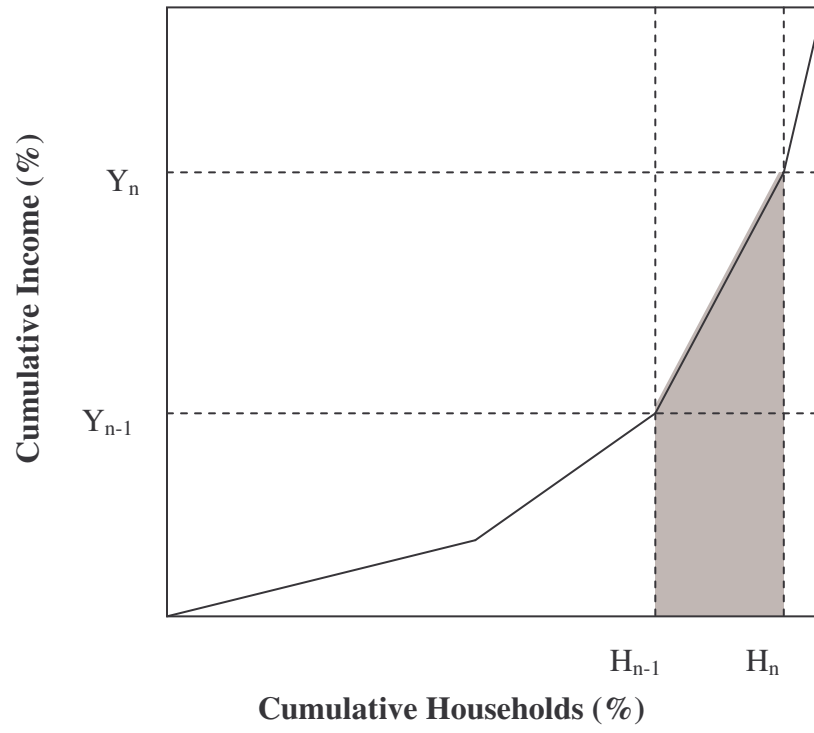


Figure 3: Economic Segregation in Metropolitan Areas by PUMS Areas
25 Largest Metropolitan Areas, 2000

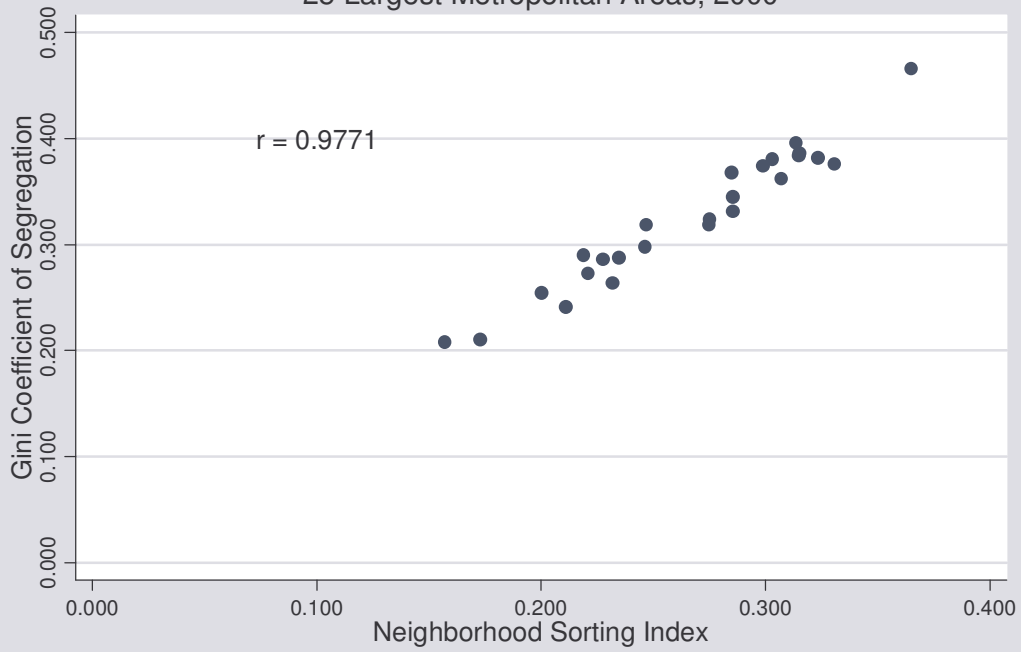


Table 1: Households, Income, and Sample Characteristics of the 25 Largest Metropolitan Areas, 2000 U.S. Census 5 Percent PUMS Sample

Metropolitan Area	Households	Sample Size	PUMS Areas	Mean Household Income (\$)
1 New York-Northeastern NJ	3,482,324	144,070	64	62,237
2 Los Angeles-Long Beach, CA	3,135,972	154,419	67	60,975
3 Chicago-Gary-Lake, IL	2,937,970	122,571	54	68,059
4 Philadelphia, PA/NJ	1,913,900	79,438	37	62,560
5 Washington, DC/MD/VA	1,777,008	86,350	33	80,813
6 Detroit, MI	1,692,431	66,836	32	63,025
7 Houston-Brazoria, TX	1,462,022	63,582	30	61,519
8 Atlanta, GA	1,460,471	64,092	32	67,331
9 Boston, MA	1,320,037	62,541	29	74,018
10 Dallas-Fort Worth, TX	1,225,886	56,568	23	66,613
11 Phoenix, AZ	1,133,243	53,493	22	59,445
12 Minneapolis-St. Paul, MN	1,096,388	38,674	21	67,598
13 Riverside-San Bernadino, CA	1,033,275	48,449	20	54,767
14 St. Louis, MO-IL	1,011,780	41,947	19	57,295
15 Tampa-St. Petersburg-Clearwater, FL	1,007,772	49,057	20	50,889
16 San Diego, CA	993,065	47,749	16	63,224
17 Baltimore, MD	958,948	44,495	22	62,919
18 Orange County, CA	936,488	47,397	17	77,994
19 Pittsburgh-Beaver Valley, PA	934,238	42,066	17	50,563
20 Seattle-Everett, WA	932,566	40,990	20	69,043
21 Nassau Co, NY	916,196	47,392	24	87,956
22 Cleveland, OH	892,471	43,189	19	55,224
23 Oakland, CA	866,144	42,740	18	77,483
24 Denver-Boulder-Longmont, CO	776,515	34,314	18	64,667
25 Miami-Hialeah, FL	768,759	36,750	19	51,955
25 Metropolitan Areas	34,665,869	1,559,169	693	64,842

Source: 2000 Public Use Microdata 5% Sample, U.S. Bureau of the Census. Tabulation by the authors.

Notes: Includes households in PUMAs with identified metropolitan areas. All computations, except sample size and PUMS areas, use the PUMS sample weight.

Table 2: Gini Coefficients and Standard Deviations of Individual and Neighborhood Income Distributions, and Economic Segregation Measures, 25 Largest U.S. Metropolitan Areas, 2000.

Metropolitan Areas	Gini Coefficients			Standard Deviation (\$)		Neighborhood Sorting Index (NSI)
	Income Inequality G^i	Neighborhood Inequality G^n	Segregation G^s	Individual Household Income	Neighborhood Mean Income	
1 New York-Northeastern NJ	0.535	0.249	0.465	79,189	28,905	0.365
2 Los Angeles-Long Beach, CA	0.496	0.185	0.374	69,626	20,806	0.299
3 Chicago-Gary-Lake, IL	0.456	0.173	0.381	69,145	20,972	0.303
4 Philadelphia, PA/NJ	0.462	0.183	0.395	63,655	19,971	0.314
5 Washington, DC/MD/VA	0.428	0.155	0.362	73,974	22,720	0.307
6 Detroit, MI	0.444	0.167	0.376	61,382	20,299	0.331
7 Houston-Brazoria, TX	0.470	0.150	0.318	65,164	16,098	0.247
8 Atlanta, GA	0.440	0.152	0.345	65,606	18,734	0.286
9 Boston, MA	0.471	0.152	0.324	76,594	21,070	0.275
10 Dallas-Fort Worth, TX	0.459	0.132	0.288	69,182	16,233	0.235
11 Phoenix, AZ	0.439	0.140	0.318	57,938	15,926	0.275
12 Minneapolis-St. Paul, MN	0.415	0.113	0.272	61,455	13,579	0.221
13 Riverside-San Bernadino, CA	0.437	0.092	0.211	51,986	8,981	0.173
14 St. Louis, MO-IL	0.440	0.169	0.384	56,637	17,825	0.315
15 Tampa-St. Petersburg-Clearwater, FL	0.455	0.095	0.208	53,873	8,459	0.157
16 San Diego, CA	0.454	0.135	0.298	64,302	15,840	0.246
17 Baltimore, MD	0.438	0.169	0.386	59,543	18,763	0.315
18 Orange County, CA	0.446	0.127	0.286	76,229	17,356	0.228
19 Pittsburgh-Beaver Valley, PA	0.462	0.117	0.254	53,833	10,777	0.200
20 Seattle-Everett, WA	0.436	0.105	0.241	67,626	14,271	0.211
21 Nassau Co, NY	0.435	0.115	0.264	83,193	19,291	0.232
22 Cleveland, OH	0.451	0.166	0.368	56,746	16,175	0.285
23 Oakland, CA	0.451	0.172	0.382	74,870	24,204	0.323
24 Denver-Boulder-Longmont, CO	0.432	0.143	0.331	62,140	17,738	0.285
25 Miami-Hialeah, FL	0.509	0.148	0.290	63,666	13,943	0.219
25 Metropolitan Areas	0.462	0.160	0.345	67,051	19,003	0.281