Senior Thesis 301
Department of Economics, Vassar College

Defining and Measuring Economic Segregation Based on What Matters for Creating Opportunity

Kai Matheson

April 1, 2019

Abstract

The growing wealth gap in the United States points to the increasing importance of determining the causes of economic segregation and its impact on economic mobility. Yet, theories of class segregation, such as the concentric zone model, have not been empirically evaluated. Aspatial methods to measure segregation are often overly simplistic, ignoring critical aspects of the economic landscape. Evaluation of the relationship between economic mobility and economic segregation is sensitive to how one measures segregation. Thus, this work aims to highlight the ways in which economic segregation would impact intergenerational mobility and to show how current aspatial measures of segregation obscure part of this relationship. I argue that these measures capture economic homogeneity rather than the spatial phenomenon we know as segregation. I employ k-means clustering to capture important information from income bin maps of cities in order to illuminate what current measures lack, comparing the R^2 of models of absolute upward mobility regressed on the various measures and clusters.

1. Introduction

Economic segregation affects our lives in many ways, including in one's ability to improve their financial standing. Living in separate, distinct neighborhoods makes it so that residents are less likely to interact with or make use of those amenities used by the wealthy, resulting in lowered economic opportunity. As economic segregation is such an important issue, it is crucial that we measure it accurately in order to capture the state of the phenomenon in different regions. In this paper, I urge researchers to work towards a better, explicitly spatial measure of economic segregation. I utilize the theoretical link to economic mobility in evaluating the performance of measures of economic segregation.

In outlining my argument, I emphasize how important it is to consider the geography or "shape" of economic segregation when considering how segregation might matter in our lives. It is crucial to consider the shape of segregation in order to understand the intensity of social exclusion among the disadvantaged and the degree of social closure achieved by the advantaged, in comparison with many methodologies which do not take into account the location of neighborhoods within a metropolitan system (Dwyer, 2010).

The purpose of this paper is to explore methods of measuring economic segregation and to investigate the relationship between these measures and economic mobility in US cities. I draw attention to the fact that the way in which segregation matters for mobility is explicitly spatial and thus measurements of segregation must take space into account. Specifically, I focus on the popular rank-order information theory index, a measure of segregation which does not take space into account. Further, I take a direct approach in measuring segregation spatially, by creating and subsequently assessing spatial information about economic segregation within metropolitan areas, relying solely on analysis of income bin maps.

2. Literature review

Throughout the literature review, I detail the ways in which economic segregation matters for economic mobility. I describe how this relates to economic models of the city and theories of urban segregation. Further, I explain why segregation must be interpreted as a spatial phenomenon and how the currently popular methods fall short of doing that.

2.1. Why the geography of economic segregation matters for economic mobility

Why would the relationship between upward mobility and segregation be an important one? Further, why might growing up in more segregated cities have long term negative consequences, even for individuals who move away? The theory behind the relationship between segregation and economic mobility suggests that the mechanisms between the two come in the form of school financing and neighborhood effects (Mayer, 2002). The way in which schools are financed in the United States typically is that they are funded by the property taxes of the local communities. Thus, a low income school district will suffer from low school spending, resulting in low school quality that leads to low educational outcomes. Educational outcomes are, of course, linked with income, and thus better educational outcomes go hand in hand with higher percentiles in the income distribution. A poor child starting out in a school with lots of concentrated poverty sees smaller educational gains and thus one would think is less upwardly mobile than a similar poor child in a school with less concentrated poverty. On the other hand, neighborhood effects are those benefits that affluent residents might generate for their neighbors, for example through providing role models, social networks, or increased neighborhood monitoring. These two gateways to opportunity are why I decide to explore economic rather than

racial segregation in its link to economic mobility. However, it is important to note that while racial segregation is an extremely important and detrimental phenomenon, the results of this paper cannot be translated to apply to it: the correlation between economic segregation and racial segregation has been measured to be more modest than one might expect, estimated to be a mere 0.240 in 1990 (citation).

Much of the empirical literature backs up the theory of school financing as the link between economic segregation and economic mobility. Orfield and Lee (2005) study the modern process of resegregation in U.S. schools, and find that the level of concentrated poverty in a school is one of the biggest predictors of educational outcomes. While that study is focused on class segregation within schools, work by Mayer (2002) highlights the effects of Census tract-level class segregation on educational outcomes. She finds that increases in economic segregation within Census tracts in the same state hardly change average educational attainment but exacerbate the inequality between high-income and low-income children, with increases in segregation resulting in high-income children's increased educational attainment corresponding directly to low-income children's decreased educational attainment. Further, Mayer (2002) finds that economic inequality *within* tracts has little effect on low-income children's educational outcomes while changes in inequality *between* tracts lessen educational outcomes for low-income children (Mayer, 2002).

Jumping off this discussion of whether *between*- or *within*-neighborhood segregation matters most, it is important to consider the society we live in, in which data availability may not reach our ideal expectations. This can help to clarify what we mean when we say we want to measure economic segregation.

In a perfect world, researchers would be able to evaluate economic segregation by accessing household-level data on income for all households in a given metropolitan area. In this scenario, we would work to develop methods in order to measure between-neighborhood

or within-neighborhood segregation—in fact, we don't even need to consider neighborhoods at all.

However, in reality, we are generally restricted to accessing only household-level data that has been aggregated over census tracts. We are working to navigate government-imposed boundaries to neighborhoods and aggregate counts, and we do not know any geographic information about what is happening within a census tract.

If segregation needs to be measured spatially, which is what I will argue throughout this work, then it is impossible to detect within-tract segregation with the data we have, by merely knowing the distribution of income within a "neighborhood" or census tract.

Thus, we must shift our focus to the measurement of *between-neighborhood segregation*. This is not to disregard within-neighborhood segregation, because it is not necessarily unimportant—however, the most we can do while lacking geographic information within census tracts is to measure relative income diversity within census tracts. Because segregation is a spatial phenomenon, this is the best we can do with the data we have.

We now draw away from the discussion on why segregation matters and more towards theories of how it manifests in cities, which have yet to be empirically evaluated. There are many contradictory theories of urban land use debating the ways in which class segregation manifests in cities. Among these are two different concentric zone models, the sector model, and the multiple nuclei model.

The first concentric zone or monocentric city model was created in 1841 by J.G. Kohl, based on the pre-industrial cities of continental Europe. The basic structure of the model is rings of populations segregated by class radiating from the city center, in which the high-income population was housed closer to the city center and the low-income communities resided farther from the city. The other and more well-known monocentric

city model, created by Ernest W. Burgess in 1925, posits that the large American city can be generalized to have a central business district surrounded by a zone of transition including other industries, followed by inner-city poor residences and then high-income residences located in the suburbs. This is contextualized within an industrial city, where the wealthy prefer the suburbs because of pollution and violence downtown, as well as the lower cost of land, while low-income individuals prioritize lower transit costs.

Studies have supported the theory that the primary reason for central city poverty is access to the public transportation system. Upward mobility, that is the capacity of increasing one's social or economic position, is currently higher in cities with less sprawl, as measured by commute times to work (Chetty et al., 2014). This suggests that job access, economic segregation, and transportation access are strongly correlated and interdependent. Thus, this builds on the argument for the relationship between economic mobility and economic segregation.

The Burgess model is still in use today but is starting to be seen as outdated as the post-industrial processes of gentrification and displacement become increasingly more common. Now, many urbanists believe American cities are undergoing demographic inversion, which refers to the process by which suburbs become the principal region where low-income individuals settle due to the increasing cost of living in central cities (Ehrenhalt, 2012). From 2000 to 2008-2012, the percentage of suburban poor in the United States increased by 139%, which is nearly three times more than within cities. By 2008-2012, 46% of all non-rural poor residents living in concentrated poverty lived in the suburbs (Kneebone, 2014). Large metropolitan suburbs house about one-third of low-income Americans, a greater share than big cities, small metropolitan areas, or rural areas. Through the 2000s, suburban poverty increased at a rate five times more than what we have seen within cities (Tomer et al., 2011). With demographic inversion occurring in many large American cities, the Kohl model becomes relevant again in such

cases.

There are other models, however, that do not assume a ring-like structure to the city. In 1937, Homer Hoyt came up with the sector model in which "growth along a particular axis of transportation usually consists of similar types of land use" (Harris and Ullman, 1941). Imagine a city that looks like a pie chart, with each of the slices termed a "sector" in which there is one dominant type of land use in each sector (Beauregard, 2007). This is what Hoyt's sector model represents. On the other hand, the multiple nuclei model by Harris and Ullman (1941) proposes that land use patterns are not built around a single city center but around several "nuclei" that could have developed at any point in the city's history. These models allow for more flexibility in the form of cities.

There is no consensus over how to evaluate which model is best fit for certain cities. Further, there has been almost no work done on directly analyzing the shape of economic segregation. The existence of this debate highlights the need for research on evaluating which models are the most accurate, or for which types of cities. There has been little work done to empirically evaluate the theories aside from individual case studies based on subjective observation alone. In using data-driven methodology, I add to the literature on urban land use by investigating whether these structures of economic segregation become apparent in the results of my analysis and whether these structures are what matter for influencing economic mobility on a larger scale.

The growing wealth gap in the United States points to the increasing importance of determining the causes of economic segregation and its impact on economic mobility. Yet, theories of class segregation, such as the concentric zone model, remain largely untested. Common measures of economic segregation are often overly simplistic, ignoring critical aspects of the economic landscape. Evaluation of the relationship between economic mobility and economic segregation is sensitive to how one measures segregation.

2.2. Measures of economic segregation: Reardon's rank-order information theory index

Past studies have mostly relied on aspatial segregation largely influenced by arbitrary political divisions and boundaries (Dwyer, 2010). Even in the few studies that consider spatial segregation, they indirectly measure it through measures of segregation termed concentration, centralization, and clustering, which each in its own constitutes an aspatial measurement.

The *rank-order information theory index* first introduced by Reardon et al. (2006) is the most commonly used measure of economic segregation today, by economists and sociologists alike. It is based on a more general segregation measure used to evaluate not just income segregation but also racial or ethnic segregation, called the *Theil Index*. The Theil Index is an entropy-based measure.

To go into more on what that means, entropy is commonly used in physics and information theory to measure the randomness of a system. Theil brought entropy to the social sciences with the Theil Index. It can also be thought of as the amount of information needed to describe a probability distribution. It is used in segregation measures, but it can only be calculated for discrete distributions (Roberto, 2015).

The equation for entropy of the population when divided into two groups is as follows. Imagine a population of income-earners and let p be a percentile of the income distribution, between 1 and 100. Let $E(p)$ denote the entropy of the population when divided into these two groups.

$$E(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p} \tag{1}$$

The Theil Index shares a similar, familiar form to the previous equation. Let x_i be

the income of groups of earners, and $\bar{x}$ be the average income. Let I denote the Theil Index.

$$I = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{\bar{x}} \log \frac{x_i}{\bar{x}} \tag{2}$$

When all incomes are equal (all individuals earn the mean income), there is no inequality so $I = 0$ (Roberto, 2015).

Now let us introduce the rank-order information theory index, denoted H^R. Again, let p be income percentile ranks in a given income distribution i.e. $p = F(Y)$ where Y is income and F is the cumulative income density function. In other words, p is the proportion of the population with incomes below a certain threshold. For any p, we can compute the residential segregation between those with income ranks less than p and those with income ranks greater than or equal to p. Let $H(p)$ denote the value of the traditional information theory index of segregation computed between the two groups.

$$H(p) = 1 - \sum_{j} \frac{t_j E_j(p)}{TE(p)} \tag{3}$$

where T is the population of the metropolitan area and t_j is the population of neighborhood j.

Because segregation of the extreme poor and the extreme wealthy can often be the most illuminating, Reardon and Bischoff (2011) define *segregation of poverty* and *segregation of affluence* to be the values of $H(p)$ at $p = 25$ and $p = 75$ respectively. The segregation of poverty can be interpreted as the uneven distribution of low- and non-low income households among neighborhoods. Likewise, the segregation of affluence can be interepreted as the uneven distribution of high- and non-high income households among neighborhoods (Reardon and Bischoff, 2011).

Then the rank-order information theory index H^R can be written as

$$H^R = 2\ln(2)\int_0^1 E(p)H(p)dp \tag{4}$$

It is important to consider what data is necessary to compute the index. In this case, we need categorical income bin data, providing the population count within each income bin or bracket. The 2000 census reported 16 income categories, which allow us to compute $H(p)$ at 15 values of p. They then approximate the function $H(p)$ over $(0, 1)$ by fitting an mth-order (i.e. 4th-order) polynomial to the values, weighting each point by the square of $E(p)$:

$$H(p) \approx \beta_0 + \beta_1 p + \beta_2 p^2 + ... + \beta_m p^m + \varepsilon_p,\ \varepsilon_p \sim N(0, \frac{\sigma^2}{E(p)^2}) \tag{5}$$

If $\hat{\beta}_k$ is the kth coefficient from this model, then

$$\hat{H^R} = \hat{\beta}_0 + \frac{1}{2}\hat{\beta}_1 + ... + (\frac{2}{(m+2)^2} + 2\sum_{n=0}^{m}\frac{(-1)^m - n({}_mC_n)}{(m-n+2)^2})\hat{\beta_m} \tag{6}$$

where ${}_mC_n \approx m!/(n!(m-n)!)$ is the binomial coefficient (the number of distinct combinations of n elements from a set of size m). Then, we can estimate other $H(p)$ at different values of p by using the fitted polynomial $H(p)$ equation.

The rank-order index can be interpreted as the ratio of within-unit (tract) income rank variation to overall (metropolitan area) income rank variation. A value of 0 indicates no income segregation, or that the income distribution in each census tract mirrors that of the region as a whole. A value of 1 indicates complete income segregation, or that there is no variation in income in any census tract and only across census tracts. H^R can be thought of as a weighted average of the binary income segregation at each point in the income distribution, where the weights are proportional to entropy $E(p)$ which is maximized when $p = 0.5$ and minimized at $p = 0$ or $p = 1$, meaning that these weights

assume that segregation between those above and below the median is more important than segregation between those above and below the 90th percentile.

2.3. Why space matters: the checkerboard problem and reshuffling neighborhoods

It is important to note that Reardon's rank-order index is aspatial– that is, it does not take into account the geography of census tracts in a metropolitan area. There is a spatial analog of the rank-order index, however it simply uses arbitrary definitions of what is considered "local" ranging from radii of 500 to 4,000 meters (Reardon, 2011). This results in different segregation estimates for regions of a metropolitan area, but the overall metropolitan area measure of segregation is still the same as before, not taking into account any spatial component.

Even if this adaptation proved to be somehow helpful, it is not likely to be utilized. In 2018, Reardon released a Stata module entitled RANKSEG that one can use to compute rank-order segregation measures with finite sample-bias correction (Reardon et al., 2018). This function does not allow any input for geographic information or distances between census tracts, meaning that Reardon himself may not be expecting those using his measure to utilize the spatial analog, and that there are significant barriers to utilizing the spatial analog compared with the original index if one wants to do so.

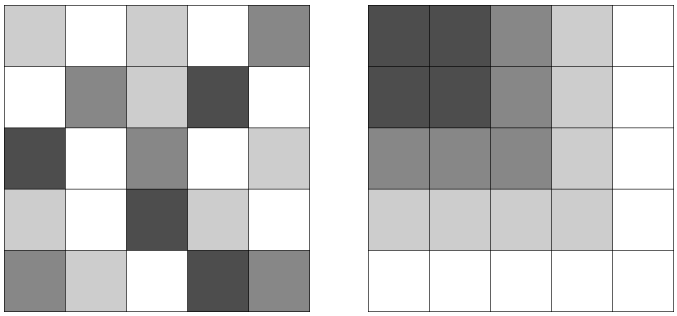
Because of how this measure completely disregards geography, I agree with the assessment by Roberto (2015) that this index measures relative diversity and not segregation, so it is problematic to interpret it as a measure of segregation. Rather, it measures relative homogeneity, comparing the diversity of local areas to overall diversity of a region.

Regardless, this segregation measure has been consistently and widely used in the fields of economics and sociology since its inception, seen as an improvement upon the now outdated *neighborhood sorting index* (explained in Appendix Section A). For example, it is the measure utilized by Chetty and Hendren (2016) in evaluating the correlation between economic mobility and economic segregation. Thus, although other measures of economic segregation have been proposed and studied, this is the main measure I dissect in this work.

Because the rank-order index does not incorporate spatial information but simply integrates over census tracts, it does not account for the spatial relationships between neighborhoods in the metropolitan area. This can have negative consequences. For example, think about the New York-Northern New Jersey-Long Island, NY-NJ-CT-PA MSA. The boundary of this MSA captures regions as disparate as Poughkeepsie, Yonkers, and Brooklyn. The rank-order index treats a census tract in Poughkeepsie in the exact same way as a census tract in Brooklyn, although they are in completely different landscapes– one a small city in a relatively rural area and the other a large borough of New York City. The problems with this approach can be further illustrated by the "checkerboard problem."

Imagine a city laid out in a 5×5 grid, where each square is a neighborhood in which every resident makes the same income. In other words, every individual's income is equal to the mean of the neighborhood. Thus, there is complete segregation *within* neighborhoods because everyone within a neighborhood has the same income, but we haven't described anything about the state of segregation *between* neighborhoods just yet. Imagine coloring an income map of this city, to portray where the wealthy and the impoverished live. It may result in a city that looks like one of the two pictured in Figure 1.

Figure 1: Checkerboard problem



(a) Integrated city (b) Segregated city

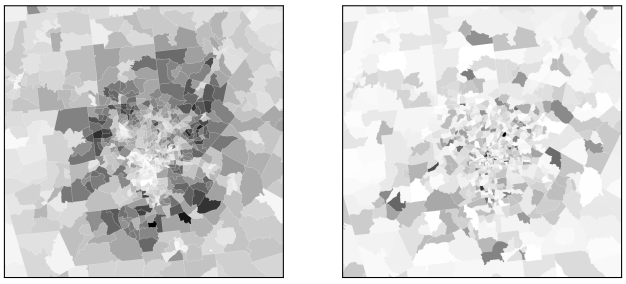
On a larger scale, the city pictured in Figure 1(a) is, to a degree, economically integrated (though there is complete segregation *within* each neighborhood). The city pictured in Figure 1(b) is segregated both within and between neighborhoods, arguably much more segregated than in Figure 1(a).

The two images shown in Figure 1 would result in the same aspatial indices. In other words, the rank-order index would not pick up on any differences between Figures 1(a) and 1(b). This is one of the major flaws in typical methods of measuring "aspatial segregation," as the rank-order index does. The "checker-board problem" highlights the fact that aspatial segregation measures ignore the spatial proximity of neighborhoods and instead only capture the composition within neighborhoods (Reardon, 2006).

The importance of this oversight can be further illustrated by viewing "reshuffled" maps of cities. The original data is of counts of how many individuals are in each income bin of the 16 reported income categories on the 2000 Census, in every census tract within the US' 276 designated Metropolitan Statistical Areas (MSAs). What is displayed in the maps is the proportion of people who are within the stated income bracket– darker areas show higher proportions of people within that income bracket, while lighter areas have smaller proportions of people who fall in that income bracket.

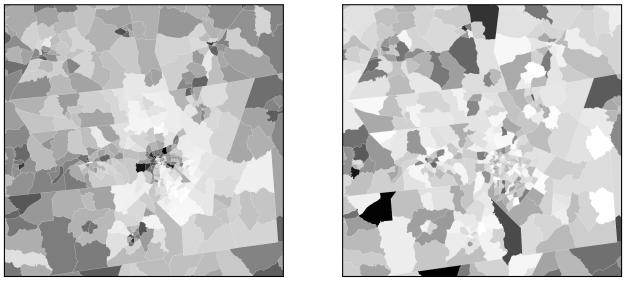
I reshuffle the data by assigning each census tract's population to a different census tract in the metropolitan area, breaking up any *between-tract* segregation as I am randomly placing populations throughout the metropolitan area. Here are some of the resulting maps.

Figure 2: Side-by-side: Atlanta, GA; $75,000 to $99,999



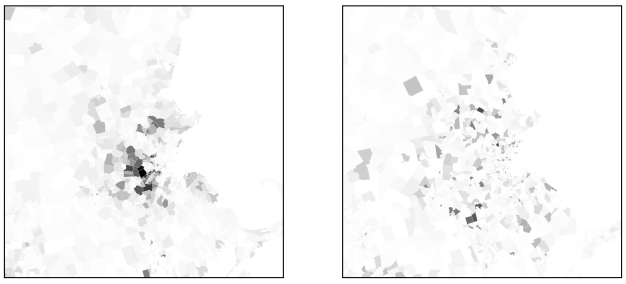
(a) Original (b) Reshuffled

Figure 3: Side-by-side: Charlotte-Gastonia-Rock Hill, NC-SC; $15,000 to $19,999



(a) Original (b) Reshuffled

Figure 4: Side-by-side: Boston-Worcester-Lawrence, MA-NH-ME-CT; $200,000 or more



(a) Original (b) Reshuffled

These images really make the checkerboard problem come to life. As is seen in the Figure 2, the original image shows stark segregation– there exists a ring of earners in the $75,000 to $99,999 income bracket just outside the center city. However, in the reshuffled map, there is no pattern to where darker census tracts appear, and thus, while there is segregation *within* tracts, there appears not to be segregation *between* census tracts. And yet, these two metropolitan areas would get the exact same value of Reardon's rank-order index.

Likewise in Figures 3 and 4, there exist concentrated areas where these other income brackets lie in these metropolitan areas in the original maps, and the reshuffled maps look quite different. The rank-order index, among all the other aspatial measures of segregation, do not take note of the differences between these original cities and their reshuffled counterparts.

2.4. Towards spatial measures of segregation

Many existing measures of segregation that are simply one index can only indicate average situations without any indication of variation around that situation or the range of contexts that individual members of groups experience (Johnston et al., 2014). Thus,

there has been a push towards identifying *dimensions* of segregation.

There are five very common categories of spatial segregation measures, which are what Dwyer (2010) terms the five "spatial dimensions" of segregation: evenness, exposure, concentration, centralization, and clustering (Dwyer, 2010). They are defined in Table 1. While these measures do consider geography, they do so implicitly or indirectly.

Dimension	Definition
Evenness	Degree to which a group is spread in equal proportions among another group across neighborhoods
Exposure	Likelihood a member of one group will come into contact with a member of the other group
Concentration	Land area taken up by one group compared to another, whether a group resides in a relatively small portion of the metro area or is spread out over more space
Centralization	Degree to which groups are located near the center of the metropolitan area versus the periphery
Clustering	Whether one group is located in neighborhoods near other neighborhoods dominated by the same group versus the other group

Table 1: The five spatial dimensions of segregation as defined by Dwyer (2010) and Reardon (2006).

Dwyer (2010) predicts the spatial form of class segregation using metropolitan factors of suburbanization and levels of class and racial inequality. She assesses the spatial form of class segregation by looking at each of the spatial dimensions in conjunction with one another. She argues that by combining multiple spatial dimensions, she is able to characterize the economic geography of cities. She defines *hypersegregation* as multiple overlapping dimensions of segregation (Dwyer, 2010). Further, she aims to assess how widely the concentric zone model of class segregation holds in metropolital areas and investigate whether there are alternative spatial forms. She looks into whether levels of class and racial inequality can predict the spatial form of class segregation. She finds

evidence for two spatial forms of affluent-poor segregation: the concentric zone model, and a more integrated type of city for which she lists examples such as Boulder, Portland, and Tampa.

That makes this paper the second-ever study to analyze the spatial dimensions of class segregation. But, rather than relying on a combination of aspatial measures of "spatial dimensions," as is done by Dwyer (2010), my methods evaluate spatial segregation in cities drawing on methods from computational science. In order to directly assess and create measures or dimensions of spatial segregation, I use actual income bin maps of cities as the bases for assessment.

It is important to take note that there is one existing measure that takes into account geography in a more meaningful and explicit way than the implementation of arbitrary radii. This is the spatial ordering index proposed by Dawkins (2007).

Dawkins (2007) explains that there is an initial "parade" of neighborhood per capita incomes (y_j) ranked such that $y_1 \leq y_2 \leq ... \leq y_j \leq y_J$, a spatial ordering is a reranking of the original parade in a way that reflects the jth neighborhood's spatial position.

There are many ways to do spatial orderings, but the two that Dawkins (2007) suggests are *Nearest Neighbor Spatial Ordering* and *Monocentric Spatial Ordering*.

Nearest Neighbor Spatial Ordering is explained as the process of holding neighborhood incomes constant while assigning each neighborhood the income parade ranking of the most spatially proximate nearby neighborhood.

Alternatively, *Monocentric Spatial Ordering* is if one defines a city center and then, holding neighborhood incomes constant, ranking the neighborhood income parade by ascending or descending values of distance to the city center from the centroid of each neighborhood.

Using these spatial orderings, we can create a *Spatial Ordering Index* (S_r), the ratio of the covariances, modifying the standardized spatial Gini index from Dawkins (2004) as follows:

$$S_r = G_r / G_B \tag{7}$$

where S_r is a spatial ordering index calculated from the rth spatial ordering, G_B is the Gini index of between-neighborhood income segregation, and G_r is a spatial Gini index calculated from either a nearest neighbor or monocentric spatial ordering. This makes S_r a measure of the degree to which a given spatial configuration of neighborhoods results in a reordering of neighborhood income distribution used to construct G_B– so it is a direct measure of the importance of the checkerboard phenomena.

G_r and G_B can be calculated using covariance-based formulas:

$$G_r = \frac{2\sum_{j=1}^{J}(\bar{R}_{j(n)} - \frac{H+1}{2})y_j}{HY} \tag{8}$$

$$G_B = \frac{2\sum_{j=1}^{J}(\bar{R}_{j} - \frac{H+1}{2})y_j}{HY} \tag{9}$$

where y_j is aggregate household income earned by residents of the jth neighborhood, Y is aggregate household income earned by residents of the region, $\bar{R}_j$ is the average rank of per capita income earned by neighborhood j within the overall neighborhood per capita income distribution, and $\bar{R}_{j(n)}$ is the average spatial rank of per capita income earned by neighborhood j within the overall neighborhood per capita income distribution.

To calculate $\bar{R}_j$, first rank neighborhoods in ascending order by per capita income. Let N_j be "the cumulative number of households within each successive position within the neighborhood income parade" (I don't really understand this). Let N_{j-1} be the cumulative total of households one rank lower within the income parade. The jth neighborhood's average rank is then equal to $(N_j + N_{j-1} + 1)/2$.

To calculate $\bar{R}_{j(n)}$ for a nearest neighbor spatial ordering, pair each neighborhood with its most spatially proximate neighbor ($j(n)$), and assign the jth neighborhood a rank equal to the rank of $j(n)$. $N_{j(n)}$ is the cumulative number of households within each successive position within the spatially ordered income parade, and $N_{j-1(n)}$ is the cumulative number of households one rank lower within the spatially ordered income parade. The jth neighborhood's average rank is then equal to $(N_{j(n)} + N_{j-1(n)} + 1)/2$.

Then, we can plug in G_B and G_r to S_r to yield:

$$S_r = \frac{\sum_{j=1}^{J}(\bar{R}_{j(n)} - \frac{H+1}{2})y_j}{\sum_{j=1}^{J}(\bar{R}_j - \frac{H+1}{2})y_j} \tag{10}$$

The interpretation of the spatial ordering index is that it can be thought of as a ratio of two covariances.

The spatial ordering index utilizes the idea of spatial autocorrelation– and is unique in doing so. Dawkins (2007) claims this makes the spatial ordering index less sensitive to presence of outliers, while it still satisfies the "principle of transfers," and is "flexible enough to quantify a variety of spatial patterns of segregation".

However, I have a major critique of this measure. In order to calculate the spatial ordering index, one needs average neighborhood income per household, which is not reported on the census but must be estimated. This measure seems only to pick up on between-neighborhood segregation and ignores the aspect that mean income ("neighborhood income") is not reflective of the income of everyone in the census tract and that income could sometimes vary greatly (or not) within neighborhoods, making means or medians not representative of the actual phenomenon, rendering the initial ranking to be somewhat uninformative.

With Dawkins' measure, there is a big loss of information due to the averaging of incomes.

Further, mean income is not a good statistic to describe income due to its typically skewed distribution.

While this measure is a step in the right direction, it ultimately obscures greater information and misrepresents the phenomenon of segregation by not accounting for economic diversity within census tracts. The monocentric spatial ordering is nice because it takes into account some urban economic theory but with the rise of polycentric cities it becomes a not so great way of capturing information when there might be multiple downtowns. It would be better to not impose a superficial structure on the city and rather just interpret it as it truly is.

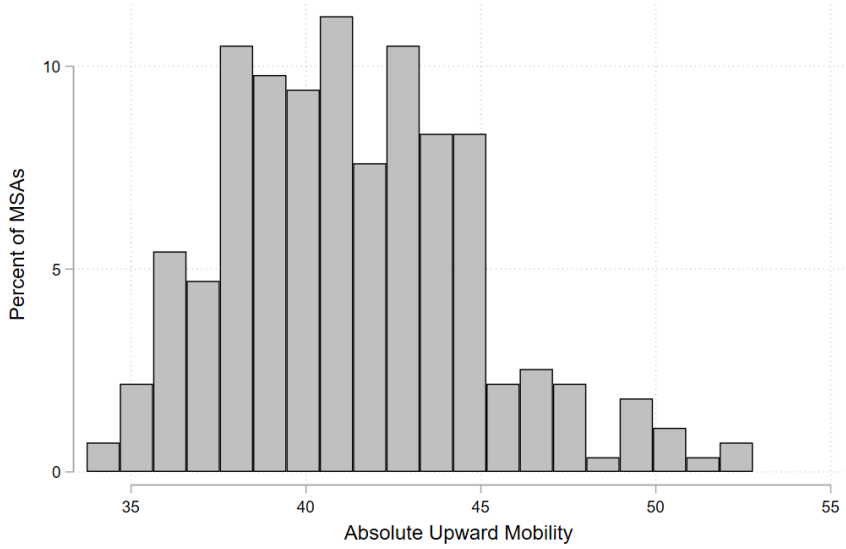
The academic community is faced with a complex problem: we must step away from aspatial measures of segregation, and yet, there does not exist an alternative that is shown to be "better" (whatever that may mean). Thus, the remainder of this paper attempts to capture some of the spatial information that may actually be important in describing the meaningful dynamics of and geography of economic segregation among US cities.

In the sections that follow, the data is first described in Section 3, and then I explain my methodology in Section 4, detailing the creation of income bin maps by city and the use of k-means clustering for variable creation. I then evaluate regressions utilizing these variables in Section 5. In Section 7, I conclude that more of the variation in economic mobility can be explained by my measures rather than by the rank-order index, which serves as a call to action for researchers to meaningfully incorporate the geography of income segregation into their estimates when considering economic segregation in their research.

3. Data description

The measure used for economic mobility consistently throughout this paper is actually termed by (Chetty et al., 2014) as *absolute upward mobility*. This measure is defined as the mean rank (in the national child income distribution) of children whose parents are in the 25th percentile of the national parent income distribution, with values possibly ranging from 1 to 100. I source data on a MSA level of this measure of absolute upward mobility from Chetty et al. (2014). In this sample, the values range from 33.72783 to 52.77465, the distribution of which can be seen in Figure 5. They predict this outcome using federal income tax records between 1996 and 2012 for approximately 10 million children who have a valid SSN, were US citizens in 2013, for whom they are able to identify parents with positive income, and were born between 1980 and 1982. The authors link the children to their parents' pre-tax household incomes from where they lived at age 16 (in 1996-1998), and measure children's pre-tax incomes in 2012, when they are ages 30-32 (Chetty et al., 2014).

Figure 5: The distribution of absolute upward mobility across MSAs



It is important to measure economic mobility like this rather than measuring relative mobility, estimates of which may be driven by worse-off circumstances of those who grew up wealthy rather than better-off circumstances of those who grew up poor Chetty et al. (2014).

Because the 2000 Census is the closest in time to the dates parents' household incomes were measured (1996-1998), I utilize data on population by income bin by census tract for all 276 MSAs designated in the 2000 Census. I source these data from NHGIS. Income is reported in 16 different bins or brackets as follows:

1. Less than $10,000
2. $10,000 to $14,999
3. $15,000 to $19,999
4. $20,000 to $24,999
5. $25,000 to $29,999
6. $30,000 to $34,999
7. $35,000 to $39,999
8. $40,000 to $44,999
9. $45,000 to $49,999
10. $50,000 to $59,999
11. $60,000 to $74,999
12. $75,000 to $99,999
13. $100,000 to $124,999
14. $125,000 to $149,999
15. $150,000 to $199,999
16. $200,000 or more

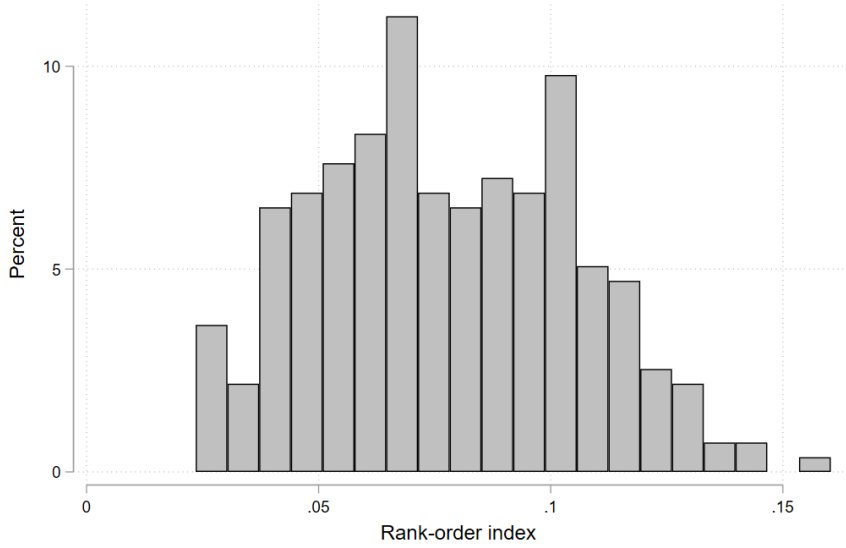
Thus, for every MSA, for every census tract within that MSA, we have the count of how many individuals are in each income bin in the year 2000. This makes up our main dataset.

Because absolute upward mobility is reported by 2013 MSA and not by 2000 MSA, we must link 2013 MSAs to 2000 MSAs by creating a crosswalk/relationship file between the MSAs in the two years.

I source shapefiles of 2000 and 2013 MSA boundaries and 2000 census tract boundaries from NHGIS and the Census. I use these in construction of the crosswalk and in the creation of the maps (shown in Section 2.3 and 4.1). I construct the crosswalk and employ it to compute estimates of absolute upward mobility corresponding to the 2000 MSA boundaries using simple weighted averages based on the proportions of area of each 2013 MSA within the 2000 MSAs.

Then, I create the maps as discussed in Section 4.1 and calculate rank-order index using the RANKSEG Stata module and the income bin data by census tract. The distribution of the rank-order index by MSA looks familiarly bimodal as is seen in Figure 6.

Figure 6: The distribution of the rank-order index across MSAs



4. Methodology

I draw from computational science in order to characterize and measure the shape of segregation. This process involves creating maps of binned income data by census tract for entire MSAs and bits of their surrounding regions, converting those maps to images, and running analyses on them. I run an unsupervised clustering algorithm (called *k-means clustering*) on the maps and then convert the results into indicator variables indicating which cluster each MSA belongs to, thus picking up on certain elements of economic segregation that these cities share, and accounting for those similarities.

Using those indicator variables, I run a regression to predict absolute upward mobility and compare those results to the regression of rank-order index predicting absolute upward mobility, specifically focusing on the different R^2 and adjusted R^2 values to point to the greater power of these indicator variables in explaining economic mobility.

As robustness checks, I use stepwise selection procedures as well as a method termed *Multiple Correspondence Analysis* (MCA) in order to reduce the dimensionality of the data and then rerun the regression on absolute upward mobility using the subset of variables and then the "dimensions" of segregation.

4.1. Map creation

I create maps in Python using the GeoPandas and Matplotlib libraries, coloring them all the same scale and allowing for areas with no data to be given the same color as areas with a value of 0. These maps are of the proportion of people within a certain income bin, for example the lowest income bin reported by the Census is for those who make less than $10,000, thus I generated 276 maps for all the 2000 MSAs displaying what proportion of people make $10,000 or less in each census tract in every MSA. I save

these maps as images of all the same pixel size, 224 by 224.

The use of maps allows us to capture spatial relationships in a new way not seen before by methods to measure economic segregation. It allows us to visualize the densities of different income classes throughout cities, and somewhat obscures the boundaries posed to be a problem in discussion of the *modifiable areal unit problem* (MAUP). MAUP arises from population data typically being collected, aggregated, and reported for spatial units based on political divisions that may have no relationship with meaningful social or spatial divisions. This method of aggregating data implicitly assumes that people within these regions are more similar than the people on the boundaries of the regions who may be geographically located nearer to one another but in separate municipalities. Unless spatial boundaries correspond to meaningful social boundaries, all measures of segregation that rely on aggregates are sensitive to the drawing of boundaries (Reardon, 2006). While I am not able to address this problem completely, my approach of using maps without drawn borders that are aggregated on the census tract level blur the boundaries more implicitly relative to other approaches that either ignore between-neighborhood segregation or handle the boundaries explicitly.

4.2. K-means clustering and variable creation

K-means clustering is a non-hierarchical clustering algorithm, a technique that falls under the umbrella of unsupervised learning models (Baumer et al., 2017). It is commonly used because of its simplicity. It essentially groups together data that is most similar to the mean of a certain cluster. It iteratively decides which cluster to place the next data point into. The algorithm's objective is to find k centroids, or means, one for each cluster of data points, that are placed as far away as possible from one another. When a data point is added to a cluster, one must recalculate the centroid of that cluster. The algorithm minimizes a "cost" or objective function, usually a sum of squared errors (Kodinariya

and Makwana, 2013).

To choose k, the number of clusters, I use the elbow method, which is the oldest method for determining the number of clusters. The idea is to calculate the cost function given a range of k values, and to display the graph of number of clusters versus cost. There may be some value k for which the cost drops dramatically, and after that it plateaus upon increasing k further. We reach an "elbow," after which the cost function decreases very slowly (Kodinariya and Makwana, 2013).

I take each set of maps corresponding to a given income bin, represent them as 224×224 matrices, and cluster them into anywhere between 5 to 12 groups using the k-means clustering algorithm implementation from the scikit-learn library in Python. I employ the elbow method, testing each set of maps on a range of 2 to 15 clusters, in order to choose the optimal number of clusters for each income bracket.

With the results of clustering by each income bin, I transform these data into indicator variables to use in the rest of my analyses.

4.3. Step-wise model selection and multiple correspondence analysis

In order to check the robustness of my results, I employ methods of step-wise model selection and dimensionality reduction. Step-wise model selection is the adding or removing of variables from the model sequentially in order to minimize the value of some information criterion, in this case the Akaike Information Criterion, an estimator of the relative quality of a statistical model on a given set of data.

There is often redudant information in the data, as there is in the case of my clusters. The irrelevant variables add noise and obscure the actual patterns in the data, so we can remove them using dimensionality reduction techniques (Baumer et al., 2017).

One such method is *multiple correspondence analysis* (MCA). MCA is an extension of correspondence analysis (CA) in which one can analyze the pattern of relationships of categorical variables. It is also a generalization of principal component analysis when the variables are categorical, not quantitative. It is obtained by using a standard CA on an indicator matrix. The eigenvalues obtained from analysis of the Burt matrix give a decent approximation of the inertia explained by each of the variables inputted into MCA. The interpretation behind MCA is based upon proximities between points in a low-dimensional map (Valentin and Abdi, 2007).

I utilize MCA to produce ten "dimensions" of spatial segregation that I then evaluate as predictors of absolute upward mobility.

5. Results

I evaluate the relationship between Reardon's rank-order index and absolute upward mobility and find a statistically significant result that has an adjusted R^2 of 0.0451.

Table 2: Relationship between rank-order index and absolute upward mobility

	Absolute upward mobility	
Rank-order index	-28.43***	(7.604)
Constant	43.53***	(0.625)
Observations	276	
R^2	0.0486	
Adjusted R^2	0.0451	
Degrees of freedom (Model)	1	
Degrees of freedom (Residual)	274	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Then, I evaluate the regression of rank-order index on my clusters which results in

a model with an adjusted R^2 of 0.1507, significantly higher than the adjusted R^2 of the model with rank-order index. See the regression results for only the statistically significant variables reported in Table 3 or see the full results in Appendix Section C.

Table 3: Full model: Absolute upward mobility on clusters (showing only statistically significant variables)

	Absolute upward mobility	
\$30,000 to \$34,999 (Cluster 3)	-1.892*	(0.949)
\$45,000 to \$49,999 (Cluster 3)	-3.004*	(1.317)
\$45,000 to \$49,999 (Cluster 5)	-2.026*	(0.970)
\$45,000 to \$49,999 (Cluster 8)	-2.310*	(1.048)
\$60,000 to \$74,999 (Cluster 6)	-3.558*	(1.397)
\$60,000 to \$74,999 (Cluster 7)	-3.289*	(1.444)
\$75,000 to \$99,999 (Cluster 2)	-3.251*	(1.491)
\$75,000 to \$99,999 (Cluster 3)	-4.641**	(1.503)
\$75,000 to \$99,999 (Cluster 6)	-2.766*	(1.240)
\$125,000 to \$149,999 (Cluster 3)	-2.150*	(0.928)
\$200,000 or more (Cluster 3)	3.407*	(1.427)
Constant	46.97***	(3.850)
Observations	276	
R^2	0.5584	
Adjusted R^2	0.1507	
Degrees of freedom (Model)	132	
Degrees of freedom (Residual)	143	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Because my full regression has so many degrees of freedom, one might be worried about what is driving the results: am I actually measuring segregation in a more meaningful way, or is it just because of the incredible amount of flexibility in a model with so many variables? Thus, I perform some robustness checks.

5.1. Robustness checks

In order to verify that the results from the full regression are due to meaningful information in the clusters rather than just the flexibility given by a model with so many parameters, I run two different types of robustness checks on the clusters: step-wise model selection procedures and Multiple Correspondence Analysis (MCA). I first employ forward and backward step-wise model selection procedures evaluating by the Akaike Information Criterion in order to find an important subset of clusters to run the model on. Second, I utilize MCA in order to reduce the dimensionality of the data. I show that using both of these procedures, the adjusted R^2 values corresponding to these models are consistently much higher than that of the model using rank-order index.

The results of the model selection procedures can be found in Table 4. With only 53 indicator variables (versus 132 in the full model), this model is still able to explain much more of the variation in mobility than that of the rank-order index. It has an adjusted R^2 of 0.3564, significantly higher than that of the rank-order index model and than that of the full model due to its smaller number of predictors.

Table 4: Relationship between stepwise-selected clusters and absolute upward mobility

	Absolute upward mobility	
Less than $10,000 (Cluster 3)	2.855**	(1.090)
$10,000 to $14,999 (Cluster 5)	-1.614**	(0.529)
$10,000 to $14,999 (Cluster 6)	-2.521	(1.651)
$10,000 to $14,999 (Cluster 7)	-3.222**	(1.154)
$10,000 to $14,999 (Cluster 11)	-1.620*	(0.652)
$10,000 to $14,999 (Cluster 12)	-2.096	(1.316)
$15,000 to $19,999 (Cluster 2)	1.609	(0.934)
$15,000 to $19,999 (Cluster 4)	1.079	(0.748)
$15,000 to $19,999 (Cluster 5)	1.215*	(0.469)
$15,000 to $19,999 (Cluster 8)	1.853**	(0.689)
$15,000 to $19,999 (Cluster 9)	1.751	(1.014)
$100,000 to $124,999 (Cluster 3)	1.258*	(0.601)
$100,000 to $124,999 (Cluster 4)	1.485*	(0.703)

Table 4: Relationship between stepwise-selected clusters and absolute upward mobility

	Absolute upward mobility	
$100,000 to $124,999 (Cluster 5)	1.116*	(0.498)
$25,000 to $29,999 (Cluster 2)	-0.921	(0.629)
$25,000 to $29,999 (Cluster 4)	-3.136**	(1.153)
$25,000 to $29,999 (Cluster 7)	-2.168*	(0.844)
$25,000 to $29,999 (Cluster 9)	-2.228**	(0.711)
$25,000 to $29,999 (Cluster 11)	-3.200*	(1.474)
$30,000 to $34,999 (Cluster 2)	1.374	(0.915)
$30,000 to $34,999 (Cluster 3)	-1.884***	(0.551)
$35,000 to $39,999 (Cluster 3)	1.254	(0.860)
$35,000 to $39,999 (Cluster 4)	1.280	(0.656)
$35,000 to $39,999 (Cluster 5)	1.091	(0.569)
$35,000 to $39,999 (Cluster 6)	3.417**	(1.154)
$35,000 to $39,999 (Cluster 8)	-2.111**	(0.744)
$40,000 to $44,999 (Cluster 2)	-1.358*	(0.621)
$40,000 to $44,999 (Cluster 3)	-1.400	(0.982)
$40,000 to $44,999 (Cluster 4)	-1.445	(0.807)
$40,000 to $44,999 (Cluster 6)	-1.279*	(0.637)
$40,000 to $44,999 (Cluster 8)	-1.704*	(0.749)
$40,000 to $44,999 (Cluster 10)	-3.519**	(1.288)
$45,000 to $49,999 (Cluster 2)	-0.971	(0.717)
$45,000 to $49,999 (Cluster 3)	-3.082***	(0.795)
$45,000 to $49,999 (Cluster 4)	-1.796**	(0.628)
$45,000 to $49,999 (Cluster 5)	-1.885***	(0.557)
$45,000 to $49,999 (Cluster 8)	-2.529***	(0.611)
$60,000 to $74,999 (Cluster 6)	-1.687**	(0.537)
$60,000 to $74,999 (Cluster 7)	-1.582**	(0.594)
$75,000 to $99,999 (Cluster 3)	-2.614***	(0.779)
$75,000 to $99,999 (Cluster 6)	-0.911	(0.597)
$100,000 to $124,999 (Cluster 3)	2.428**	(0.861)
$100,000 to $124,999 (Cluster 8)	2.536**	(0.872)
$125,000 to $149,999 (Cluster 3)	-1.994***	(0.591)
$150,000 to $199,999 (Cluster 4)	-8.657*	(3.361)
$150,000 to $199,999 (Cluster 8)	-5.391	(3.158)
$150,000 to $199,999 (Cluster 9)	1.111*	(0.507)
$200,000 or more (Cluster 2)	2.455	(1.315)
$200,000 or more (Cluster 3)	2.976**	(0.893)
$200,000 or more (Cluster 6)	4.339*	(2.152)
$200,000 or more (Cluster 7)	1.440*	(0.646)
$200,000 or more (Cluster 11)	0.863	(0.555)
$10,000 to $14,999 (Cluster 8)	-0.762	(0.581)

Table 4: Relationship between stepwise-selected clusters and absolute upward mobility

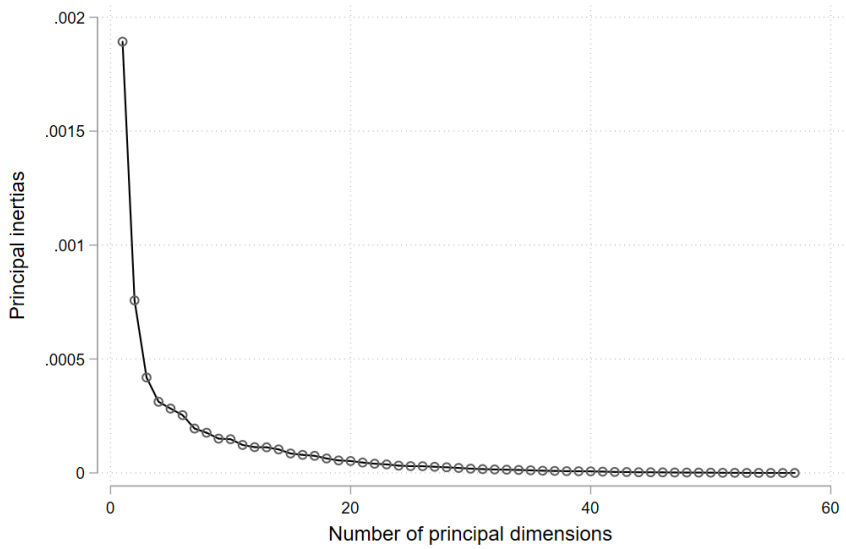
	Absolute upward mobility	
Constant	43.35***	(0.693)
Observations	276	
R^2	0.4805	
Adjusted R^2	0.3564	
Degrees of freedom (Model)	53	
Degrees of freedom (Residual)	222	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In MCA, we can produce a scree plot in order to showcase the percentages of inertia explained by each MCA principal dimension. The first dimension explains the most inertia, and it continues on in descending order. Figure 7 showcases these results.

Figure 7: Scree plot of principal inertias after MCA



What you can gleam from Figure 7 is that the "elbow" of the curve (the same idea as the elbow method described in Section 4.2) is found when the number of principal

dimensions is roughly equal to ten. Thus, using the results of MCA, I run the regression on ten dimensions to showcase the robustness of the results.

Table 5: Relationship between ten dimensions and absolute upward mobility

	Absolute upward mobility	
Dimension 1	0.461*	(0.201)
Dimension 2	-0.881***	(0.201)
Dimension 3	0.0309	(0.201)
Dimension 4	0.365	(0.201)
Dimension 5	0.179	(0.201)
Dimension 6	0.229	(0.201)
Dimension 7	0.141	(0.201)
Dimension 8	-0.208	(0.201)
Dimension 9	-0.276	(0.201)
Dimension 10	-0.699***	(0.201)
Constant	41.33***	(0.201)
Observations	276	
R^2	0.1469	
Adjusted R^2	0.1147	
Degrees of freedom (Model)	10	
Degrees of freedom (Residual)	265	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

With ten dimensions of segregation, the adjusted R^2 remains relatively high compared to that of the rank-order index regression, at 0.1147.

Then, I run the regression on a single dimension of segregation, to showcase how even a single number created using geographic information can be a more powerful predictor than the currently used rank-order index.

Table 6: Relationship between Dimension 2 and absolute upward mobility

	Absolute upward mobility	
Dimension 2	-0.881***	(0.207)
Constant	41.33***	(0.207)
Observations	276	
R^2	0.0622	
Adjusted R^2	0.0588	
Degrees of freedom (Model)	1	
Degrees of freedom (Residual)	274	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

With only a single dimension of segregation, I can explain still considerably more of the variation in economic mobility than the rank-order index is able to, with an adjusted R^2 of 0.0588.

6. Discussion

As captured by the higher R^2 and adjusted R^2 values of my models in comparison with that of the model involving rank-order index, there is a stronger relationship between my spatial measures of segregation and economic mobility than there is between the aspatial rank-order index and economic mobility. It is important to ask, why would this be the case?

As has been discussed previously, the rank-order index fails to capture any information about the geographic proximity of census tracts. Returning to the checkerboard problem, individuals residing in a city like Figure 1(a) are able to travel short distances outside of their neighborhoods to interact with individuals from other income groups, share and exchange resources, attend schools, etc. For those in a city like Figure 1(b), this class

mixing becomes less likely, and economic mobility stemming from both neighborhood effects and school financing is theoretically lowered. And yet, these two cities are not recognized as different by the rank-order index.

Thus, it is likely that since the rank-order index fails to capture *how* segregation matters for mobility, it leads to an underestimate of the relationship between segregation and mobility. This obscures part of the way that segregation might impact mobility.

Meanwhile, my methodology theoretically does capture the difference between these two cities– the k-means clustering algorithm would be unlikely to group cities like Figures 1(a) and 1(b) into the same cluster, therefore more correctly distinguishing between these scenarios. My best result showcases the ability of a subset of these indicator variables to explain 48.05% of the variation in mobility, with an adjusted R^2 of 0.3564, as shown in Table 4.

Much more of the variation in economic mobility can be explained by economic segregation compared to what researchers have found in the past. This conclusion is extremely important for seeking out opportunities to improve mobility for poor families.

7. Conclusion

Using the information I have generated from the clustering of income bin maps, I am able to explain a greater portion of the variation in economic mobility than is done using the rank-order index– something not attributed to the number of variables in the regression, as robustness checks involving multiple correspondence analysis and step-wise model selection processes have shown us.

There are definitely many limitations to my methodology. I do not control for college

towns, where many people might make $10,000 or less a year in what are in actuality very wealthy neighborhoods. Throughout my use of k-means and MCA, my variables and coefficients are not readily interpretable compared with many of the indices that exist on a range of 0 to 1 or -1 to 1. There is no value for "complete segregation" or "no segregation," as if those are things that even exist in theory.

There are some potential drawbacks of using k-means on the raw data, but it seems to not be a problem. It would have been good to transform the data. Yet, this seems to not be too great of an issue because my results are still extremely significant.

I am not suggesting for my methodology to become a new way of measuring economic segregation, for it does not have the properties that many researchers are concerned with, for example the "principle of transfers." However, even though I have not proposed a new spatial segregation index, I have created a process by which to generate spatial information or dimensions of economic segregation, an important contribution to the broader conversation about how to measure economic segregation and why economic segregation is important.

I do not think that segregation can be condensed into one single measure on the real number line– it is not something that is either high or low, but it is a phenomenon that can manifest in many different ways that are not able to be definitively ranked or ordered. Segregation manifests in varying forms as theorized by urban studies and economics scholars alike for centuries. These structures likely have impact. A more nuanced, sensitive approach to measuring spatial segregation must not try to summarize estimates into a single term, and it must incorporate the geography of wealth, or the shape of segregation.

We need a better measure for segregation. I urge the academic community to stop thinking of the rank-order index as one that measures the phenomenon known as segregation,

and rather to think of it as one measuring economic homogeneity. This paper can serve as a call to action for researchers to work towards a better, explicitly spatial measure of economic segregation.

References

Baumer, B. S., N. J. Horton, and D. T. Kaplan (2017). *Modern Data Science with R*. New York, NY: CRC Press.

Beauregard, R. (2007). More Than Sector Theory: Homer Hoyt's Contributions to Planning Knowledge.

Chetty, R. and N. Hendren (2016). The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates *. *The Quarterly Journal of Economics*.

Chetty, R., N. Hendren, P. Kline, and E. Saez (2014). Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States. *Quarterly Journal of Economics 129*(4).

Dawkins, C. J. (2007). Space and the measurement of income segregation. *Journal of Regional Science 47*(2), 255–272.

Dwyer, R. E. (2010). Poverty, prosperity, and place: the shape of class segregation in the age of extremes. *Social Problems 57*(1), 114–137.

Ehrenhalt, A. (2012). *The great inversion and the future of the American city*. Knopf.

Harris, C. D. and E. L. Ullman (1941). The Nature of Cities. Technical report.

Jargowsky, P. A. (1996). Take the money and run: Economic segregation in U.S. metropolitan areas. *American Sociological Review 61*, 984–998.

Jargowsky, P. A. and J. Kim (2005). A Measure of Spatial Segregation: The Generalized Neighborhood Sorting Index.

Johnston, R., M. Poulsen, and J. Forrest (2014). Segregation matters, measurement matters. *Social-spatial segregation: Concepts, processes and outcomes*, 13–44.

Kneebone, E. (2014). The growth and spread of concentrated poverty, 2000 to 2008-2012. *The Brookings*.

Kodinariya, T. M. and P. R. Makwana (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies 1*(6).

Mayer, S. E. (2002). How economic segregation affects children's educational attainment. *Social forces 81*(1), 153–176.

Orfield, G. and C. Lee (2005). Why segregation matters: Poverty and educational inequality. *Civil Rights Project at Harvard University (The)*.

Reardon, S. F. (2006). A conceptual framework for measuring segregation and its association with population outcomes. In *Methods in social epidemiology*, pp. 169–192.

Reardon, S. F. (2011). Measures of Income Segregation.

Reardon, S. F. and K. Bischoff (2011). Income Inequality and Income Segregation. *American Journal of Sociology 116*(4), 1092–1153.

Reardon, S. F., K. Bischoff, A. Owens, and J. B. Townsend (2018). Has Income Segregation Really Increased? Bias and Bias Correction in Sample-Based Segregation Estimates. *Demography 55*, 2129–2160.

Reardon, S. F., G. Firebaugh, D. O'Sullivan, and S. Matthews (2006). A new approach to measuring socio-spatial economic segregation. In *29th general conference of the International Association for Research in Income and Wealth, Joensuu, Finland*.

Roberto, E. (2015). The Divergence Index: A Decomposable Measure of Segregation and Inequality. Technical report.

Tomer, A., E. Kneebone, R. Puentes, and A. Berube (2011). Missed opportunity: Transit and jobs in metropolitan America.

Valentin, D. and H. Abdi (2007). Multiple Correspondence Analysis.

Appendix

A. Neighborhood Sorting Index

The *Neighborhood Sorting Index* (NSI) is a measure of segregation that was created by Jargowsky (1996) and was very commonly used in sociology before Reardon et al.

(2006) came out with the rank-order index, after which it began to fade out of importance (Reardon and Bischoff, 2011). Its equation (seen below in Equation 11 can be interpreted as the ratio of the standard deviation of subarea mean incomes (weighted by subarea population) to the standard deviation of income in the regional population (Reardon, 2006).

$$NSI = \frac{\sigma_N}{\sigma_H} = \frac{\sqrt{(\sum_{n=1}^{N} h_n(\bar{y_n} - \bar{y}))/H}}{\sqrt{(\sum_{i=1}^{H}(y_i - \bar{y})^2)/H}} \tag{11}$$

where y is income, i indexes households, n indexes neighborhoods, h_n is the number of households in neighborhood n, and H and N are the total number of households and neighborhoods respectively.

It can be thought of as the proportion of income variation that lies between subareas, or the square root of the correlation ratio of between-tract income variance over the total income variance (Reardon and Bischoff, 2011; Jargowsky, 1996). Between-tract variance can be thought of as the household-weighted variance of the neighborhood means. In order to compute total variance of income, one needs access to individual-level sample data or income bin data can be used but in that case assumptions must be made about the distribution of households within income brackets– based on testing and comparison with PUMS estimates, Jargowsky (1996) assumes linear distributions in lower brackets and Pareto distributions in the brackets above the MSA mean, described in his Appendix A (Jargowsky, 1996).

Because NSI does not satisfy the property of being independent from income inequality, it may confound changes in residential sorting by income with differences in income distributions across time, place, and groups, which is why the rank-order index has been viewed as an improvement upon the NSI (Reardon and Bischoff, 2011).

There is also a spatial analog of the NSI termed the *Generalized Neighborhood Sorting Index* (GNSI) which simply incorporates a "distance-decay" effect (Reardon, 2006). The

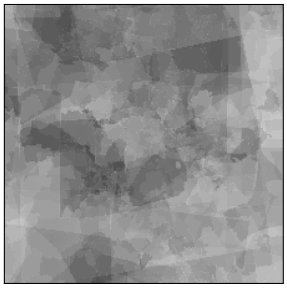
key difference from NSI is that the numerator of GNSI incorporates a flexible moving window for the calculation of a neighborhood's economic level which is larger than a the neighborhood itself, e.g. all neighborhoods whose centroids are within a certain distance r of the given neighborhood's centroid (Jargowsky and Kim, 2005). This again is based on arbitrary radii and falls prey to the failures caused by that. However, if you want to read more about it, please reference Jargowsky and Kim (2005).

B. Visualizing the clustering results

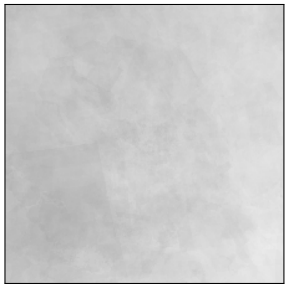
What is also interesting to look at is what is actually contained in these clustered groups of maps– these can be quite illuminating in describing what actually matters for economic segregation.

Here are a couple examples of clusters that were statistically significant in the full model. What is visualized here is, given each pixel value of each 224×224 image, take the average pixel value and create a new image that is the average of all the images within a cluster.

Figure 8: Statistically significant clusters: \$30,000 to \$34,999

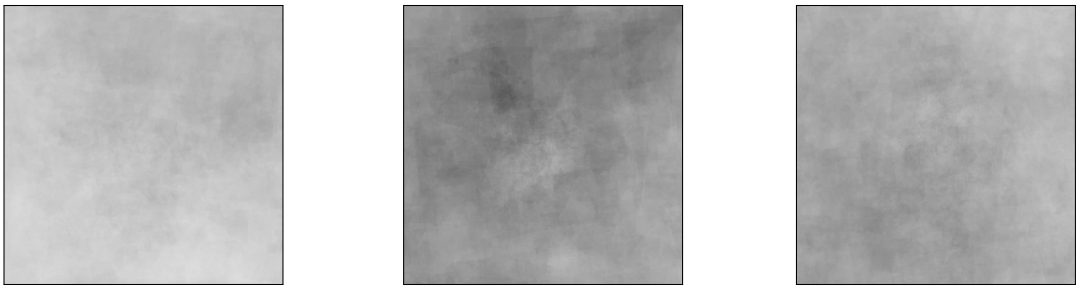


(a) Cluster 1: reference group



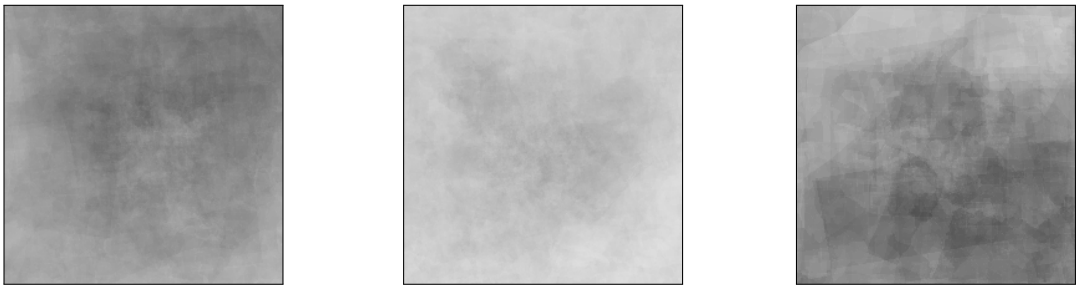
(b) Cluster 3: negative effect

Figure 9: Statistically significant clusters: $45,000 to $49,999



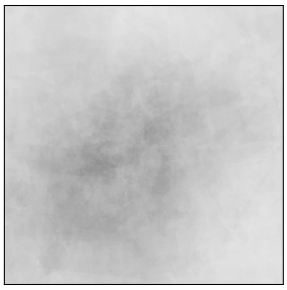
(a) Cluster 1: reference group (b) Cluster 5: negative effect (c) Cluster 8: negative effect

Figure 10: Statistically significant clusters: $60,000 to $74,999

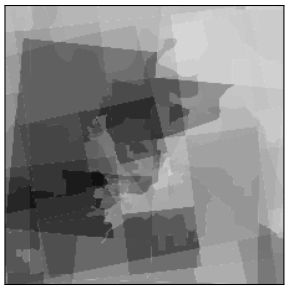


(a) Cluster 1: reference group (b) Cluster 6: negative effect (c) Cluster 7: negative effect

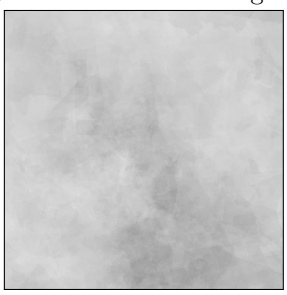
Figure 11: Statistically significant clusters: $75,000 to $99,999



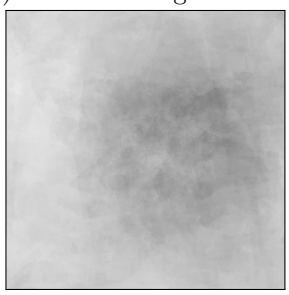
(a) Cluster 1: reference group



(b) Cluster 2: negative effect



(c) Cluster 3: negative effect



(d) Cluster 6: negative effect

C. Full regression results: full table

Here are the full results from my full regression with all 132 indicator variables.

Table 7: Full model: Absolute upward mobility on clusters

	Absolute upward mobility	
Less than $10,000 (Cluster 2)	-0.359	(1.651)
Less than $10,000 (Cluster 3)	1.262	(1.857)
Less than $10,000 (Cluster 4)	-0.369	(1.392)
Less than $10,000 (Cluster 5)	-0.636	(0.764)
Less than $10,000 (Cluster 6)	-0.751	(0.980)
Less than $10,000 (Cluster 7)	0.521	(3.103)
Less than $10,000 (Cluster 8)	-0.286	(2.037)
$10,000 to $14,999 (Cluster 2)	0.510	(1.302)
$10,000 to $14,999 (Cluster 3)	1.143	(1.710)

Table 7: Full model: Absolute upward mobility on clusters

	Absolute upward mobility	
$10,000 to $14,999 (Cluster 4)	0.198	(1.488)
$10,000 to $14,999 (Cluster 5)	-1.703	(1.011)
$10,000 to $14,999 (Cluster 6)	-1.761	(2.341)
$10,000 to $14,999 (Cluster 7)	-1.946	(1.904)
$10,000 to $14,999 (Cluster 8)	-0.604	(1.008)
$10,000 to $14,999 (Cluster 9)	-0.994	(2.613)
$10,000 to $14,999 (Cluster 10)	-1.441	(1.914)
$10,000 to $14,999 (Cluster 11)	-1.331	(1.046)
$10,000 to $14,999 (Cluster 12)	-2.492	(2.242)
$15,000 to $19,999 (Cluster 2)	2.478	(2.585)
$15,000 to $19,999 (Cluster 3)	2.016	(2.587)
$15,000 to $19,999 (Cluster 4)	2.246	(2.495)
$15,000 to $19,999 (Cluster 5)	1.635	(2.325)
$15,000 to $19,999 (Cluster 6)	1.693	(2.619)
$15,000 to $19,999 (Cluster 7)	-0.0482	(2.335)
$15,000 to $19,999 (Cluster 8)	2.792	(2.465)
$15,000 to $19,999 (Cluster 9)	3.039	(2.671)
$15,000 to $19,999 (Cluster 10)	0.907	(2.423)
$100,000 to $124,999 (Cluster 2)	-0.159	(0.798)
$100,000 to $124,999 (Cluster 3)	1.057	(1.035)
$100,000 to $124,999 (Cluster 4)	1.191	(1.135)
$100,000 to $124,999 (Cluster 5)	0.998	(0.774)
$25,000 to $29,999 (Cluster 2)	-1.304	(1.224)
$25,000 to $29,999 (Cluster 3)	-0.390	(1.230)
$25,000 to $29,999 (Cluster 4)	-2.718	(2.049)
$25,000 to $29,999 (Cluster 5)	-1.018	(1.247)
$25,000 to $29,999 (Cluster 6)	-0.499	(1.188)
$25,000 to $29,999 (Cluster 7)	-2.787	(1.488)
$25,000 to $29,999 (Cluster 8)	-0.576	(2.152)
$25,000 to $29,999 (Cluster 9)	-2.270	(1.487)
$25,000 to $29,999 (Cluster 10)	-0.299	(1.225)
$25,000 to $29,999 (Cluster 11)	-3.933	(2.262)
$25,000 to $29,999 (Cluster 12)	-0.889	(1.069)
$30,000 to $34,999 (Cluster 2)	2.176	(1.448)
$30,000 to $34,999 (Cluster 3)	-1.892*	(0.949)
$30,000 to $34,999 (Cluster 4)	-0.318	(0.986)
$30,000 to $34,999 (Cluster 5)	-0.599	(2.271)
$30,000 to $34,999 (Cluster 6)	0.0465	(0.921)
$30,000 to $34,999 (Cluster 7)	-1.417	(1.180)
$30,000 to $34,999 (Cluster 8)	-0.367	(0.995)

Table 7: Full model: Absolute upward mobility on clusters

	Absolute upward mobility	
$30,000 to $34,999 (Cluster 9)	0.878	(1.539)
$35,000 to $39,999 (Cluster 2)	0.0738	(0.820)
$35,000 to $39,999 (Cluster 3)	1.814	(1.424)
$35,000 to $39,999 (Cluster 4)	1.468	(1.089)
$35,000 to $39,999 (Cluster 5)	1.403	(1.050)
$35,000 to $39,999 (Cluster 6)	2.753	(1.681)
$35,000 to $39,999 (Cluster 7)	-1.349	(2.826)
$35,000 to $39,999 (Cluster 8)	-2.044	(1.251)
$40,000 to $44,999 (Cluster 2)	-1.433	(1.034)
$40,000 to $44,999 (Cluster 3)	-1.413	(1.923)
$40,000 to $44,999 (Cluster 4)	-1.893	(1.286)
$40,000 to $44,999 (Cluster 5)	-1.129	(2.300)
$40,000 to $44,999 (Cluster 6)	-1.146	(1.025)
$40,000 to $44,999 (Cluster 7)	0.287	(1.097)
$40,000 to $44,999 (Cluster 8)	-1.716	(1.212)
$40,000 to $44,999 (Cluster 9)	-0.349	(0.799)
$40,000 to $44,999 (Cluster 10)	-1.991	(2.080)
$45,000 to $49,999 (Cluster 2)	-1.293	(1.410)
$45,000 to $49,999 (Cluster 3)	-3.004*	(1.317)
$45,000 to $49,999 (Cluster 4)	-2.017	(1.204)
$45,000 to $49,999 (Cluster 5)	-2.026*	(0.970)
$45,000 to $49,999 (Cluster 6)	-0.163	(1.434)
$45,000 to $49,999 (Cluster 7)	-0.484	(1.759)
$45,000 to $49,999 (Cluster 8)	-2.310*	(1.048)
$50,000 to $59,999 (Cluster 2)	0.916	(1.337)
$50,000 to $59,999 (Cluster 3)	-0.340	(0.993)
$50,000 to $59,999 (Cluster 4)	-0.228	(1.167)
$50,000 to $59,999 (Cluster 5)	-0.170	(1.139)
$50,000 to $59,999 (Cluster 6)	-0.00260	(0.988)
$50,000 to $59,999 (Cluster 7)	-0.680	(1.096)
$60,000 to $74,999 (Cluster 2)	-2.006	(1.355)
$60,000 to $74,999 (Cluster 3)	-1.909	(1.443)
$60,000 to $74,999 (Cluster 4)	-1.544	(1.595)
$60,000 to $74,999 (Cluster 5)	-2.310	(1.606)
$60,000 to $74,999 (Cluster 6)	-3.558*	(1.397)
$60,000 to $74,999 (Cluster 7)	-3.289*	(1.444)
$60,000 to $74,999 (Cluster 8)	-2.273	(1.391)
$60,000 to $74,999 (Cluster 9)	2.147	(4.674)
$75,000 to $99,999 (Cluster 2)	-3.251*	(1.491)
$75,000 to $99,999 (Cluster 3)	-4.641**	(1.503)

Table 7: Full model: Absolute upward mobility on clusters

	Absolute upward mobility	
$75,000 to $99,999 (Cluster 4)	-3.768	(2.155)
$75,000 to $99,999 (Cluster 5)	-1.988	(1.258)
$75,000 to $99,999 (Cluster 6)	-2.766*	(1.240)
$75,000 to $99,999 (Cluster 7)	-2.419	(1.723)
$75,000 to $99,999 (Cluster 8)	-1.876	(1.347)
$75,000 to $99,999 (Cluster 9)	-1.418	(1.069)
$100,000 to $124,999 (Cluster 2)	0.256	(2.020)
$100,000 to $124,999 (Cluster 3)	3.019	(2.305)
$100,000 to $124,999 (Cluster 4)	5.591	(4.152)
$100,000 to $124,999 (Cluster 5)	0.577	(2.054)
$100,000 to $124,999 (Cluster 6)	0.952	(2.185)
$100,000 to $124,999 (Cluster 7)	0.103	(1.990)
$100,000 to $124,999 (Cluster 8)	3.692	(2.342)
$100,000 to $124,999 (Cluster 9)	1.712	(1.987)
$100,000 to $124,999 (Cluster 10)	2.238	(2.435)
$125,000 to $149,999 (Cluster 2)	-0.441	(1.637)
$125,000 to $149,999 (Cluster 3)	-2.150*	(0.928)
$125,000 to $149,999 (Cluster 4)	0.536	(1.745)
$125,000 to $149,999 (Cluster 5)	-2.939	(4.257)
$125,000 to $149,999 (Cluster 6)	-0.531	(1.154)
$125,000 to $149,999 (Cluster 7)	-0.312	(1.310)
$125,000 to $149,999 (Cluster 8)	3.770	(3.801)
$125,000 to $149,999 (Cluster 9)	-3.138	(3.542)
$125,000 to $149,999 (Cluster 10)	-0.0902	(0.896)
$150,000 to $199,999 (Cluster 2)	-3.766	(3.583)
$150,000 to $199,999 (Cluster 3)	-1.650	(1.804)
$150,000 to $199,999 (Cluster 4)	-7.716	(5.246)
$150,000 to $199,999 (Cluster 5)	-2.951	(3.629)
$150,000 to $199,999 (Cluster 6)	-0.353	(1.230)
$150,000 to $199,999 (Cluster 7)	-0.0742	(2.056)
$150,000 to $199,999 (Cluster 8)	-5.853	(5.991)
$150,000 to $199,999 (Cluster 9)	0.863	(0.875)
$150,000 to $199,999 (Cluster 10)	-2.501	(2.991)
$200,000 or more (Cluster 2)	1.572	(2.210)
$200,000 or more (Cluster 3)	3.407*	(1.427)
$200,000 or more (Cluster 4)	-1.101	(1.608)
$200,000 or more (Cluster 5)	0	(.)
$200,000 or more (Cluster 6)	5.896	(4.165)
$200,000 or more (Cluster 7)	1.835	(1.106)
$200,000 or more (Cluster 8)	4.632	(6.035)

Table 7: Full model: Absolute upward mobility on clusters

	Absolute upward mobility	
$200,000 or more (Cluster 9)	1.448	(4.556)
$200,000 or more (Cluster 10)	-0.182	(4.513)
$200,000 or more (Cluster 11)	1.055	(0.861)
$200,000 or more (Cluster 12)	-1.912	(5.201)
Constant	46.97***	(3.850)
Observations	276	
R^2	0.5584	
Adjusted R^2	0.1507	
Degrees of freedom (Model)	132	
degrees of freedom (Residual)	143	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$