The Measurement and Social Consequences of U.S. Income Inequality

by

Andrew Carr

Department of Sociology
Duke University

Date: ______________________

Approved:

_________________________

Martin Ruef, Dissertation Advisor

_________________________

Lisa Keister

_________________________

Christopher Bail

_________________________

Scott Lynch

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Sociology in the Graduate School of Duke University

2021
ABSTRACT

The Measurement and Social Consequences of U.S. Income Inequality

by

Andrew Carr

Department of Sociology
Duke University

Date:_______________________

Approved:

___________________________
Martin Ruef, Supervisor

___________________________
Lisa Keister

___________________________
Christopher Bail

___________________________
Scott Lynch

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Sociology in the Graduate School of Duke University

2021
Abstract

Researchers have long sought to understand how income disparities contribute to other forms of inequality. Inquiries into this subject, however, are constrained by data limitations. These constraints create challenges both for the measurement of income inequality and for the identification of geographically fine-grained economic data. In the chapters that follow, I address these issues separately. To begin, I propose a new statistical method for estimating U.S. income inequality and use Census data to demonstrate the accuracy of this method. I then draw from two data sources to examine how income differences contribute to other social disparities. First, I use mobile phone data to explore how income and occupational differences explain varying social distancing levels in the early weeks of the COVID-19 pandemic. Next, I draw on mortgage data to examine how the composition of incomes within and between neighborhoods shapes opportunities for homeownership. The methodological contributions of the first chapter open new possibilities for research into how income inequality generates other forms of social stratification. The subsequent chapters illustrate ways in which rising income inequality in the U.S. may produce to health and wealth disparities.
Dedication

To Sara, Poppy, and Claude.
Contents

Abstract ........................................................................................................................................ iv

List of Tables ................................................................................................................................. viii

List of Figures ................................................................................................................................. ix

Acknowledgements ......................................................................................................................... x

1. Introduction .................................................................................................................................. 1

2. Lorenz Interpolation: A Method for Estimating Income Inequality from Grouped Income Data ......................................................................................................................... 4

   2.1 Background ............................................................................................................................. 8

   2.2 Methods ................................................................................................................................ 12

       2.2.1 Fitting a Piecewise Linear Function to Grouped Income Data ................................... 12

       2.2.2 Interpolating the Lorenz Curve and Creating a Distribution of Exact Incomes .......... 16

   2.3 Data and Measures .................................................................................................................. 19

       2.3.1 Evaluating Lorenz Interpolation .................................................................................. 20

   2.4 Results .................................................................................................................................. 23

       2.4.1 Estimating tract-level and school district-level inequality measures ....................... 28

   2.5 Discussion and Conclusion ..................................................................................................... 31

3. Using Mobile Phone Data to Estimate the Effects of Income and Occupation on Social Distancing during the Early Response to Covid-19 ........................................................................ 36

   3.1 Data and Measures .................................................................................................................. 38

   3.2 Analytical Strategy .................................................................................................................. 39

   3.3 Trends in Time at Home and Full-Time Employment ............................................................ 40
List of Tables

Table 1: Gini, Theil, Atkinson, and SD Error Terms for PUMAs. ............................................ 24
Table 2: Error Terms for Tracts and School Districts. ................................................................. 29
Table 3: Spatial Autocorrelation Models of Change in Proportion Employed Full-Time 46
Table 4: Descriptive Statistics of Measures for Homeownership Models. ......................... 65
Table 5: Fixed Effects Models of Income Segregation, Concentrated Poverty, and Gentrification .................................................................................................................. 68
Table 6: Error Terms with Sampling Variation for School Districts. ................................. 84
Table 7: Error Terms with Sampling Variation for School Districts (Large Samples). ..... 85
List of Figures

Figure 1: Lorenz Interpolation in Four Steps................................................................. 14

Figure 2: Comparison of Fitted Lorenz Curves to True Lorenz Curve...................... 18

Figure 3: Comparison of Lorenz Curve Estimates from CDF and Lorenz Interpolation. 26

Figure 4: Residuals from Atkinson Estimates............................................................... 28

Figure 5: Trends in Daily Time Home and Full Time Employment by Income. ........... 41

Figure 6: Relationship Between Income and Hourly at Home Rate in Philadelphia........ 42

Figure 7: Effect of Median Income on Change in Daily Minutes Spent at Home and Change in Proportion Full-Time Employment (Largest 100 MSAs)................................. 44

Figure 8: Effects of County-level Occupation Proportions on Change in Full-Time Employment. ................................................................................................................. 48

Figure 9: Homeownership Rates for All Families, 1990-2018....................................... 53

Figure 10: MSA-level Contributions to Gentrification and Concentrated Poverty........... 69

Figure 11: The Distribution of Mortgage Originations Among Low and High-Income Households in L.A. ........................................................................................................... 71

Figure 12: PDF, CDF, and Lorenz Curve......................................................................... 77

Figure 13: Densities of Lorenz Interpolation Functions.................................................. 80

Figure 14: PDF Based on Lorenz Interpolation............................................................... 81
Acknowledgements

First and foremost, I would like to thank Martin Ruef for his guidance at every stage in the development of these chapters. I would also like to thank Colin Birkhead, Lisa Keister, Christopher Bail, Scott Lynch, Josh Bruce, and other members of the Duke Sociology department whose ideas have helped improve the quality of this research. For his thoughts on the methods section of chapter 1, I thank Peter Hoff of the Duke Statistics Department. This chapter also benefited from the recommendations of participants from the Duke Economic Sociology Workshop. Finally, I owe a debt of gratitude to my partner, Sara, without whom I could not have completed my graduate studies.
1. Introduction

Income inequality has been rising in the United States since the late 1970s (Piketty and Goldhammer 2014). This trend has created methodological challenges for researchers and raises questions about the interrelationship between U.S. economic stratification and other facets of social life. Methodologically, some of the problems faced by income inequality researchers are caused by the limitations of publicly available income data. The increasing right-skewness of the income distribution makes these limitations all the more constraining. Furthermore, given the small samples upon which income data is based, reliable inequality estimates are difficult to obtain at the levels of analysis that are of interest to sociologists. This is particularly problematic for those interested in the consequences of income inequality for local communities.

Concerns about income inequality have prompted sociologists to explore how this phenomenon contributes to other forms of social stratification (Wilkinson and Pickett 2009). Many of these studies operate at the societal level and use cross-national comparisons to reveal connections between income inequality and social problems such as violent crime and poor health and educational outcomes (Hsieh and Pugh 1993; Pickett and Wilkinson 2007). More recently, a growing body of research has concentrated on the effects of income inequality within the U.S. (Owens 2016, 2019). Many of these studies take advantage of the panel structure of U.S. Census data to build fixed-effects models linking changes in income inequality to shifts in other economic and
social domains (Reardon and Bischoff 2011). Still, development in this area of research is hampered by the paucity and inaccessibility of the geographically fine-grained economic data needed to analyze how changes in income inequality influence other areas of society.

In the chapters that follow, I contribute to sociological research on income inequality measurement and the relationship between income inequality and social stratification in the U.S. In the absence of individual-level income data, the chapters devoted to the latter subject focus on differences in median income levels between census block groups, neighborhoods, and metropolitan areas. While focusing on regional differences prevents one from drawing causal inferences from the findings in these chapters, the results highlight several possible ramifications of the growing problem of income inequality.

In the next chapter, I propose a new method, which I call Lorenz interpolation, for estimating income statistics using the tabular data published in the Census. The method consists of plotting this data as points on a Lorenz curve and fitting a quadratic-cubic spline to these points. I pay particular attention to the upper tail of the income distribution, which is interpolated using set of slope constraints. Drawing public ACS data and a restricted Census dataset to which I have been granted access through a Federal Statistical Research Data Center (FSRDC), I demonstrate the utility of the
method for measuring income inequality at the PUMA (public use microdata area) and school district levels.

The following chapter analyzes how income and occupational inequities shaped differences in social distancing levels in the early response to the COVID-19 pandemic. Using mobile phone dataset that tracks people’s hourly movements, I show that areas with higher levels of employment in lower-income occupational categories experienced smaller increases in social distancing during the month of March 2020. I attribute these discrepancies to differences in the employment conditions of low and high-paying jobs and the greater ability of people in higher-paying occupational groups to transition to working from home.

The final chapter analyzes the relationship between economic segregation – specifically, the proportion of income inequality that can be attribute to between-neighborhood income differences – and low-income homeownership. Tracing mortgage origination rates among low-income households over the last two decades, I show that cities with declining levels of income segregation experienced decreasing levels of low-income homeownership. I relate these findings to the California housing crisis, making the argument that the disappearance of lower-income neighborhoods has constrained homeownership opportunities among low-income residents.
2. Lorenz Interpolation: A Method for Estimating Income Inequality from Grouped Income Data

Income inequality has increased substantially in many countries around the world. In the United States, income inequality is higher than it has been since the beginning of the Great Depression (Galster and Sharkey 2017). Rising income inequality has been implicated in a range of negative public health consequences (Pickett et al. 2005; Pickett and Wilkinson 2007), and research has provided support for a connection between income inequality and economic segregation (Reardon and Bischoff 2011), which in turn has implications for educational stratification (Owens 2016). Growing income inequality has also driven increasing inequities in the intergenerational transmission of wealth (Chetty et al. 2017), which has led to decreased levels of social mobility (Grusky and MacLean 2016).

Amidst growing concern around these and other possible consequences of rising income inequality, researchers have sought ways to measure the income distribution more precisely to trace how and where it is changing over time. They are constrained in their efforts, however, by the format in which most countries provide their income data. To protect respondent confidentiality, censuses generally publish income data in tabular format, as counts in income brackets ($0 to 9,999, $10,000 to 14,999, …, $200,000+). Researchers attempting to use this data to estimate income inequality must make assumptions about how incomes are distributed within these brackets. Moreover, the highest income bracket in this data is always unbounded at the top, which makes
estimating the upper tail of the income distribution especially difficult. For some income inequality measures, even small errors in the estimation of this upper tail can lead to large errors in the estimation of income inequality.

A number of studies have proposed methods to estimate income statistics using grouped income data (Quandt 1966; Gastwirth and Glauberman 1976; Kakwani 1976; Hajargasht et al. 2012; Minoiu and Reddy 2012; Tillé and Langel 2012; von Hippel, Scarpino, and Holas 2016; von Hippel, Hunter, and Drown 2017; Jargowsky and Wheeler 2018). Two recent studies have shown that income distribution means, which are provided in the public portions of many national censuses, can be used to produce more accurate estimates of income inequality (von Hippel et al. 2017, Jargowsky and Wheeler 2018). Still, two problems remain. First, even these methods have difficulty with inequality measures that rely heavily on the upper tail of the income distribution. This is unfortunate because some of these measures have desirable properties. For instance, the Theil coefficient can be decomposed into groups to determine how different kinds of households or people contribute to income inequality. This makes Theils useful to researchers interested in understanding how changing levels of income inequality between geographic subregions or racial groups has contributed to overall changes in income inequality.

More importantly, the existing methods for estimating inequality from grouped data have only been tested on the income distributions of relatively large geographic
regions, such as metropolitan areas and counties. For understanding the social consequences of income inequality, it often makes more sense to focus on a lower geographic level. For instance, a researcher interested in the relationship between income inequality and educational inequality may want to work at the level of the school district, which determines the portion of a school’s budget that comes from property taxes, or the school attendance zone, which delineates which schools a child is eligible to attend. Alternatively, if a researcher wants to understand how income inequality is experienced in one’s community, the relevant geography might be the neighborhood, which is often operationalized by the census tract. Given that geographies like these cover areas with smaller populations, their income distributions may vary more than those belonging to metropolitan areas and counties. These distributions may also violate the parametric assumptions upon which many methods of estimating income distributions from grouped data depend. These issues suggest that income inequality estimates based on small regions could be overly biased or unreliable. But while some researchers have admonished against estimating income inequality with grouped income data for smaller geographic regions (von Hippel et al. 2016), the performance of these methods for smaller geographies has yet to be empirically tested.

In this chapter, I outline a new method for estimating income inequality from grouped income data. I demonstrate the utility of this method for producing income inequality measures that are sensitive to the upper tail of the income distribution and
that are based on smaller geographic regions. The method, which I call Lorenz interpolation, consists of using grouped income data to estimate points along a Lorenz curve and fitting a piecewise function to these points. The primary advantage of Lorenz interpolation over prior methods stems from the way in which the method captures the variance of incomes in the upper tail of the income distribution. Using public microsample data from the 2011-2015 American Community Survey to estimate income statistics for PUMAs (public use microsample areas), I show that Lorenz interpolation outperforms the best alternative methods at estimating many measures of income inequality. Next, I evaluate the performance of Lorenz interpolation at two lower geographic levels, the census tract and the school district, using restricted census data to which I have been granted access through a Federal Statistical Research Data Center. Results indicate that neither Lorenz interpolation nor the other methods produce accurate estimates of certain income inequality measures at the tract level. However, Lorenz interpolation yields fairly accurate estimates of income inequality at the school-district level, suggesting that geographies consisting of small groups of census tracts are sufficiently large for the accurate estimation of income inequality. I conclude with a discussion of the use cases of Lorenz interpolation and some unresolved issues surrounding the measurement of income inequality.

---

1 Information about the locations of FSRDCs can be found on the Census website (https://www.census.gov/fsrdc).
2.1 Background

A common method for deriving income inequality statistics from grouped income data is to set the incomes in the closed income brackets to their midpoints and estimate a Pareto distribution for the open bracket at the top of the income distribution (Henson 1967; Jargowsky 1996). Fitting a Pareto distribution to the top bracket requires that the researcher estimate a Pareto distribution shape parameter, $\alpha$, which can be computed using the counts of households in the top two brackets of the income distribution (Jargowsky and Wheeler 2018:346). This technique, known as the Pareto-linear procedure, is based on Pareto’s observation that the relationship between the populations and incomes of the income brackets in the upper tail of the income distribution tends to be linear (Nielsen and Alderson 1997). Once $\alpha$ is estimated, the mean of the Pareto distribution for the top bin can be computed with the following formula.

$$\mu = \beta \frac{\alpha}{\alpha - 1}$$

Where $\beta$ is the lower bound of the top income bracket. Incomes in this bracket are then set to the estimate of $\mu$.

2 There are small differences in how sociologists employ this approach. For instance, Nielsen and Alderson (1997) assigned bracket midpoints to every income bin below the bin containing the median income and used Pareto distributions to estimate bracket means for the median income bin and every bin above it. Owens (2019) used the robust Pareto midpoint estimator developed by von Hippel et al. (2016), which uses $\alpha$ to estimate the harmonic rather than arithmetic mean of the Pareto distribution.
Researchers have identified some issues with this approach. von Hippel et al. (2016) observed that mean estimates derived using the Pareto-linear procedure are not robust to even small errors in the estimation of $\alpha$. This is particularly true when this procedure underestimates $\alpha$ because the Pareto distribution is undefined when $\alpha$ is less than 1 and its mean approaches infinity as $\alpha$ approaches 1 from above.\(^3\) Also, by imputing midpoints for the closed brackets and a Pareto distribution mean for the open bracket, the method ignores the within-bracket variation among incomes. This omission will lead to downwardly biased income inequality estimates if a significant source of the variation among incomes lies within the income brackets.

Recently, researchers have developed ways to get around these issues.\(^4\) Proposing a method called cumulative density function (CDF) interpolation, von Hippel et al. (2017) exploit the fact that grouped income data can be used to plot a set of points along the CDF of the income distribution. Using a cubic spline function, the authors interpolate these points to estimate a piecewise probability density function (PDF) of the underlying distribution. When this distribution is constrained by the distribution mean, \(^5\)

---

\(^3\) To get around this problem, von Hippel and coauthors (2016) proposed a robust Pareto mean estimator, which uses $\alpha$ to estimate a harmonic rather than arithmetic mean for the Pareto distribution of the top income bin. As the name suggests, this method is robust to errors in the estimation of $\alpha$.

\(^4\) Researchers have employed a wide variety of techniques to estimate the income distribution from grouped data, and a comprehensive review of these methods is beyond the scope of this article. Several studies have used parametric techniques, usually with distributions from the generalized beta family, to estimate the income distribution from income groups (McDonald and Xu 1995; von Hippel et al. 2016; Kakamu 2016). Additionally, researchers have attempted to approximate the income distribution with semiparametric and non-parametric methods, such as with generalized method of moments estimation (Hajargasht et al. 2012) and kernel density estimation (Minoiu and Reddy 2012).
CDF interpolation estimates Gini coefficients within 1-2% of their true values (von Hippel et al. 2017:651).\(^5\)

Jargowsky and Wheeler (2018) also propose a method that takes advantage of the income distribution mean. Instead of fitting a function to the CDF of the income distribution, the authors apply a set of linear and uniform functions directly to the grouped income data. Their technique, which they call mean-constrained integration over brackets (MCIB), consists of estimating a piecewise linear function for the closed income brackets, fitting a Pareto distribution to the top income bracket, and integrating over the resulting distribution to produce various income statistics.\(^6\) This technique uses the mean of the income distribution to estimate the mean of the top income bracket. After fitting a function to the closed income brackets, the total closed-bracket income is subtracted from the total income, which is then divided by the number of households in the top bin to produce an estimate of the top bracket mean. The authors show that MCIB outperforms estimation techniques that do not incorporate the income distribution mean.

\(^5\) The version of CDF interpolation that is the focus of this paper is implemented by the splinebins function from the binsmooth package in R. This package also includes a function, stepbins, that fits a step function to the closed bins of the income distribution and a uniform, exponential, or Pareto distribution to the top bin. A preliminary analysis suggests that splinebins may produce more accurate income inequality estimates than stepbins, although stepbins appears to produce more slightly accurate estimates at the PUMA-level when a Pareto distribution is selected for the top bin.

\(^6\) This is only a viable technique for income statistics that can be decomposed into contributions from each income bin. Income statistics that cannot be disaggregated into contributions from each income bin, such as the Gini, must be estimated through different methods, such as numerical integration or by applying a bound to the top of the income distribution.
Although MCIB and CDF interpolation have been shown to produce more accurate income inequality estimates than earlier techniques, these methods struggle to approximate the upper tail of the income distribution. The method that I propose, which I call Lorenz interpolation, uses slope constraints to capture the variance at the upper tail of the income distribution. After plotting the income groups as points on a Lorenz curve, this method interpolates these points in a way that slows the upward trajectory of this curve. As I will demonstrate, this technique corrects for the tendency of other grouped income interpolation methods to underestimate of the upper tail variance.

While Lorenz curves are rarely found in sociological research, economists have developed techniques to estimate inequality statistics by interpolating the Lorenz curve (Gastwirth 1971; Gastwirth and Glauberman 1976; Cowell and Mehta 1982; Gastwirth et al. 1986; Tillé and Langel 2012). Many of these studies were developed for kinds of data that contain more information about the income distribution than the data in most national censuses (e.g. Cowell and Mehta 1982). My approach differs from these studies by imposing on the data the same limitations as most publicly available income data. My method also pays particular attention to the top bin of the income distribution, which has disproportionately contributed to recent increases in income inequality (Piketty and Goldhammer 2014:315).
2.2 Methods

The method proposed in this paper, visualized in Figure 1, consists of several steps. First, following the method described in Jargowsky and Wheeler’s (2018) paper, I fit a piecewise linear function to the closed brackets of the grouped income data. This function is used to estimate the closed income bracket means, from which points along a Lorenz curve are estimated. Next, the points of the Lorenz curve are connected with a set of quadratic functions and one cubic function. The cubic function approximates the open-ended bracket at the top of the distribution and is estimated using the final two points of the Lorenz curve and a pair of slope constraints. Interpolating the Lorenz curve in this way produces a piecewise quadratic-cubic function, from which a weighted sample of exact incomes is derived. These incomes are then used to compute inequality statistics from the underlying income distribution.

2.2.1 Fitting a Piecewise Linear Function to Grouped Income Data

A piecewise linear function is defined for the closed brackets of the income distribution by the following equation.

\[ f_b(x) = m_b x + c_b \]

Where \( m_b \) and \( c_b \) are the slope and y-intercept of the linear function over the range of incomes in bracket \( b \). \( m_b \) is computed by plotting a histogram of the grouped income data and connecting the midpoints of each bin. Each bin that lies between a higher bin and a lower bin – in other words, on a rising or descending part of the income
distribution – is given a linear density, the slope of which is computed by averaging the slopes of the lines connecting the midpoint of the focal bin to the midpoints of the neighboring bins. Bins that lie between two higher or lower bins are given uniform densities. These represent the peaks and valleys of the income distribution. Once \( m_b \) is computed, the y-intercept \( c_b \) can be calculated by solving the equation above, plugging in the bin midpoint for \( x \) and the bin height for \( f_b(x) \). Finally, these parameters are rescaled so that the function they define is a PDF. The area under the resulting function should equal the proportion of total households accounted for by the closed income brackets.

Steps 1 and 2 in Figure 1 show how this process works. The plot at the top left of the figure is a step function where each income bin is represented by a uniform density. The dashed line segments connect the midpoints of each bin. For bins on upward and downward inclines of the income distribution, the slopes of these line segments are averaged to create slopes for linear distributions. Finally, distributions must be chosen for the bottom bin and top closed bin (the second to last bin), each of which has only one neighboring closed bin. Following Jargowsky and Wheeler (2018), I assign a uniform distribution to the bottom bin and a linear distribution to the top closed bin.\(^7\) The slope

\(^7\) The bin height, referred to as the relative frequency in Jargowsky and Wheeler (2018), can be computed by dividing the bin width, or the difference between the bin’s upper and lower bounds, by the number of households in a given bin.

\(^8\) The latest version of the lorenz R package includes a parameter that allows for the selection of either a uniform or linear distribution for the bottom bin. If one chooses the linear distribution option, the slope of
of the latter equals the slope of the final line segment connecting the last and second-to-last closed bins. The upper right plot displays the piecewise linear function fit to the grouped income data.

Figure 1: Lorenz Interpolation in Four Steps.

the bottom bin will be determined by interpolating between the midpoints and relative frequencies of the bottom two bins. In practice, a uniform distribution provides a better fit of the bottom bin of ACS income data. One reason for this is that this data includes negative incomes. While assigning a linear distribution to the bottom bin will usually result in an estimated bin mean greater than the bin midpoint, the true bin mean of the bottom bin is likely below the bin midpoint. However, one may wish to use a linear distribution for the bottom bin when working with income data that does not include negative incomes. * For some income distributions, the slopes given to the final closed bracket produce lines that intersect the x-axis. When this happens, I adjusted the density so that it intersected the axis at the upper bound of the top closed bracket. This produced triangular densities, as seen in the example in Figure 1. My understanding is that this is how Jargowsky and Wheeler fit distributions to their data, given their claim that they constrained their linear distributions to produce only non-negative densities (p. 349).
Next, Lorenz interpolation uses this piecewise linear function to estimate means for the closed income bins. The means of the bins to which uniform densities have been assigned are the bin midpoints. For bins with linear densities, bin means can be computed by rescaling each bin to have an area of 1 and finding the expectation, as shown below.

\[ E[x] = \int_{l_b}^{u_b} x f_b(x) \, dx \]

\[ = \int_{l_b}^{u_b} m_b x^2 + c_b x \, dx \]

\[ = \left. \frac{m_b x^3}{3} + \frac{c_b x^2}{2} \right|_{l_b}^{u_b} \]

Once closed bracket means have been calculated, these means are used to plot points on a Lorenz curve. The coordinates of these points are calculated using the following equations.

\[ x_b = \frac{\sum_{i=1}^{b} n_i}{N} \]

\[ y_b = \frac{\sum_{i=1}^{b} \mu_i n_i}{N} \]

Where \( x_b \) and \( y_b \) represent the cumulative share of households and the cumulative share of income up to and including bin \( b \), \( n_b \) is the household frequency of bin \( b \), \( \mu_b \) is the mean associated with bin \( b \), and \( N \) is the total population of the income distribution.
2.2.2 Interpolating the Lorenz Curve and Creating a Distribution of Exact Incomes

Having used the grouped income data to plot points on a Lorenz curve (step 3 in the figure), the next step is to interpolate these points. To begin, a series of quadratic functions are fit to the points associated with the closed income brackets. Specifically, a quadratic function is fit to each set of three consecutive points, which is then used to interpolate the first two of these points. For example, to interpolate the points associated with the first income bin, a quadratic function is fit to the first three points on the Lorenz curve – \((0, 0), (x_1, y_1), (x_2, y_2)\). This function is used to connect \((0,0)\) and \((x_1, y_1)\). This process is repeated for each closed income bin except for the top closed bin, to which a quadratic function is fit using the last three points associated with the closed bins.

To interpolate the portion of the Lorenz curve associated with the open income bin, the space between \((x_B, y_B)\) and \((1,1)\) where \(B\) is the number of closed income brackets, I use a cubic function with two slope constraints. The absence of an upper bound on the top income bin makes fitting a distribution to this bin particularly challenging. Cubic functions fit to the top four points of the Lorenz curve tend to underestimate the concentration of incomes at the top of the distribution (Kakwani 1976:489). Imposing slope constraints on the cubic remedies this problem. The first constraint sets the starting slope of the cubic function to the slope of the line tangent to the function for the previous bin at \((x_B, y_B)\). This ensures that the function is continuous as it passes through \((x_B, y_B)\) and makes the starting slope lower than it would be if a cubic were fit to the last four points of the Lorenz curve. The second constraint adjusts
the coefficient on the first term of the cubic function so that the slope increases more gradually as the function moves from \((x_B, y_B)\) to \((1,1)\). First, I fit a quadratic function to the top three points of the Lorenz curve and find the slope of the function at the midpoint between the top two points. Then I constrain the cubic function to have a slope that is 90% of the quadratic slope at this midpoint.

The technique of fitting a cubic function to two points with two slope constraints can be represented by the following system of equations.

\[
\begin{bmatrix}
    x_B^3 & x_B^2 & x_B & 1 \\
    1 & 1 & 1 & 1 \\
    3x_B^2 & 2x_B & 1 & 0 \\
    3x_B^{2\text{mid}} & 2x_{B\text{mid}} & 1 & 0
\end{bmatrix}
\begin{bmatrix}
    a \\
    b \\
    c \\
    d
\end{bmatrix} =
\begin{bmatrix}
    y_B \\
    1 \\
    m_B \\
    .9m_{B\text{mid}}
\end{bmatrix}
\]

Where \(a, b, c,\) and \(d\) are the coefficients of the cubic function. The first two rows of the system constrain these coefficients so that the function passes through \((x_B, y_B)\) and \((1,1)\). The last two rows represent the slope constraints, where \(m_B\) is the slope of the line tangent to the quadratic function fit to the last closed bin at \((x_B, y_B)\), and \(m_{B\text{mid}}\) is the slope of the line tangent to a quadratic function fit to the top bin at \(x_{B\text{mid}}\), the midpoint between \((x_B, y_B)\) and \((1,1)\).
Figure 2: Comparison of Fitted Lorenz Curves to True Lorenz Curve.

The slope constraints imposed on the cubic function of the top bin may seem arbitrary, but they ensure that the Lorenz curve captures the spread of income at the top of the income distribution. Figure 2 demonstrates this by comparing two estimated Lorenz curves, one where the top bin is fitted with a cubic function interpolating the last four points on the Lorenz curve and one where the top bin is fitted with slope constraints. The solid line represents the true Lorenz curve, while the dotted and dashed lines are the estimated Lorenz curves. The line based on the slope constraints curves upward more gradually and has a flatter slope at the beginning of the bin, the $200,000 cutoff point, than the cubic function fitted to the four points. These features allow the
function to better approximate the true Lorenz curve (see Appendix A for derivations of the PDF and CDF based on this curve).

Having estimated a Lorenz curve, the final step is to draw a sample of exact incomes based on this curve. A computationally efficient way to approximate a sample from the Lorenz curve is to plot equidistant points along the interpolated curve and multiply the slopes of the line segments connecting these points by the grand mean. This generates samples from the underlying income distribution, which can be weighted using the relative frequencies provided in the grouped income data. Finally, this weighted sample is plugged into an income inequality formula to produce an income inequality estimate.

2.3 Data and Measures

The data for this study comes from the 2011-2015 five-year pooled American Community Survey (ACS). The ACS contains social and economic data for the U.S. population and is administered on a rolling basis by the Census. To provide more reliable estimates, the Census publishes ACS data in five-year groupings, which cover roughly 5% of the U.S. population (Ruggles et al. 2021). To compute income inequality at the PUMA level, I used household income data from the public use microdata sample (PUMS) component of the ACS, which contains households’ exact incomes.¹⁰ Estimates

¹⁰ The household income measure consists of individuals’ total pre-tax income or losses from all income sources during the previous year. These include wages and salaries, self-employment income, interests,
were calculated for 1,185 public use microdata areas (PUMAs), which are the smallest geographies for which PUMS data is available.

To calculate income inequality for tracts, I used restricted Census data to which I have been granted access through a Federal Statistical Research Data Center. This data contains exact incomes with geographic information down to the block level. To produce income inequality estimates for school districts, I used a crosswalk between census tracts and school districts from the National Center for Education Statistics (Geverdt 2019). Tract-level incomes were assigned to school districts based on the distribution of each tract’s land area across school districts. For example, if 60% of a tract’s area is in district A and 40% of its area is in district B, 60% of its incomes were assigned to district A and the remaining 40% were assigned to district B. The school district data is based on boundaries from 2013, which falls in the middle of the study period (2011-2015). These analyses were based on 69,675 tracts and 13,360 school districts.

2.3.1 Evaluating Lorenz Interpolation

To assess the performance of Lorenz interpolation, I created two datasets, an exact income dataset and a grouped income dataset. The latter was produced by
converting each region’s exact income data into counts in income brackets, which were
determined by the income bounds used in the U.S. Census since 2000. Next, each
region’s “true” level of income inequality was calculated by plugging the exact income
data into an income inequality formula. Lorenz interpolation was then applied to the
grouped income data to generate an income inequality estimate. Finally, the error
statistics were calculated by comparing true and estimated income inequality.

I evaluated Lorenz interpolation with four measures of income inequality. First, I
looked at Gini coefficients, which can be computed with the following equation.

\[ G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n\bar{x}} \]

The Gini is the mean absolute difference among the incomes divided by twice the
aggregate income. Next, I computed the Theil coefficient. Based on information theory,
Theils account for the level of entropy (i.e., unpredictability) in a dataset. Income
distributions with low entropy have higher Theils, which indicate more inequality.
Theils are calculated with the following equation.

\[ \text{Theil} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{x_i - x_j}{2n\bar{x}} \]

One limitation of this approach is that the exact income data in both the public and restricted ACS data are
top-coded. In this analysis, I choose to ignore top-coding, which may downwardly bias income inequality
estimates from the exact income data. One justification for this decision is that the restricted exact income
data that was used for the tract-level and school district-level estimates of income inequality have
significantly less top-coding than the public data. In the public data, wage and salary income above a state-
determined threshold is assigned the mean of incomes above that threshold (Ruggles et al. 2021). The exact
income data in both the public and restricted ACS may also suffer from mismeasurement (e.g., rounding of
reported income) and incomplete coverage of high-income earners. Although the Census attempts to correct
for the latter issue, undersampling of those at the top of the income distribution may lead to downwardly
biased estimates of income inequality.
\[
T = \frac{1}{n} \sum_{i=1}^{n} \log \frac{x_i x_i}{\bar{x} \bar{x}}
\]

One nice feature of Theils is that they can be decomposed to determine contributions from different groups. This means that Theil contributions can be estimated for each bracket of an income distribution. Performing such a decomposition shows that most of the Theil coefficient can often be attributed to the top bracket of the income distribution. I also computed the Atkinson Index (Atkinson 1970). This measure includes a parameter that determines the relative influence of the lower and upper tails of the income distribution. It is defined by the following equation.

\[
A = \begin{cases} 
1 - \frac{1}{\mu} \left( \frac{1}{n} \sum_{i=1}^{n} x_i^{1-\varepsilon} \right)^{1/(1-\varepsilon)} & \text{for } 0 \leq \varepsilon \neq 1 \\
1 - \frac{1}{\mu} \left( \prod_{i=1}^{n} x_i \right)^{1/n} & \text{for } \varepsilon = 1
\end{cases}
\]

Where \( \varepsilon \), the inequality aversion parameter, determines the relative influence of the income distribution tails. I chose to set \( \varepsilon \) to .2, which increases the influence of the upper tail on income inequality. Finally, I computed the income distribution standard deviation.

I compared the numbers yielded from Lorenz interpolation to estimates from two other methods: von Hippel et al.’s (2017) CDF interpolation method and Jargowsky and Wheeler’s (2018) MCIB method. To implement CDF interpolation, I used the
binsmooth package in R (Hunter and Drown 2016). To implement MCIB, I used Jargowsky’s MCIB module (Jargowsky 2019), which is available in Stata.

2.4 Results

Table 1 compares error terms from Gini, Theil, Atkinson, and standard deviation estimates based on MCIB, CDF interpolation, and Lorenz curve interpolation (see Appendix B for error terms associated with quantile and income share estimates). The top three rows show the errors associated with these methods when the income distribution mean is not provided, while the bottom three rows display errors when the distribution mean is provided. Following von Hippel et al. (2017), I calculated percent relative bias and percent relative root mean squared error (RMSE) terms. These measures are based on the percent estimation error, \(e_j\), a function of the ratio of the error to the true value (100 * \(\frac{\hat{\theta} - \theta}{\theta}\)). The percent relative bias is the mean of the percent estimation error, and the percent RMSE is the square root of the mean of the squared percent estimation errors. An important advantage of these metrics is that they are invariant to the different scales of the inequality measures. This makes accuracy comparisons between the measures possible. All estimates have been rounded to the nearest hundredth.
Table 1: Gini, Theil, Atkinson, and SD Error Terms for PUMAs.

<table>
<thead>
<tr>
<th></th>
<th>Gini Bias</th>
<th>Gini RMSE</th>
<th>Theil Bias</th>
<th>Theil RMSE</th>
<th>Atkinson Bias</th>
<th>Atkinson RMSE</th>
<th>SD Bias</th>
<th>SD RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCIB (no $\mu$)</td>
<td>2.10</td>
<td>4.04</td>
<td>.83</td>
<td>6.98</td>
<td>14.23</td>
<td>36.05</td>
<td>11.74</td>
<td>19.10</td>
</tr>
<tr>
<td>CDF (no $\mu$)</td>
<td>-1.48</td>
<td>3.27</td>
<td>-4.71</td>
<td>8.76</td>
<td>-4.98</td>
<td>9.32</td>
<td>-4.86</td>
<td>12.18</td>
</tr>
<tr>
<td>Lor (no $\mu$)</td>
<td>-.91</td>
<td>2.96</td>
<td>-2.27</td>
<td>7.75</td>
<td>-2.52</td>
<td>7.13</td>
<td>-1.32</td>
<td>10.08</td>
</tr>
<tr>
<td>MCIB ($\mu$)</td>
<td>-.76</td>
<td>.97</td>
<td>-5.48</td>
<td>6.36</td>
<td>-1.88</td>
<td>5.39</td>
<td>-1.04</td>
<td>3.93</td>
</tr>
<tr>
<td>CDF ($\mu$)</td>
<td>-1.25</td>
<td>1.38</td>
<td>-4.79</td>
<td>5.11</td>
<td>-4.83</td>
<td>6.55</td>
<td>-4.92</td>
<td>7.11</td>
</tr>
<tr>
<td>Lor ($\mu$)</td>
<td>-.71</td>
<td>.88</td>
<td>-2.48</td>
<td>3.06</td>
<td>-2.56</td>
<td>3.42</td>
<td>-1.94</td>
<td>2.95</td>
</tr>
</tbody>
</table>

N = 1,185

Looking at the first three rows of the table, which compare inequality estimates when the distribution mean is not provided, CDF interpolation outperformed MCIB at estimating the Gini, the Atkinson, and the standard deviation, while MCIB outperformed CDF interpolation at estimating the Theil. Lorenz interpolation outperformed both methods according to all error metrics except for the Theil relative RMSE. Overall, MCIB does a poor job at estimating these inequality measures when the distribution mean is not provided. This may be a reflection of the fact that MCIB uses the Pareto-linear procedure to estimate the top bin mean when a distribution mean is not given. As pointed out by von Hippel et al. (2016), underestimation of the Pareto’s $\alpha$ parameter can produce large overestimates of the top bin mean.

How do these methods compare when the income distribution mean is provided? Turning to the bottom three rows of the table, Lorenz interpolation produced estimates with lower relative RMSEs for the four inequality measures and lower relative
bias for all the measures except for the standard deviation. For Theil coefficients, Atkinson measures, and standard deviations, Lorenz interpolation produced considerably more accurate estimates, while the estimates from the other two methods had significantly higher relative RMSEs. For instance, while the Theil estimates generated by MCIB were 6.36% off according to the relative RMSE, the Lorenz Theils had a relative RMSE of 3.72%, amounting to an error reduction of more than 40%. While Lorenz interpolation produced significantly better Theil, Atkinson, and standard deviation estimates, the estimation methods performed comparably well at estimating the Gini. All three methods produced Gini estimates with relative RMSEs of about 1%.

The improvements with which Lorenz interpolation estimates income inequality can be attributed to the way that this method estimates the upper tail of the income distribution. Figure 3 compares the upper ends of the Lorenz curves of income distributions estimated by CDF interpolation and Lorenz interpolation. This figure is based on the income distribution of a PUMA located in the San Francisco-Oakland-Hayward metropolitan statistical area. The solid line represents the actual Lorenz curve of the income distribution, the dashed line is the Lorenz curve estimated by Lorenz interpolation, and the dotted line is the Lorenz curve based on CDF interpolation. According to this figure, CDF and Lorenz interpolation both provide accurate estimates of the lower part of the Lorenz curve, the region corresponding to the closed bins of the income distribution. However, beginning at the 75th percentile of the income
distribution, the Lorenz curve based on CDF interpolation starts to diverge from the true Lorenz curve. The curve produced by Lorenz interpolation, on the other hand, breaks away from the true Lorenz curve more gradually, at around the 90\textsuperscript{th} percentile of the income distribution. This accounts for the improvement of Lorenz interpolation over CDF interpolation, which produces downwardly biased inequality estimates.

![Figure 3: Comparison of Lorenz Curve Estimates from CDF and Lorenz Interpolation.](image)

While CDF interpolation estimates were more biased than Lorenz interpolation estimates, MCIB estimates had larger RMSEs. This may be a result of MCIB’s use of the Pareto distribution to approximate the upper tail of the income distribution. Figure 4
shows scatterplots for the residuals of estimates from the three estimation methods. The solid lines denote the bias associated with an estimation method, and the dashed lines represent a one standard-deviation distance from the solid lines. Despite having the lowest bias, MCIB produced some of the largest residuals of the three methods. As can be discerned by comparing the space between the dashed lines in the MCIB and other plots, the improvement of Lorenz interpolation over MCIB is due to a reduction in variance. Finally, for all three plots, the points with the largest residuals were those belonging to regions with higher levels of inequality. This is unsurprising, given that these regions have more variance among incomes in the top bracket of the income distribution. The grouped data associated with these regions provides insufficient information for accurately estimating this variance.
2.4.1 Estimating tract-level and school district-level inequality measures

As von Hippel et al. (2016) point out, estimates of income inequality based on grouped data are less accurate for small regions than for large regions. This may reflect the greater heterogeneity of income distributions associated with smaller regions. Due in part to economic segregation, the income distributions of neighborhoods or municipalities likely vary more than those of larger regions such as metropolitan areas, which encompass entire regional economies and tend to resemble each other more. These considerations raise the question of whether estimation techniques like Lorenz interpolation can be used to produce valid income inequality estimates for small areas.
In this section, I compare the accuracy of tract and school district-level estimates of income inequality produced by MCIB, CDF interpolation, and Lorenz interpolation.

Table 2 shows the error terms for tract and school district-level estimates. First, note the size of the relative RMSEs. With some exceptions, the RMSEs were higher for tracts than for school districts and for school districts than for PUMAs, a pattern that reflects the greater heterogeneity of income distributions in smaller areas and the fact that the relative frequencies of income groups in these areas provide less information about the income group means. Comparing MCIB and CDF interpolation, the latter method produced significantly more accurate estimates of the Theil and the Atkinson at both the tract and school district levels. CDF interpolation also outperformed MCIB at estimating the standard deviation at the tract-level, but the two methods performed comparably well at the school district-level. Most importantly, while MCIB estimates tended to have lower bias than CDF estimates, their relative RMSEs indicate that this difference in bias is outweighed by the residual variance of MCIB estimates.

<table>
<thead>
<tr>
<th></th>
<th>Gini</th>
<th>Theil</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tract</td>
<td>School District</td>
<td>Tract</td>
<td>School District</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
</tr>
<tr>
<td>MCIB</td>
<td>-1.00</td>
<td>2.33</td>
<td>-1.01</td>
<td>1.41</td>
<td>-2.32</td>
<td>23.6</td>
</tr>
<tr>
<td>CDF</td>
<td>-1.18</td>
<td>2.32</td>
<td>-1.44</td>
<td>1.76</td>
<td>-4.65</td>
<td>7.82</td>
</tr>
<tr>
<td>Lorenz</td>
<td>-0.06</td>
<td>2.35</td>
<td>-.624</td>
<td>1.27</td>
<td>-1.17</td>
<td>7.42</td>
</tr>
<tr>
<td>Tract</td>
<td>School District</td>
<td>Tract</td>
<td>School District</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>----------------</td>
<td>-------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
</tr>
<tr>
<td>MCIB</td>
<td>-1.20</td>
<td>15.60</td>
<td>-3.19</td>
<td>13.47</td>
<td>-.31</td>
<td>18.33</td>
</tr>
<tr>
<td>CDF</td>
<td>-4.62</td>
<td>9.01</td>
<td>-5.99</td>
<td>7.96</td>
<td>-5.02</td>
<td>10.50</td>
</tr>
<tr>
<td>Lorenz</td>
<td>-1.70</td>
<td>7.23</td>
<td>-3.17</td>
<td>5.41</td>
<td>-1.59</td>
<td>8.36</td>
</tr>
<tr>
<td>N</td>
<td>69,675</td>
<td>13,360</td>
<td>69,675</td>
<td>13,360</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lorenz interpolation performed comparably to MCIB and CDF interpolation at estimating the Gini, while producing significantly more accurate estimates of the Theil, Atkinson, and standard deviation. At the school district level, estimates based on Lorenz interpolation had 25-32% lower relative RMSEs than those based on the next best method, CDF interpolation. The gap between MCIB and Lorenz interpolation was even larger: Atkinson and Theil estimates based on Lorenz interpolation were about 60% lower than those based on MCIB. Finally, while tract-level error metrics from Lorenz interpolation were also lower than those based on the other methods, the size of the relative RMSEs at this level of analysis suggest that researchers should refrain from estimating certain inequality measures for census tracts and smaller geographies. At the tract-level, even the Lorenz estimates had relative RMSEs of 7-9%, and these numbers do not account for sampling variation. For income data based on the five-year pooled ACS, the influence of sampling variation on tract-level estimates is substantial (see Appendix
B for an analysis of how sampling variation affects income inequality estimates based on the ACS).

In sum, Lorenz interpolation, CDF interpolation, and MCIB produced accurate Gini coefficient estimates of both tract and school district-level income distributions. These estimation methods seem to be up to the task of estimating income inequality measures that depend less on the upper tail of the income distribution, such as the Gini. Although CDF produced more accurate estimates of the Theil, Atkinson, and standard deviation than MCIB, Lorenz interpolation yielded more accurate estimates of these measures than CDF interpolation. This reflects the downward bias of CDF interpolation estimates of the upper tail of the income distribution.

2.5 Discussion and Conclusion

I conclude with a discussion of the use cases of Lorenz interpolation, some contexts in which the method should not be applied, and some ways in which the method could be extended. Starting with use cases, Lorenz interpolation is a valuable method for estimating income inequality at geographic levels that are not accounted for in Census microdata but for which Census summary table data is available. For instance, PUMS data only covers a subset of counties in the U.S.\textsuperscript{12} Researchers interested in examining income inequality for all U.S. counties must rely on grouped income data

\textsuperscript{12} County-level PUMS data is only available for counties that can be identified using PUMAs or other lower-level geographic areas (Ruggles et al. 2021).
provided in Census summary tables. For geographies like these, including counties and
census-designated places, using Lorenz interpolation on summary Census data is a
viable way to estimate income inequality.

Lorenz interpolation can also be used to estimate income distributions for
geographies that are not provided in Census data but can be approximated by
aggregating incomes from a geographic level that is provided in the Census. For
example, Ann Owens’ (2016) recent work on U.S. economic segregation, which is
organized around school-district boundaries, uses von Hippel et al.’s (2016) robust
Pareto midpoint estimator, an improved version of the technique of imputing bracket
midpoints for incomes in closed brackets and assigning a Pareto distribution mean to
incomes in the top bracket. For a study like this one, Lorenz interpolation would be a
preferable method for estimating income inequality. Alternatively, researchers
interested in the implications of income inequality for disparities in the quality of local
public services may wish to approximate municipal income distributions by aggregating
tract-level data to the municipal level. For such an analysis, Lorenz interpolation would
yield more accurate estimates of municipal income inequality and should be used in lieu
of Pareto-midpoint estimators and the other interpolation methods discussed in this
paper.

While there are many use cases for Lorenz interpolation, there are also situations
where this method should not be used, either because the PUMS data is sufficient or
grouped data is inadequate. For example, researchers that require income statistics at the MSA or state level should simply use PUMS data, which include geographical information for large regions like MSAs and states. On the other hand, while grouped income data is the only publicly available resource for studying incomes at lower geographic levels like census tracts and blocks, researchers should refrain from estimating certain inequality measures from this data. The errors associated with tract-level Theil, standard deviation, and Atkinson estimates produced in this analysis were large enough to cast doubt on the reliability of tract-level estimates of certain income inequality measures.

The large residuals of these estimates are even more concerning when one considers that Census income data is sample data. Until 2000, this data was collected in the long-form portion of the decennial Census, which is based on a sample covering approximately 18% of the U.S. population (Logan et al. 2018). Since then, income data has been collected in the ACS, which in its five-year form is based on a sample of about 5% of the population. Although the ACS provides error margins that can be used to construct 90% confidence intervals around the frequency estimates of each bin of grouped income data, researchers have yet to develop methods for estimating income inequality that make use of these margins. The general approach has been to ignore
them and work at a high enough geographic level that the error caused by sampling variation is negligible.\textsuperscript{13}

To improve upon the method put forth in this paper, researchers may want to consider using Bayesian methods to produce more reliable income inequality estimates for small areas. For instance, Empirical Bayes might be a viable method to supplement tract-level income data with information from neighboring regions. It should be noted, however, that income data from the U.S. Census is already reweighted to incorporate demographic and other information from neighboring areas. This suggests that shrinking estimates from sparsely populated tracts toward the inequality levels of surrounding areas may have a limited effect on improving the reliability of estimates based on this data. Nonetheless, studies have successfully employed Bayesian methods to improve small area estimates using Census data from other countries’ national censuses (Assuncao et al. 2005; Schmertmann and Gonzaga 2018). These methods may yield better estimators for small regions, particularly for unweighted income data.

\textsuperscript{13} For researchers interested in estimating income statistics for neighborhoods, block groups, or even smaller geographies that are not publicly available, the best course of action is to apply for access to the exact income data through a Federal Statistical Research Data Center (FSRDC). The roughly 30 FSRDCs around the U.S. have resources for handling confidential data, enabling researchers to work directly with Census records. Although this data remedies the problem of measurement error introduced by grouped data, even this approach does not resolve the issue of sampling variation, which can be large for sparsely-populated regions.
In this chapter, I proposed a new method, Lorenz interpolation, for estimating income inequality from grouped income data. I showed that this method produces more accurate and reliable estimates of income inequality and that these performance improvements can be attributed to how the method estimates the upper tail of the income distribution. I also estimated income inequality for tracts and school districts and found that Lorenz interpolation outperforms other methods in estimating income inequality for smaller regions. Finally, I provided some scope conditions for the use of Lorenz interpolation. Although Lorenz interpolation produces accurate Gini coefficient estimates at the tract-level, the method yields insufficiently accurate tract-level estimates of the other inequality measures. However, Lorenz interpolation generated more accurate estimates of the Theil, Atkinson coefficients, and standard deviations at the school-district level. As the influence of the income distribution upper tail on income inequality continues to grow (Piketty and Goldhammer 2014), methods that capture the variance of incomes in this upper tail will become increasingly important.
3. Using Mobile Phone Data to Estimate the Effects of Income and Occupation on Social Distancing during the Early Response to Covid-19

Following the discovery and rapid spread of coronavirus disease 2019 (COVID-19), governments around the world have implemented social distancing measures to prevent people from leaving their homes and to limit human interaction. In the United States, the federal government released the 15 Days to Slow the Spread program, a set of public health guidelines that was developed by the CDC and includes recommendations to physically distance and stay at home whenever possible. Although this initiative coincided with a dramatic rise in social distancing across the country, few studies have examined about how increases in social distancing varied across social and demographic groups during this time.\(^\text{14}\)

One sign that group-level differences in social distancing had emerged following the COVID-19 outbreak is the evidence of demographic and social disparities in COVID-19 prevalence rates (Hooper, Nápoles, and Pérez-Stable 2020; Koh 2020). Although these disparities are partly a result of comorbidity patterns, they also point to differences in how people have altered their movement through society in response to this disease (Courtemanche et al. 2020). Differences in social distancing levels may be especially pronounced in the United States, where high levels of economic inequality contribute to disparities in employment arrangements and economic resources that allow for social distancing. For instance, lower-income workers are less likely to hold jobs that provide paid sick leave, and workers without sick leave are less likely to remain home from work when they are sick (Derigne, Stoddard-Dare, and Quinn 2016).

\(^{14}\) An important exception is a paper by Weill and coauthors (2020), which found a strong connection between income and social distancing levels in a county-level analysis of data tracking people’s daily movements.
Moreover, many jobs held by lower-income workers consist of tasks that must be carried out in person and in close proximity to others, while higher-paying jobs can often be performed remotely (Dingel and Neiman 2020). As a result, the widespread transition to working from home that started in March 2020 may have been a white collar phenomenon. Finally, low-income households are less likely to have savings, second homes, and other economic resources that facilitate changing one’s daily routine in response to new events. One sign of the relative ease with which the rich have socially distanced is the large number of residents from higher-income neighborhoods who out-migrated from Manhattan at the outset of the COVID-19 crisis in New York City (Quealy 2020).

These considerations suggest that occupational status and income were particularly important dimensions along which people’s levels of social distancing varied following the outbreak of COVID-19. In this chapter, I examine how these factors influenced people’s everyday movements during this period. Using a dataset that tracks people’s physical movements through their mobile phone activity, I estimate several models evaluating the effects of income and occupational status on stay-at-home patterns. I find a large and positive effect of neighborhood-level median income on social distancing, including diminished levels of leaving the home to go to work. I also estimate models that include a set of occupational measures based on job categories from The Bureau of Labor Statistics. According to these models, the lowest paying occupations were associated with smaller increases in social distancing, while the highest paying occupations were associated with smaller increases in social distancing, while the highest

---

15 The paucity of lower-paying jobs that can be performed remotely also has implications for the relationship between occupation and COVID-19 prevalence, especially for companies that prioritize assuaging customers’ fears above protecting their employees. This is suggested by a recent spike in OSHA lawsuits, in which many companies – including Target, Amazon, McDonald’s, REI, Cargill, Smithfield Foods, Delta Airlines, Urban Outfitters, and General Electric – have been accused of pressuring employees that have been diagnosed with COVID-19 to hide their diagnoses from their coworkers (Eidelson 2020).
paying jobs experienced the largest jumps in social distancing. This finding is particularly striking given that unemployment in the early stages of the COVID-19 pandemic increased more among people in lower paying occupations (BLS 2020). These results suggest that the relative prevalence of COVID-19 among lower-income households may be attributable in part to an occupational structure that has failed to accommodate lower-income workers. In the absence of employment arrangements enabling working from home or subsidizing reduced levels of work, lower-income workers face a choice between risking exposure to COVID-19 or losing employment. I conclude with a discussion of the policy implications of these economically patterned gaps in social distancing levels.

3.1 Data and Measures

The data in this chapter come from SafeGraph, a company that tracks foot-traffic patterns from over 45 million mobile devices in the United States. This data is based on an opt-in panel of mobile users whose devices are pinged throughout the day. To protect the anonymity of these users, SafeGraph aggregates this data to census block groups (CBG), which are subdivisions of census tracts that contain about 1,400 people on average (U.S. Census 2019). The social distancing patterns analyzed in this study are based on a sample of 212,099 CBGs.

I used two measures to capture changes in social distancing levels. The first, median daily time at home, tracks the amount of time spent at home among devices whose homes fall within the boundaries of a CBG. Mobile devices are assigned to a home if they have been geolocated to a common nighttime location (within a 153 square-meter rectangle) over a six-week period. The second measure follows the proportion of devices exhibiting full-time work behavior. This measure is based on the number of devices that were away from home for a period lasting longer than six hours between 8am and 6pm. Averaging across these daily measures, I computed social
distancing levels for two time periods: the two weeks preceding March 16 (the start of the 15 Days to Slow the Spread initiative), and the three weeks following March 16. To focus on changes in workweek social movement patterns, I chose to exclude weekend social distancing levels from these averages.16

The explanatory and control variables in these models come from the 2014-2018 five-year pooled American Community Survey (ACS). To account for demographic differences among CBGs, I controlled for proportions of male, Black, Asian, and non-white Hispanic residents. I also included several economic measures, including median income, the proportion of residents over 25 with a college degree, the poverty rate, and the unemployment rate. Finally, I added measures for occupational groups (5 categories) and occupational subgroups (25 categories).17

3.2 Analytical Strategy

The effect of median income on social distancing was estimated with several spatial error models (Ward and Gleditsch 2019). The spatial error model is a kind of mixed model that uses a random effect to account for the dependency structure of the residuals associated with spatially clustered observations. The model has the following structure:

\[ y_i = x_i \beta + \varepsilon_i + \lambda w_i \xi_i \]

\[ \varepsilon \sim N(0, \sigma^2 I) \]

16 The exclusion of weekend social distancing data does not change the substantive interpretation of the results presented in this paper.

17 Occupational subgroup controls are based on the 23 major occupational groups from the 2018 Standard Occupational Classification (SOC) system (U.S. Census Bureau 2018). Two of these groups, Healthcare Practitioners and Technical Occupations and Protective Service Occupations, have been further disaggregated to create the 25 occupational subgroups used in this study. The 5 occupational group controls are based on the 6 high-level aggregation titles from the SOC system. One of the 6 high-level titles, Military Specific Occupations, was excluded from the present study.
Where the error term consists of a spatial component $\lambda w_i \xi_i$ and a spatially uncorrelated component $\epsilon_i$. In this equation, $\lambda$ denotes the level of spatial correlation among the residuals. Given that CBGs often resemble neighboring CBGs, this will likely be positive in the models estimated here. A likelihood ratio test giving a low probability that $\lambda = 0$ indicates the necessity of using a model that accounts for spatial clustering. $w_i$ indexes the CBGs that neighbor CBG $i$.

Neighboring CBGs are determined using Queen’s criterion, which groups regions based on whether they share a border or a vertex (Anselin 2020).

Maximum likelihood estimates of the model parameters were computed using the *spdep* package in R (Bivand 2020). The results are organized into three sets of models. The first two sets consist of CBG-level estimates of the effect of median income on time at home and full-time work behavior (spending at least six hours of the day in a single location outside the home) for each of the 100 most populous metropolitan statistical areas (MSAs) in the U.S. For the third set of models, I aggregated the CBG data to the county level and estimated the effect of median income on full-time work behavior for the entire U.S. This third set of models introduces several occupation and occupation group variables. I include these measures to evaluate whether the effect of income on social distancing can be attributed to differences in the jobs held by lower and higher-income mobile device users.

### 3.3 Trends in Time at Home and Full-Time Employment

Figure 5 shows trends in social distancing by income groups, which are divided by lower, middle, and upper income tertiles. Looking first at the plot at the top of the figure, daily minutes spent at home increased for all three groups during the month of March 2020. However, time at home increased more in higher-income CBGs: the gap between the income groups grew as the days passed. The bottom plot displays changes in rates of full-time work behavior, proxied by the
proportion of residents that were away from home for at least six hours of the day. Before March 16th, a greater proportion of devices in higher-income CBGs engaged in full-time work behavior. After the 15 Days to Slow the Spread initiative began, rates of full-time work behavior in higher-income and middle-income CBGs dropped sharply. Full-time work rates from this point onward were comparable across the income groups.

To provide a more detailed picture of how changes in social distancing varied over the course of the day, Figure 6 focuses on changes in hourly stay-at-home patterns in the Philadelphia MSA. The map on the left shows the geographic distribution of lower, middle, and upper income CBGs across the region. This map shows a moderate amount of segregation by income: many

![Graph showing Trends in Daily Time Home and Full Time Employment by Income.](image)

**Figure 5: Trends in Daily Time Home and Full Time Employment by Income.**
Figure 6: Relationship Between Income and Hourly at Home Rate in Philadelphia.
lower-income CBGs can be found on the eastern part of the map (closer to downtown), while higher-income CBGs are in the central and western areas of the map (the suburbs). This pattern is representative of many large MSAs, suggesting that income differences in social distancing may produce social distancing disparities at larger geographic levels, such as census designated places. The plot on the right shows hourly fluctuations in at-home rates. For every 24-hour period, the trend lines peak in the middle of the night and reach their lowest points in the early afternoon. Before social distancing measures were enacted, a lower proportion of residents of high-income CBGs were at home in the middle of the day. This changed in the middle of the month: daytime stay-at-home rates for high-income CBGs went from the lowest to the highest of the three income groups. Notably, income differences in nighttime stay-at-home rates, the peaks in the plot, experienced little change. Low-income CBGs had lower nighttime stay-at-home rates over the entire study period. This suggests the broadening gap in social distancing between high and low-income CBGs reflected in Figure 5 can be attributed to changes in daytime behaviors.

### 3.4 Models of Time at Home and Full-Time Employment

Figure 7 shows regression lines for the effects of income on changes in time at home and full-time employment. Spatial error models were estimated for 100 MSAs, each of which is represented by a grey regression line. The mean predicted value of the outcome across these MSAs is indicated by the blue line, and the two standard-deviation spread of predicted values is represented by the grey shaded region. Looking first at the plot on the left, the positive y-intercepts of the grey regression lines indicate that time at home increased for all 100 MSAs. The positive slopes of these lines show that higher-income CBGs experienced larger increases in time at home for every MSA in the sample. The slope of the blue regression line is 44.23, meaning that
the average effect of a one standard-deviation increase in median income across the 100 MSAs is a roughly forty-four minute increase in the median time spent at home.\textsuperscript{18}

Figure 7: Effect of Median Income on Change in Daily Minutes Spent at Home and Change in Proportion Full-Time Employment (Largest 100 MSAs).

The blue regression line in the plot at the right of the figure further supports a general pattern in which higher-income CBGs experienced larger increases in social distancing during the study period. For most, but not all, MSAs in the sample, higher median income levels were associated with greater declines in full-time employment behavior. That some MSAs show a positive association between income and full-time work behavior may reflect occupational

\textsuperscript{18} This statistic represents the unweighted average effect of income across the MSAs. Weighting by MSA population or number of CBGs does not change the substantive interpretation of the results.
differences among MSAs. In cities with higher employment in the service sector, for instance, layoffs may have countered the effect of transitioning to working from home, resulting in a neutral or even positive relationship between median income and change in full-time employment behavior. Still, the slope of the blue line is negative, indicating an average negative association between neighborhood income and full-time employment behavior across the sample of 100 MSAs.

While the Figures 5-7 show that neighborhood-level income differences explain varying levels of social distancing following the COVID-19 outbreak, this relationship may be driven by occupational differences between high and low-income neighborhoods. Table 3 shows the results of county-level models testing this possibility. Models 1 and 2 include the variables from the previous models, while models 3 and 4 introduce the occupation measures. Specifically, model 3 includes five measures of occupational groups – managerial occupations; service occupations; sales and office occupations; natural resources, construction, and maintenance occupations; and production, transportation, and material moving occupations. Model 4 leaves out the occupational group measures and instead includes measures of the twenty-five occupations within these groups.

According to the models’ Akaike Information Criterion (AIC) scores, each model fits the data better than the previous model. Particularly large drops in the AIC resulted from adding median income to the model (Model 2) and incorporating the occupation measures (Model 4). Conversely, the occupational group measures, which provide less detail than the occupation measures, did little to improve model fit. Comparing models 2 and 4, the inclusion of occupation attenuated the coefficient on median income by 50%, from -.006 to -.004. Still, the effect of income was larger than any of the other coefficients in the final model.
Table 3: Spatial Autocorrelation Models of Change in Proportion Employed Full-Time

<table>
<thead>
<tr>
<th></th>
<th>Change in Proportion Employed Full-Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>Percent Male</td>
<td>.0005**</td>
</tr>
<tr>
<td>Percent Black</td>
<td>.002***</td>
</tr>
<tr>
<td>Percent Asian</td>
<td>.001***</td>
</tr>
<tr>
<td>Percent NW Hispanic</td>
<td>.0002</td>
</tr>
<tr>
<td>Percent College</td>
<td>-.002***</td>
</tr>
<tr>
<td>Poverty Rate</td>
<td>.004***</td>
</tr>
<tr>
<td>Unemployment Rate</td>
<td>.0001</td>
</tr>
<tr>
<td>Median Income</td>
<td>-.006***</td>
</tr>
<tr>
<td>Constant</td>
<td>-.030***</td>
</tr>
<tr>
<td>Occupation Grp Controls</td>
<td>No</td>
</tr>
<tr>
<td>Occupation Controls</td>
<td>No</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>.123</td>
</tr>
<tr>
<td>AIC</td>
<td>-18,779</td>
</tr>
</tbody>
</table>

N=3,217. *p < .1, **p < .01, ***p < .001
Figure 8 looks at the coefficients estimated for each occupation in Model 4. Positive coefficients on the occupation measures indicate occupations associated with less social distancing (smaller decreases in full-time work behavior). The three occupations with positive and significant coefficients are Arts, Design, Sports, and Media; Food Preparation and Serving; and Community and Social Service. According to the Bureau of Labor Statistics, the mean salaries for these occupations were $61,960, $26,670, and $50,480 in 2019 (BLS 2020). The three occupations with the largest significant and negative coefficients – the jobs associated with the largest drops in full-time employment behavior – are Computer and Mathematical; Business and Financial Operations; and Management. These occupations had annual wages of $93,760, $78,130, and $122,480, respectively, in 2019. These numbers suggest that neighborhoods with greater employment in higher paying occupations experienced larger declines in full-time employment behavior, despite the fact that fewer people holding these jobs were laid off during the study period.
3.5 Discussion

The beginning of the COVID-19 outbreak in the United States marked the enactment of a public health campaign to limit the spread of the disease through social distancing. As this study shows, trajectories of social distancing during this time varied widely depending on neighborhood income levels. Increases in time at home were larger among residents of higher-income neighborhoods, while transitioning to working from home was more common among these
residents as well. Furthermore, the relationship between income and working from home is partly a consequence of occupational differences among the residents of high and low-income neighborhoods. Neighborhoods with higher proportions of certain high-paying jobs experienced the largest drops in working outside the home, while neighborhoods with higher rates of the lowest paying jobs witnessed the smallest drops in working outside the home.

The effects of the occupation measures on social distancing levels may reflect differences in the activities associated with high and low-paying jobs. While many lower paying jobs – those in Food Preparation and Serving, for instance – consist of tasks that need to be performed on-site, many higher paying jobs can be completed remotely. As a result, transitioning to working from home has been more common among people with higher-paying positions. This arrangement has forced workers in low-paying positions to choose between risking exposure to COVID-19 and losing their jobs, while enabling employees in higher paying occupations to insulate themselves from the disease without losing employment. Further, many lower paying jobs do not include benefits such as paid sick leave and health insurance that are often provided to those in better paying positions. Without paid sick leave, workers are motivated to conceal concerns about having contracted COVID-19 from their employers. The absence of health insurance also deters workers who fear they have contracted COVID-19 from getting tested for the disease. These factors create dangerous incentives for workers to ignore the signs that they might be sick and continue their usual everyday activities.

Encouragingly, this study suggests that social distancing levels in lower-income communities may be increased by policies directed toward the companies that employ lower-income workers. A mandate requiring that these companies provide paid sick leave to employees with a COVID-19 diagnosis, along with initiatives to facilitate access to free testing in these communities, would go a long way in redressing the institutional factors that contribute to the
social distancing gaps that I have identified here. Alternatively, the government should consider subsidizing businesses that cannot operate remotely. The U.S. took a step in this direction in the early months of the pandemic, providing forgiveness on Payment Protection Program (PPP) loans to businesses that spent a portion of this money on their employees’ salaries. A more aggressive action would be to provide money to small businesses (both employees and their employers) to close their doors temporarily. Particularly for non-essential businesses that cannot follow social distancing guidelines – restaurants, bars, gyms, and concert venues, for example – such temporary subsidies would preserve jobs and businesses while preventing the spread of COVID-19 through lower-income communities and the public at large.

The outbreak of COVID-19 in the U.S. has produced new forms of social inequality, many of which are perpetuated by an economy that distributes money and other resources unevenly through society. I have shown that social distancing is associated with higher levels of income and that this relationship is a function of the jobs held by people at different income levels. These findings contribute to a larger literature documenting the ways in which economic inequality produces negative public health consequences (Wilkinson and Pickett 2009). Finally, I have put forth some policy ideas to promote social distancing in lower-income areas, but more far-reaching reforms are needed to ensure that Americans have health insurance, employment benefits, savings, and other economic resources necessary to adopt new behaviors in the event of a future pandemic.
4. Gentrification, Concentrated Poverty, and Low-Income Homeownership

Homeownership has been cited as one of the most important mechanisms of wealth inequality in the U.S. (Alba and Logan 1992). Houses constitute the largest source of wealth among American families and bring many economic benefits to homeowners. Homeowners with mortgages build savings through monthly mortgage payments, broadening the wealth gap between themselves and renters. Homeowners also receive tax benefits that renters do not, including the ability to deduct mortgage insurance payments on their taxes. In 2017, U.S. homeowners saved a total of $71 billion through such deductions (Desmond 2017). While this number has recently gone down due to changes in the tax code, mortgage deductions still account for tens of billions of dollars today.

Concerned by growing inequality between owners and renters, researchers and policy analysts have sought to identify factors that facilitate homeownership among low-income households. Many studies have shown that having access to economic resources – a decent income, a substantial savings, a good credit history – increases one’s chances of becoming a homeowner (Acolin et al. 2016). Fewer analyses have considered how the spatial distribution of these resources across the urban environment may matter for homeownership. Given that living near housing that is available for purchase increases one’s chances of becoming a homeowner, the spatial distribution of household economic resources likely plays a role in promoting homeownership access among low-income households. However, cities where lower-income families live closer to owner-occupied housing will only foster low-income homeownership if this accessible housing is also affordable. Thus, the spatial arrangements that foster low-income homeownership are not self-evident.
Two relevant areas of research offer contrasting accounts of how space reinforces inequality in urban areas. On one hand, a large body of literature on concentrated poverty, a consequence of economic segregation, has documented how high-poverty neighborhoods deplete lower-income households of valuable economic and social resources (Dwyer 2007; Reardon and Bischoff 2011). Such deprivations may obstruct the path to homeownership for lower-income residents. On the other hand, studies of gentrification have shown that the migration of higher-income households into lower-income neighborhoods, a form of “economic desegregation”, can hurt lower-income residents’ homeownership prospects by raising housing prices (Atkinson 2000; Vigdor 2002; Newman and Wyly 2006). Particularly in cities with a limited supply of land, gentrification may depress low-income homeownership rates by depriving potential homeowners of affordable housing options (Matlack and Vigdor 2008).

In this chapter, I attempt to reconcile these perspectives on space and inequality. Using U.S. Census data from 1990 to 2018 and Home Mortgage Disclosure Act (HMDA) mortgage data covering the same period, I show that cities where income segregation declined experienced decreases in low-income homeownership. This relationship was driven in part by increases in gentrification, measured as the proportion of mortgages purchased by higher-income homeowners in lower-income neighborhoods. This trend has accelerated since 2010 and is particularly prevalent among metropolitan areas in California, where lower-income households have been shut out of mortgage markets almost entirely. At the same time, metropolitan areas with increasing rates of concentrated poverty also had declining levels of low-income homeownership. Ultimately, these findings support a more nuanced view of how the spatial distribution of income shapes housing inequality. While income segregation can depress low-income homeownership by creating regions that are bereft of certain place-based resources, income integration in the form of “urban renewal” can also hurt low-income families’ homeownership prospects.
4.1 U.S. Homeownership from 1990 to 2018

As shown in Figure 9, the U.S. housing market since 1990 can be characterized by two periods. In the first period, which ended with the housing market collapse in 2007, homeownership rates for U.S. families rose from roughly 64 percent to 68 percent. In the second period, homeownership rates dropped, reaching 63 percent in 2016. The rate has gradually climbed since then, reaching 65 percent in the middle of 2018. Homeownership rates for higher-income (above the median) and lower-income (below the median) families experienced similar shifts in homeownership, but a sizeable gap persists between them. Specifically, homeownership rates ranged between 78.5 and 84.5 percent for higher-income families and between 48 and 54 percent for lower-income families.

Figure 9: Homeownership Rates for All Families, 1990-2018.

These periods can be characterized by distinct policy regimes governing the underwriting of mortgages to would-be homebuyers. In the first period, the federal government passed several laws that increased the availability of mortgage credit to low-income and minority households (Bhutta 2012). The 1992 Federal Housing Enterprise Financial Safety and Soundness Act
required that government-sponsored enterprises (GSEs) Fannie Mae and Freddie Mac, which operated the secondary mortgage market and provided credit to potential homebuyers, meet certain goals in making mortgage credit accessible to lower-income families and neighborhoods (HUD 1998). To further expand mortgage credit access to underserved populations, GSEs relaxed underwriting standards for families with incomes below the median (Quercia, McCarthy, and Wachter 2003). The new standards included reduced down payment requirements, income-to-mortgage payment ratios, and total debt-to-mortgage payment ratios. On the industry side, banks made similar adjustments, offering mortgages with reduced or even no down payments beginning in the early 2000s (Rosenthal 2002). These changes were made possible by the securitization of mortgage credit, which protected banks from the risk associated with these new standards (McCoy, Pavlov, and Wachter 2009).

By the mid-2000s, lenient housing constraints were offset by rising housing prices, causing low-income homeownership rates to plateau. When the housing bubble burst in 2007, large numbers of low-income families foreclosed on their homes, and tighter mortgage underwriting standards were re-introduced (Acolin et al. 2016). These changes caused the low-income homeownership rate to decline to mid-1990s levels. In recent years, however, low-income homeownership has rebounded despite mortgage underwriting standards remaining more stringent than they were during the housing market boom of the late 1990s and early 2000s.

Some scholars have interpreted the volatility of the housing market over the past several years as evidence that homeownership is a poor vehicle for wealth accumulation among lower-income households (Shlay 2006). However, Freeman and Quercia (2014) have shown that lower-income families that owned homes during the housing market collapse accumulated more wealth than lower-income families that rented. Studies of the conditions that facilitate access to homeownership have focused on the household-level characteristics that shape homeownership
opportunities (Acolin et al. 2016). A key finding from this research is that household wealth is a more significant constraint on low-income homeownership than savings or credit history (Linneman and Wachter 1989). Scholars have also examined demographic predictors of homeownership. Haurin, Hendershott, and Wachter (1997) found that younger and minority households are more likely to have financial constraints that limit homeownership. Moreover, financially constrained minority households have fewer homeownership opportunities than similarly constrained non-minority households (Gyourko, Linneman, and Wachter 1999).

In contrast to the large literature on household-level predictors of homeownership, research on the environmental factors that constrain and facilitate homeownership is limited. One exception is the handful of studies on the supply of affordable housing (Mayer and Somerville 1996; Collins, Crowe, and Carliner 2002). Downward filtering is a key concept in this body of research (Grigsby 1963). Also referred to as “trickle-down housing”, downward filtering occurs when higher-income residents move into new homes, leaving behind older houses that “filter down” to lower-income households at more affordable prices. Some research suggests that macro-level developments such as rising income inequality (Matlack and Vigdor 2008) and the rate of household formation (Galster 1996) can interfere with the process of downward filtering, causing affordable housing shortages and obstructing homeownership access for some communities (Malpezzi and Green 1996).

Aside from a few studies on the supply of affordable housing, researchers have paid little attention to how low-income homeownership is structured by the spatial distribution of economic resources across the urban environment. Two areas of research, one on the effects of living in high-poverty neighborhoods and one on the consequences of gentrification, have implications for this relationship. In the next section, I draw from these areas to build hypotheses.
4.2 How Income Segregation Constrains Low-Income Homeownership: Concentrated Poverty, Spatial Mismatch, and Land-Use Regulations

How are opportunities for low-income homeownership constrained by levels of income segregation within metropolitan areas? Evidence indicates that high-poverty neighborhoods, which are endemic to segregated areas, reduce opportunities for upward mobility (Chetty and Hendren 2018). Studies of the economic effects of the Moving to Opportunity experiment, in which 4,604 randomly-selected families were relocated from high-poverty to low-poverty neighborhoods, have found that children who moved to better neighborhoods had higher college attendance rates and adulthood earnings (Chetty, Hendren, and Katz 2016). These factors influence residents’ chances of becoming homeowners, suggesting that living in a high-poverty neighborhood obstructs the path to homeownership.

Evidence of a relationship between neighborhood disadvantage and mobility raises the question of how high-poverty neighborhoods limit opportunities for economic advancement. One important mechanism is the spatial mismatch between these neighborhoods and key economic resources, many of which are important for homeownership. According to the spatial mismatch hypothesis, the geographic isolation of lower-income black households in the central city and the movement of low-skilled jobs into the suburbs has produced persistent black unemployment in inner-city neighborhoods (Kain 1968; Mouw 2000; Gobillon, Selod, and Zenou 2007). To the extent that these trends operate similarly along class lines, economic segregation may limit lower-income residents’ access to employment opportunities. Furthermore, economic segregation may contribute to an unevenness in the spatial distribution of key housing market actors, including mortgage lenders. This makes gaining access to mortgage credit more difficult for residents of lower-income areas (Bhutta 2012).
In addition to depriving lower-income areas of important resources, economic segregation concentrates resources in areas whose residents have an interest in solidifying their place-based advantage (Logan and Molotch 1987). Residents of higher-income suburbs in economically segregated metropolitan areas frequently support restrictive zoning laws to maintain property values and limit free riding on public amenities (Rothwell and Massey 2010; Trounstine 2020). These measures suppress low-income homeownership rates by confining lower-income households to parts of metropolitan areas that have fewer owner-occupied housing units. They also interfere with the process of downward filtering by preventing the construction of less expensive housing (Malpezzi and Green 1996). This produces communities in which the homes that filter remain prohibitively expensive to lower-income households.

In sum, economic segregation can limit lower-income homebuying by depriving lower-income communities of resources that promote homeownership. Economic segregation may also prompt higher-income communities to pass laws that limit the construction and downward filtering of affordable housing. These considerations form the basis of my first hypothesis, that spatially concentrated poverty confines lower-income residents to regions where homeownership opportunities are scarce.

Hypothesis 1. Metropolitan areas with rising numbers of high-poverty neighborhoods experienced declining rates of low-income homeownership.

4.3 How Income Segregation Protects Low-Income Homeownership: Gentrification and Increased Housing Costs

In a study of the relationship between racial segregation and black homeownership, Freeman (2005) pointed out that racial segregation may foster black homeownership by depressing housing values in segregated neighborhoods. Although Freeman found mixed evidence for this, his insight suggests one way that economic segregation may promote low-
income homeownership. Namely, economic segregation may keep housing prices in segregated neighborhoods low enough that lower-income residents can afford to buy houses in these neighborhoods. This effect will be especially pronounced in cities where the supply of land is limited (Matlack and Vigdor 2008). In these cities, land restrictions will lead to affordable housing shortages unless conditions exist that insulate lower-income homebuyers from higher-income buyers’ greater ability to pay for housing. Economic segregation may create such conditions by concentrating lower-income households in neighborhoods into which higher-income residents are unwilling to move. Similarly, rising income inequality may reduce low-income homeownership opportunities by prompting developers to cater more to higher-income households (Dwyer 2007). In the context of rising income inequality, neighborhood-level differences in housing quality and value may enable lower-income households to remain in the market for owner-occupied housing.

Further support for the idea that economic segregation prevents low-income homebuyers from being priced out of the housing market can be found in the literature on gentrification and residential displacement. Multiple studies have shown that gentrification increases housing costs for families in gentrifying neighborhoods (Vigdor 2002; Freeman and Braconi 2004). One reason for this is that few lower-income households out-migrate in response to gentrification. In his study of gentrification in Boston, Vigdor (2002) reasoned that the improvements that higher-income residents brought to gentrifying neighborhoods, along with the costs associated with moving to more affordable neighborhoods, combined to keep incumbent residents in gentrifying neighborhoods. Taken together, these studies suggest that gentrification can make homeownership a less attainable goal for incumbent residents, even if the migration of higher-income residents depresses housing prices in other regions of the metropolis.
In sum, research on gentrification suggests that economic integration may constrain low-income homeownership by equalizing housing prices across neighborhoods, attenuating neighborhood-level differences that may keep housing affordable to lower-income homebuyers. Furthermore, gentrification creates a spatial mismatch between the location lower-income households and affordable housing. Specifically, residents of gentrifying neighborhoods may accept increased housing costs rather than incurring the costs of out-migrating to more affordable regions. By interfering with the process of downward filtering and driving up housing prices for incumbent residents, gentrification is expected to have a constraining impact on low-income homeownership.

Hypothesis 2: Metropolitan areas with increasing rates higher-income homebuyers migrating into lower-income neighborhoods experienced declining rates of low-income homeownership.

4.4 Data and Measures

The homeownership data used in this chapter come from the Home Mortgage Disclosure Act (HMDA) datasets. Originally enacted in 1975, the HMDA requires most U.S. financial institutions to disclose certain loan-level details about mortgages. Relevant loan-level data include the census tract, county, and state in which the mortgage was created. Additionally, a 1989 amendment was passed requiring that institutions release the household incomes of mortgage holders. HMDA data were used for four time periods: 1990 (based on the 1990-1992 datasets), 2000 (the 2000-2002 datasets), 2010 (the 2010-2012 datasets), and 2017 (the 2017-2019 datasets). For each period, the data was limited to include only originated mortgages associated with owner-occupied home purchases. The final sample consists of 3,643,384

Data for sociodemographic features of U.S. metropolitan areas were collected from the U.S. Census. The 1990 and 2000 datasets come from the 1990 and 2000 Decennial censuses, the 2010 data is from the 2010 Decennial Census and the 2008-2012 American Community Survey, and 2017 data comes from the 2014-2018 American Community Survey. One challenge in using Census data for longitudinal analyses is that the geographic entities in the Census have boundary definitions that change over time. For consistency across time periods, tract definitions were harmonized to the 2010 boundary definitions using the Longitudinal Tract Database (Logan, Xu, and Stults 2014). Metropolitan area boundaries are based on 2003 definitions from the Office of Management and Budget (Bureau of the Census 2016). For consistency, a crosswalk was used linking counties to metropolitan areas. These counties, the boundaries of which mostly remained static from 1990 to 2017, were used to represent a given metropolitan area for the entire period. I limited the final sample to include only the 100 largest metropolitan areas in 2010.

4.4.1 Dependent and Independent Variables

The dependent variable, lower-income homeownership, was computed by dividing the number of mortgages belonging to lower-income households by the number of lower-income households in a metropolitan area and multiplying this by 100,000 to get the number of mortgages per 100,000 lower-income households. Lower-income households are defined as

19 The Census publishes changes to county boundaries on their website (https://www.census.gov/programs-surveys/geography/technical-documentation/county-changes.html). Only one boundary change mattered for defining the metropolitan areas in the sample: between 1990 and 2000, Miami and Dade counties were merged into Miami-Dade county.
households making less than the median household income of the metropolitan area. This follows the definition used by the Department of Housing and Urban Development (Shlay 2006).

The first independent variable, income segregation, was measured using the rank-order information theory index, $H$ (Reardon and Bischoff 2011). $H$ measures the nonrandom sorting of household incomes across neighborhoods of a metropolitan area. One nice feature of $H$ is that it captures changes in the ranks of incomes – defined in this case by the income brackets provided in the Census – making this measure less sensitive to changes in the shape of the underlying income distribution. This makes $H$ less responsive to shifts in levels of income inequality, which ensures that income segregation estimates are not confounded by increasing income inequality. Researchers have recently expressed concern that measures like $H$ that are based on variance proportions may be upwardly biased when estimated from small samples (Logan et al. 2018). To address these concerns, I developed a tool for producing sample-adjusted $H$ estimates in R, following a recent paper by Reardon and coauthors that provides a method for computing bias-corrected $H$ estimates (Reardon et al. 2018).20

The next independent variable, the concentrated poverty rate, was measured as the proportion of lower-income residents living in a high-poverty neighborhood. Following Jargowsky (2003), high-poverty neighborhoods were defined as census tracts with at least 40% of households living below the poverty line. While the poverty line is typically defined as half the national median income, I instead defined poverty as half the metropolitan area median income. This distinction accounts for differences in the cost of living across metropolitan areas.

20 R code for the implementation of the bias-corrected rank-ordered information theory index can be found at https://github.com/andrewcarr24/rank_info/blob/master/rank_order_information_index_function.R.
The third independent variable, gentrification, was measured as the proportion of houses purchased by higher-income homebuyers that are located in lower-income neighborhoods. This can be interpreted as the propensity of higher-income homebuyers to move into a lower-income neighborhood. Higher and lower-income households were defined as being above or below the median income of the metropolitan area. Measuring gentrification in reference to a proportion of homebuyers controls for changes in the total number of mortgage purchases over time, which prevents gentrification from being confounded by fluctuations in the housing market. One limitation of this measure is its insensitivity to income differences among households defined as having lower and higher incomes. To account for this, I ran several models using gentrification measures based on different income thresholds – for example, using below 25 percent and above 75 percent splits to define lower and higher income households. The results of these analyses were substantively identical to those based on the income thresholds used in this paper.

4.4.2 Control Variables

Studies have shown that racially segregated metropolitan areas tend to have lower levels of minority homeownership (Flippen 2001; Freeman 2005). To isolate the effect of income segregation from trends in racial segregation, I included a measure of black-white segregation in the full model. I also controlled for several demographic characteristics. Studies of the subprime mortgage crisis indicate that racial and ethnic distributions influence metropolitan area homeownership patterns (Fischer and Lowe 2015). To account for this, I controlled for proportions of Black, Asian, and Hispanic residents. I also included a control for proportions of foreign-born residents across metropolitan areas (Kuebler and Rugh 2013). Additionally, a variable for sex (proportion male) was included in the models. Given that married couples are more likely to become homeowners, a measure of the proportion of the adult population that is married was also included in the models. Finally, to account for differences in age distributions
across metropolitan areas, I included two controls, one for the proportion of residents under 18 and one for the fraction of residents over 69.

Along with demographic controls, I also included several economic measures. Matlack and Vigdor (2008) argued that income inequality can limit low-income homeownership in cities with limited housing. I included a measure of the metropolitan-level Gini coefficient to account for this. To generate Gini estimates, I used the Lorenz interpolation discussed in chapter 2. This technique estimates income inequality from a Lorenz curve, which is approximated from the grouped income data provided in the Census. Lorenz interpolation was implemented using the lorenz R package, which is available on CRAN (Carr 2020). In addition to income inequality, I added controls for metropolitan area population, logged median income, proportion of residents with a bachelor’s degree or higher, and the unemployment and poverty rates.

### 4.4.3 Analytical Strategy

The effects of income segregation, concentrated poverty, and gentrification on low-income homeownership were estimated with a series of fixed effects models of metropolitan area-level changes in homeownership. Fixed effects models are insensitive to time invariant differences across metropolitan areas that may confound the estimated effects of the independent variables in a pooled-OLS model (Halaby 2004). Furthermore, given that metropolitan boundaries are fairly static over time, the data has a panel structure that lends itself naturally to a fixed effects specification. To account for changes in the housing market across metropolitan areas, I also included dummy variables for each time period in these models.

I made some robustness checks of my models. To account for heteroskedasticity, I used Huber White standard errors. To check for multicollinearity, I computed the variance inflation factors (VIFs) associated the key independent variables. The VIFs associated with these coefficients did not exceed 2.5, suggesting that multicollinearity was not a concern (Allison
Finally, studies have shown that part of the rise in low-income and minority homeownership in the 1990s and early 2000s resulted from the predatory mortgage lending practices that targeted these groups (Rugh and Massey 2010). Households holding subprime mortgages that proliferated during this time experienced higher foreclosure rates during the housing market collapse, raising the question of whether homeownership improves economic outcomes for lower-income households. I address this concern by checking the robustness of the results to the omission of subprime mortgages from the HMDA data. In this analysis, mortgages are defined as subprime if they have an interest rate that exceeds 3% of the rate on a comparable mortgage. Finding no meaningful effect from removing subprime mortgages from the analysis, I present results based on the full data including subprime mortgages here.

4.5 Results

Table 4 provides descriptive statistics for each of the four time periods. Low-income homeownership rose from 321 to 761 mortgages per 100,000 lower-income households between 1990 and 2000. Following the trends in Figure 9, this number dropped between 2000 and 2010 and rose again between 2010 and 2017. Income segregation also seesawed from one decade to the next, dropping between 1990 and 2000, rising between 2000 and 2010, and then dropping again from 2010 to 2017. These trends mirror the results from Reardon et al.’s (2018) analysis of bias-corrected $H$, though the exact numbers differ slightly.

Turning to concentrated poverty estimates, these numbers show an increase in concentrated poverty from 1990 to 2000. This disagrees with Jargowsky’s (2003) finding that the

---

21 The HMDA requires that lenders include this information in the mortgage data.
22 This is likely a result of differences in the underlying sample. Reardon et al.’s (2018) sample consists of 113 metropolitan areas that had populations of at least 500,000 in 2007.
Table 4: Descriptive Statistics of Measures for Homeownership Models.

<table>
<thead>
<tr>
<th></th>
<th>1990</th>
<th>2000</th>
<th>2010</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DV</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low-Inc</td>
<td>321.0</td>
<td>761.0</td>
<td>506.0</td>
<td>769.0</td>
</tr>
<tr>
<td>Homeownership</td>
<td>(131)</td>
<td>(306)</td>
<td>(167)</td>
<td>(302)</td>
</tr>
<tr>
<td><strong>IVs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income Segregation</td>
<td>0.094</td>
<td>0.092</td>
<td>0.098</td>
<td>0.093</td>
</tr>
<tr>
<td>Rate</td>
<td>(0.024)</td>
<td>(0.021)</td>
<td>(0.020)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>Concentrated Poverty Rate</td>
<td>0.130</td>
<td>0.140</td>
<td>0.187</td>
<td>0.172</td>
</tr>
<tr>
<td>Gentrification</td>
<td>0.470</td>
<td>0.488</td>
<td>0.428</td>
<td>0.453</td>
</tr>
<tr>
<td></td>
<td>(0.076)</td>
<td>(0.069)</td>
<td>(0.057)</td>
<td>(0.065)</td>
</tr>
<tr>
<td><strong>CVs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>1.56e6</td>
<td>1.78e6</td>
<td>1.96e6</td>
<td>2.07e6</td>
</tr>
<tr>
<td>BW Segregation</td>
<td>0.603</td>
<td>0.569</td>
<td>0.526</td>
<td>0.533</td>
</tr>
<tr>
<td>Gini</td>
<td>0.424</td>
<td>0.450</td>
<td>0.453</td>
<td>0.461</td>
</tr>
<tr>
<td>Log Median Income</td>
<td>10.4</td>
<td>10.7</td>
<td>11.0</td>
<td>11.1</td>
</tr>
<tr>
<td>College Prop</td>
<td>0.210</td>
<td>0.251</td>
<td>0.294</td>
<td>0.324</td>
</tr>
<tr>
<td>Unemployment Rate</td>
<td>0.063</td>
<td>0.058</td>
<td>0.095</td>
<td>0.059</td>
</tr>
<tr>
<td>Poverty Rate</td>
<td>0.108</td>
<td>0.116</td>
<td>0.145</td>
<td>0.136</td>
</tr>
<tr>
<td>Male Prop</td>
<td>0.480</td>
<td>0.490</td>
<td>0.490</td>
<td>0.490</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.008)</td>
<td>(0.007)</td>
<td>(0.007)</td>
</tr>
</tbody>
</table>
number of high-poverty neighborhoods dropped 25 percent between 1990 and 2000. One reason for this discrepancy is that Jargowsky’s measure was based on the national poverty rate, while the measure used in this study is based on metropolitan area median income. Concentrated poverty estimates also increased significantly since 2000, a result that parallels prior research on this topic (Kneebone 2017). Finally, the gentrification measure agrees with recent research on macro-level shifts in U.S. gentrification, peaking in the 1990s, declining in the 2000s, and then rebounding in the 2010s (Martin n.d.).

Table 5 displays the results from five fixed effects models: a baseline model with only year fixed effects, a model adding income segregation to the baseline model, two models...
incorporating the concentrated poverty and gentrification measures from hypotheses 1 and 2, and a full model with controls. As shown in models 2-5, income segregation had a positive and substantial effect on low-income homeownership. According to the standardized coefficient on income segregation, a one standard deviation increase in income segregation was associated with a 105 mortgage per 100,000 household increase in low-income homeownership. The coefficients on income segregation were statistically significant in all models, which fit the data better than the baseline model, as evidenced by the adjusted $R^2$ associated with each model.

Model 3, which adds a measure of concentrated poverty to the previous model, provides evidence supporting the first hypothesis. Rising concentrated poverty had an attenuating effect, significant at the .001-level, on low-income homeownership. One way to interpret this effect is that a 1% increase in concentrated poverty, the proportion of lower-income households living in a neighborhood where at least 40% of households made less than half the metropolitan area median income, was associated with a reduction in lower-income homeownership of about 11 mortgages per 100,000 households. Alternatively, a one standard deviation increase in concentrated poverty was associated with a decrease in low-income homeownership of roughly 72 mortgages per 100,000 households. Considering the scale of the homeownership measure – average low-income homeownership ranged from 321 to 769 between 1990 and 2017 – this represents a substantial change.

Model 4 shows the effects of gentrification and income segregation on low-income homeownership. The results provide weak support for the second hypothesis. Although gentrification exerts a negative effect, this effect narrowly fails to reach statistical significance at the .05-level. However, the coefficient on gentrification becomes significant in Model 5, which includes both the concentrated poverty measure and control variables. According to Akaike’s Information Criterion, the combination of variables that best fit the data includes gentrification,
Table 5: Fixed Effects Models of Income Segregation, Concentrated Poverty, and Gentrification.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income Segregation</td>
<td>2868*</td>
<td>4,272***</td>
<td>2,427*</td>
<td>5,706***</td>
<td>4,908***</td>
</tr>
<tr>
<td>Concentrated Poverty</td>
<td>-1,129***</td>
<td>-946***</td>
<td>-924***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gentrification</td>
<td>-443</td>
<td>-491*</td>
<td>-518*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=400. *p &lt; .1, **p &lt; .01, ***p &lt; .001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Income segregation, concentrated poverty, and four control variables: logged median income and proportions of male residents, foreign-born residents, and residents over 69.

To identify the metropolitan areas that contributed most to the effects of concentrated poverty and gentrification, I reran the best fitting model with different resamples of the underlying data. For each metropolitan area, I omitted the row of the data associated with that metropolitan area and drew 100 samples with replacement from the remaining 99 rows. Next, I fit a model to the resampled data and another model that included the omitted metropolitan area. I calculated the differences between the coefficients on concentrated poverty and gentrification across these two models to get estimates of the effects of omitting the metropolitan area on these coefficients. Finally, I repeated these steps 200 times to compute bootstrapped confidence intervals of these effects.
Figure 10 shows the 15 metropolitan areas that had the largest absolute effects on gentrification and concentrated poverty. The metropolitan areas that mattered most for gentrification are in the left plot and those that mattered most for concentrated poverty are in the right plot. The error bars in these plots are 95 percent confidence intervals based on the bootstrapped samples of coefficient differences. Looking first at the gentrification plot on the left, ten of the fifteen metropolitan areas that contributed most to the negative effect of gentrification are located in the West, and seven of these are located in California. Each of the top fifteen metropolitan areas contributed to the coefficient on gentrification. These contributions make up between 4 and 20 percent of gentrification’s effect.

Figure 10: MSA-level Contributions to Gentrification and Concentrated Poverty.
One reason that California is so well represented on this ranking may be the mounting housing crisis in the Bay State. According to a 2015 report, California ranked 49th in the nation in number of housing units per capita and meeting the demand for housing would require the construction of 3.5 million new homes by 2025 (McKinsey 2016, p. 3). Housing shortages are most severe in California’s largest metropolitan areas, where growing income inequality and limited land have prompted developers to cater increasingly to higher-income families, even in neighborhoods that these families had previously avoided.

Figure 11 visualizes this transformation with two maps comparing the Los Angeles housing markets in 2010-2012 and 2017-2019. The red dots on the map represent mortgages belonging to lower-income households, and the blue dots on the map are mortgages belonging to higher-income households. In the underlying data, mortgages have been geolocated to the tract level. Each mortgage was plotted by selecting a random point inside the census tract in which the mortgage is located.

Comparing the map on the left to the map on the right, certain regions in which mostly lower-income households had purchased mortgages in 2010 had been crowded out by mortgages bought by higher-income households by 2017. This change is especially pronounced in southeast Los Angeles. In this region, cities like Inglewood and Compton, where many lower-income households had purchased mortgages in 2010, have seen shifts in their housing markets toward higher-income homebuyers. As this figure suggests, low-income homeownership in Los Angeles declined from 234 in 2010 to 106 in 2017. During this time, the gentrification measure increased from .51 to .63. These trends are representative of changes among metropolitan areas in California, where average levels of homeownership dropped by 115 and gentrification increased by .09 between 2010 and 2017. Moreover, these changes run counter to trends in the nation as a whole, where mean low-income homeownership increased by 210 and gentrification increased by
only .025 between 2010 and 2017. After rerunning the parsimonious model with California metropolitan areas removed, the gentrification measure lost statistical significance, suggesting that the effect of gentrification is driven by these metropolitan areas.

To summarize, the results of these models provide strong support for the first hypothesis, that concentrated poverty constrained low-income homeownership, and weaker support for the second hypothesis, that gentrification constrained low-income homeownership. Despite the small effect size of the coefficient on gentrification in the best fitting model, gentrification substantially reduced low-income homeownership access in California’s cities. This finding makes sense given the growing housing crisis in the Bay state. Finally, the largest standardized effect in the parsimonious model was income segregation. This may be due to a gentrification effect that is not
captured by the gentrification measure, or it may reflect other mechanisms through which income segregation facilitates low-income homeownership.

4.6 Discussion

This chapter provides empirical support for two processes through which the organization of economic resources across urban areas has been theorized to reinforce social inequality. The analysis showed that gentrification depressed homeownership for low-income households, particularly in California where the supply of affordable housing is increasingly limited. I also found evidence of a connection between concentrated poverty and low-income homeownership, which surged in cities where concentrated poverty declined in the 1990s and 2010s.

The results presented here offer a counterpoint to a spate of recent studies showing that income segregation perpetuates social inequality (Reardon and Bischoff 2011; Owens 2016). One form of declining income segregation, gentrification, attenuates low-income homeownership by threatening segments of the housing market that cater to low-income homebuyers. Many cities in the analysis also experienced low-income homeownership growth in less developed regions at the periphery of the metropolitan area. Particularly in the context of rising income inequality, it appears that low-income homeownership flourishes in metropolitan areas that have regions that higher-income households were unwilling to move. While rising low-income homeownership in these cities may reflect place-based inequities that exacerbate certain forms of social inequality (Owens 2016), one advantage of these inequities is that they keep lower-income households in the market for owner-occupied housing. Researchers should therefore refrain from conceptualizing income segregation as synonymous with social inequality. In cases where income segregation reflects the presence of working and middle-class neighborhoods or the availability of land for development, this construct may reflect the existence of lower-income communities that are insulated from the relative buying power of higher-income households.
While the overall effect of income segregation was to preserve the lower-income segment of the housing market, one form of income segregation – the concentration of the poorest households in a subset of the neighborhoods in a metropolitan area – had the opposite effect. This supports classic hypotheses of how spatial mismatch and the political economy of urban spaces can reinforce the disadvantage of living in a poor neighborhood (Kain 1968; Logan and Molotch 1987). Taken together, these findings support a more nuanced understanding of how space structures inequality. Although some forms of economic separation enhance homeownership opportunities for lower and middle-class households, the segregation of the poor isolates residents from the resources necessary to obtain a mortgage. This is another mechanism through which concentrated poverty contributes to cycles of intergenerational disadvantage, as documented in the neighborhood effects literature (Chetty and Hendren 2018).

Although these results provide suggestive evidence that the effects estimated here may reflect causal relationships, one cannot make valid causal inferences from these models. First, in the absence of a valid instrument for income segregation, I cannot rule out the possibility that the effects estimated in this paper reflect the effect of homeownership on income segregation. While the use of lagged variables would remedy this issue somewhat, lagged variables have been shown to bias effect estimates in panel models that rely on only a few time periods (Vaisey and Miles 2017). Second, the measurements of concentrated poverty and gentrification used here have some necessary shortcomings. Measuring concentrated poverty in relation to metropolitan area median income ignores neighborhoods in poor cities that might otherwise be defined as high-poverty in relation to the national median income; conversely, this measure may falsely label non-poor neighborhoods in rich cities as high-poverty neighborhoods. The advantage of defining concentrated poverty relative to the metropolitan area median income, however, is that doing so makes the measure unresponsive to rising differences in the cost of living across metropolitan areas. Finally, the gentrification measure has some issues. The thresholds designating high-income households and low-income neighborhoods used to construct the gentrification measure
are necessarily arbitrary. That said, the results presented here are robust to setting these thresholds at varying levels. Gentrification scholars will also recognize that the measure used here captures only one aspect of what is typically viewed as a multidimensional construct. Other operationalizations of gentrification incorporate data on educational attainment, racial, age, and occupational distributions, proximity to the central city, and the poverty rate to identify gentrifying neighborhoods (Wyly and Hammel 1998; Freeman 2005). Given that the analysis is motivated by the expectation that certain forms of income desegregation are harmful to low-income households, I have opted to concentrate only on the income dimension of gentrification here.

The results presented here have implications for recent policy efforts directed at increasing the supply of affordable housing. When state-led initiatives to increase affordable housing come into conflict with local zoning restrictions, an unintended consequence may be the construction of housing that exacerbates the issue of concentrated disadvantage. Efforts at removing local zoning restrictions, on the other hand, run the risk of making lower-income communities vulnerable to gentrification, a possibility that derailed the enactment of Senate Bill 827, a bill requiring cities to allow midrise-apartment construction near transit lines, in California (Bliss 2019). Ultimately, policies aimed at expanding affordable housing to low-income households must thread a needle between reinforcing patterns of spatial inequality and protecting lower-income communities from a deregulated housing market.
5. Conclusion

In this work, I developed a method for estimating income inequality from publicly available income data from the Census. I demonstrated that this method produces more accurate estimates of several income statistics, and that reliable measures of certain income inequality estimates can be obtained at the school district level. Then, using Census data, mobile phone tracking records, and mortgage loan data, I estimated several models in which I analyzed the effect of income on two phenomena related to social stratification: social distancing and homeownership. In chapter 3, I showed that income and occupational differences were salient predictors of social distancing levels in the early stages of the response to COVID-19. This illustrates one way that economic inequality may contribute to health disparities, creating an environment in which those with better paying jobs have exclusive access to the resources necessary to socially distance. In final chapter, I analyzed how the geographic distribution of income relates to rates of low-income homeownership. Results from fixed-effects models showed that increases in within-neighborhood inequality may depress rates of low-income homeownership by rising housing prices in the neighborhoods into which higher-income homebuyers migrate.

Moving forward, I plan to continue researching the spatial dynamics of income and other economic resources influence economic opportunities for lower-income individuals and families. The final chapter of this dissertation illustrates one way that rising inequality on a smaller geographic scale may produce effects that are impossible
to identify in studies that examine inequality at higher geographic levels of analysis. By implementing the method that I proposed in Chapter 2, I hope to study how rising inequality within the metropolis, municipality, school district, and neighborhood has shaped the distribution of social, educational, and economic resources within these places. Tracing these lower-level effects is crucial to understanding the whole story of how rising economic inequality has reshaped the economic and cultural divisions among the social classes.
Appendix A. Probability Density Function of Lorenz Interpolated Income Data

Figure 12 illustrates the relationship between the probability density function, the cumulative density function, and the Lorenz curve. These functions are based on the income distribution of the Birmingham, Alabama metropolitan area. \( P(x) \) is the probability density, \( F(x) \) is the cumulative density (the probability that a randomly-selected household has an income that is less than or equal to a given income), and \( L(x) \) is the cumulative aggregate income share.

![Figure 12: PDF, CDF, and Lorenz Curve.](image)

This term is computed by aggregating incomes below a certain quantile of the income distribution and dividing this by the total income. The CDF of the income distribution can be derived from the Lorenz curve in two steps. First, the derivative of the function for the Lorenz curve is computed and multiplied by the distribution mean. This converts the Lorenz curve into an inverse cumulative density function. Next, this
function is inverted to generate the cumulative density function, the derivative of which is then computed to get the probability density function of the income distribution.

Lorenz interpolation fits quadratic functions to the points associated with the closed bins of the grouped income data and a cubic function to the bin at the top of the income distribution. The steps outlined above show that using quadratic functions to interpolate the Lorenz curve results in an income distribution that is a series of step functions. This can be seen from the following calculations.

\[ f(x) = ax^2 + bx + c \]
\[ f'(x) = 2ax + b \]

Setting \( f'(x) \) to \( g(x) \) and inverting.

\[ g(x) = 2a \mu x + b \mu \]
\[ x = 2a \mu g^{-1}(x) + b \mu \]
\[ g^{-1}(x) = \frac{x - b \mu}{2a \mu} \]

Setting \( g^{-1}(x) \) to \( h(x) \) and taking the derivative to get the PDF.

\[ h(x) = \frac{x - b \mu}{2a \mu} \]
\[ h'(x) = \frac{1}{2a \mu} \]

Applying this technique to the cubic yields a square root function, as can be seen from the following steps.

\[ f(x) = ax^3 + bx^2 + cx + d \]
\[ f'(x)\mu = (3ax^2 + 2bx + c)\mu \]
\[ g(x) = (3ax^2 + 2bx + c)\mu \]
\[ \frac{g(x)}{\mu} = 3ax^2 + 2bx + c \]
\[ \frac{x}{\mu} = 3ag^{-1}(x)^2 + 2bg^{-1}(x) + c \]
\[ 0 = 3ag^{-1}(x)^2 + 2bg^{-1}(x) + c - \frac{x}{\mu} \]

Using the quadratic formula to solve for \( g^{-1}(x) \) where \( a^* = 3a, b^* = 2b \), and \( c^* = c - \frac{x}{\mu} \):

\[ g^{-1}(x) = \frac{-2b + \sqrt{4b^2 - 12a(c - \frac{x}{\mu})}}{6a} \]

Setting \( g^{-1}(x) \) to \( h(x) \) and taking the derivative with respect to \( x \) to get the PDF.

\[ h'(x) = \frac{1}{12a\sqrt{4b^2 - 12a(c - \frac{x}{\mu})}} \cdot \frac{12a}{\mu} \]
\[ h'(x) = \frac{1}{\mu \sqrt{4b^2 - 12a(c - \frac{x}{\mu})}} \]

The PDF that results from Lorenz interpolation is a set of step functions followed by this square root function. In some places along the income distribution, these functions may overlap. This is evident in the figure below, which displays the probability densities associated with the functions derived from Lorenz interpolation. Overlaps in the estimated PDF will occur if the slope of the line tangent to the Lorenz function at the lower bound of an income bin is less than the slope of the tangent at the upper bound of the preceding bin. To gain some intuition for why this is the case, recall
that the product of the slope of a line tangent to any point along the Lorenz curve and the income distribution mean equals the exact income estimate at that point.

![Figure 13: Densities of Lorenz Interpolation Functions.](image)

Note that the final step function and the square root function at the top of the distribution do not overlap. This is a consequence of the slope constraints used to fit the cubic function to the top of the Lorenz curve. Because the slope at the beginning of the cubic function is set to equal the slope at the end of preceding function, the density associated with this function ends where the density of the cubic function begins. Although one could apply similar constraints to ensure that the step functions do not overlap, in practice this does not improve the accuracy with which Lorenz interpolation estimates income inequality.
Figure 14 shows the PDF associated with the set of functions derived through Lorenz interpolation. This resembles the previous figure, except that the areas where the step functions overlap have been added. This density shows that the probability distribution implied by Lorenz interpolation does not resemble a real income distribution. For this reason, this method should not be used to computing certain statistics of the income distribution, such as the mean incomes of specific income brackets. The method is best suited for estimating income distribution variance parameters, which makes it an ideal technique for estimating inequality.

![Figure 14: PDF Based on Lorenz Interpolation.](image)
Appendix B. Incorporating Sampling Variation into Inequality Estimates

The analyses in this study consist of comparisons between income statistics estimated from Census grouped data and exact income data. This microdata has been treated as representing the population of a given region. In actuality, income data from the Census are based on a sample. Samples from the five-year pooled American Community Surveys cover about 5% of the U.S. population. To increase the reliability of estimates based on small areas, the Census reweights this sample using a complex weighting procedure that considers demographic and social characteristics of the regions from which samples are collected. These weights are provided with Census microdata and were used in this study to construct the “true” income distributions against which the estimates based on grouped income data were compared.

Although this approach allows for comparison between methods of estimating income statistics from grouped data, it does not provide a clear picture of the accuracy of Census income statistics at different geographic levels. To assess the viability of Lorenz interpolation and other methods for estimating income statistics for smaller geographies, one must incorporate sampling variation into the exact income estimates.

To add sampling variation into the estimates, I first estimated standard errors for the exact-income statistics. The Census provides a set of 80 replicate weights for the household-level components of the ACS. These weights are produced using successive difference replication (Fay and Train 1995). Standard errors can be computed for any
derived measure from the Census by plugging the replicate weights into the following equation.

\[
SE(X) = \sqrt{\frac{4}{80} \sum_{r=1}^{80} (X_r - X)^2}
\]

Where \(X\) is some measure (e.g., the Gini coefficient) derived using the main weights, and \(X_r\) is the same calculation based on a set of replicate weights.

This analysis was conducted at the school district level. I computed standard error estimates for the Gini, the Theil, the standard deviation, and the Atkinson index. To reduce computation time, I drew a random sample of 199 school districts from the full dataset and computed the standard errors associated with the four inequality measures for each district.

Next, I ran a series of simulations. For every combination of inequality measure and school district, I took 1000 random draws from a normal distribution with mean equal to the income inequality estimate based on the main set of weights and standard deviation equal to the standard error computed using the replicate weights. For each iteration, an error term was calculated by taking the difference between the MCIB/CDF/Lorenz interpolation estimate and the true estimate sample. The relative bias and RMSE of the 1000 iterations were then computed to produce error metrics for each school district-income inequality pair. Finally, these measures were averaged across school districts for each inequality measure.

The results of this analysis are provided in Table 6 below. Comparing these to the school district-level results in Table 2, Gini estimates that incorporate sampling
variation are about 3 times larger, Theils are 2 times larger, and standard deviations and Atkinson coefficients are 1.4 and 1.8 times larger, respectively. These are substantial increases. In relative terms, the Gini errors remain small, with a relative RMSE of 3.79 percent. The relative RMSE of the Atkinson index is larger (10.58 percent), while the relative RMSEs associated with the Theil and standard deviation are quite large, 16.62 percent and 18.14 percent respectively. As in the main analyses, inequality estimates produced from Lorenz interpolation had lower bias and RMSE than those derived from MCIB and CDF interpolation. For Gini coefficients, Lorenz interpolation yields slightly better estimates, while for the Theil and Atkinson these improvements are substantial.

Still, the relative RMSEs suggest that these measures are insufficiently reliable for making fine-grained comparisons between school districts.

### Table 6: Error Terms with Sampling Variation for School Districts.

<table>
<thead>
<tr>
<th></th>
<th>Gini</th>
<th>Theil</th>
<th>Atkinson</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
<td>RMSE</td>
</tr>
<tr>
<td>MCIB</td>
<td>-.91</td>
<td>3.86</td>
<td>-7.70</td>
<td>19.32</td>
</tr>
<tr>
<td>CDF</td>
<td>-1.42</td>
<td>3.99</td>
<td>-6.54</td>
<td>17.39</td>
</tr>
</tbody>
</table>

N = 199

One way to improve the reliability of these estimates is to restrict the analysis to districts based on larger samples. To demonstrate this, I restricted the analysis to school districts based on samples of more than 1000 households. The remaining districts comprise 50 percent of the original sample. Limiting the sample to larger districts had
little effect on the standard deviation RMSE, but RMSEs associated with the other inequality measures were reduced by between 25 and 50 percent, as shown in Table 7.

**Table 7: Error Terms with Sampling Variation for School Districts (Large Samples).**

<table>
<thead>
<tr>
<th></th>
<th>Gini</th>
<th>Theil</th>
<th>Atkinson</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
</tr>
<tr>
<td>MCIB</td>
<td>-1.04</td>
<td>2.82</td>
<td>-8.13</td>
<td>13.10</td>
</tr>
<tr>
<td>CDF</td>
<td>-1.54</td>
<td>3.02</td>
<td>-7.45</td>
<td>10.85</td>
</tr>
<tr>
<td>Lorenz</td>
<td>-0.763</td>
<td>2.74</td>
<td>-3.59</td>
<td>8.91</td>
</tr>
</tbody>
</table>

N = 98
References


Freeman, Allison and Roberto G. Quercia. 2014. Low- and Moderate-Income Homeownership and Wealth Creation. Chapel Hill, NC.


Ruggles, Steven, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek. 2021. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS.


Biography

Andrew Carr attended Wesleyan University in 2008, where he received a BA in East Asian Studies and History. He went on to receive an MA in Social Science from the University of Chicago, where he focused on social science research methods. Andrew started his PhD in Sociology at Duke University as a doctoral student in the Sociology Department in August 2015. He graduated with his doctorate in May 2021. He also holds an MS, which he received in May 2021, from the Statistics Department at Duke.