Representative Neighborhoods of the United States

Alejandro Badel

Many metropolitan areas in the United States display substantial racial segregation and substantial variation in incomes and house prices across neighborhoods. To what extent can this variation be summarized by a small number of representative (or synthetic) neighborhoods? To answer this question, U.S. neighborhoods are classified according to their characteristics in the year 2000 using a clustering algorithm. The author finds that such classification can account for 37 percent of the variation with two representative neighborhoods and for up to 52 percent with three representative neighborhoods. Furthermore, neighborhoods classified as similar to the same representative neighborhood tend to be geographically close to each other, forming large areas of fairly homogeneous characteristics. Representative neighborhoods seem a promising empirical benchmark for quantitative theories involving neighborhood formation. (JEL R2, D31, D58, J24)

Federal Reserve Bank of St. Louis *Review*, Second Quarter 2014, *96*(2), pp. 147-72.

Racial segregation is a striking trait of U.S. cities. Iceland, Weinberg, and Steinmetz (2002) report that 64 percent of the black population would have needed to change residence for all U.S. neighborhoods to become fully integrated in the year 2000. Income differences across neighborhoods have also been well documented. Wheeler and La Jeunesse (2007) report that between-neighborhood inequality in 2000 represented around 20 percent of overall annual household income inequality in Census data. The variation in housing prices across neighborhoods has also been the focus of a large literature.[1]

This article attempts to summarize the landscape of U.S. cities using a small number of representative neighborhoods. The motivation for this effort is twofold. On the one hand, a clear and concise characterization of the American urban landscape may be useful in the construction of theories involving neighborhood formation. On the other hand, a simple representation can be used to impose empirical discipline on quantitative models with a small number of locations. These types of models are important since they can address complex dynamic issues such as the interaction between neighborhood formation and human capital accumulation without becoming computationally infeasible (see, for example, Fernandez and

Alejandro Badel is an economist at the Federal Reserve Bank of St. Louis. The author thanks Christopher Martinek, Brian Greaney, and Brian Bergfeld for research assistance and Juan Sánchez for useful comments.

Rogerson, 1998). Here it is important to highlight that another part of the urban landscape consists of household heterogeneity *within* neighborhoods (see Ioannides, 2004). This article focuses exclusively on variation *across* neighborhoods.

The empirical strategy consists of applying a clustering algorithm to Census 2000 data describing U.S. neighborhoods. The *K*-means clustering algorithm is used here. This algorithm attempts to classify neighborhoods in such a way that neighborhoods within a cluster are similar to one another and dissimilar with respect to neighborhoods in other clusters. The aggregate of all neighborhoods in each cluster is interpreted as a representative neighborhood.

The rest of the article proceeds as follows. The next section describes the data, and the following section explains the clustering algorithm. Subsequent sections describe the clustering results and the representative-neighborhood characterization. These descriptions are followed by a section with robustness exercises. The final section provides conclusions and closing remarks.

DATA

Data for this study are from the 2000 Census of Population and Housing Summary File 3 (SF3) (U.S. Census Bureau, 2000). The SF3 contains geographically coded summary statistics at various levels of spatial aggregation.

This study focuses on the Census-tract level of geographic aggregation. Census tracts are small geographic subdivisions of the United States. According to the Census Bureau, tract boundaries are defined with the goal of obtaining areas containing demographically and economically homogeneous populations of about 4,000 people. These tract features are obviously desirable for classifying neighborhoods into distinct types.

The set of variables used as Census counterparts for income, racial composition, and house prices is described next. Table 1 defines these variables in terms of SF3 variable names.

Variables

Income (*Y*). Two measures of a tract's income are used. First, a tract's *labor* income (earnings hereafter) is measured as the log of average household earnings in the tract. Second, a tract's total income is measured as the log of average household income in the tract.

Racial Composition (*R*). The measure of racial composition used here is the fraction of white households in the tract. This fraction is obtained as the number of non-Hispanic white households divided by the total number of households in a Census tract.

Price of Housing Services (*P*). Three variables in the dataset can be used to construct measures of the price of housing services: median gross rent, median house value, and median owner costs (see the appendix for details). These variables are measures of housing expenditures. Since expenditures are the product of quantity and price, log expenditures equal the sum of a log price component and the log number of units consumed. The price component is isolated here by regressing the log of the median expenditure measure against a set of house

Table 1

Variable Definitions

Variable	Definition (Census code)
Fraction of black HHs	p151b001/(p151a001+…p151g001)
Fraction of non-Hispanic white HHs	p151i001/(p151a001+…p151g001)
Average tract HH earnings	p067001/p058001
Average tract HH income	p054001/p052001
Average white HH income	p153i001/p151i001
Average black HH income	p153b001/p151b001
Average both races HH income	(P153a001+...P153g001-p153i001-p153b001)
	/(P151a001+...P151g001-p151i001-p151b001)
Median gross rent	H063001
Median value (owner-occupied)	H085001
Median selected monthly owner costs	H091001 (owner-occupied with mortgage)
Median number of rooms in unit	H027002 (owner), H027003 (renter)
Distribution of number units in structure	H032003-012/H032002 (owner),
	H032014-023/H032013 (renter)
Median year structure built	H037002 (owner), H037003 (renter)
Distribution of number of bedrooms	H042003-008/H042002 (owner),
	H042010-015/H042009 (renter)
Fraction with telephone service	H043003/H043002 (owner),
	H043020/H043019 (renter)
Fraction with plumbing facilities	H048003/H048002 (owner),
	H048006/H048005 (renter)
Fraction with kitchen facilities	H051003/H051002 (owner),
	H051006/H051005 (renter)
Distribution of heating fuel	HCT010003-011 (owner),
	HCT0010013-021 (renter)
Distribution of time to work	P031003-014/P031002
Fraction of population in group quarters	P009025/P0009001

NOTE: HH, household.

SOURCE: Census Bureau. 2000 Census of Population and Housing—Summary File 3, Technical Documentation, released September 2002.

Table 2

Definition of Variable Configurations

Name	Income	Racial composition	Price of housing services
ben	Log household earnings	% Non-Hispanic whites	Clean IRV (owners)
inc	Log household income	% Non-Hispanic whites	Clean IRV (owners)
prent	Log household earnings	% Non-Hispanic whites	Clean rent (renters)
pcost	Log household earnings	% Non-Hispanic whites	Clean owner's cost (owners)

NOTE: Implicit rental value (IRV) is defined as a percentage of a home's market value.

characteristics and using the residual from this regression as the measure of the price of housing services.[2]

The benchmark measure of (Y, R, P) is composed of the log of mean earnings, the fraction of white households, and the "clean" log value of housing for owners. Results for alternate configurations after replacing one of the variables with an alternative measure are also presented. This changing-one-variable-at-a-time strategy results in three additional sets of variables. Each configuration is denoted by the name of the variable that changes with respect to the benchmark configuration (see Table 2 for the variables in each variable configuration).

Sample Selection

The baseline sample aims to provide a comprehensive picture of the distribution of income, racial composition, and house prices in U.S. metro areas. Metropolitan statistical areas (MSAs) with populations of at least 1 million are considered. Since the focus is on the black and non-Hispanic white populations, the sample is further restricted to MSAs where at least 10 percent of the population is black.

Within each selected MSA, the sample is restricted to Census tracts where less than 50 percent of the population reports being neither black nor non-Hispanic white. To guarantee the exclusion of rural areas, only the Census tracts with at least 100 people per square kilometer are retained.[3] Attention is also restricted to tracts with at least 200 households and no more than 25 percent of the population living in group quarters.[4]

Application of these sample selection criteria results in a set of 28 MSAs in 25 states including 80.7 million people and 17,815 Census tracts. The largest MSA in the sample is New York-Northern New Jersey-Long Island, with 3,850 tracts; the smallest is Raleigh-Durham-Chapel Hill, with 157 tracts. Table 3 presents the number of observations deleted by each criterion. Table 4 lists some summary statistics of the final sample. The section on robustness compares the results obtained under the baseline sample with those obtained under four variations of the sample selection criteria.

Table 3

Sample Selection Criteria

Criteria	Observations dropped	Total observations
Initial without missing values		50,167
MSA population less than 1 million	14,397	35,770
MSA with less than 10% black HHs	14,244	21,526
Population density less than 100/sq km	1,785	19,741
Other race more than 50%	1,421	18,320
Tract with less than 200 HHs	226	18,094
Institutionalized population more than 25%	279	17,815

NOTE: Each observation corresponds to a Census tract. HH, household.

Table 4

Descriptive Statistics: Main Variables

Variable	Mean	SD	5th Percentile	95th Percentile
Fraction black HHs	0.23	0.32	0	0.95
Fraction white HHs	0.66	0.32	0.01	0.97
Fraction other race HHs	0.10	0.11	0.01	0.35
Average tract HH income ($)	63,921	33,981	28,178	125,278
Average tract HH income ($, blacks)	55,927	43,712	20,508	117,500
Average tract HH income ($, whites)	66,413	36,393	26,702	130,605
Average tract HH income ($, other races)	61,015	37,956	21,550	124,834
Median IRV* ($)	13,372	5,593	6,748	22,571
Median gross rent* ($)	8,806	2,413	5,782	12,743
Median selected owner costs* ($)	15,415	3,795	10,323	21,772
Tract population	4,427	2,265	1,536	8,403
Number of HHs in tract	1,701	890	573	3,300
Population density (population/sq km)	3,391	6,150	192	13,605
Fraction of population in group quarters	0.01	0.03	0	0.08

NOTE: HH, household; IRV, implicit rental value. *Statistics reported controlling for certain factors via linear regression; see Data section for details.

CLUSTERING ALGORITHM

Cluster analysis attempts to classify a large set of objects into a small number of groups (clusters). A perfect classification is obtained if the large set is composed of a small number of groups of identical objects. For example, a dataset composed of only zeros and ones can be perfectly classified with two clusters.

A common clustering method consists of minimizing a square error (SE) criterion. The method used is known as the K-means algorithm and creates a partition of a set containing I objects into K mutually exclusive subsets (where $I \geq K$). How? Suppose each element $i \in I$ is described by the vector x_i. The algorithm searches for a partition of I into subsets $\{C_1, C_2, C_3,\ldots, C_k,\ldots, C_K\}$ that minimizes the within-cluster variation of x_i (or the SE) around each group's centroid c_k. The centroid c_k of cluster C_k is usually taken to be the vector of averages of x_i taken over all elements i belonging to the cluster C_k:

$$SE = \sum_{k} \sum_{i \in C_k} \omega_i \left(\mathrm{x}_i - c_k\right) \cdot \left(\mathrm{x}_i - c_k\right),$$

where ω_i is a weighting factor equal to the number of households in each tract.

Conceptually, the optimal partition could be found by computing the SE for every possible partition of I and then choosing the one that produces the smallest SE. In practice, the search needs to be conducted with a heuristic algorithm known as "iterative relocation."[5] A cluster resulting from iterative relocation has two desirable properties. First, each cluster has a centroid, which is the mean of the objects in that cluster. Second, each object belongs to the cluster with the nearest centroid. On the downside, this type of algorithm does not guarantee finding the optimal partition, and its outcome depends on the initial partition. For all clustering exercises reported here, the clustering procedure is applied 10 times using random starting values, and the cluster that minimizes the SE is reported.

Normalization of Data

A cluster's outcome is sensitive to the relative scaling of variables that describe each tract. One solution to this problem is to normalize each component of x_i to have a sample mean of 0 and a sample variance of 1. This method is referred to as z score normalization in what follows.[6] An alternative normalization method, based on the Mahalanobis transformation, accounts for the correlations across components of x_i. This method normalizes the data by the inverse of the sample covariance matrix of x_i, $\hat{\Omega}^{-1}$. In this case, SE becomes the standard error of the mean (SEM):

$$SEM = \sum_{k} \sum_{i \in C_k} \omega_i \left(\mathrm{x}_i - \mathrm{c}_k\right)' \hat{\Omega}^{-1} \left(\mathrm{x}_i - \mathrm{c}_k\right).$$

A comparison of selected results using z score and Mahalanobis normalizations is provided.

Table 5

Cluster Compactness (percent)

Variable configuration	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
z Score normalization					
ben	37	50	57	62	66
inc	37	52	58	63	68
prent	34	52	58	63	67
pcost	36	50	57	62	66
Mahalanobis normalization					
ben	26	43	53	58	62
inc	26	44	54	58	63
prent	26	43	54	59	62
pcost	26	43	54	58	62
Average	31	47	56	60	65

NOTE: This statistic corresponds to the percentage of (Y, R, P) sum of variances explained by between-cluster variation.

RESULTS

This section describes the results of the clustering exercise. The main results are obtained by applying the clustering algorithm once to the full sample of neighborhoods from all MSAs. Alternate results obtained by applying the algorithm separately to each MSA are reported in the "Regional Stability" subsection.

Cluster Validity

How much of the variance can be captured by a clustering representation? One way to address this question is to assess the "compactness" of a cluster.[7] I use an intuitive indicator of compactness to address the validity of the clusters obtained and complement it with a visual summary of the distribution of (Y, R, P) within and between clusters.

The compactness indicator compares the SE from the clustering algorithm with the overall variability of x_i with respect to the vector of sample means, c.[8] In what follows, this measure is referred to as R^2 because of its mechanical similarity to the familiar concept from standard econometric analysis:

$$R^2 = 1 - \frac{SE}{\sum_k \sum_{i \in C_k} \omega_i (\mathrm{x}_i - \mathrm{c}) \cdot (\mathrm{x}_i - \mathrm{c})}.$$

A value of $R^2 = 1$ means that the data consist of K types of identical elements. Table 5 presents the R^2 values obtained for $K = 2, 3, 4, 5, 6$ and each of the selected variable configurations using z score and Mahalanobis normalizations.

Not surprisingly, compactness increases with *K*. For $K = 2$, R^2 averages 31 percent across all variable configurations and normalizations and its maximum is 37 percent. The average increases to 47 percent when $K = 3$ and increases further to 65 percent as *K* increases from 3 to 6. Thus, most of the gains in explanatory power occur at $K = 2$ and $K = 3$. These clusterings provide a reasonable degree of compactness while maintaining an acceptable level of complexity. With $K \geq 4$, the complexity becomes substantially greater without a significant increase in explanatory power.

Figures 1 and 2 show a variety of statistics regarding the distribution of (Y, R, P) within and between clusters for $K = 2$ and $K = 3$, respectively. These plots reflect a large degree of similarity across different variable configurations measuring (Y, R, P) and large differences in the distributions of each variable across clusters.

For instance, consider the first column of Figure 1. The blue boxes depict the interquartile range of the distribution of racial configuration (i.e., the fraction of white households in the neighborhood) in each cluster. For all rows, the interquartile ranges of each cluster do not intersect. Average incomes show a similar result (see the second column). In contrast, all interquartile ranges for the average price of housing services of the two clusters intersect, although the central tendency is the same as for income.

The brackets in each plot in Figure 1 represent the range between the 5th and 95th percentiles of each distribution. For the fraction of white residents, the 95th percentile of neighborhoods of type *I* is below the median of neighborhoods of type *II* (depicted as the center of the corresponding blue box) and is also below the mean (depicted by a vertical solid line). For all variable configurations, the 5th percentile of neighborhood *II* is above the mean and the median of neighborhood *I*.

Figure 2 shows the $K = 3$ case. As shown in the first column, the separation of the distributions of racial configuration across clusters becomes larger between neighborhoods of type *A* and neighborhoods of type *B* and *C* than it is between neighborhoods of type *I* and *II* for $K = 2$. In turn, the distributions in neighborhoods of type *B* and *C* overlap substantially. The second column shows a different picture for income: Income distribution in neighborhood *C* is separated from those in neighborhoods *A* and *B*, while those in neighborhoods *A* and *B* exhibit significant overlap. The third column shows that patterns for distributions of house prices behave more like the patterns for income than those for race.

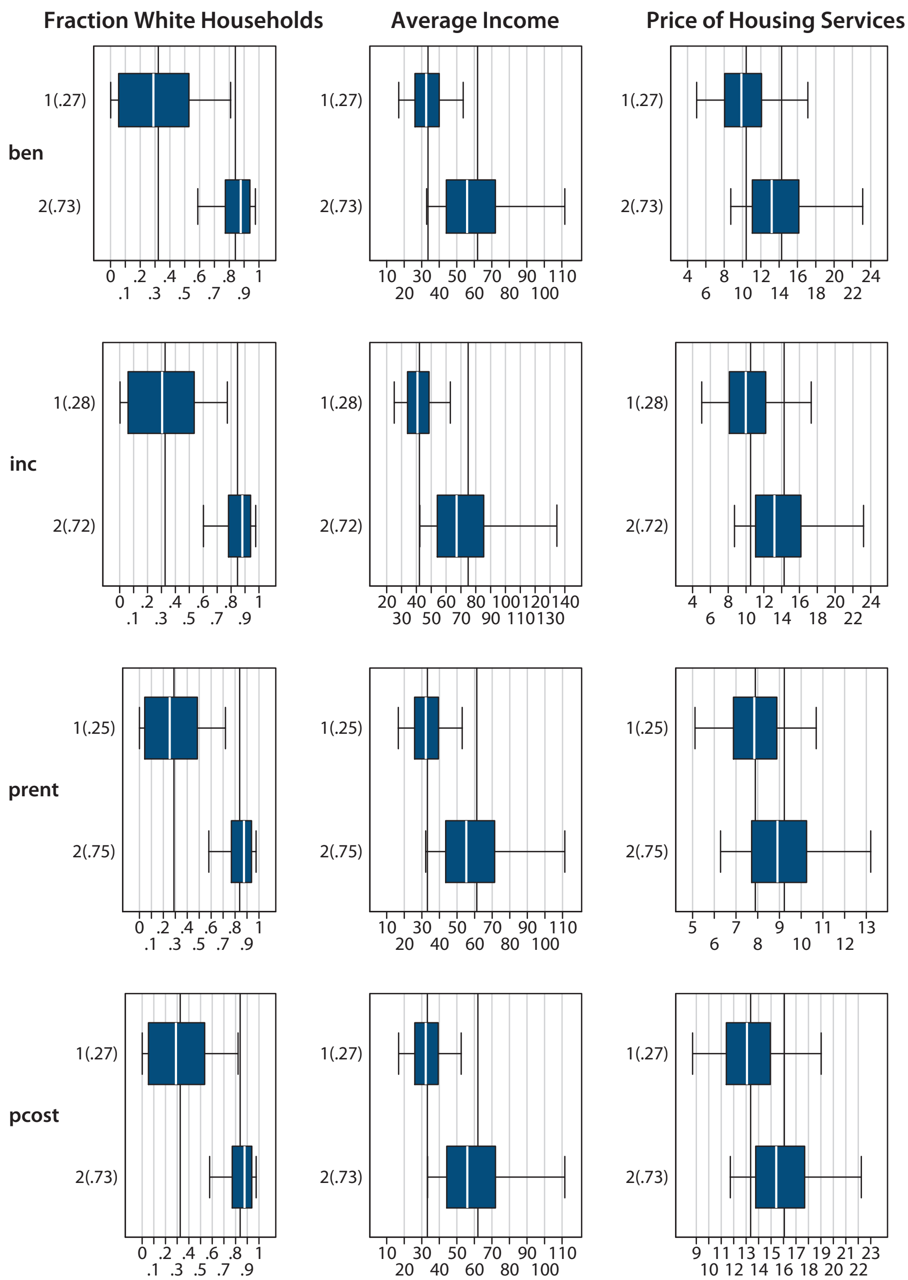
In summary, an off-the-shelf clustering procedure can be used to capture (i) up to 37 percent of the variation in income, racial configuration, and housing service prices across U.S. neighborhoods using only two representative neighborhoods and (ii) up to 52 percent of the variation using three representative neighborhoods.

Spatial Contiguity

Spatial theories of human capital formation emphasize spillovers across geographic locations. A common view states that the strength of these interactions declines with geographic distance. Therefore, the degree to which the tracts in each cluster are spatially contiguous suggests that the classification is potentially consistent with theories featuring spatial spillovers. In contrast, a low degree of contiguity would imply that each cluster is composed of scattered

Figure 1

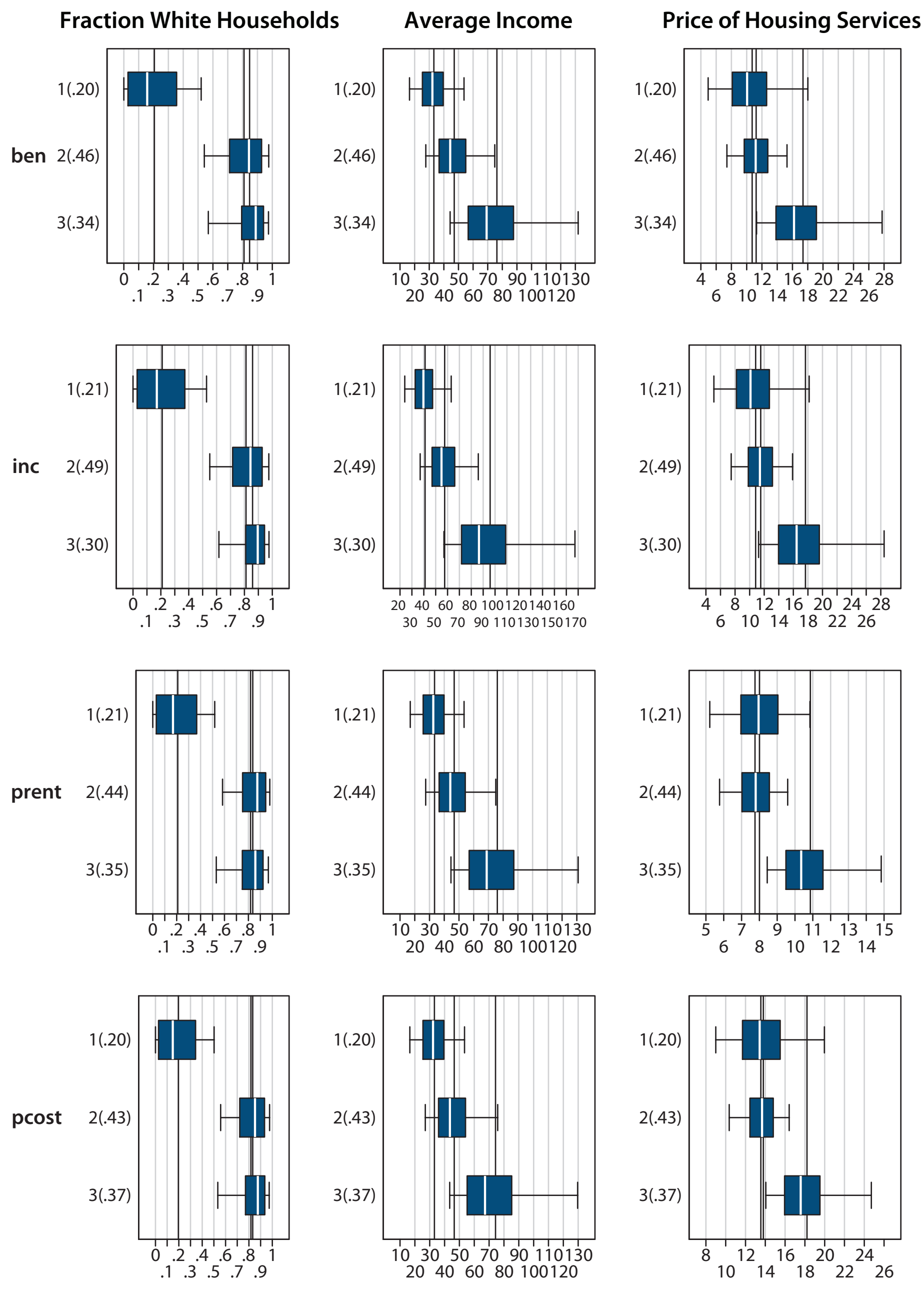
Within-Cluster Distribution of Neighborhood Characteristics: $K = 2$



NOTE: Columns show the plots for (i) racial configuration, (ii) earnings, and (iii) price of housing services measures. Rows correspond to each variable configuration. Within each plot, neighborhood classes (1 and 2 stand for types *I* and *II*, respectively) are listed in the vertical axis (the fraction of households is listed in parentheses). Vertical lines indicate neighborhood means (or centroid c_k). Boxes indicate the range between the 25th and 75th percentiles. Lines within boxes indicate medians. Brackets indicate the range between the 5th and 95th percentiles. All statistics are weighted by the number of households in each tract.

Figure 2

Within-Cluster Distribution of Neighborhood Characteristics: *K* = 3



NOTE: Columns display the plots for (i) racial configuration, (ii) earnings, and (iii) price of housing services measures. Rows correspond to each variable configuration. Within each plot, neighborhood classes (1, 2, and 3 stand for types *A*, *B*, and *C*, respectively) are listed in the vertical axis (the fraction of households is listed in parentheses). Vertical lines indicate neighborhood means (or centroid c_k). Boxes indicate the range between the 25th and 75th percentiles. Lines within boxes indicate medians. Brackets indicate the range between the 5th and 95th percentiles. All statistics are weighted by the number of households in each tract.

Table 6

Cluster Contiguity: $\kappa = 2.5$

	$K = 2$		$K = 3$		
MSA	*I*	*II*	*A*	*B*	*C*
z Score normalization					
Contiguity (%)	40	64	43	32	50
Adjacency	5.6	7.2	5.6	5.7	5.9
Mahalanobis normalization					
Contiguity (%)	41	68	41	28	50
Adjacency	5.6	7.4	5.6	5.7	6.1

NOTE: For a randomly chosen tract *i* of cluster C_k, contiguity equals the expected fraction of tracts of cluster C_k that are connected to *i*. Adjacency equals the expected number of cluster C_k tracts that are directly adjacent to *i*.

geographic areas, so that potential spatial spillovers would not have a large scope of action.

In this article, spatial contiguity is not imposed in any way.[9] However, spatial contiguity serves here as an additional measure of cluster adequacy.

Two strategies are used to assess spatial contiguity. The first computes a simple indicator that measures the fraction of neighborhoods of a class C_k to which the average neighborhood in C_k is "connected." The second strategy presents a few maps indicating the location of each class of neighborhoods in selected MSAs.

To measure contiguity, I begin with a pair of neighborhoods *A* and *B*. The Census Bureau provides the geographic coordinates at one internal point of each neighborhood, denoted as p_A and p_B. A neighborhood's radius can be defined as the radius (r_A, r_B) of a circle with the same geographic area as the corresponding neighborhood. Then, say that neighborhoods *A* and *B* are adjacent if $distance(p_A, p_B) \geq \kappa \max(r_A, r_B)$, where $\kappa \geq 1$ is an arbitrary constant. A *connected* set of neighborhoods is defined as a set of neighborhoods that cannot be separated into two subsets without separating at least one pair of adjacent neighborhoods.

The adjacency parameter of the contiguity indicator is set to $\kappa = 2.5$. This means that two tracts in the same cluster are considered adjacent if the distance between their Census-assigned internal points is less than 2.5 times the larger of their neighborhood radiuses.

Table 6 shows that, using the z score normalization and $K = 2$, type *I* neighborhoods are connected to 40 percent of their own type within an MSA and type *II* neighborhoods are connected to 64 percent of neighborhoods of their own type within an MSA. Thus, representative neighborhoods defined by the clustering procedure describe large geographic areas with homogeneous characteristics. Also, the expected number of same-type tracts adjacent to a randomly selected neighborhood lies between 5.6 and 7.2. Similar results hold using the Mahalanobis normalization.

For $K = 2$, type *I* neighborhoods tend to be substantially less connected than type *II* neighborhoods. This obeys the fact that type *I* neighborhoods form "islands" in a "sea" of type *II* neighborhoods (see the "MSA Maps" subsection below). Connectedness is lower because several MSAs contain more than one "island." For $K = 3$, type *B* neighborhoods tend to be less connected than the other two types.

MSA Maps

Figures 3 to 8 are maps corresponding to selected areas of three MSAs in the sample. For each MSA, the $K = 2$ and $K = 3$ characterizations are depicted in different shades of blue.

The two-neighborhood characterization exhibits a striking degree of contiguity. In the selected MSA, type *I* (low-income) neighborhoods form a small number of large areas, which are surrounded by type *II* (high-income) neighborhoods. This is remarkable given that (i) no geographic location information was used in the clustering procedure and (ii) the number of tracts within each "island" is large. For example, the Washington-Baltimore-Arlington MSA contains 1,453 tracts, of which 378 are type *I*. Almost all of these tracts are grouped into three islands (see Figure 7).

Finally, the three-neighborhood characterization is consistent with the two-neighborhood characterization. The type *I* (low-income) cluster of the two-neighborhood characterization is basically the same as the type *A* cluster in the three-neighborhood characterization. Therefore, the degree of contiguity for type *A* areas is also remarkable in the three-neighborhood characterization (Table 6). Roughly, the type *II* (high-income) neighborhoods of the two-neighborhood characterization are split into two new types (labeled *B* and *C*) when proceeding from $K = 2$ to $K = 3$. In the three-neighborhood characterization, type *B* neighborhoods exhibit the lowest degree of contiguity. These types of neighborhoods appear to the eye as transition areas between the clearly defined "islands" of type *A* and the "sea" of type *C* neighborhoods.

Regional Stability

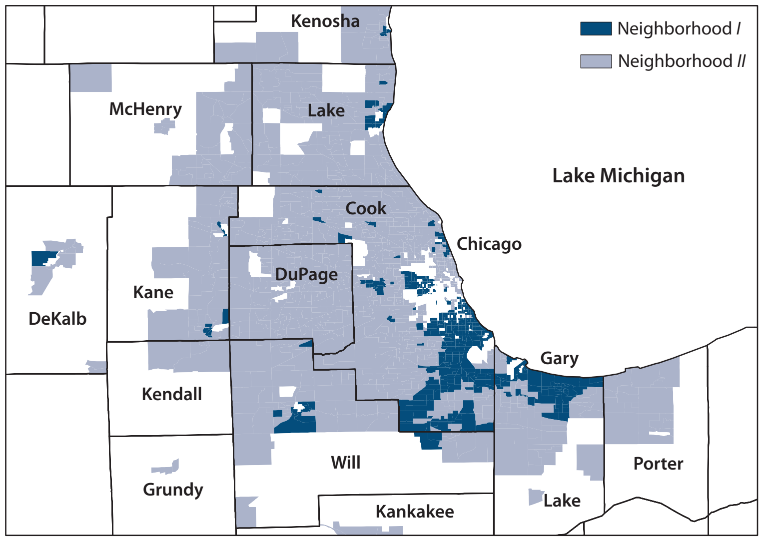
Recall that so far all results correspond to applying the clustering algorithm once to all neighborhoods in the sample. This subsection addresses whether the characterization of neighborhood is meaningful at the MSA level for $K = 2, 3$. The question is approached at two levels. First, do all MSAs contain a roughly similar fraction of each type of neighborhood, or are neighborhoods of each type concentrated in particular MSAs? In other words, are the fractions of each type stable across MSAs? Second, would the classification of neighborhoods differ substantially if centroids were allowed to vary across MSAs? The answers are yes and no, respectively.

First, Table 7 presents the percentage of each neighborhood class by MSA for $K = 2$ and $K = 3$. Each class of neighborhood exists in each MSA in roughly the same percentages, with standard deviations close to 8 percent for $K = 2$ and between 7.6 and 13.7 percent for $K = 3$.

Second, to allow for different centroids across MSAs, I cluster neighborhoods independently for each MSA using the benchmark variable configuration and *z* score normalization for $K = 2, 3, 4$. Then I use the cluster similarity measure to compare these clustering results with those for all MSAs pooled. Table 8 shows the percentage of neighborhoods classified in

Figure 3

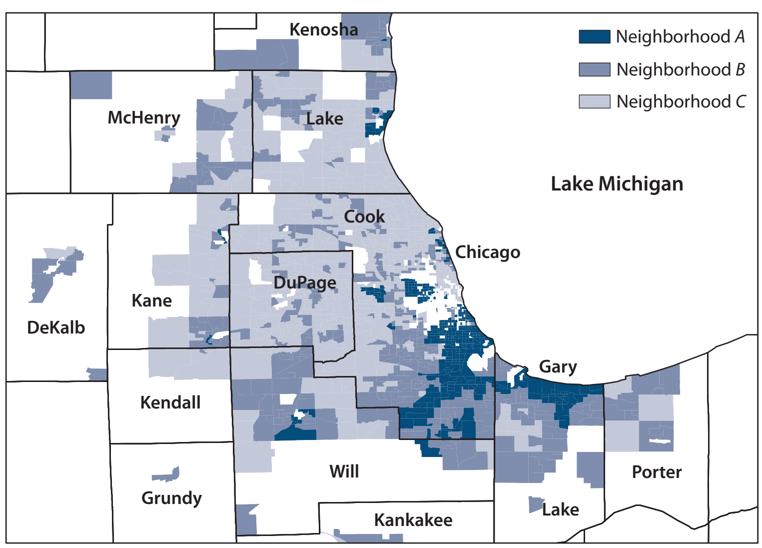
Chicago-Gary-Kenosha MSA: *K* = 2



NOTE: The figure shows selected neighborhoods of the corresponding MSA. Names of counties within the MSA are also listed.

Figure 4

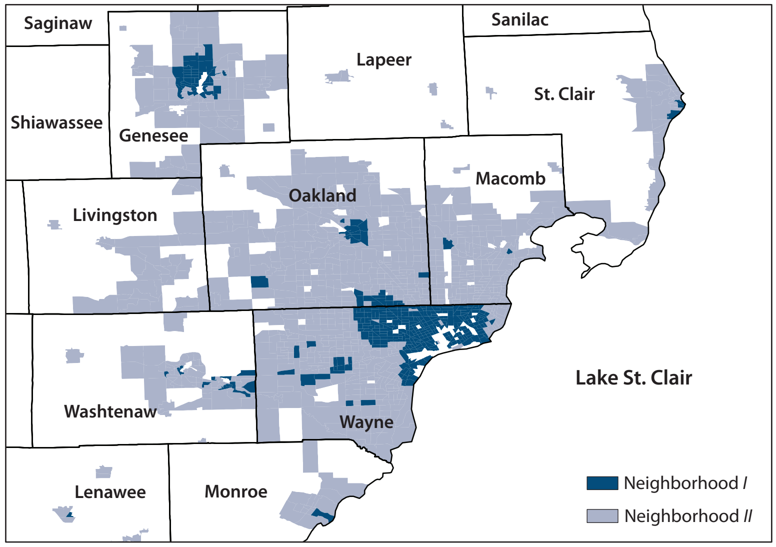
Chicago-Gary-Kenosha MSA: *K* = 3



NOTE: The figure shows selected neighborhoods of the corresponding MSA. Names of counties within the MSA are also listed.

Figure 5

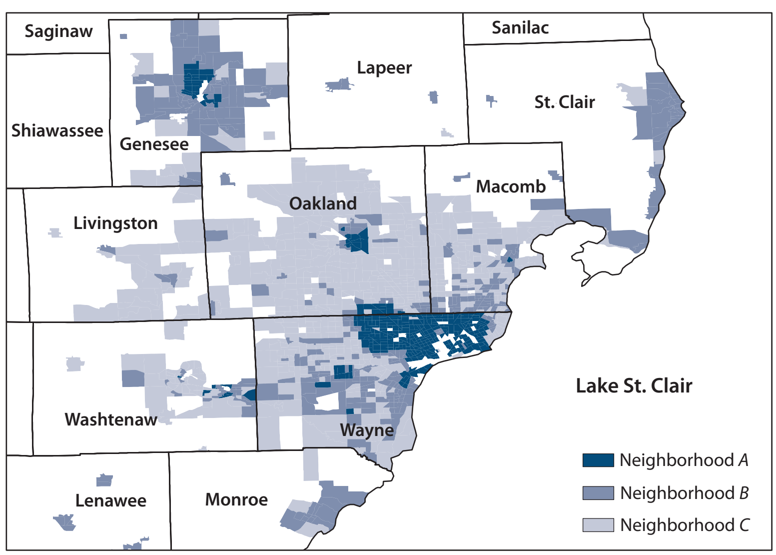
Detroit-Ann Arbor MSA: *K* = 2



NOTE: The figure shows selected neighborhoods of the corresponding MSA. Names of counties within the MSA are also listed.

Figure 6

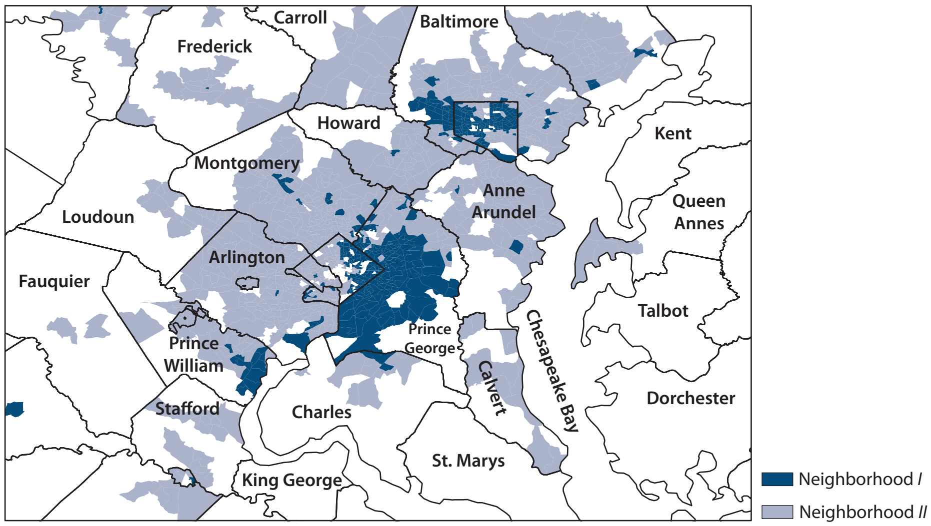
Detroit-Ann Arbor MSA: *K* = 3



NOTE: The figure shows selected neighborhoods of the corresponding MSA. Names of counties within the MSA are also listed.

Figure 7

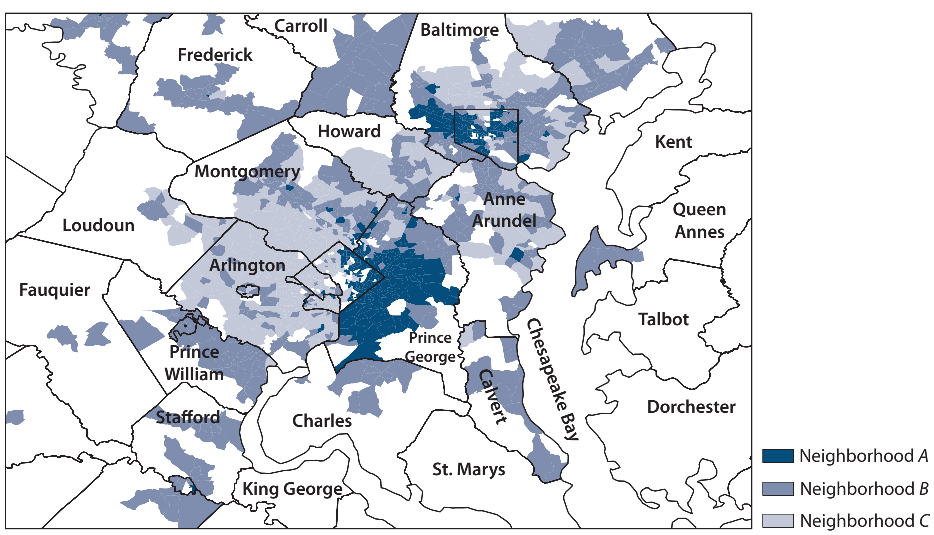
Washington-Baltimore-Arlington MSA: *K* = 2



NOTE: The figure shows selected neighborhoods of the corresponding MSA. Names of counties within the MSA are also listed.

Figure 8

Washington-Baltimore-Arlington MSA: *K* = 3



NOTE: The figure shows selected neighborhoods of the corresponding MSA. Names of counties within the MSA are also listed.

the same group for each MSA using MSA-by-MSA versus pooled clustering. The classifications are virtually identical for $K = 2$. The results for $K = 3$ are quite satisfactory with some exceptions: For example, for the Miami-Fort Lauderdale, the Norfolk-Virginia Beach-Newport News, and the Dallas-Fort Worth MSAs, clustering matches up for just 63, 61, and 65 percent of neighborhoods, respectively. I interpret these results as suggesting that the representative neighborhoods obtained reflect general economic and social forces common to most of the selected MSAs and specific regions or MSAs.

THE NATURE OF REPRESENTATIVE NEIGHBORHOODS

Two-Neighborhood Clustering

Two-neighborhood clustering provides the following characterization: Type *I* neighborhoods contain 27 percent of all households, have 4,800 residents per square kilometer, and cover about 4,600 square kilometers (Table 9). The population density of type *I* neighborhoods is about twice the density of type *II* neighborhoods, while the land area for type *I* neighborhoods is about one-fifth that of type *II* neighborhoods.

The $K = 2$ characterization reflects strong segregation by race. Of the households residing in type *I* neighborhoods, 32 percent are white, while 84 percent of households in type *II* neighborhoods are white (Table 10).

The $K = 2$ characterization also exhibits strong segregation by income. Household earnings average $33,591 in type *I* neighborhoods, representing 54 percent of average earnings in type *II* neighborhoods ($61,889). Household income averages $41,747 in type *I* neighborhoods, representing 56 percent of average income in type *II* neighborhoods ($74,577) (see Table 10).

Among black households, the average income for those in type *I* neighborhoods is 70 percent of the income for those in type *II* neighborhoods. For white households, type *I* neighborhood income is 58 percent of type *II* neighborhood income; for households in other racial categories the number is 62 percent. In type *I* neighborhoods, the average income of black households is $40,076, which is 90 percent of the average income of white households in that type of neighborhood ($44,727), while it is 74 percent in type *II* neighborhoods. Finally, the price of a unit of housing services is $10,405 in type *I* neighborhoods, representing 73 percent of the price in type *II* neighborhoods ($14,268) (see Table 10). This ratio is higher than the ratio observed for income, meaning that prices vary less than incomes across the two neighborhoods. This observation echoes a finding from the cross-MSA literature. Davis and Ortalo-Magne (2011) present evidence that the share of housing expenditures in income is constant in the United States. They show that a model with constant expenditure shares (i.e., with Cobb-Douglas preferences for housing and nonhousing consumption) and identical agents implies that prices should disproportionately reflect differences in incomes across MSAs. As in our two-neighborhood representation, the price measures provided by Davis and Ortalo-Magne vary less than incomes across MSAs. They find this observation puzzling viewed through the lens of their model.

Table 7

Percentage of Households by Neighborhood Class within Each MSA*

	K = 2		*K* = 3			
MSA	*I*	*II*	*A*	*B*	*C*	No. of tracts
Atlanta	32	68	27	53	20	568
Buffalo-Niagara Falls	34	66	16	79	5	250
Charlotte-Gastonia-Rock Hill	23	77	15	46	38	246
Chicago-Gary-Kenosha	22	78	19	35	47	1,658
Cincinnati-Hamilton	19	81	11	68	21	391
Cleveland-Akron	26	74	18	64	18	738
Columbus, OH	20	80	13	64	23	310
Dallas-Fort Worth	29	71	18	53	28	833
Detroit-Ann Arbor-Flint	24	76	20	29	50	1,335
Greensboro-Winston-Salem-High Point	29	71	20	58	23	196
Houston-Galveston-Brazoria	41	59	29	56	15	630
Indianapolis	22	78	13	66	21	278
Jacksonville, FL	32	68	15	65	20	162
Kansas City	24	76	15	66	19	400
Louisville	20	80	13	73	14	207
Memphis	49	51	46	32	22	203
Miami-Fort Lauderdale	49	51	33	42	25	409
Milwaukee-Racine	22	78	17	49	33	392
New York-Northern New Jersey-Long Island	23	77	19	24	57	3,850
Nashville	18	82	13	54	33	186
New Orleans	44	56	34	45	21	339
Norfolk-Virginia Beach-Newport News	36	64	26	66	8	309
Orlando	33	67	18	65	17	287
Philadelphia-Wilmington-Atlantic City	25	75	19	58	23	1,356
Raleigh-Durham-Chapel Hill	20	80	14	34	52	157
St. Louis	26	74	18	65	18	429
West Palm Beach-Boca Raton	33	67	16	57	27	243
Washington-Baltimore	29	71	23	46	31	1,453
Total	27	73	20	46	34	17,815
SD	8.4	8.4	7.6	13.7	12.4	
Tracts	5,649	12,166	4,458	7,456	5,901	

NOTE: *Benchmark variable configuration, *z* score normalization.

Table 8

Cluster Similarity: Pooled Versus MSA by MSA Clustering*

MSA	$K = 2$	$K = 3$	$K = 4$
Atlanta	96	89	80
Buffalo-Niagara Falls	91	74	79
Charlotte-Gastonia-Rock Hill	88	87	65
Chicago-Gary-Kenosha	98	79	84
Cincinnati-Hamilton	98	76	71
Cleveland-Akron	97	85	72
Columbus, OH	76	78	67
Dallas-Fort Worth	85	65	87
Detroit-Ann Arbor-Flint	99	91	65
Greensboro-Winston-Salem-High Point	98	75	61
Houston-Galveston-Brazoria	96	92	86
Indianapolis	82	71	62
Jacksonville, FL	98	78	70
Kansas City	99	82	64
Louisville	98	74	58
Memphis	97	77	76
Miami-Fort Lauderdale	93	63	72
Milwaukee-Racine	98	73	60
New York-Northern New Jersey-Long Island	83	91	54
Nashville	92	86	63
New Orleans	95	69	59
Norfolk-Virginia Beach-Newport News	93	61	63
Orlando	80	70	55
Philadelphia-Wilmington-Atlantic City	95	83	69
Raleigh-Durham-Chapel Hill	96	70	67
St. Louis	97	70	67
West Palm Beach-Boca Raton	95	86	93
Washington-Baltimore	86	66	55
Average	93	77	69

NOTE: The reported statistic corresponds to the percentage of neighborhoods classified in the same group by applying the clustering algorithm to the pooled dataset (all MSAs) versus applying it separately to each MSA. *Benchmark variable configuration, z score normalization.

Table 9

Population Density and Area*

	K = 2		*K* = 3		
Variable	***I***	***II***	***A***	***B***	***C***
Population density	4,837	2,251	5,333	2,070	2,674
Area (1,000 sq km)	4.59	25.32	3.19	17.07	10.0

NOTE: **z* Score normalization.

Table 10

Characteristics of Representative Neighborhoods: *K* = 2

Neighborhood	*I*	*II*	*I*/(*I* + *II*)
Number of households (thousands)			
Black	4,451	1,359	0.77
White	2,662	18,577	0.13
Other	1,152	2,150	0.35
Total	8,265	22,085	0.27
White/Total	0.32	0.84	

Neighborhood	*I*	*II*	*I*/*II*
Average income ($)			
Black	40,076	57,124	0.70
White	44,727	76,711	0.58
Other	41,320	67,166	0.62
Total	41,747	74,577	0.56
Average earnings ($)	33,591	61,889	0.54
Price of housing services*	10,405	14,268	0.73

NOTE: *Units are normalized to match the value of the original IRV measure (see the appendix).

Three-Neighborhood Clustering

The three representative neighborhoods are denoted by *A*, *B*, and *C*. The three-neighborhood clustering generates the following characterization. Type *A* neighborhoods cover 3,200 square kilometers, while type *B* neighborhoods cover 17,000 square kilometers and type *C* neighborhoods cover 10,000 square kilometers. The population density of type *A* neighborhoods is about 5,300 residents per square kilometer, while the density is much lower in the other two neighborhoods: 2,100 per square kilometer in type *B* and 2,700 per square kilometer in type *C* (see Table 9).

Table 11

Characteristics of Representative Neighborhoods: *K* = 3

Neighborhood	*A*	*B*	*C*	*A*/(*A* + *B* + *C*)	*B*/(*A* + *B* + *C*)
Number of households (thousands)					
Black	4,074	1,244	492	0.70	0.21
White	1,268	11,244	8,727	0.06	0.53
Other	821	1,396	1,084	0.25	0.42
Total	6,163	13,884	10,303	0.20	0.46
White/Total	0.24	0.90	0.95		
Neighborhood	***A***	***B***	***C***	***A*/*C***	***B*/*C***
Average income ($)					
Black	39,949	49,059	65,481	0.61	0.75
White	43,955	58,651	94,982	0.46	0.62
Other	40,363	53,483	77,620	0.52	0.69
Total	40,899	57,696	93,407	0.44	0.62
Average earnings ($)	33,142	47,106	76,303	0.43	0.62
Price of housing services*	10,715	11,238	17,377	0.62	0.65

NOTE: *Units are normalized to match the value of the original IRV measure (see the appendix).

In terms of racial configuration, there is a strong concentration of black households in type *A* neighborhoods, while type *B* and *C* neighborhoods contain similarly large percentages of white households. Only 21 percent of households in type *A* neighborhoods are white, while 81 percent and 85 percent of households in type *B* and *C* neighborhoods, respectively, are white (Table 11).

Percentage differences in income between type *A* and *B* neighborhoods and between type *B* and *C* neighborhoods are similar, generating three approximately equally spaced strata. Average earnings are $33,142, $47,106, and $76,303 in type *A*, *B*, and *C* neighborhoods, respectively. Thus, the ratio of average earnings of *A* with respect to *B* is 0.70, while the ratio of *B* to *C* is 0.62. The picture is similar for average income. Incomes in type *A*, *B*, and *C* neighborhoods average $40,899, $57,696, and $93,407, respectively (see Table 11).

For black households, the ratio of average income for those in type *A* neighborhoods to those in type *C* neighborhoods is 0.61. This between-neighborhoods ratio is 0.46 for white households and 0.52 for households in other racial categories. These ratios of average income by race in neighborhoods type *B* with respect to *C* are 0.75 for black households, 0.62 for white households, and 0.69 for households of other races. This shows that, in terms of averages, the sorting of households in ascending order of income into neighborhoods *A*, *B*, and *C* holds not only for aggregate populations but also for each race separately.

The black-to-white ratio of average income is 0.91, 0.84, and 0.69 in type *A*, *B*, and *C*, neighborhoods, respectively, while the overall ratio is 0.61.[10] The fact that the within-neighborhood ratios are above 0.61 suggests that within-neighborhood racial inequality is smaller than overall racial inequality for every neighborhood. This was also a feature of the two-neighborhood characterization (see Tables 10 and 11). Also, it is interesting to note that, as in the $K = 2$ case, cross-neighborhood differences are less marked for the price of housing services than for income; the ratio of the price of housing services in *A* with respect to *B* is 0.95, while the *B* to *C* ratio is 0.65.

ROBUSTNESS

This section determines the degree to which Census tracts in the sample are classified in the same way under (i) several variable configurations and variable normalization strategies and (ii) several variations of the sample selection criteria.

Variable Configuration/Normalization

First, the clustering procedure is applied under each possible (*variable configuration*, *normalization strategy*) combination. Then, the resulting clusterings are compared and a measure of clustering similarity is applied to determine whether the results are similar.

There is a natural measure of similarity in the literature that works well when the number of clusters *K* is small. The measure takes two different clusterings, say $C^1 = \{C_1^1 \ldots C_K^1\}$ and $C^2 = \{C_1^2 \ldots C_K^2\}$, and counts the fraction of objects that are classified into the same group in both clusterings.[11]

The results are striking for $K = 2$. In the worst case, 90 percent of neighborhoods are classified in the same group. On average, 94 percent of neighborhoods are classified in the same group. In many cases, the classification is identical. The results for $K = 3$ are less robust so they are provided in Table 12. In the worst case, 76 percent of neighborhoods are classified in the same group, but in most cases, more than 80 percent are classified in the same way.

The similarity across these clusterings suggests that racial configuration, income, and price of housing services provide a meaningful characterization of neighborhoods. Regardless of the diverse measures and normalizations used, the neighborhoods are similarly classified.

Sample Selection Criteria

Sample selection criteria are varied to examine the robustness of the representative-neighborhood characterizations presented in the previous two subsections. I consider the following four variations of sample selection criteria:

1. including MSAs with populations above 250,000 (versus 1 million in the baseline sample);
2. including MSAs with a 5 percent (or more) black population (versus 10 percent in the baseline sample);
3. including neighborhoods with 90 percent or less of "other race" households (versus 50 percent or less in the baseline sample); and

Table 12

Robustness to Variable Configuration/Normalization: Cluster Similarity (percent)*

	Normalization							
	z Score				Mahalanobis			
Configuration/normalization	ben	inc	prent	pcost	ben	inc	prent	pcost
z Score normalization								
ben	100	94	80	89	76	77	78	77
inc		100	81	87	76	78	78	77
prent			100	81	80	80	90	80
pcost				100	77	77	80	77
Mahalanobis normalization								
ben					100	92	88	98
inc						100	86	92
prent							100	88
pcost								100

NOTE: The reported statistic corresponds to the percentage of neighborhoods classified in the same group under two alternative variable configurations. Variable configurations are described in Table 2. *$K = 3$, all variable configurations and normalizations.

4. excluding neighborhoods with average earnings above $150,000 (versus no upper limit in the baseline sample).

The clustering procedure is applied to each sample variation. Table 13 presents the values of the centroids for (Y, R, P), as well as other important characteristics of the two-neighborhood characterization for the baseline sample and sample variations 1 through 4.

Sample variations 1 and 2 result in a dataset containing neighborhoods from 56 and 41 MSAs, respectively (compared with 28 MSAs in the baseline sample). Table 13 shows that sample variation 1 leaves the two-neighborhood characterization virtually unchanged with respect to the baseline sample.[12] Sample variation 2 implies changes in the racial composition of the sample. The overall fraction of black households in the sample falls from 0.19 to 0.16 with respect to the baseline. This change is reflected in the neighborhood characterization. The fraction of white households in type *I* neighborhoods moves from 0.32 to 0.40. This is the only appreciable change in the neighborhood characterizations imposed by the sample variation 2. Sample variation 3 implies the addition of 1,098 tracts to the sample (the number of MSAs remains 28). This change leaves the neighborhood characterization virtually unchanged. Finally, sample variation 4 implies the deletion of 212 observations, with no appreciable effects on the two-neighborhood characterization. Therefore, the results obtained in the baseline sample for the high-earnings neighborhood (type *II*) are not affected by the presence of a small fraction of neighborhoods with very large average earnings.

Table 13

Varying Sample Selection Criteria*

Statistic	Baseline	Sample variation 1	2	3	4
Neighborhood I					
Average earnings ($)	33,591	32,656	33,606	33,795	33,402
Fraction of white HHs	0.32	0.33	0.40	0.31	0.32
Price of housing services ($)	10,405	9,976	10,577	10,716	10,063
Neighborhood II					
Average earnings ($)	61,889	60,222	62,911	61,930	60,311
Fraction of white HHs	0.84	0.84	0.83	0.84	0.84
Price of housing services	14,268	13,604	16,562	14,228	13,768
Aggregate					
Fraction of HHs living in *II*	0.73	0.71	0.69	0.67	0.71
Overall fraction of white HHs	0.70	0.70	0.70	0.67	0.69
Overall fraction of black HHs	0.19	0.20	0.16	0.19	0.20
Number of MSAs	28	56	41	28	28
Number of tracts	17,815	20,148	24,054	18,913	17,603

NOTE: HH, household. Sample variations 1 through 4 correspond to the following sample selection criteria: (1) including MSAs with population above 250,000 (versus 1 million in baseline sample); (2) including MSAs with 5 percent or more black population (versus 10 percent in baseline sample); (3) including neighborhoods with 90 percent or less of "other race" households (versus 50 percent or less in baseline sample); (4) excluding neighborhoods with average earnings above $150,000 (versus no upper limit in the baseline sample). *Benchmark variable configuration, z score normalization, $K = 2$.

REMARKS AND CONCLUSION

This article explores the existence of a suitable representative-neighborhood characterization of metropolitan U.S. data. Such a characterization allows complex neighborhood-level data to be simplified. A simple characterization permits a transparent interpretation of data through models featuring a small number of neighborhoods with the advantage that the characterization has a direct geographic counterpart (see Figures 3 to 8).

One potential use of this characterization is to impose empirical discipline on quantitative models with a small number of locations. The main advantages for quantitative models calibrated to match a representative-neighborhood characterization are simplicity and clarity, yet such calibration has another appealing feature. The K-means clustering algorithm, as applied here, provides a partition of neighborhoods that minimizes a sum of squares criterion. Therefore, if the representative neighborhoods are reproduced by locations in a quantitative model, such a model will achieve the best possible fit to neighborhood-level data under the sum of squares criterion. This feature provides a rationale for fitting more-complex models to match aspects of the characterization developed in this article. ■

APPENDIX

Price of Housing Services

The data contain three sources of information regarding expenditures for housing services. The first source is the *median gross rent* variable. This is the median rent paid by renter households in a tract. The measure is designed to include the cost of utilities and fees, such as condo fees, when applicable, in addition to rent. The second source is the *median house value* variable. This measure is computed by the Census Bureau using market values of housing units reported by home-owning households. The housing literature uses these values to construct an expenditure measure or implicit rental value (IRV). The third source is the *median selected monthly owner costs* variable. This measure is constructed by the Census Bureau to estimate the monthly cost of housing for homeowners.[13]

Median tract house values are converted into median annual IRVs using a procedure based on Poterba (1992). This procedure consists of applying an annual user-cost factor to house values. A factor of 8.93 percent of the house value is used.[14]

NOTES

[1] See Calabrese et al. (2006, footnote 4) for a list of examples.

[2] The set of housing characteristics for each tract is composed of (i) the median number of rooms in the unit, (ii) a distribution of the number of units in the housing structure (10 categories), (iii) a distribution of the number of bedrooms (6 categories), (iv) the fraction of units with telephone service, (v) the fraction of units with complete plumbing facilities, (vi) the fraction of units with complete kitchen facilities, and (vii) the distribution of travel time to work (12 categories).

[3] This is a standard threshold in the housing literature above which an area is considered urban.

[4] Correctional institutions, nursing homes, juvenile detention facilities, college dormitories, military quarters, and group homes are considered group quarters.

[5] Iterative relocation proceeds as follows: (i) Assign elements arbitrarily into an initial partition consisting of *K* clusters and calculate the centroid of each cluster. (ii) Generate a new partition by reassigning each element to the nearest cluster centroid. If no objects were reassigned, terminate. (iii) Compute new centroids using the partition obtained in step (ii). Return to step (ii).

[6] The choice of normalization is not necessarily innocuous. For example, Jain and Dubes (1988, p. 25) provide a case in which *z* score normalization destroys the cluster structure in a particular dataset.

[7] See Jain and Dubes, 1988, section 4.5, for an extensive discussion of the concept of cluster validity.

[8] Since x_i and c are vectors, the "overall variability" is defined as the sum of each component's variation (see the denominator in the expression for R^2).

[9] A branch of classification analysis deals with the clustering of objects that are described by a vector of variables (x_i) and also by their position on a plane. In some cases, it may be desirable that objects in the same class are also spatially contiguous. In the problem of digital image segmentation, it is usually desirable that adjacent pixels belong to the same class. See, for example, Theiler and Gisler (1997). In the extreme, one could restrict all elements in a given class to be contiguous. This constrained clustering problem is known as regionalization (see, for example, Duque, Church, and Middleton, 2006). A simpler approach (i) includes the spatial coordinates of each object in the vector of characteristics x_i and (ii) applies an unconstrained clustering algorithm. The algorithm will tend toward generating clusters that are "close" in the plane.

[10] These ratios are not provided in the tables but are easily calculated from income entries in Table 11.

[11] This task is complicated by the fact that the subindexes labeling each cluster can be assigned arbitrarily (i.e., there is no way to decide which cluster in C^1 corresponds to any particular cluster in C^2). Therefore, one should examine all possible permutations of the cluster subindexes and choose the one yielding the maximum fraction of coincidences. If *P* is the set of all possible permutations *p*(*k*) of the indexes (1,2,3...*k*...*K*), then the measure can be expressed as

$$\max_{p \in P} \frac{\sum_{k=1}^{K} \left| C_k^1 \cap C_{p(k)}^2 \right|}{N},$$

where the "absolute value" denotes the number of elements in a cluster C_k.

[12] The results shown in Table 13 compare only the (*Y*, *R*, *P*) averages across Census tracts (centroids). However, analysis of higher moments of (*Y*, *R*, *P*) in each representative neighborhood shows these are remarkably stable across different samples as well. Details are available from the author upon request.

[13] The selected monthly owner costs variable includes reported payments of mortgages, deeds of trust, contracts to purchase, or similar debts on the property (including payments for the first mortgage, second mortgage, home equity loans, and other junior mortgages); real estate taxes; fire, hazard, and flood insurance on the property; utilities (electricity, gas, water, and sewer); and fuels (oil, coal, kerosene, wood, and so on). It also includes monthly condominium fees or mobile home costs (installment loan payments, personal property taxes, site rent, registration fees, and license fees).

[14] Calabrese et al. (2006) use this approach. The user cost of housing for homeowners is calculated by letting implicit rental values *IRV* be given by $IRV = \kappa_p V$, where *V* is the market value of the home. The annual user-cost factor is given by

$$\kappa_p = (1 - t_y)(i + t_v) + \psi,$$

where t_y is the income tax rate, i is the nominal interest rate, t_v is the property tax rate, and ψ contains the risk premium for housing investments, maintenance and depreciation costs, and the inflation rate.

REFERENCES

Calabrese, Stephen; Epple, Dennis; Romer, Thomas and Sieg, Holger. "Local Public Good Provision: Voting, Peer Effects, and Mobility." *Journal of Public Economics*, August 2006, *90*(6-7), pp. 959-81.

Davis, Morris A. and Ortalo-Magne, Francois. "Household Expenditures, Wages, Rents." *Review of Economic Dynamics*, April 2011, *14*(2), pp. 248-61.

Duque, Juan Carlos; Church, Raúl and Middleton, Richard. "Exact Models for the Regionalization Problem," in Western Regional Science Association Annual Meetings, Santa Fe, 2006.

Fernandez, Raquel and Rogerson, Richard. "Public Education and Income Distribution: A Dynamic Quantitative Evaluation of Education-Finance Reform." *American Economic Review*, September 1998, *88*(4), pp. 813-33.

Iceland, John; Weinberg, Daniel H. and Steinmetz, Erika. "Racial and Ethnic Residential Segregation in the United States: 1980-2000." U.S. Census Bureau, Series CENSR-3. Washington, DC: U.S. Government Printing Office, 2002; http://www.census.gov/prod/2002pubs/censr-3.pdf.

Ioannides, Yannis M. "Neighborhood Income Distributions." *Journal of Urban Economics*, November 2004, *56*(3), pp. 435-57.

Jain, Anil K. and Dubes, Richard C. *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall, 1988.

Poterba, James M. "Taxation and Housing: Old Questions, New Answers." *American Economic Review*, May 1992, *82*(2), pp. 237-42.

Theiler, James and Gisler, Galen. "A Contiguity-Enhanced K-Means Clustering Algorithm for Unsupervised Multispectral Image Segmentation." Proceedings of the Society of Optical Engineering, October 1997, *3159*, pp. 108-118.

U.S. Census Bureau. "2000 Census of Population and Housing—Summary File 3." Last revised October 13, 2011; http://www.census.gov/census2000/sumfile3.html.

Wheeler, Christopher H. and La Jeunesse, Elizabeth. A. "Neighborhood Income Inequality." Working Paper No. 2006-039B, Federal Reserve Bank of St. Louis, June 2006, revised February 2007; http://research.stlouisfed.org/wp/2006/2006-039.pdf.