**INTRODUCTION**

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.

Second, researchers have only tested their methods for estimating inequality from grouped data 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, someone 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, existing methods for estimating income inequality may be insufficiently reliable when deployed at lower levels of analysis. While some researchers have admonished against using these methods to estimate inequality for smaller geographic regions (von Hippel et al. 2016), how these methods fare at producing income inequality estimates for these regions has yet to be empirically tested.

In this paper, I outline a new method for estimating income inequality from grouped income data. The method, which I call Lorenz interpolation, consists of using income quantiles derived from this data to estimate the Lorenz curve of the underlying distribution as a cubic spline function. Parameters from the income distribution can be estimated from this approximated Lorenz curve. A key advantage of Lorenz interpolation over other methods comes from the accuracy with which the method estimates the upper tail of the income distribution.[[1]](#endnote-1)

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 income inequality, especially inequality measures that are sensitive to the upper tail of the income distribution. Then, I evaluate the performance of Lorenz interpolation at two lower geographic levels, the census tract level and the school district level, using restricted census data to which I have been granted access through a Federal Statistical Research Data Center.[[2]](#endnote-2) Results indicate that Lorenz interpolation produces more accurate estimates of income inequality at all three geographic levels. However, tract-level estimates are insufficiently reliable for most analyses. Finally, Lorenz interpolation yields fairly accurate estimates of the Gini, Theil, and Atkinson 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.

**BACKGROUND**

A common method for deriving income inequality statistics from grouped income data is to set the incomes in the closed income bins to their midpoints and estimate a Pareto distribution for the open bin at the top of the income distribution (Henson 1967; Jargowsky 1996).[[3]](#endnote-3) Fitting a Pareto distribution to the top bin requires that the researcher estimate a Pareto distribution shape parameter, , which can be computed using the counts of households in the top two bins 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 bins in the upper tail of the income distribution tends to be linear (Nielsen and Alderson 1997). Once is estimated, the mean of the Pareto distribution for the top bin can be computed with the following formula.

( 1 )

where is the lower bound of the top income bin. Incomes in this bracket are then set to the estimate of .

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 . This is particularly true when this procedure underestimates as the Pareto distribution is undefined when is less than 1 and its mean approaches infinity as approaches 1 from above.[[4]](#endnote-4) Also, by imputing midpoints for the closed bins and a Pareto distribution mean for the open bins, the method ignores the within-bracket variation among incomes. This omission will produce biased income inequality estimates if a significant source of the variation among incomes lies within the income bins.

Recently, researchers have developed ways to get around these issues.[[5]](#endnote-5) 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. After interpolating the CDF, the authors use this function along with the income distribution mean to determine the upper bound of the CDF.[[6]](#endnote-6) Income statistics are derived from the interpolated CDF through numerical integration. When tested on county-level income distributions, the authors show that CDF interpolation estimates Gini coefficients within 1-2% of their true values (von Hippel et al. 2017:651).[[7]](#endnote-7)

Jargowsky and Wheeler (2018) also propose a method that takes advantage of the income distribution mean. Rather than fitting a function to points along the CDF of the income distribution, the authors estimate a PDF directly from 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 bins, fitting a Pareto distribution to the top bin, and integrating over the resulting distribution to produce various income statistics.[[8]](#endnote-8) This technique uses the mean of the income distribution to estimate the mean of the top income bin. After using estimating the total closed-bin income from the estimated PDF, the authors subtract this quantity is from the total income and divide this by the number of households in the top bin to produce an estimate of the top bracket mean.[[9]](#endnote-9) The authors show that MCIB outperforms many estimation techniques that do not incorporate the income distribution mean.

Although CDF interpolation and MCIB produce more accurate estimates of income inequality by making use of the income distribution mean, these methods are less successful at estimating income statistics that rely heavily on the upper tail of the income distribution. This is a challenge for all methods for estimating inequality from grouped income data, which provides limited information on the distribution of incomes at the top of the income distribution. The method put forth in this paper provides more accurate estimates of these inequality measures than MCIB or CDF interpolation. The method, which I call Lorenz interpolation, uses the income quantiles contained in the grouped income data to approximate a Lorenz curve of the income distribution.

The Lorenz curve visually represents the level of inequality in an income distribution. To create a Lorenz curve from a set of incomes, one sorts these incomes in ascending order and computes a running sum. This is divided by the total income to produce a cumulative income share, which is plotted against the cumulative population share. The cumulative population share is plotted on the x axis, and the cumulative income share is plotted on the y axis.

Much of the information about the shape of the Lorenz curve is provided by the income quantiles in the grouped income data. The connection between the income quantiles and the Lorenz curve is reflected by the following equation.[[10]](#endnote-10)

( 2 )

Where is the derivative of the Lorenz function taken with respect to , is the cumulative population share (the proportion of households that have an income less than or equal to ), is the income distribution mean, and is an income quantile. Equation 2 says that the slope of the Lorenz curve can be calculated at any point by dividing the income quantile at that point by the income distribution mean. Given that the income data provided in the Census includes both the distribution mean and the several income quantiles, this data can be used to determine the slope of the Lorenz curve at several points along the x axis. For example, if 10% of the population belongs to the bottom income bin, and this bin has an upper bound of $10,000, then the Lorenz curve has a slope of at = .1, where is the cumulative population share and is plotted on the x axis. In addition to using each bin’s income quantiles to compute slopes along the Lorenz curve, one can also use the quantiles of neighboring bins to influence the upward trajectory of the Lorenz curve. This technique determines the bin means associated with each bin of the grouped income data.

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 many 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. Lorenz interpolation also pays particular attention to the upper tail of the income distribution, which has disproportionately contributed to recent increases in income inequality (Piketty and Goldhammer 2014:315).

**METHODS**

Lorenz interpolation consists of computing quantiles from grouped data and using these to build a cubic spline function that approximates the Lorenz curve of the underlying distribution. Each segment of this spline is built in ascending order, starting with the first bin of the grouped data. The cubic function associated with each segment is estimated using the point on the Lorenz curve associated with the lower bound of its corresponding bin and a set of slope constraints to ensure that the slope of the Lorenz curve at each boundary equals the quotient of the boundary and the distribution mean. Finally, several points are plotted on the estimated Lorenz curve to produce a weighted sample of exact incomes, which is then used to compute inequality statistics from the underlying distribution.

*Building a Lorenz Curve from Grouped Income Data*

Figure 1 illustrates the steps though which Lorenz interpolation approximates a Lorenz curve from grouped data. The vertical dashed lines represent the cumulative population shares provided by the grouped income data. Although the cumulative *income* shares are unknown, one can use the income boundaries to determine the slopes of the Lorenz curve at each of the vertical lines. These slopes are represented by the red line segments in Figure 1.

[Insert Figure 1 About Here]

Lorenz interpolation begins by estimating a cubic function to represent the portion of the Lorenz curve associated with the lowest bin of the grouped income data. This step is shown in the left plot of Figure 1. This cubic function is constrained to pass through (0, 0), the point where the Lorenz curve begins. The slope of this function is constrained to equal at (0, 0), at , and at , where indexes the bins of the grouped income data and is set to 1, is the number of households in bin , is the total population, is the lower bound of the first bin, and are the upper bounds of the first and second bins, and is the cumulative population share plotted on the x axis. These four constraints – one that determines the starting point of the curve and three that determine its slope – define a unique cubic function, which is applied to the portion of the Lorenz curve associated with the first bin of the grouped income data.

After defining the first part of the Lorenz curve, the y-coordinate for the second portion of the Lorenz curve is estimated by plugging into the cubic function defined for the first bin. The index is then incremented by 1, and the Lorenz curve for the second bin is estimated using the same technique as the first (see the center plot in Figure 1). Specifically, a cubic function is estimated that passes through and with slopes of , , at , , and respectively. These steps are repeated for the remaining bins except for the final closed bin and the top bin.

The coefficients of each cubic function defined by these constraints can be calculated by solving the following system of equations.

= (3 )

Where , , , and are the coefficients of the cubic function applied to each bin. The first row of the system ensures that the function passes through , the x and y-coordinates associated with the bin’s lower bound. The next three rows represent the slope constraints: the expressions in the left matrix show the first derivative of the cubic function with the function input set to the bin lower bound, the bin upper bound, and the upper bound of the following bin.

At this stage, the Lorenz curve has been defined for all but the last two bins of the grouped income data. The cubic function for the final closed bin is defined using three of the four constraints used for the other closed bins: the cubic is constrained to pass through the point and where is indexed to the final closed bin, and the slopes of this function are based on the lower and upper bounds of the final closed bin. However, in place of a third slope constraint the cubic function is constrained to pass through (1, 1), the point associated with the upper bound of the top bin.

Finally, a special cubic function is defined for the top bin of the grouped data. Given the absence of an upper bound for this bin, one cannot use group income data to compute the slope at the top of the Lorenz curve. This makes defining a cubic function for the top bin particularly challenging. Cubic functions fit to points along the Lorenz curve potentially underestimate the variance of incomes at the top of the income distribution (Kakwani 1976:489). To remedy this issue, Lorenz interpolation uses a slope constraint to reduce the upward trajectory of the cubic function applied to the top bin. Specifically, a quadratic function is defined for the top bin based on the point and slope associated with the lower bound of the top bin and the point associated with the upper bound of this bin (1,1). Next, the slope of this quadratic function is computed at the bin midpoint. This slope is then incrementally reduced. At each iteration, the convexity of the resulting cubic function is evaluated by checking whether the second derivative of the function is negative at either of the bin boundaries. This iterative process is repeated until the function is no longer convex, at which point the method selects the smallest slope that does not violate the convexity check for the bin midpoint. This results in a gradually increasing cubic function that captures that variance at the upper tail of the income distribution.[[11]](#endnote-11)

Having estimated a Lorenz curve, the final step is to create 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 estimated curve and multiply the slopes of the line segments connecting these points by the income distribution mean. This generates samples from the underlying income distribution, which can be weighted using the frequencies provided in the grouped income data. This weighted sample is plugged into an income inequality formula to produce an income inequality estimate.

*The Probability Density Function Implied by Lorenz Interpolation*

The PDF implied by Lorenz interpolation is a piecewise function consisting of several square root functions. This can be recognized by deriving the PDF from the Lorenz curve. Equation 4 shows a segment of the estimated Lorenz function, , as a function of the cumulative population share .

(4)

Where is the set of incomes in bin and , , , and are the coefficients of the cubic function defined for the portion of the Lorenz curve associated with bin . The domain of this function is bounded by and , the cumulative population shares at the lower and upper bounds of bin .

To derive the PDF from the Lorenz function, one must first compute the CDF from this function. One way to do this is to use the following rule noted by Gastwirth (1971).

(5)

This is the definition of the Lorenz curve as a function of the cumulative population share Taking the derivative of both sides of this equation with respect to and multiplying by yields the following.

(6)

It follows that taking the derivative of the cubic function from Equation 4 with respect to and multiplying by the distribution mean will produce an equation that gives income, , as a function of.

(7)

Everything can be moved to the right side of the equation by dividing both sides by and subtracting .

(8)

The function for the CDF, , can now be computed using the quadratic formula.

(9)

Taking the derivative of both sides with respect to yields the PDF of the income distribution.

(10)

Finally, the domain is rewritten in terms of .

(11)

Note that the slope constraints used to compute , , , and ensure that the domain of this function is bounded at the income boundaries of bin We can compute these boundaries from the approximated Lorenz curve by plugging and into (Equation 2), which is also the expression used here to derive the CDF from the Lorenz function.

Figure 2 shows an example of a PDF based on Lorenz interpolation.

[Insert Figure 2 About Here]

The spikes along the upper tail of the income distribution pictured above are a result of the square root function Lorenz interpolation uses to approximate each bin of the income distribution. While these spikes indicate that Lorenz interpolation can overestimate the clustering of incomes at the lower bounds of bins along the upper tail of the income distribution, the results show that Lorenz interpolation is capable of producing comparatively accurate estimates of local income statistics, including bin means, income quantiles, and income shares.

**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.[[12]](#endnote-12) Estimates 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.

*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 that have been used by 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.[[13]](#endnote-13) 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.

( 12 )

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, or unpredictability, in a dataset. Income distributions with low entropy have higher Theils, which indicate more inequality. Theils are calculated with the following equation.

( 13 )

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 bin 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.

( 14 )

Where , the inequality aversion parameter, determines the relative influence of the income distribution tails. I chose to set 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 generated 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.

**RESULTS**

Table 1 compares error terms from Gini, Theil, Atkinson, and standard deviation estimates based on MCIB, CDF interpolation, and Lorenz interpolation. The top three rows show the errors produced by running these methods without the income distribution means, and the bottom three rows display errors based on estimates that incorporate the income distribution means.[[14]](#endnote-14) 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*, , a function of the ratio of the error to the true value (). 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. These metrics have the advantage of being invariant to the different scales of the inequality measures, which makes accuracy comparisons between the measures possible. All estimates have been rounded to the nearest hundredth.

[Insert Table 1 About Here]

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 and the standard deviation, while MCIB outperformed CDF interpolation at estimating the Theil and the Atkinson. Lorenz interpolation outperformed both methods at estimating the Gini, Atkinson, and standard deviation and performed worse than MCIB at estimating the Theil. Turning to errors metrics for estimates that incorporate the income distribution means, Lorenz interpolation estimates had lower relative RMSEs for all four inequality measures and lower relative bias for all measures except for the Atkinson coefficient. For Theil coefficients, Atkinson measures, and standard deviations, Lorenz interpolation produced considerably more accurate estimates as measured by relative RMSE. For instance, while the Theil estimates based on MCIB had a relative RMSE of 6.36%, Theils from Lorenz interpolation had a relative RMSE of 2.49%, amounting to an error reduction of more than 60%. Standard deviation estimates based on CDF interpolation had a relative RMSE of 6.46%, while standard deviation estimates from Lorenz interpolation had a relative RMSE of 2.84%. This represents a reduction of about 56%. Although Lorenz interpolation also outperformed the other methods at estimating the Gini, all three methods produced highly accurate Gini estimates, with relative RMSEs ranging between .75 and 1.38%.

The greater accuracy of income inequality estimates based on Lorenz interpolation can be attributed to the way that the method estimates the bin means of the grouped income data. Table 2 compares the relative bias and RMSE terms of bin mean estimates based on MCIB, CDF interpolation, and Lorenz interpolation. This table also includes the error terms for each method’s estimate of the total closed bin income, which determines the top bin mean estimates.

[Insert Table 2 About Here]

The three methods performed comparably at estimating individual bin means. Among the 16 bin mean estimates shown in Table 2, MCIB had the lowest bias for 1 bin and the lowest RMSE for 4, CDF interpolation had the lowest bias for 5 bins and the lowest RMSE for 1, and Lorenz interpolation had the lowest bias for 10 estimates and the lowest RMSE for 11. Lorenz interpolation also had the lowest relative bias and RMSE for the total closed bin income. Both MCIB and CDF interpolation produced positively biased estimates of the total closed bin income. This results from the tendency of these methods to overestimate closed bin means, which produces negatively biased estimates of the top bin mean.

While the greater accuracy with which Lorenz interpolation estimates the total closed bin income partially explains why the method produces more accurate inequality estimates than MCIB or CDF interpolation, this does not tell the whole story. For example, although Lorenz interpolation produced Atkinson estimates with higher relative bias than Atkinson estimates produced from MCIB, the relative RMSE of MCIB Atkinson estimates was significantly higher. This may reflect the MCIB’s use of the Pareto distribution to approximate the upper tail of the income distribution. Figure 3 shows scatterplots for the residuals of Atkinson estimates from all three estimation methods. The solid lines denote the bias associated with each estimation method, and the dashed lines represent a one standard-deviation distance from the solid lines. As can be recognized by comparing the space between the dashed lines in each plot, the improvement of Lorenz interpolation over MCIB is partly due to the greater reliability of Lorenz interpolation estimates. The difference in reliability between MCIB and the other methods is even greater at the school district and the census tract levels.

[Insert Figure 3 About Here]

*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. This is chiefly due to sampling variation, but it may also reflect the heterogeneity of income distributions associated with smaller regions. The income distributions of neighborhoods or municipalities may vary more than those of larger regions such as metropolitan areas, which encompass entire regional economies and may resemble each other more. This raises the question of whether estimation techniques like Lorenz interpolation can be used to produce valid income inequality estimates for small areas.

Table 3 shows the error terms for tract and school district-level estimates. Note the size of the relative RMSEs: these metrics are larger for tracts than for school districts, and they are larger for school districts than for PUMAs. Comparing MCIB and CDF interpolation, the latter method produced 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 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 variance of MCIB estimates.

[Insert Table 3 About Here]

Lorenz interpolation produced slightly more accurate Gini estimates than those produced by MCIB and CDF interpolation and significantly more accurate estimates of the Theil, Atkinson, and standard deviation. At the school district level, Theil, Atkinson, and standard deviation estimates based on Lorenz interpolation had 31-46% lower relative RMSEs than those based on the next best method, CDF interpolation. The gap between MCIB and Lorenz interpolation was even larger: Theil, Atkinson, and standard deviation estimates based on Lorenz interpolation were about 33-67% lower than those based on MCIB. Tract-level error metrics from Lorenz interpolation were also lower than those based on the other methods. However, the magnitudes of the relative RMSEs at this level of analysis are also a bit larger. At the tract-level, Lorenz interpolation estimates had relative RMSEs of about 6.5-8.5%. Also, 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.

In sum, Lorenz interpolation, CDF interpolation, and MCIB produced fairly accurate Gini coefficient estimates from both tract and school district-level income distributions. These 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 significantly more accurate estimates of these measures than both CDF interpolation and MCIB.

*Quantile and Income Share Estimates from MCIB, CDF Interpolation, and Lorenz Interpolation*

Table 4 shows relativebias and RMSE terms associated with MCIB, CDF interpolation, and Lorenz interpolation estimates of income quantiles and income shares. Starting with the income quantiles, estimates were produced for the incomes associated with the 20th, 40th, 60th, 80th, and 95th percentiles of the income distribution at the PUMA level. Lorenz interpolation estimates had the lowest relative bias and RMSE for all quantiles except at the 80th percentile, where MCIB produced the most accurate estimates. Moving on to income shares, statistics were estimated for each quintile of the income distribution. Lorenz interpolation produced the most accurate estimates of all income shares except for the bottom quintile. MCIB generated the most accurate estimates of the bottom quintile. In short, Lorenz interpolation generated slightly better estimates of income quantiles and income shares. However, performance improvements were small (a few hundredths of a percent in many cases). Although Lorenz interpolation can be used to estimate income quantiles, income shares, and other local statistics, its main utility over MCIB and CDF interpolation is for estimating income inequality.

[Insert Table 4 About Here]

**DISCUSSION**

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 viable 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.[[15]](#endnote-15) Researchers interested in examining income inequality for all U.S. counties must rely on grouped income data provided in Census summary tables. For geographies like these, including counties and census-designated places, using Lorenz interpolation on summary Census data is a valid 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 be wary of estimating certain inequality measures from this data. The errors associated with tract-level Theil, standard deviation, and Atkinson estimates produced in this analysis may be too large for some analyses.

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 from the 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.[[16]](#endnote-16)

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.

Researchers should also evaluate the utility of Lorenz interpolation for estimating statistics from other kinds of grouped data. This study looked exclusively at income data from the U.S. Census. Grouped data from this source is lower bound inclusive and upper bound exclusive. As a result, incomes that fall directly on an income boundary are assigned to the higher of the two bins subdivided by that boundary. This feature reduces the true bin means associated with the income groups and is one reason that Lorenz interpolation outperforms the other methods. For data that applies different rules for handling incomes falling on the income boundaries, the improvement of Lorenz interpolation may be smaller or non-existent. Researchers could determine this by comparing the performance of Lorenz interpolation, MCIB, and CDF interpolation using data from other national censuses.

**CONCLUSION**

In this paper, 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 improvements can be attributed to how the method estimates the closed bin means of grouped income data. I also estimated income inequality for tracts and school districts. The results indicate that Lorenz interpolation outperforms MCIB and CDF interpolation at estimating income inequality for smaller geographic regions. Conversely, a comparison of the accuracy with which these methods estimate income quantiles and income shares showed that none of the methods consistently produced more accurate estimates of these measures. Finally, I provided some scope conditions for the use of Lorenz interpolation. Although Lorenz interpolation produced more accurate inequality estimates at the tract-level, these estimates may be insufficiently reliable for some purposes. Lorenz interpolation yielded more reliable estimates at the school-district level. These estimates were also more accurate than those generated by MCIB and CDF interpolation. Lorenz interpolation especially outperforms the other methods at estimating income inequality measures like the Theil that are sensitive to the upper tail of the income distribution. As the contribution of this upper tail to income inequality continues to grow (Piketty and Goldhammer 2014), methods for estimating these measures will become increasingly important.

**REFERENCES**

Atkinson, Anthony B. 1970. “On the Measurement of Inequality.” *Journal of Economic Theory* 2(3):244–63.

Chetty, Raj, David Grusky, Maximilian Hell, Nathaniel Hendren, Robert Manduca, and Jimmy Narang. 2017. “The Fading American Dream: Trends in Absolute Income Mobility since 1940.” *Science* 356:398–406.

Cowell, Frank A., and Fatemeh Mehta. 1982. “The Estimation and Interpolation of Inequality Measures.” *The Review of Economic Studies* 49(2):273.

Fay, Robert, and George Train. 1995. “Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties.” Alexandria, VA: Proceedings of the Section on Government Statistics, American Statistical Association: 154-159. Retrieved February 15, 2021 (https://usa.ipums.org/usa/resources/repwt/FayTrain95.pdf).

Galster, George, and Patrick Sharkey. 2017. “Spatial Foundations of Inequality: A Conceptual Model and Empirical Overview.” *Rsf* 3(2):1–33.

Gastwirth, Joseph L. 1971. “A General Definition of the Lorenz Curve.” *Econometrica* 39(6):1037.

Gastwirth, Joseph L., and Marcia Glauberman. 1976. “The Interpolation of the Lorenz Curve and Gini Index from Grouped Data.” *Econometrica* 51(3):49–51.

Gastwirth, Joseph L., Tapan K. Nayak, and Abba M. Krieger. 1986. “Large Sample Theory for the Bounds on the Gini and Related Indices of Inequality Estimated from Grouped Data.” *Journal of Business and Economic Statistics* 4(2):269–73.

Geverdt, Douglas E. 2019. “Education Demographic and Geographic Estimates Program

(EDGE): School District Geographic Relationship Files User’s Manual.” Washington, DC: U.S. Department of Education. National Center for Education Statistics. Retrieved September 7, 2020 (https://nces.ed.gov/pubsearch/).

Grusky, David B., and Alair MacLean. 2016. “The Social Fallout of a High-Inequality Regime.” *Annals of the American Academy of Political and Social Science* 663(1):33–52.

Hajargasht, Gholamreza, William E. Griffiths, Joseph Brice, D. S. Prasad. Rao, and Duangkamon Chotikapanich. 2012. “Inference for Income Distributions Using Grouped Data.” *Journal of Business and Economic Statistics* 30(4):563–75.

Henson, Mary F. 1967. ﻿ “Trends in the Income of Families and Persons in the United States,

1947–1964.” Washington, DC: U.S. Department of Commerce, Bureau of the Census.

Hunter, David J., and McKalie Drown. 2016. *Binsmooth: Generate PDFs and CDFs from Binned Data.* R Package Version 0.2.2 Comprehensive R Archive Network. Retrieved May 28, 2020 (https://cran.r-project.org/web/packages/binsmooth/).

Jargowsky, Paul A. 1996. “Take the Money and Run: Economic Segregation in U.S. Metropolitan Areas.” *American Sociological Review* 61(6):984–98.

Jargowsky, Paul A., and Christopher A. Wheeler. 2018. “Estimating Income Statistics from Grouped Data: Mean-Constrained Integration over Brackets.” *Sociological Methodology* 48(1):337–74.

Kakamu, Kazuhiko. 2016. “Simulation Studies Comparing Dagum and Singh–Maddala Income Distributions.” *Computational Economics* 48(4):593–605.

Kakwani, Nanak. 1976. “On the Estimation of Income Inequality Measures from Grouped Observations.” *The Review of Economic Studies* 43(3):483–92.

Logan, John R., Andrew Foster, Jun Ke, and Fan Li. 2018. “The Uptick in Income Segregation: Real Trend or Random Sampling Variation?” *American Journal of Sociology* 124(1):185–222.

McDonald, James B., and Yexiao J. Xu. 1995. “A Generalization of the Beta Distribution with Applications.” *Journal of Econometrics* 66(1–2):133–52.

Minoiu, Camelia, and Sanjay G. Reddy. 2012. “Kernel Density Estimation on Grouped Data: The Case of Poverty Assessment.” *Journal of Economic Inequality* 12(2):163–89.

Nielsen, Francois, and Arthur S. Alderson. 1997. “The Kuznets Curve and the Great U-Turn: Income Inequality in U.S. Counties, 1970 to 1990.” *American Sociological Review* 62(1):12–33.

Owens, Ann. 2016. “Inequality in Children’s Contexts: Income Segregation of Households with and without Children.” *American Sociological Review* 81(3):549–74.

Owens, Ann. 2019. “Building Inequality: Housing Segregation and Income Segregation.” *Sociological Science* 6:497–525.

Pickett, Kate E., Shona Kelly, Eric Brunner, Tim Lobstein, and Richard G. Wilkinson. 2005. “Wider Income Gaps, Wider Waistbands? An Ecological Study of Obesity and Income Inequality.” *Journal of Epidemiology and Community Health* 59(8):670–74.

Pickett, Kate E., and Richard G. Wilkinson. 2007. “Child Wellbeing and Income Inequality in Rich Societies: Ecological Cross Sectional Study.” *British Medical Journal* 335(7629):1080–85.

Piketty, Thomas, and Arthur Goldhammer. 2014. *Capital in the Twenty-first Century.* Cambridge, MA: The Belknap Press of Harvard University.

Quandt, Richard E. 1966. “Old and New Methods of Estimation and the Pareto Distribution.” *Metrika* 10(1):55–82.

Reardon, Sean F., and Kendra Bischoff. 2011. “Income Inequality and Income Segregation.” *American Journal of* Sociology 116(4):1092–1153.

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.

Tillé, Yves, and Matti Langel. 2012. “Histogram-Based Interpolation of the Lorenz Curve and Gini Index for Grouped Data.” *American Statistician* 66(4):225–31.

U.S. Census Bureau. 2007. “PUMS Accuracy of Data (2007).” Retrieved May 28, 2020

(https://www2.census.gov/programs-surveys/acs/tech\_docs/pums/accuracy/2007AccuracyPUMS.pdf).

U.S. Census Bureau. 2015. “American Community Survey and Puerto Rico Community Survey

2015 Subject Definitions.” Retrieved May 28, 2020 (https://www2.census.gov/programs-surveys/acs/tech\_docs/subject\_definitions/2015\_ACSSubjectDefinitions.pdf).

U.S. Census Bureau. 2015. “Gini Index of Money Income and Equivalence-Adjusted Income:

1967 to 2014.” Retrieved May 28, 2020 (https://www.census.gov/content/dam/Census/library/visualizations/2015/demo/gini-index-equil-adj-inc.pdf).

von Hippel, Paul T., Samuel V. Scarpino, and Igor Holas. 2016. “Robust Estimation of Inequality from Binned Incomes.” *Sociological Methodology* 46(1):212–52.

von Hippel, Paul T., David J. Hunter, and McKalie Drown. 2017. “Better Estimates from Binned Income Data: Interpolated CDFs and Mean-Matching.” *Sociological Science* 4:641–55.

**ENDNOTES**

1. The reason for Lorenz interpolation’s empirical advantage over other methods is not entirely clear. Paul von Hippel has suggested that the Lorenz curve is smoother and may be easier to estimate with a cubic spline than other representations of the income distribution such as the CDF. The rules according to which Lorenz interpolation approximates the Lorenz curve may also contribute to the method’s accuracy improvement over other methods. As the results show, the method produces less biased estimates of the bin mean of the open bin at the top of the income distribution. [↑](#endnote-ref-1)
2. Information about the locations of FSRDCs can be found on the Census website (https://www.census.gov/fsrdc). [↑](#endnote-ref-2)
3. 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 to estimate the harmonic rather than arithmetic mean of the Pareto distribution. [↑](#endnote-ref-3)
4. To get around this problem, von Hippel and coauthors (2016) proposed a robust Pareto mean estimator, which uses 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 . [↑](#endnote-ref-4)
5. 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). [↑](#endnote-ref-5)
6. There are actually two versions of CDF interpolation that von Hippel et. al (2017) describe in their paper. In addition to the version of CDF interpolation discussed here, the authors also propose a version of the method in which the points on the CDF are interpolated using a quadratic spline. In the implementation of this version of the method, the user has the option to choose among a Pareto, exponential, or a uniform distribution for the top bin. [↑](#endnote-ref-6)
7. 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 suggested 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. [↑](#endnote-ref-7)
8. 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. [↑](#endnote-ref-8)
9. The authors of MCIB also use a form of Lorenz interpolation to estimate the Gini coefficient, which unlike other inequality measures such as the Theil and variance is not additively separable into the income bins and therefore cannot be computed through integration. The authors’ solution to this issue is to plot points on a Lorenz curve based on the bin means estimated from MCIB. For the closed bins, the spaces between these points are divided into five equidistant segments, and additional points are plotted on the Lorenz curve based on the linear function defined for the PDF. The top bin is divided into five equal probability segments, on which points are plotted based on a Pareto distribution. Finally, the area under the Lorenz is computed as a sum of trapezoids by simply connected these dots. [↑](#endnote-ref-9)
10. This equation is easily derived from an equation in Gastwirth 1971. [↑](#endnote-ref-10)
11. This convexity check is also applied to the other segments of the cubic spline function that Lorenz interpolation estimates using the closed bins of grouped income data. Specifically, any cubic function that has a negative second derivative at either of its boundaries is adjusted so that it is as close to the original cubic function as possible without violating this convexity rule. [↑](#endnote-ref-11)
12. 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, dividends, net rental income, royalty income, or income from estates and trusts, social security income, supplemental security income, public assistance income, retirement, survivor, or disability income, unemployment compensation, worker’s compensation, Department of Veteran’s Affairs (VA) payments, and alimony and child support (U.S. Bureau of the Census 2015). [↑](#endnote-ref-12)
13. 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. [↑](#endnote-ref-13)
14. When the income distribution mean is not provided, Lorenz interpolation simply estimates the income distribution mean as the sum of the bin midpoints weighted by their relative frequencies. The open-ended bin at the top of the income distribution is assigned $300,000 as its midpoint. [↑](#endnote-ref-14)
15. 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). [↑](#endnote-ref-15)
16. 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. [↑](#endnote-ref-16)