Senior Thesis 301
Department of Economics, Vassar College

Measuring Economic Segregation Based on What Matters for Creating Opportunity

Kai Matheson

May 2019

Abstract

Economic opportunity is considerably lower in US cities that endure more extreme economic segregation. I argue that this relationship, between economic mobility and economic segregation, has been underestimated due to mismeasurement of economic segregation. The commonly-used rank order information theory index is "aspatial" in that it does not take into account the geography of neighborhoods. I use publicly-available household income data on US metropolitan areas in order to compare the rank-order index with my own findings. In my methodology, I directly assess the spatial dimensions of economic segregation by clustering income density maps of cities into visually similar groups and then employing lasso and multiple correspondence analysis. I utilize the theoretical link between economic segregation and economic mobility in order to evaluate the predictive power of these two approaches. The regression with rank-order index as a predictor results in an adjusted R^2 of 0.1450, compared with 0.3067 via my approach. I find that economic segregation explains more variation in economic mobility than previously realized, and I illuminate connections between economic geography and economic mobility. These results serve as a call to action for researchers to work towards explicitly spatial and multidimensional methods of capturing economic segregation.

Contents

Acknowledgements

I am filled with gratitude for each and every individual who has helped me through the process of dreaming up and creating this work.

Professor Pearlman, you made this process not just bearable but enjoyable. It was a pleasure to work with you, and to learn together. Thanks for believing in me.

Professor Frye and Professor Pradhan, thanks for the guidance and hours of conversations. Professor An, Professor Koechlin, Professor de Leeuw, and Professor Morin, thanks for giving me the skills to achieve this. Professor Cunningham, Neil Curri, Professor Lisa Lowrance, Professor Adam Lowrance, and Professor McCleary, thank you for helping me troubleshoot and problem-solve.

Mom, thank you for raising me, loving and accepting me, and being there for me. I hope I can not only make an impact on this unfair system but also make you proud. Dad and Nathan, thanks for always listening to me ramble, and being proud of me.

Ruth, thanks for being such a supportive mentor, friend, and godparent. Suze, I am proud to follow in your footsteps as I cross the stage in May. Derek, thank you for inspiring me to love numbers. Granny Cecily, your unwavering support and pride fill me with warmth.

April, Paige, Katie, Eli, Lauren, Mercy, Colby, Katherine, Max, and Kyle, thank you for the late nights, the long talks, and the encouragement. Es and Scott, thank you for helping me amidst your own chaos. To all my friends, thanks for being there to support me.

Monique and Saira, thank you for helping me cope.

Operation Understanding DC, thank you for shaping me.

Opportunity Insights, I am so excited to join you all and contribute to the work that you do. Thanks for being passionate for my passions.

I could not have done it without you all.

Love,

Kai

1. Introduction

The American Dream tells a story that hard work pays off, and anyone can achieve their goals if they are driven enough. But for decades, economists and sociologists have tugged at this cultural assumption, illuminating the ways in which the experience of economic opportunity is far from uniformly distributed.

Recent work by Chetty et al. (2014) adds to a growing body of literature highlighting the role of place in economic opportunity, measured via economic mobility. Chetty et al. (2014) find that their measure of *absolute upward mobility* is strongly negatively correlated with economic segregation, which they quantify by using the *rank-order information theory index*, an economic segregation measure proposed by Reardon et al. (2006) that ranges from zero to one.

While this insight by Chetty et al. (2014) is important, I argue that it suffers from issues of mismeasurement. The rank-order information theory index is "aspatial" in that it does not take into account the geography of neighborhoods, or what I refer to as the "shape" of segregation. While its claim is to capture economic segregation, it seemingly only captures economic homogeneity. This measure fails to distinguish between extremely different scenarios, leading to an underestimate of the presence of economic segregation. This, in turn, could potentially lead to an underestimate of the relationship between economic segregation and economic mobility.

As economic segregation is such an important issue, it is crucial that it be measured spatially and multidimensionally in order to accurately capture the state of the phenomenon. In capturing economic homogeneity rather than segregation, policymakers miss key insights into the impact of segregation, as it is left as an omitted variable in all the studies that employ the rank-order information theory index. This causes policymakers to draw the wrong conclusions about how segregation truly impacts our lives, seemingly minimizing the true impacts of segregation on outcomes that matter.

There are disagreements in the literature about how to define economic segregation. I define economic segregation as the geographic separation of different economic classes into distinct and separate neighborhoods. I argue that if geography is inherent in the definition of segregation, then the measurement of this phenomenon must be considered spatially. Thus, it cannot be measured in the absence of any spatial information.

In this paper, I use 2000 Census data on household income by census tract in order to explore competing methods of measuring economic segregation, while being mindful of typical public data availability. I present my own method of evaluating spatial segregation in multiple dimensions, and compare it to the commonly-used rank-order information theory index (Reardon et al., 2006). In the context of US cities, I utilize the theoretical link between economic segregation and economic mobility to evaluate the predictive power of these two methodologies. I find that, using my methodology, economic segregation explains more variation in economic mobility than has been previously realized by other researchers, showing segregation holds even more weight in

economic outcomes for families than had been previously thought. I illuminate some of the geographic structures that may cause heightened or lowered opportunity. I conclude with a call to action for researchers to work towards better, explicitly spatial measures of economic segregation.

To my knowledge, this paper serves as the second-ever analysis of the spatial dimensions of economic segregation. But, rather than relying on combinations of aspatial measures of "spatial dimensions" as is done by Dwyer (2010), my methods more directly evaluate spatial segregation in metropolitan areas using income maps of cities as the bases of assessment, which are then analyzed via methods from computational science.

Specifically, I employ the k-means clustering algorithm to cluster income maps of cities into visually similar groups. I utilize the lasso regression analysis method as well as multiple correspondence analysis in stripping the most important information from these groupings to predict economic mobility.

Section 2 frames my work within the context of the literature. I begin by first describing how economic segregation matters for economic mobility, and then introducing the rank-order information theory index and its limitations, emphasizing that geographic shape matters. The data and methodology are described in Section 3. In Section 4, I evaluate regressions utilizing the lasso method and multiple correspondence analysis, and I interpret the results. Section 5 discusses the implications of these results. I conclude in Section 6 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.

2. Literature review

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

Why might growing up in more segregated cities have long term negative consequences, even for individuals who move away? Theory suggests that economic segregation impacts economic mobility via school financing and neighborhood effects (Mayer, 2002). Schools in the United States are typically financed by property taxes of local communities. Thus, a low income school district that suffers from lower tax revenues will experience lower school quality, resulting in poorer educational outcomes. Educational achievement is a strong predictor of future income. A poor child starting out in a school with concentrated poverty may see smaller educational gains and thus would be less upwardly mobile than a similarly 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.

Much of the empirical literature backs up the theory of school financing as the link between economic segregation and economic mobility, while there is less research on neighborhood effects. 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 largest predictors of educational outcomes. While that study is focused on class segregation within schools, 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).

In this context, it is important to consider patterns in the shape of economic segregation to better understand how the spatial distribution of income affects opportunity in a city.

2.1.1. Urban land use models: Theories of economic segregation

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. These models are theorized in urban economic literature, but have yet to be rigorously evaluated on US cities.

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 consists of 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 inner-city poor residences and then followed by 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.

The Burgess model is still in use today but is becoming 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 has been seen within cities (Tomer et al., 2011). With demographic inversion occurring in many large American cities, the Kohl model becomes relevant again.

There are other models, however, that do not assume a ring-like structure. 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, 1945). This can be envisioned as a pie chart, with each slice termed a "sector" in which there is one dominant type of land use (Beauregard, 2007). On the other hand, the multiple nuclei model by Harris and Ullman (1945) 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 flexible layouts of cities.

There has been almost no work done on directly analyzing the impacts of different shapes of economic segregation, highlighting the need for research on evaluating the outcomes of cities of various socioeconomic-spatial forms. Thus, an added contribution of this work is to investigate whether these structures are what matter for influencing economic mobility on a larger scale.

2.2. Measuring economic segregation: Reardon's rank-order index

Now that I have motivated why the spatial form of economic segregation matters for economic mobility, I explain how it is typically measured aspatially. First introduced by Reardon et al. (2006), the *rank-order information theory index* is the most commonly used measure of economic segregation today, by economists and sociologists alike. 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 measure I decide to dissect. In order to do this work, the equations of the rank-order index must first be introduced and understood.

Let p be an income percentile rank in a given local income distribution such that $p = F(Y)$ 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. The rank-order index divides the population into two groups, those who are above percentile

p and those who are below. It utilizes the following equation, which is the typical formula for entropy of a population when divided into these two groups.[1]

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

Denote T as the population of the metropolitan area and t_j as the population of neighborhood j.[2] I can then compute $H(p)$, the traditional information theory index of segregation, representing the level of segregation between the two groups.

$$H(p) = 1 - \sum_j \frac{t_j E_j(p)}{T E(p)} \tag{2}$$

Note here that this index simply sums over neighborhoods, failing to capture the geographic relationships between them.

Then the rank-order information theory index H^R can be written as follows (Reardon et al., 2006).

$$H^R = 2 \ln(2) \int_0^1 E(p) H(p) \, dp \tag{3}$$

The rank-order index can be interpreted as the ratio of within-neighborhood 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 neighborhood 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 neighborhood in the metropolitan area and only across neighborhoods. It can be thought of as a weighted average of the binary income segregation at each point in the income distribution, weighted under the assumption that segregation between those above and below the median is more important than segregation between those above and below the 90th percentile or above and below the 10th percentile. The rank-order index can be computed using income bin data as explained in Appendix A.

In the sections that follow, I discuss three theoretical shortcomings of the rank-order index. First, in Section 2.2.1 I consider the loss of interpretability due to the choice to use locally-ranked income. Second, I describe three different scenarios in which the index equals zero which must be distinguished from one another in Section 2.2.2. Third, since the rank-order index integrates the traditional information theory index over the range of p, it fails to capture geographic relationships between neighborhoods, an idea that is returned to in Section 2.3.

[1] Entropy can be thought of as the "randomness of a system," and is commonly used in physics and information theory as the amount of information needed to describe a discrete probability distribution (Roberto, 2015).

[2] Typically, these neighborhoods are represented as census tracts, the implications of which are discussed in Section 3.1.1.

2.2.1. Implications of using locally-ranked income

The rank-order index is seen as an improvement upon the now outdated *neighborhood sorting index* (NSI) mainly because it employs percentile ranks with respect to a local metropolitan income distribution rather than making use of absolute dollar amounts. Reardon et al. (2006) argue that this is a net benefit when measuring economic segregation due to the different costs of living across cities. However, this also creates problems with interpretability.

Say there exist two theoretical cities of 100 people each, where cost of living is constant throughout both cities. In City A, there are 90 people who make \$1 and 10 people who make \$100. In City B, there are 90 people who make \$1 and 10 people who make \$2.

In both cities, by the nearest-rank method, those who make \$1 would be considered to be in the 1st through 89th percentile. The 90th through 100th percentile ranks would be made up of those who make \$100 in City A and those who make \$2 in City B, since I utilize the *local* metropolitan income distribution.

These two cities are extremely different– City A is made up of exclusively extremely rich people and extremely poor people. There is a lot of economic diversity, and thus opportunity for poor families to interact with and share resources with wealthier families, depending on how segregated the city is. City B, on the other hand, is made up of entirely low-income people with very little economic diversity present throughout the city, so the poor community of this city does not have the same chance to engage with wealthier families. Despite these differences, if those who make \$1 are distributed the same way in both cities, and if those who make \$2 are distributed across City B in the same way as those who make \$100 in City A, then the computed rank-order index for these two cities would be equal.

Because the rank-order index is based on income percentile ranks rather than absolute dollar amounts, it stretches out a very narrow distribution like \$1 to \$2 as if it were the same as \$1 to \$100. Cities with very little economic diversity would get the same values of the rank-order index as cities with much more economic diversity.

The rank-order index fails to distinguish between a city that is very economically diverse and a city that is so segregated that it is largely only made up of one economic class. This is a major shortcoming of the measure, and leads to lower interpretability of the rank-order index.

Thus, economic segregation must be measured as a function of economic diversity. Otherwise, researchers miss the opportunity to evaluate cities as macro-segregated– cities that are either high- or low- income as a whole which do not exhibit much income diversity.

2.2.2. Conceptualizing "zero economic segregation"

It is theoretically unclear what it might mean to have "zero segregation." Nevertheless, the rank-order index can equal zero, and in many disparate situations.

When asked to picture a city with "zero segregation," one might immediately envision a diverse city where there is perfect integration of economic classes. When the rank-order index equals zero, this is one of the possible scenarios it could be describing, as this occurs when the distribution of income in each individual neighborhood reflects the distribution of income in the metropolitan area as a whole. However, other situations can also result in the rank-order index equaling zero in theory.

Picture City A in which everyone makes the same amount of money, and they are universally poor. That would have the rank order index equaling zero, but is it because there is no segregation, or because there is macro level segregation?

Then picture City B in which there are lots of different income levels and there are the same amount of people in each income bin in each tract throughout the city, where each tract is just as economically diverse as the city as a whole.

City A and City B would both have a rank-order index equaling zero. Looking at the Gini coefficient G_i of each city, $G_A = 0$ while $G_B \gg 0$. Upward mobility would likely be higher in City B than in City A because low-income people in City B are surrounded by more affluent people in their own neighborhoods, while in City A, low-income people are surrounded exclusively by other low-income people.

This is in stark contrast to what the outcomes of the measures might suggest theoretically: Segregation is viewed as bad, so having a low rank-order index would be viewed as good. Income inequality is also bad, so having a lower Gini coefficient would be viewed as good as well. Thus, City B, with rank-order index equaling zero and a higher Gini coefficient, would theoretically seem like a place with worse opportunity for poor households compared to City A.

With that said, it does not make sense to have a rank-order index of zero if everyone makes the same income. Then, researchers must ask: Theoretically, what does it mean to have economic segregation equal zero? Can this phenomenon of no segregation occur when there is no one present to be integrated with?

I argue that economic segregation cannot exist in the absence of economic inequality. If there is no economic inequality, then there is no variation in income, thus it is impossible for individuals to be geographically separated by income.

Further, there is always economic inequality if one increases their viewpoint to a larger geographic scale, so the absence of economic inequality truly reflects macro-segregation.

A zero-valued rank-order index captures at least three disparate situations: the entire city could be poor, the entire city could be rich, *or* the entire city could be truly integrated. This one metric describes conflicting situations that have different implications

for mobility. A better measure of economic segregation must not attempt to describe the phenomenon in a single number ranging from low to high– it just does not make sense in theory to do so.

2.3. Space matters: The checkerboard problem and reshuffling neighborhoods

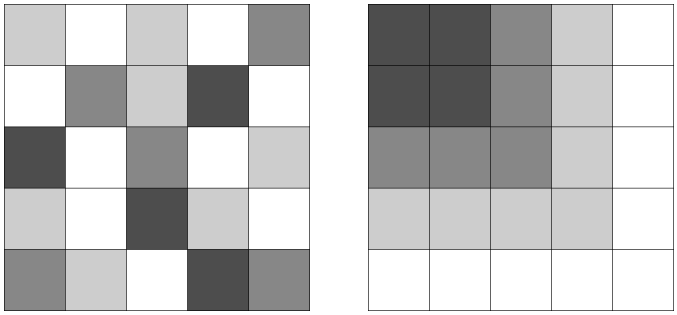
It is important to note that Reardon's rank-order index is aspatial– that is, it does not take into account the geography of neighborhoods in a metropolitan area. There is a spatial analog of the rank-order index, however it simply uses arbitrary definitions of what is considered "local" ranging from 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.[3]

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

Because the rank-order index does not incorporate spatial information but simply sums over neighborhoods, it does not account for the spatial relationships between neighborhoods in the metropolitan area. The problems with this approach can be 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 I have not 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.

[3]Even if this adaptation proved to be informative, 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, presenting researchers with significant barriers to utilizing the spatial analog instead of the original index.



(a) Integrated city (b) Segregated city

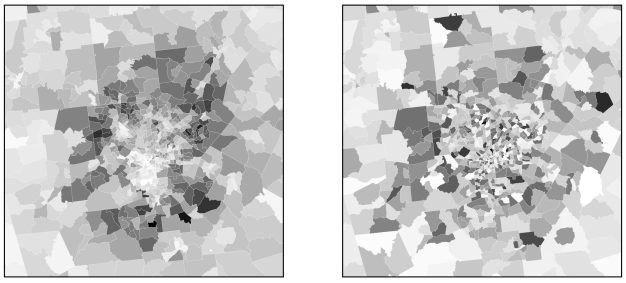
Figure 1: Checkerboard problem

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).

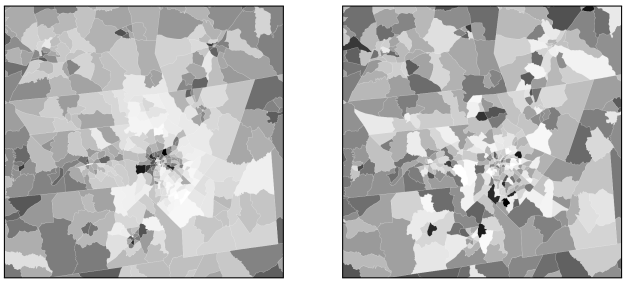
The importance of this oversight can be further illustrated by viewing "reshuffled" maps of cities in Figures 2-4. The original data contains counts of how many households are in each income bin of the 16 reported income categories on the 2000 Census, in every census tract within the United States' 276 designated Metropolitan Statistical Areas (MSAs). The maps display the proportion of households within a stated income bracket– darker areas show higher proportions of people within that income bracket, while lighter areas have smaller proportions of households 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. Here are some of the resulting maps.



(a) Original (b) Reshuffled

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

These images make the checkerboard problem come to life. As seen in Figure 2, the original image (a) 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 (b), 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 the 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 other aspatial measures of segregation, does 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 dimension 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 segregation measures, which are what Dwyer (2010) terms the five "spatial dimensions" of segregation: evenness, exposure, concentration, centralization, and clustering (Dwyer, 2010). These categories are defined in Table 1. While these measures do consider geography, they do so implicitly or indirectly.

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

Dimension	Definition
Evenness	Degree to which a group is spread in equal proportions 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

Dwyer (2010) assesses the spatial form of class segregation by looking at each of the spatial dimensions in conjunction with one another. Her analysis ultimately relies on combinations of aspatial measures of segregation, deciding based on theory what these different combinations mean.

The academic community is faced with a complex problem: researchers must step away

from aspatial measures of segregation, and yet, there does not exist an alternative that is shown to be "better."[4] The remainder of this paper attempts to capture some of the spatial information that may be important in describing the geography of economic segregation among US cities.

3. Data description and methodology

In this section, I describe the data and methodology. I first orient the reader with the data in Section 3.1. Then, I quantify economic segregation using two different approaches in Section 3.2. In Section 3.3, I describe my outcome of interest, absolute upward mobility. Finally, I lay out my main analysis plan in Section 3.4.

3.1. Getting oriented with the data

3.1.1. The geographic units of analysis: Census tracts and metropolitan areas

In this work, I utilize two different geographies: census tracts and Metropolitan Statistical Areas (MSAs). For the purposes of my research, census tracts can be thought of as proxies for neighborhoods while MSAs serve as proxies for cities.

Core based statistical areas (CBSAs) fall within the nation and regularly exist across multiple states (Rossiter, 2014). MSAs are one of two types of CBSAs, the other being Micropolitan Statistical Areas (μSAs). MSAs must be at or above the population threshold cutoff of 50,000 individuals, while μSAs range in population from 10,000 to 49,999. Figure 5 shows all the boundaries of the MSAs in the United States excluding Hawaii and Alaska so the reader can get a sense of the scale of MSAs.

[4] 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). While the spatial ordering index is a step in the right direction, it ultimately obscures income information and misrepresents the phenomenon of segregation by failing to account for economic diversity within census tracts.

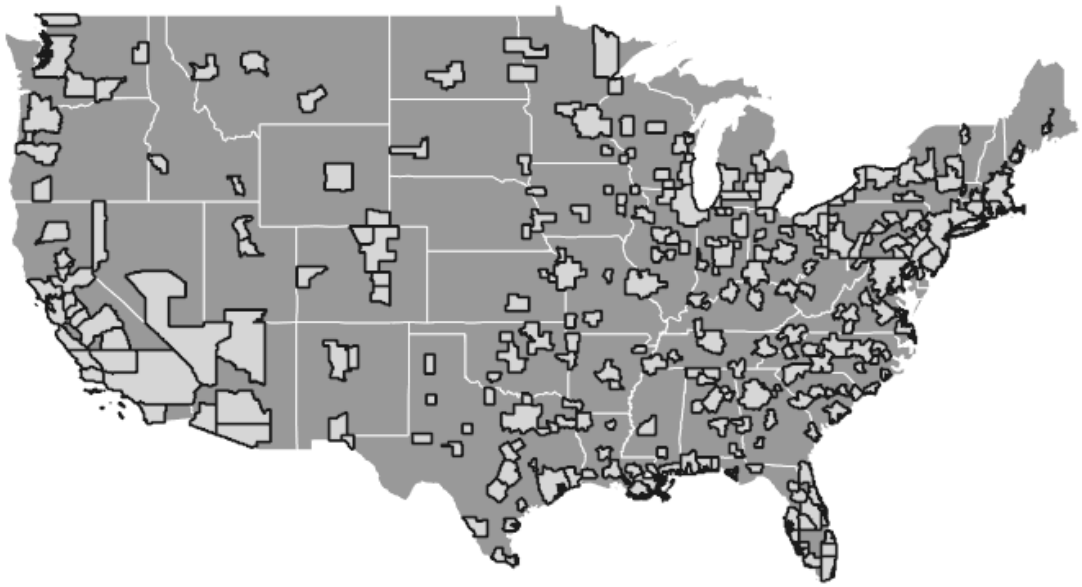


Figure 5: MSA boundaries (2000)

See Table 2 for summary statistics on MSAs. The total population of an MSA ranges from anywhere between around 58,000 people to around 21 million people, and the area in square miles also ranges a lot. Table 2 illuminates the great amount of variation in MSAs in terms of their size in both population and land area, their levels of inequality, their wealth, and their economic mobility.

Table 2: MSA Summary Statistics

	Mean	Standard Deviation	Minimum	Maximum
Total population	818774.2	1968621.3	57813	21199865
Area (square miles)	2607.21	3803.15	257.55	39719.04
Gini coefficient	0.444	0.023	0.385	0.528
Median household income	39331.45	5920.87	24863	62024
Absolute upward mobility	41.33	3.54	33.73	52.77

MSAs are drawn so that census tracts are nested within them. A census tract is a relatively permanent area nested within a county that normally follows visible features, but may follow governmental unit boundaries and other invisible features. Census tracts are designated to have roughly the same population, averaging about 4,000 inhabitants but in reality varying between 2,500 and 8,000 people. They are confined to particular counties, and they do not necessarily coincide within any other geography designated by the Census Bureau (Rossiter, 2014).

To get a sense of the size of a census tract, look to Figure 6 which visualizes the bound-

aries of census tracts in white within the New York-Northern New Jersey-Long Island, NY-NJ-CT-PA MSA. MSA boundaries are shown in black.

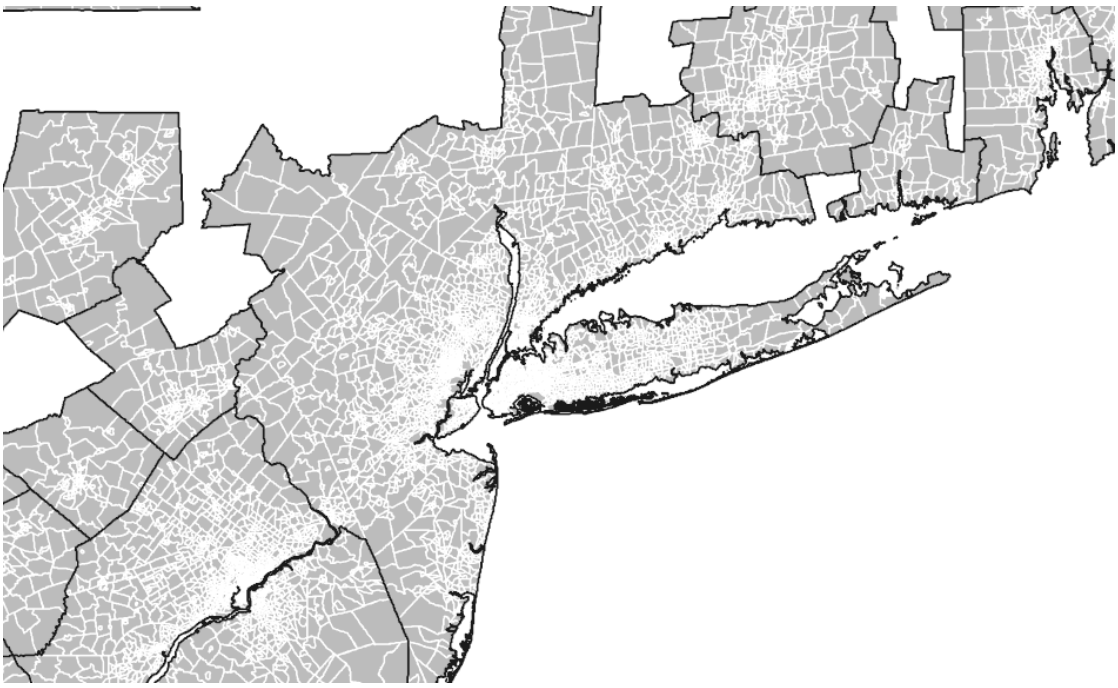


Figure 6: New York-Northern New Jersey-Long Island, NY-NJ-CT-PA MSA

Note that other neighboring MSAs are also displayed in the image but the MSA that includes New York City is in the center of the image, spanning all of Long Island, Northern New Jersey, and into the Hudson Valley. This encompasses quite a large area, although some of the neighboring MSAs in the image are much smaller. Also take note of the varying sizes of census tracts– the census tracts north of New York City are large and visible on the map, but the census tracts in Manhattan and Long Island are so small and so close together that none of them can actually be seen. Look to Figure 7 for a closer look at Manhattan, Brooklyn, and the surrounding area.

Figure 7: New York City, NY

In a high-density place like Manhattan, one can consider a census tract to be around a block, while in a lower-density place like in the New Jersey suburbs, a census tract represents a much larger area.

3.1.2. Working within the confines of publicly-available data

There are two considerations to be made concerning the restrictions imposed by publicly-available data. First is the question of which government-imposed boundary most closely corresponds to the idea of a "neighborhood," and whether data on that entity is widely available. Second is the geographic level of granularity available on households.

Because of census tracts' small size in high-density areas, many fields of research utilize census tracts as proxies for neighborhoods, although there is a growing body of literature that shows this is not the wisest assumption to make. School districts may be a better approximation. School districts control school financing and enrollment. Designated in the Standard Hierarchy of Census Geographic Entities, school districts are boundaries that fall within each state which may or may not cross county boundaries (Rossiter, 2014). Census tracts may not be drawn to encompass or lie within school districts– rather, they could be intersecting so that a school district is divided into multiple census tracts, or a census tract is divided into multiple school districts.

If school financing is a mechanism through which economic mobility is connected to economic segregation, then the boundaries of school districts may prove more meaningful than the boundaries of census tracts. However, data on school districts is not typically available and thus researchers are limited by public data availability. In this work, I resort to census tracts, the next best geography made widely available.

Publicly-available data on census tracts is aggregated so that there is no geographic information on a level smaller than the census tract. Researchers do not know what is happening within a census tract, as it is the finest level of geographic detail with readily available and accessible public data. Because segregation must be measured spatially, it is then not possible to identify segregation below this level. As a consequence, researchers must shift their focus to measuring relative income diversity within census tracts and segregation between census tracts.

3.1.3. Measuring economic segregation: Household income bins

Now that I have chosen the geographic level of analysis, MSAs and the census tracts in and around them, I next describe the data that is used on that level. I utilize data on population by household income bin by census tract for all census tracts in the United States, including all census tracts within the 276 MSAs designated by the 2000 Census.[5] I source these data from the National Historical Geographic Information System (NHGIS). Household 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

For every census tract, my main dataset contains the count how many households are in each income bin in the year 2000.

[5] I describe in Section 3.3 my reason for choosing the year 2000.

3.2. Quantifying economic segregation

3.2.1. Maps of greater metropolitan areas

Using the data described in Section 3.1.3, I create income density maps of greater metropolitan areas in Python using the GeoPandas and Matplotlib libraries. For all of a greater metropolitan area's census tracts, these maps visualize the proportion of households within a certain income bin, computed as the number of households in that income bin in the census tract divided by the number of total households within the census tract. For census tract c, the proportion p_{ci} of households in income bin i is computed as follows.

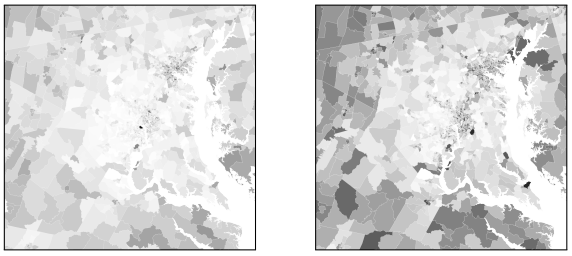
$$p_{ci} = \frac{HH_{ci}}{\sum_{i=1}^{16} HH_{ci}}. \tag{4}$$

The maps are grayscale, with black representing areas with 100% of households in income bin i, while white represents areas with either 0% of households in income bin i or no data.[6]

For example, the lowest income bin reported by the Census is for those households who make less than $10,000. Thus a single map for a given greater metropolitan area will show areas that have 100% of income earners making less than $10,000 as black while areas that have 0% of income earners making less than $10,000 as white.

Figure 8 previews some of these maps, specifically four different income bins for the Washington-Baltimore, DC-MD-VA-WV MSA. One can see the apparent clustering of different income groups across the MSA that conforms to the monocentric city model discussed earlier.

[6]The approach of coloring the missing values a distinct color from 0 results in the k-means clustering algorithm picking up on the differences between maps in where there is missing data, resulting in clusters based on where the coastline is, which is not what I wanted to achieve. Thus, I color missing data the same as 0 although there are obvious tradeoffs and consequences of that decision.



(a) $15,000-$19,999

(b) $25,000-$29,999

(c) $50,000-$59,999

(d) $100,000-$124,999

Figure 8: Washington-Baltimore, DC-MD-VA-WV MSA

Notice that all of the maps in Figure 8 are squares. This is because I intentionally construct the maps to be the same pixel size. I do this by first drawing a bounding box around an MSA. If the rectangle is not a square, then I make the shorter side longer so that it is a square with the center of the MSA in the center of the square. The data pictured in the map is of those tracts within the MSA as well as the tracts in the surrounding area that is captured by this square, which is what I refer to as the "greater metropolitan area." I save these maps as .PNG images of size 224×224 pixels. Note that there are 16 income brackets, and 276 MSAs, thus 16×276 total maps.

I take these maps and cluster them into visually similar groups, as explained in Section 3.2.2. The use of maps allows me to capture spatial relationships in a new way not seen before in other methods to measure economic segregation. It allows the densities of different income classes to be visualized throughout cities, and somewhat obscures the boundaries explicitly posed to be a problem in discussion of the MAUP (see Appendix B).

3.2.2. Variable creation via k-means clustering

K-means clustering is a non-hierarchical clustering algorithm (Baumer et al., 2017). It is commonly used because of its simplicity, operating on many types of data. The algorithm groups data that are similar to each other into clusters, and then locates the clusters within the data. It iteratively decides which cluster to place the next observation

into. The algorithm's objective is to find k centroids, or means, one for each cluster of observations, that are placed as far away as possible from one another. When an observation 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).

I take each set of 276 maps corresponding to a given income bin i and unroll them into vectors of length $224 \times 224 \times 3$. I then cluster them using the k-means clustering algorithm implementation from the scikit-learn library in Python.

To choose k, the number of clusters, I use the elbow method, which is the oldest method for determining the number of clusters. I calculate values of the cost function testing $k = 2$ to $k = 15$ clusters, and plot the number of clusters k versus the cost. There may be some value k for which the cost drops dramatically, and after that it plateaus upon increasing k further. An "elbow" is reached, after which the cost function decreases very slowly (Kodinariya and Makwana, 2013). I find k values ranging from 5 to 12 as the optimal number of clusters for a given income bin i.

I convert the results of the clustering process into indicator variables that equal 1 for an MSA whose corresponding map is in income bin i cluster j, and 0 for MSAs not in cluster j. The goal of this clustering process is to identify which maps are most visually similar to one another. The indicator variables are akin to fixed effects, in that they capture variation in the data that remains constant across the set of maps within a cluster. These indicator variables are the bases for the rest of my analysis.

3.2.3. Computing the rank-order index twice

I calculate the rank-order index using the RANKSEG Stata module and the income bin data by census tract for each of the 276 MSAs. The distribution of the rank-order index by MSA looks bimodal as is seen in Figure 9.

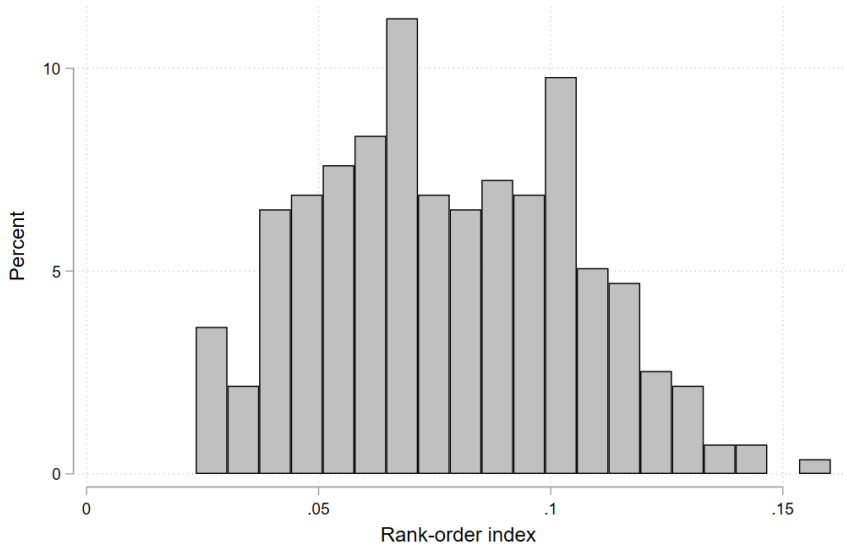


Figure 9: The distribution of the rank-order index across MSAs

Because the maps include census tracts not included in the actual bounds of an MSA, I also calculate the rank-order index over the greater metropolitan area, the same set of census tracts for each MSA. I do this by taking the proportion of land area of a census tract that is within the greater metropolitan area and multiplying the populations of individuals in each income bin by this proportion before including those tracts in my analysis. The distribution of the rank-order index by greater metropolitan area can be seen in Figure 10, and does not appear to be much different than Figure 9.

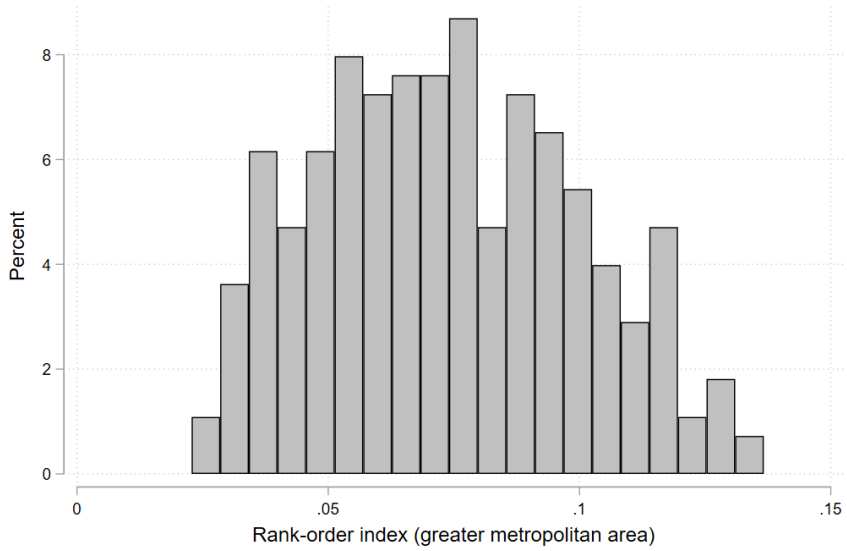


Figure 10: The distribution of the rank-order index across greater metropolitan areas

As a check, Figure 11 displays the relationship between rank-order index by MSA and rank-order index by greater metropolitan area. This relationship appears to be linear.

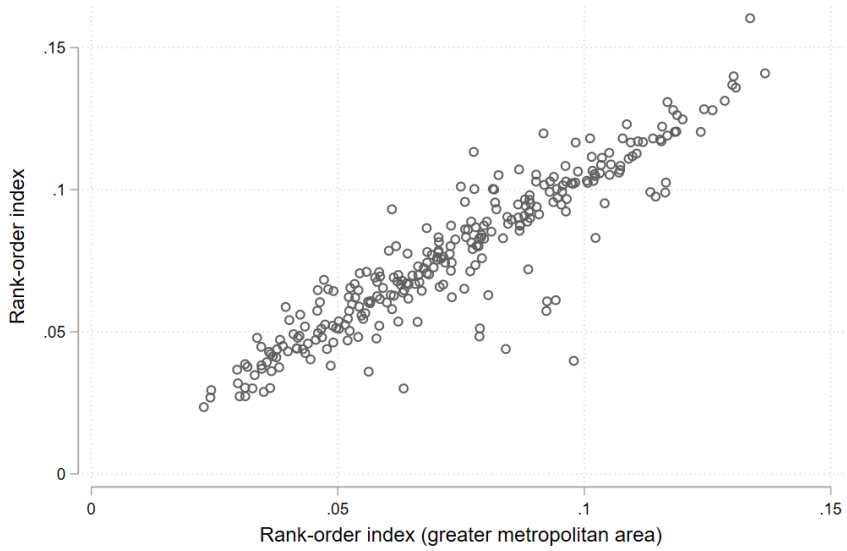


Figure 11: The relationship between rank-order index at different geographies

Due to the very linear relationship between the two variables and the geographic align-

ment between the maps and the rank-order index on the greater metropolitan area, I solely utilize the rank-order index on the greater metropolitan area in my analysis.

3.3. Measuring opportunity: Absolute upward mobility

The measure used for economic mobility throughout this paper is termed by Chetty et al. (2014) as *absolute upward mobility*. This measure is defined as the mean percentile rank in the national child income distribution of children whose parents are in the 25th percentile of the national parent income distribution, with possible values ranging from 1 to 100. I source data on an MSA level of this measure of absolute upward mobility from Chetty et al. (2014). In this sample, the values of absolute upward mobility range from 33.73 to 52.78, the distribution of which can be seen in Figure 12.

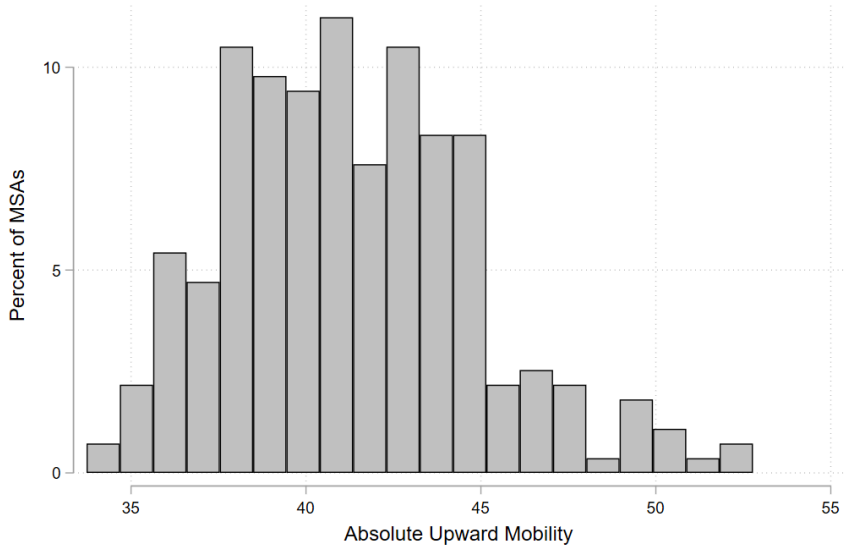


Figure 12: The distribution of absolute upward mobility across MSAs

Chetty et al. (2014) predict absolute upward mobility 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).[7]

[7]Returning to the discussion from Section 3.1.3, I choose to use 2000 Census data because the 2000 Census is the closest decennial census to the time that parents' household incomes were measured (1996-1998).

It is important to measure economic mobility absolutely rather than relatively, because the latter 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 absolute upward mobility is reported by 2013 MSA and not by 2000 MSA, I must link 2013 MSAs to 2000 MSAs. 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 that links 2000 MSAs to 2013 MSAs.

The 2000 MSA is the geographic entity of my master file. It is possible that portions of multiple 2013 MSAs fall inside one 2000 MSA. Thus, For each 2000 MSA, I calculate the proportion of the area of each 2013 MSA that intersects with the 2000 MSA:

$$p_j = \frac{MSA2013_j \cap MSA2000_i}{MSA2000_i}.$$

I then calculate the absolute upward mobility of each 2000 MSA by taking the sum of the absolute upward mobility of the overlapping 2013 MSAs, weighting by the proportion of area of each 2013 MSA that is within the 2000 MSA, as follows:

$$Mobility(MSA2000_i) = p_1 Mobility(MSA2013_1) + p_2 Mobility(MSA2013_2) + ...$$

To summarize, my final dataset consists of segregation cluster indicator variables, absolute upward mobility, and rank-order index computed over the MSA and over the greater metropolitan area. I also include the Gini coefficient and median household income. All of these variables exist in the 2000 time period, except absolute upward mobility which was measured using a cohort of people who were in their late teens in the year 2000. Each observation in the final dataset can be thought to represent an MSA.

3.4. Main analysis plan

Because the process described in Section 3.2.2 generates so many cluster indicator variables, I must find ways to tackle the curse of dimensionality. In order to efficiently assess the spatial information contained in these cluster indicator variables, my main analysis follows two paths. First, I utilize a regression analysis method called least absolute shrinkage and selection operator (lasso), which performs variable selection according to what is most important in predicting the outcome variable (in this case, absolute upward mobility). Second, I run a multiple correspondence analysis (MCA) on the cluster indicator variables before analyzing their relationship with absolute upward mobility in order to reduce the dimensionality while minimizing information lost. MCA produces ten continuous variables that capture a portion of the variation from the original cluster indicator variables. I then assess these new continuous variables as predictors of absolute upward mobility. In the process, I sometimes include various controls and I compare the

results of both of these methodologies to regressions of the rank-order index on absolute upward mobility.

Lasso is the same as ordinary least squares (OLS) when its tuning parameter λ equals 0. However, when $\lambda > 0$, the lasso coefficients minimize the residual sum of squares plus a penalty shrinkage term which is a function of λ and the magnitudes of the lasso coefficients. This causes the lasso method to perform variable selection, yielding sparse models (James et al., 2017). I utilize cross-validation techniques to choose the λ for which the cross-validation error is smallest, and then refit the lasso model using that λ value to see which variables are kept in the model. This set of variables are the focus of my analysis. The mechanisms behind the lasso shrinkage term cause the estimated coefficients to be biased towards zero and inconsistent. Thus, a common approach to reduce this bias is to run lasso to identify the set of nonzero coefficients, and then fit an unrestricted linear model to the selected set of features, what I refer to as a "lasso-OLS regression" (Hastie et al., 2009).

MCA is an extension of correspondence analysis (CA) in which one can analyze the pattern of relationships of categorical variables. MCA is also a generalization of principal component analysis when the variables are categorical, not quantitative (Valentin and Abdi, 2007). I am left with ten "dimensions" of spatial segregation that I then evaluate as predictors of absolute upward mobility. The idea behind these dimensions relates back to Table 1 in that it is possible that certain dimensions resulting from the MCA would mirror phenomena such as evenness, exposure, concentration, centralization, and clustering, while the rank-order index only successfully describes evenness.

The results of lasso and MCA are used to predict absolute upward mobility. I compare their predictive power with that of the rank-order index by juxtaposing the corresponding models' adjusted R^2 values.

It is important to consider whether the variation captured by the clusters is the same or different from variation captured by the rank-order index. I argue that the rank-order index captures economic homogeneity while the segregation clusters capture economic segregation. Thus, while related, they should account for different aspects of what might affect absolute upward mobility. If I included them both in the same regression, I would expect to see the adjusted R^2 considerably increase with respect to other regressions. Thus, I include them both in the same regression.

Since the rank-order index uses percentiles based on the local income distribution rather than income, it is thus theoretically independent from income inequality. However, my maps are not independent from income inequality. Therefore, I add the Gini coefficient, a commonly used measure of income inequality, to my regressions to ensure that my results are not driven by the accounting of income inequality. Inclusion of median household income as a predictor can help differentiate a completely high-income city from a completely low-income city, as discussed in Section 2.2.2. Thus, I step in controls for the Gini coefficient and median household income.

4. Results

As discussed in Section 3.4, my analysis plan follows two paths: model selection via the lasso method, and dimensionality reduction via MCA. These two tracks of my analysis are done in tandem and compared with the results of the regression involving the rank-order index. I show that when using both of these procedures, the adjusted R^2 values are consistently much higher than that of the model using rank-order index.

First, I evaluate the relationship between Reardon's rank-order index, calculated over the greater metropolitan area, and absolute upward mobility and find a statistically significant result, as seen in Table 3 Specification 1. For a one standard deviation or 0.026 unit increase in the rank-order index, there is an associated 0.615 percentile rank decrease in the expected absolute upward mobility. Thus, more "segregated" MSAs that score higher rank-order index values tend to be associated with lower economic mobility. This model of rank-order index alone on absolute upward mobility has an adjusted R^2 of 0.0266.

Once accounting for the Gini coefficient, the significance of the rank-order index goes away in Specifications 2 and 3 of Table 3. As seen in Specification 3 of Table 3, the inclusion of median household income makes little difference to the magnitude or significance of the results.

Table 3: Relationship between rank-order index and absolute upward mobility

	Absolute upward mobility		
	(1)	(2)	(3)
Rank-order index	-23.61**	-4.629	-12.74
(greater metropolitan area)	(8.088)	(8.191)	(10.20)
Gini coefficient		-57.33***	-50.15***
		(9.293)	(10.73)
Median household income			0.0000572
			(0.0000430)
Constant	43.07***	67.12***	62.29***
	(0.633)	(3.944)	(5.359)
Observations	276	276	276
R^2	0.0302	0.1488	0.1543
Adjusted R^2	0.0266	0.1426	0.1450
Degrees of freedom (Model)	1	2	3
Degrees of freedom (Residual)	274	273	272

Standard errors in parentheses

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

In Table 4, I evaluate the regression of absolute upward mobility on the nineteen "lasso-OLS" cluster indicator variables selected by lasso regression, which results in a model with an adjusted R^2 of 0.2510, significantly higher than the adjusted R^2 of the model with rank-order index.

With the inclusion of the Gini coefficient and median household income in Specifications 2 and 3 in Table 4, many of my cluster indicator variable coefficients remain statistically significant, indicating that they are picking up on variation that is distinct from what is captured by the Gini coefficient and median household income. Table 4 Specification 3, with both the Gini coefficient and median household income included, has an adjusted R^2 value of 0.3067, significantly higher than the adjusted R^2 of Specification 3 in Table 3 with an adjusted R^2 value of 0.1450.

Any of the cluster indicator variable coefficients seen in Table 4 can be interpreted similarly. I interpret Cluster 2 of the \$30,000 to \$34,999 income bin in Specification 1 as an example: Keeping all other income bin clusters constant, there is an average 2.022 percentile rank increase in the expected absolute upward mobility of a poor individual who grows up in an MSA that is in Cluster 2 of the \$30,000 to \$34,999 income bin compared to if that individual's MSA had been in a different cluster of that income bin. In short, membership to Cluster 2 of the \$30,000 to \$34,999 income bin is associated with higher absolute upward mobility. An analysis of the statistically significant coefficients in Specification 3 of Table 4 can be found in Section 4.1.

Table 4: Relationship between lasso-OLS clusters and absolute upward mobility

	Absolute upward mobility		
	(1)	(2)	(3)
\$10,000 to \$14,999 (Cluster 3)	2.318*	2.226*	2.253*
	(0.978)	(0.942)	(0.942)
\$20,000 to \$24,999 (Cluster 2)	-1.718***	-1.348**	-1.239**
	(0.459)	(0.449)	(0.457)
\$25,000 to \$29,999 (Cluster 4)	-2.160	-1.702	-1.578
	(1.133)	(1.096)	(1.099)
\$25,000 to \$29,999 (Cluster 7)	-1.723*	-1.333	-1.305
	(0.829)	(0.803)	(0.802)
\$30,000 to \$34,999 (Cluster 2)	2.022*	1.963*	2.009*
	(0.925)	(0.891)	(0.891)
\$30,000 to \$34,999 (Cluster 3)	-1.236*	-1.243*	-1.241*
	(0.543)	(0.523)	(0.523)
\$35,000 to \$39,999 (Cluster 5)	1.199*	1.122*	0.979
	(0.539)	(0.520)	(0.532)
\$35,000 to \$39,999 (Cluster 8)	-0.951	-0.592	-0.847
	(0.671)	(0.651)	(0.682)
\$45,000 to \$49,999 (Cluster 6)	0.962	0.496	0.574

Table 4: Relationship between lasso-OLS clusters and absolute upward mobility

	Absolute upward mobility		
	(1)	(2)	(3)
	(0.656)	(0.640)	(0.643)
$45,000 to $49,999 (Cluster 8)	-0.525	-0.858	-0.919
	(0.513)	(0.499)	(0.501)
$60,000 to $74,999 (Cluster 4)	1.423*	1.247*	1.228*
	(0.643)	(0.620)	(0.620)
$75,000 to $99,999 (Cluster 3)	-1.774*	-1.233	-1.285
	(0.793)	(0.773)	(0.773)
$100,000 to $124,999 (Cluster 2)	-1.302*	-1.119	-1.111
	(0.593)	(0.573)	(0.572)
$100,000 to $124,999 (Cluster 7)	-0.460	-0.219	-0.201
	(0.461)	(0.447)	(0.447)
$125,000 to $149,999 (Cluster 3)	-1.385*	-1.216*	-1.287*
	(0.573)	(0.553)	(0.555)
$150,000 to $199,999 (Cluster 9)	0.591	0.352	0.370
	(0.508)	(0.492)	(0.492)
$200,000 or more (Cluster 3)	2.052*	2.113*	2.095*
	(0.854)	(0.823)	(0.822)
$200,000 or more (Cluster 7)	1.445*	1.234	1.231
	(0.657)	(0.634)	(0.634)
$200,000 or more (Cluster 11)	0.963	0.983	1.020
	(0.542)	(0.522)	(0.523)
Gini coefficient		-39.29***	-36.31***
		(8.580)	(8.896)
Median household income			0.0000435
			(0.0000348)
Constant	41.69***	59.03***	56.01***
	(0.422)	(3.808)	(4.505)
Observations	276	276	276
R^2	0.3027	0.3557	0.3596
Adjusted R^2	0.2510	0.3052	0.3067
Degrees of freedom (Model)	19	20	21
Degrees of freedom (Residual)	256	255	254

Standard errors in parentheses

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

Then, I evaluate the regression of absolute upward mobility on the ten dimensions produced by MCA. While these continuous variables do not have the benefit of direct interpretability, they share the same message as the lasso-OLS approach in that much

more of the variation in absolute upward mobility is captured by using this information compared with the rank-order index. Specification 3 of Table 5 has an adjusted R^2 value of 0.2206, still considerably higher than that of the rank-order index (0.1450 in Specification 3 of Table 3).

Table 5: Relationship between ten dimensions and absolute upward mobility

	Absolute upward mobility		
	(1)	(2)	(3)
Dimension 1	0.461*	0.233	0.338
	(0.201)	(0.192)	(0.218)
Dimension 2	-0.881***	-0.629**	-0.540*
	(0.201)	(0.193)	(0.212)
Dimension 3	0.0309	-0.0634	-0.0138
	(0.201)	(0.189)	(0.195)
Dimension 4	0.365	0.607**	0.672**
	(0.201)	(0.192)	(0.203)
Dimension 5	0.179	-0.0473	-0.0562
	(0.201)	(0.192)	(0.192)
Dimension 6	0.229	0.0733	0.0655
	(0.201)	(0.190)	(0.190)
Dimension 7	0.141	0.144	0.151
	(0.201)	(0.188)	(0.188)
Dimension 8	-0.208	-0.242	-0.219
	(0.201)	(0.188)	(0.190)
Dimension 9	-0.276	-0.258	-0.255
	(0.201)	(0.188)	(0.188)
Dimension 10	-0.699***	-0.576**	-0.577**
	(0.201)	(0.189)	(0.189)
Gini coefficient		-54.92***	-52.48***
		(9.034)	(9.346)
Median household income			0.0000440
			(0.0000431)
Constant	41.33***	65.72***	62.90***
	(0.201)	(4.016)	(4.872)
Observations	276	276	276
R^2	0.1469	0.2517	0.2547
Adjusted R^2	0.1147	0.2205	0.2206
Degrees of freedom (Model)	10	11	12
Degrees of freedom (Residual)	265	264	263

Standard errors in parentheses

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

Side-by-side in Table 6, it is easy to see that there is a considerably higher adjusted R^2 on the model that includes rank-order index and the lasso-OLS clusters compared to that which includes rank-order index alone. Note that all three specifications in Table 6 include the Gini coefficient and median household income as controls. In all three specifications, the rank-order index is not statistically significant.

Table 6: Main results

	Absolute upward mobility		
	(1)	(2)	(3)
Rank-order index	-12.74	-4.326	-2.780
(greater metropolitan area)	(10.20)	(9.680)	(10.58)
Gini coefficient	-50.15***	-34.04**	-51.04***
	(10.73)	(10.25)	(10.84)
Median household income	0.0000572	0.0000530	0.0000486
	(0.0000430)	(0.0000409)	(0.0000466)
Constant	62.29***	54.96***	62.29***
	(5.359)	(5.092)	(5.410)
Lasso-OLS Clusters	No	Yes	No
Ten Dimensions	No	No	Yes
Observations	276	276	276
R^2	0.1543	0.3602	0.2548
Adjusted R^2	0.1450	0.3045	0.2179
Degrees of freedom (Model)	3	22	13
Degrees of freedom (Residual)	272	253	262

Standard errors in parentheses

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

As a final check, I illuminate non-parametric relationships between the rank-order index and absolute upward mobility, allowing for a more flexible model of rank-order index on absolute upward mobility. This check makes sure that the difference in explanatory power is not due to a lack of flexibility in the model of rank-order index on absolute upward mobility. For brevity, this is explored in Appendix C.

4.1. A visual analysis of the segregation clusters

I construct averaged maps of each cluster by taking the average of each pixel value in every map in the cluster. For example, the averaged pixel value in position $(1, 1)$ is equal to the average of all the pixel values in position $(1, 1)$ in all of the maps within

that cluster. This process produces an averaged, somewhat representative visual of the maps within a given cluster.

As seen in Specifications 1 through 3 of Table 4, membership to certain clusters results in higher or lower average mobility for poor individuals. To connect these results back to theories of urban land use, I analyze averaged maps corresponding to the seven clusters that have statistically significant indicator variables in all three specifications, while interpreting the coefficient estimates given in Specification 3.

MSAs that are part of Cluster 3 of the $10,000 to $14,999 income bin have higher mobility than the reference group, MSAs that are not part of this cluster. On average, membership to this cluster is associated with a 2.253 percentile rank increase in absolute upward mobility, statistically significant at the 5% level.

As seen in Figure 13, the map visualizes cities where there are high proportions of low-income people on the outskirts of the city, followed by a lower proportion of low-income people in the center of the city.

This cluster includes cities like Baton Rouge, LA and Santa Fe, NM.

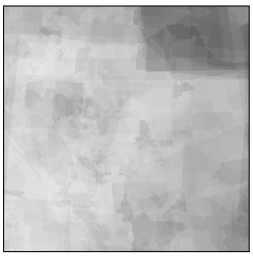


Figure 13: $10,000 to $14,999 (Cluster 3)

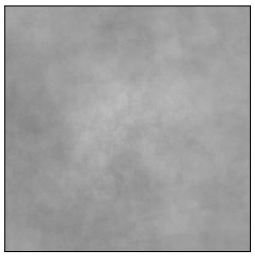


Figure 14: $20,000 to $24,999 (Cluster 2)

Membership to Cluster 2 of the $20,000 to $24,999 income bin is associated with a 1.239 percentile rank decrease in absolute upward mobility, statistically significant at the 1% level. In other words, MSAs in this cluster experience lower levels of economic mobility than MSAs outside of this cluster.

Pictured in Figure 14, this cluster seems to be made up of cities that have a significantly high proportion of people in the $20,000 to $24,999 income bin. Thus, it would make sense that with low-income people spread almost uniformly throughout the city, there would be lower economic mobility for poor children.

This cluster includes cities such as Memphis, TN-AR-MS, Charlottesville, VA, and Birmingham, AL.

The averaged map for Cluster 2 of income bin $30,000 to $34,999 appears to be dark, though it is not uniform– there are patches of lighter and darker areas in the center of the city. This income range is interesting as the purchasing power of $30,000 varies greatly from one state to the next.

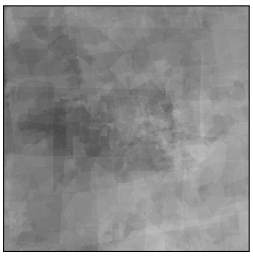


Figure 15: $30,000 to $34,999 (Cluster 2)

Membership to this cluster is nevertheless associated with increased economic mobility. Absolute upward mobility is, on average, 2.009 percentile ranks higher in MSAs that are part of this cluster compared with MSAs that are not, statistically significant at the 5% level. Some of the MSAs included in Cluster 2 are Albany-Schenectady-Troy, NY, Allentown-Bethlehem-Easton, PA, and Tallahassee, FL.

Figure 16: $30,000 to $34,999 (Cluster 3)

Compared with Cluster 2 pictured in Figure 15, the portrayal of Cluster 3 in Figure 16 is much lighter, signaling a smaller proportion of people in the $30,000 to $34,999 income range throughout the cities included in this cluster. Membership to this cluster associated with a 1.241 percentile rank decrease in absolute upward mobility, statistically significant at the 5% level.

Without more information about other income bins, it is hard to know why this may be the case. It could be because there is an absence of middle-income people in these cities, reflective of the hollowing out of the middle class, that contributes to lower economic mobility for poor individuals.

This cluster includes many larger cities such as Raleigh-Durham-Chapel Hill, NC, Miami-Fort Lauderdale, FL, Washington-Baltimore, DC-MD-VA-WV, and Boston-Worcester-Lawrence, MA-NH-ME-CT.

Figure 17 is very dark throughout, portraying cities where a very high proportion of households are in the $60,000 to $74,999 income range. This cluster is associated with a 1.228 percentile rank increase in absolute upward mobility, statistically significant on the 5% level.

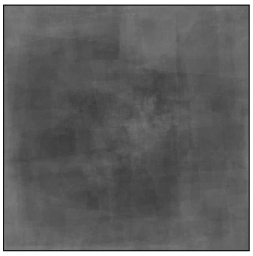


Figure 17: $60,000 to $74,999 (Cluster 4)

The income bracket represented is relatively wealthy, so it makes sense then that a higher proportion of wealthy people contributes to higher economic mobility for poorer people.

Some of the MSAs included in this cluster are Bloomington, IN, Lancaster, PA, and Iowa City, IA.

Figure 18: $125,000 to $149,999 (Cluster 3)

Cluster 3 of the $125,000 to $149,999 income bin looks like it is made up of MSAs that have wealth concentrated in the inner suburban rings, as pictured in Figure 18. Membership to this cluster is associated with a 1.287 percentile rank decrease in absolute upward mobility, statistically significant at the 5% level.

This cluster includes MSAs such as Providence-Fall River-Warwick, RI-MA, Santa Fe, NM, Gainesville, FL, Chicago-Gary-Kenosha, IL-IN-WI, San Antonio, TX, and Boston-Worcester-Lawrence, MA-NH-ME-CT.

Cluster 3 of the $200,000 or more income bin appears visually very similar to Cluster 3 of the $125,000 to $149,999 income bin, in that there is concentration of wealth in a very small area on the map. However, my results suggest a very different story: Rather than lower economic mobility, membership to this cluster is associated with a 2.095 percentile rank increase, statistically significant at the 5% level.

Figure 19: $200,000 or more (Cluster 3)

Surprisingly, the cities that make up this cluster are not the US mega-cities one might expect when they first think of places with the most concentrated wealth. Rather, other vibrant metropolitan areas such as Miami-Fort Lauderdale, FL, and Fayetteville-Springdale-Rogers, AR, the headquarters of Walmart and Tyson Foods, make up this cluster.

Cluster 3 of the $125,000 to $149,999 bin and Cluster 3 of the $200,000 or more bin present an interesting comparison, worthy of further thought. They both closely resemble the Kohl model of the concentric city. Yet, they have opposite outcomes for economic mobility. One possible explanation may be that the cities represented in the latter are cities that have one or a few dominant industries wherein people seek jobs in that town, compared with cities in the former which align more with the idea of the creative class and highly skilled labor in large ways. Whatever the reason, this certainly presents a need for further research regarding the effects of concentrated wealth in cities, regardless of their mega-city status.

5. 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 captures 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.

Table 6 showcases my most concise result: The Gini coefficient, median household income, and the rank-order index combined explain 15.43% of the variation in absolute upward mobility, with an adjusted R^2 of 0.1450, as seen in Specification 1. With the addition of a subset of the segregation cluster indicator variables in Specification 2, this number jumps to 36.02% of variation in absolute upward mobility explained, with an adjusted R^2 of 0.3045. Specification 3 of Table 4 shows that even without rank-order index, the segregation cluster indicators combined with the Gini coefficient and median household income explain 35.96% of the variation in absolute upward mobility, with an adjusted R^2 of 0.3067, negligibly different from the previously discussed result.

Much more of the variation in economic mobility can be explained by economic segregation compared to what researchers have found in the past. Without the proper tools to measure economic segregation, it becomes impossible to examine its profound impacts. My results suggest that policymakers have been overlooking the true impacts of segregation by underestimating the relationship between economic segregation and outcomes that matter to our lives such as economic mobility. This conclusion is important for both research and policy oriented towards improving mobility for poor families.

6. 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 explained by the rank-order index. This explanatory power is not attributed to the number of

variables in the regression, as I have addressed high-dimensionality problems through my use of lasso and MCA.

There are some 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." This very much connects to the idea of dimensions of segregation cited by Dwyer (2010). In building on my analysis, future work could make the ten dimensions generated by MCA more interpretable by connecting them with the dimensions presented in Table 1. This could be done by exploring correlations between measures of economic segregation that capture certain "dimensions" and the actual dimensions generated via the MCA output.

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. 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 it is important.

Segregation cannot 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 various forms as theorized by urban studies and economics scholars alike for centuries. My results suggest that these structures 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.

Quantitative social scientists 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 measuring economic homogeneity. This paper can serve as a call to action for researchers to work towards an explicitly spatial, multidimensional approach to capturing 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. *Journal of Planning History 6*(3), 248–271.

Chetty, R. and N. Hendren (2016). The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates. *The Quarterly Journal of Economics 133*(3), 1163–1228.

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), 1553–1623.

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 (1945). The Nature of Cities. *The Annals of the American Academy of Political and Social Science 242*, 7–17.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2 ed.). Verlag: Springer.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2017). *An Introduction to Statistical Learning* (7 ed.). New York: Springer.

Johnston, R., M. Poulsen, and J. Forrest (2014). Segregation matters, measurement matters. In *Social-spatial segregation: Concepts, processes and outcomes*, pp. 13–44. The Policy Press Bristol, UK.

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

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., 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.

Rossiter, K. (2014). Understanding Geographic Relationships: Counties, Places, Tracts and More.

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.

Appendices

A. Estimating the rank-order index using binned income data

It is important to consider what data is necessary to compute the index. In this case, I 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 $E(p)$:

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

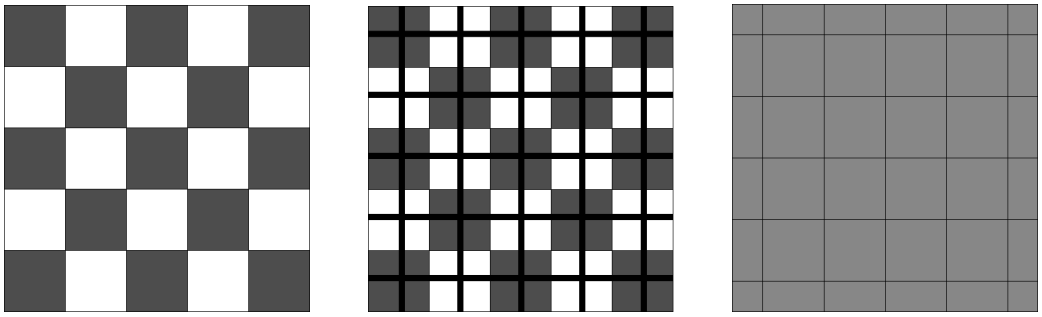
If $\hat{\beta_k}$ is the kth coefficient from this model, then

$$\hat{H^R} = \hat{\beta_0} + \frac{1}{2}\hat{\beta_1} + ... + (\frac{2}{(m+2)^2} + 2\sum_{n=0}^{m} \frac{(-1)^m - n({}_mC_n)}{(m-n+2)^2})\hat{\beta_m} \tag{6}$$

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, I can estimate other $H(p)$ at different values of p by using the fitted polynomial $H(p)$ equation.

B. Discussion of the modifiable areal unit problem

One can imagine a city that is perfectly segregated by neighborhood like a checkerboard, as pictured in Figure 20(a). Then, imagine redrawing the boundaries of neighborhoods so that each new boundary slices a preexisting neighborhood in half, as pictured in Figure 20(b). The resulting city according to the new boundaries is pictured in Figure 20(c). However, all three diagrams represent the same city with the same microdata. Is Figure 20(a) or Figure 20(c) an accurate representation of this city? Neither seem to capture the underlying phenomenon. This problem with arbitrarily drawn boundaries is known as the *modifiable areal unit problem* (MAUP).



(a) Segregated city (b) Redraw boundaries (c) Integrated city

Figure 20: Modifiable areal unit problem

MAUP typically 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.

C. Non-parametric model of rank-order index on absolute upward mobility

In Table 7, I create a more flexible model of rank-order index on absolute upward mobility by making decile indicator variables for values of the rank-order index, allowing non-parametric relationships to be illuminated.

Table 7: Relationship between absolute upward mobility and rank-order index by decile

	Absolute upward mobility	
	(1)	(2)
Decile 2	0.831	0.00224
	(0.924)	(0.927)
Decile 3	0.292	1.346
	(0.933)	(0.936)
Decile 4	-0.154	0.625
	(0.924)	(0.927)
Decile 5	-0.0196	0.337
	(0.933)	(0.936)
Decile 6	0.830	0.118
	(0.924)	(0.927)
Decile 7	-1.553	-2.158*
	(0.924)	(0.927)
Decile 8	-0.614	-1.113
	(0.933)	(0.936)
Decile 9	-1.590	-0.882
	(0.924)	(0.927)
Decile 10	-2.133*	-0.667
	(0.933)	(0.937)
Constant	41.74***	41.57***
	(0.653)	(0.656)
Greater metropolitan area	No	Yes
Observations	276	276
R^2	0.0775	0.0710
Adjusted R^2	0.0463	0.0396
Degrees of freedom (Model)	9	9
Degrees of freedom (Residual)	266	266

Standard errors in parentheses

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

Table 7 shows that even with a more flexible framework, the rank-order index is only able to explain a very limited amount of variation in absolute upward mobility, regardless of the level of geography used. It is interesting that most of the coefficients on the deciles of rank-order index are not statistically significant, except for Decile 10 for rank-order index on MSAs and Decile 7 for rank-order index on greater metropolitan areas. This suggests that there might not be such a straightforward relationship between the rank-order index and absolute upward mobility, and that the negative effects of economic homogeneity on mobility are only seen sometimes, and only in comparing extremes.