Using the Pareto Distribution to Improve Estimates of Topcoded Earnings

Philip Armour
Cornell University

Richard V. Burkhauser
Cornell University

Jeff Larrimore
Joint Committee on Taxation

Draft 10/30/13

The public-use March Current Population Survey (CPS) is the primary source of data for tracking levels and trends in United States labor earnings, in labor earnings inequality, and explaining their causes. This literature has especially focused on whether the rise in earnings inequality in the 1980s was part of a long-run secular trend or an episodic event (Autor, Katz, & Kearney, 2008; Card & DiNardo, 2002; Juhn, Murphy & Pierce, 1993. See Acemoglu, 2002, for a review of this literature). However, this public-use CPS-based literature has been hampered by its attenuated view of the right tail of the labor earnings distribution due to the topcoding of high earnings in these data.[1]

To correct for topcoding biases, CPS-based researchers have generally pursued one of three paths: (1) ignoring the topcoding problem; (2) making an ad-hoc adjustment to topcoded values earnings values, or (3) using a Pareto distribution to estimate earnings at the top of the distribution. For example, a common ad-hoc technique, based on estimates from Pareto imputations of top earnings, is to replace topcoded earnings with a multiple of the topcode threshold, so all individuals with topcoded earnings in a year are assumed to have earnings at 1.3, 1.4, or 1.5 times the topcode threshold (Autor, Katz, & Kearney, 2008; Katz & Murphy, 1992; Juhn, Murphy, & Pierce, 1993; Lemieux, 2006)**.** However, such an approach may misstate top earnings if the wrong multiple is used or if the appropriate multiple changes over time. Similarly, researchers using a Pareto imputation of top earnings may misstate those earnings if they are unable to obtain a reasonable fit for the Pareto distribution when using available public-use data.

Making use of internal March CPS files with their much higher censoring levels, we show that previous ad-hoc estimates and Pareto estimations of top earnings based on public-use

[1] Some earnings inequality research focuses on the wage questions in the May Outgoing Rotation Group (ORG) sample of the CPS, which is also subject to topcoding of high wages.

data understate mean earnings at the top of the earnings distribution and hence earnings inequality. Then, using a continuous maximum likelihood estimator along with internal CPS data, we produce a series of more accurate estimates of top earnings in the CPS data. These estimates start with the actual top earnings from the internal CPS data, only using a Pareto estimation for internally censored observations. With this hybrid approach, we create an enhanced cell-mean series that allows researchers who only have access to the public-use data to more accurately capture top earnings levels and trends.

To show the value of our new measure, we use it together with the public-use CPS to replicate the level and trend in labor earnings inequality from 1975 to the present found by Kopczuk, Saez, and Song (2010) using Social Security (SSA) administrative records.

II. Data

The March CPS survey contains a comprehensive set of questions on sources of household earnings, including labor earnings.[2] Figure 1 provides an overview of the public topcode and internal censoring levels for annual wage earnings from 1975-1986 and for primary labor earnings, which are primarily wages, from 1987-2007. Both the public topcode level and the internal censoring level (left y-axis) increase on an irregular, ad-hoc basis. As a result the percentage of individuals with earnings above the public topcode (right y-axis) rises steadily when topcodes are held nominally constant and quickly falling when they are raised.

III: Estimating Top Earnings

Most researchers interested in measuring long-term trends in earnings in the CPS have adopted ad-hoc techniques to correct for topcoding, such as imputing topcoded earnings as a

[2] The March CPS asks about income in the previous year, so the income year is always one year prior to the survey year. All references to years in this paper refer to the income year. Because of Census Bureau changes in their aggregation techniques we use wage and salary earnings for years prior to income year 1987 and all primary labor earning thereafter. Since the vast majority of primary earnings are from wages and salaries, this break does not appear to have a noticeable impact on our results.

fixed multiple above the topcode point, although there is no consensus on which multiple to use with most researchers using a multiple between 1.3 and 1.5 (Autor, Katz, & Kearney, 2008; Juhn, Murphy, & Pierce, 1993; Lemieux, 2006). Implicit in this approach, regardless of the multiplier, is an assumption that the multiple is constant across years and across changes in the threshold level.

The multiples in this fixed multiple approach were partially derived from attempts to fit top earnings to a Pareto distribution. In particular, following the long-standing assumption that top earnings can be described by the Pareto distribution, numerous researchers have imputed the top of the earnings distribution based on those fit by a Pareto distribution (Bishop, Chiou, & Formby, 1994; Fichtenbaum & Shahidi, 1988; Heathcote, Perri, & Violante, 2010; Mishel, Bernstein, & Shierholz, 2013; Piketty & Saez, 2003; Schmitt, 2003).

The Pareto distribution is defined by the CDF:

$$P(X < x) = 1 - \left(\frac{x_c}{x}\right)^{\alpha} \quad (1)$$

where x is a given value of earnings (weakly) larger than x_c, x_c is the scale or cutoff parameter, and α is the shape parameter of the distribution. Since the Pareto distribution is scale-free, the mean above any threshold y is given as:

$$M(y) = \left(\frac{\alpha}{\alpha-1}\right) y \quad (2)$$

This provides a simple link to the fixed multiple concept. By setting y as the topcode threshold, M(y) is the Pareto-imputed mean income above the threshold.

In order to use the Pareto distribution to estimate top earnings, one must first estimate the appropriate shape parameter. The most common approach is to assume that the distribution is Pareto above some lower cutoff point (x_c) and simply choose a second cutoff point above that

point – which typically is the topcode threshold itself (x_t) (Parker & Fenwick, 1983; Quandt, 1966; Shyrock & Siegel, 1975; Saez, 2000). The Pareto shape parameter is then given by:

$$\alpha = \frac{\ln(\frac{C}{T})}{\ln(\frac{X_T}{X_C})} \quad (3)$$

where C represents the number of individuals with earnings above the lower cutoff and T represents the number of individuals with earnings above the topcode threshold. Juhn, Murphy and Pierce (1993) report that their choice of cutoff points in the public-use CPS did not substantially impact their results. However, Schmitt (2003) using more recent public-use CPS data found that the choice of cutoff point could matter greatly, depending on the frequency of topcoding in the empirical distribution.

As we will illustrate below, this approach has failed to provide reasonable estimates of top incomes in public-use CPS data. This is partially because the income distribution may not be Parato far enough below the public-use topcode threshold (if at all) to obtain reasonable estimates of the scale parameter. Additionally, it may be partially because, by virtue of using only two distribution points, this estimation technique poorly measures the parameter. We address the first of these concerns by estimating the shape of the Pareto distribution using the internal data with its less restrictive censoring. This allows us to reduce the portion of the distribution over which earnings must fit the Pareto distribution, since it only requires that 1 or 2 percent of the distribution be approximated by the Pareto rather than the ~~10 or~~ 20 percent commonly ~~that were previously~~ used to estimate Pareto distributions with the public-use data (Mishel et al. 2013). To further improve the estimate, we use actual internal data when available for estimating top earnings, and only using the Pareto imputation for internally censored observations where the true value is unknown.

To address the second concern, we adapt an alternate, but rarely used, approach to estimating the Pareto scale parameter—applying a maximum likelihood formula to the empirical distribution. Polivka (2000) used this approach to analyze categorical weekly earnings data but to our knowledge, it has not been applied to continuous annual earnings data. Under this approach, the continuous, closed-form solution for estimating the Pareto parameter is:

$$\hat{\alpha} = \frac{M}{T\ln(X_T)+\sum_{x_m \le x_i < x_T}\ln(x_i)-(M+T)\ln(x_m)} \quad (4)$$

Where *M* is the number of individuals with earnings between the lower cutoff and censoring point, *T* is the number of individuals with earnings at or over the topcode or censoring point, and x_i is the earnings of an individual. Using this formula allows individuals between the cutoff and censoring points to contribute to the PDF with their actual earnings, while those at the censoring point contribute to the CDF with the information that they have earnings at least as high as the censoring point.

In Figure 2 we compare the relative accuracy of the standard proportional and our maximum likelihood Pareto imputation approaches, along with the fixed-multiple approach from Lemieux (2006) and Katz and Murphy (1992) in capturing the top part of the earnings distribution censored in the public-use CPS. We do so by comparing the mean earnings of the top 5 percent of the distribution for each of these series with those in the Larrimore et al. (2008) cell-mean series based on the internal CPS data. The Larrimore *et* al. (2008) cell mean series uses the internal CPS data to provide the mean source-level income for each source of income for any individual whose income from that source is topcoded. But it is not designed to correct for internal censoring and treats income at or above the internal censoring point as if it were equal to the censoring point. As a result, it is consistent with Census Bureau's official income

statistics, but both this series and the official Census Bureau statistics are known to represent an underestimate of the true top earnings of the population.

Since the Pareto cutoff point matters for both approaches, when using the public-use data we follow the approach of Mishel et al. (2013) and assume the distribution is Pareto above the 80th percentile of the distribution.[3] For the estimation using our Maximum likelihood technique, since we are using internal data we can use a much higher cutoff, and assume the distribution is Pareto above the 99th percentile.[4]

Figure 2 shows that the estimated mean earnings of the top 5 percent of the distribution are similar when using the fixed-multiple approach or when using the standard Pareto imputation using the proportional method with public-use data. However, while the Pareto imputation slightly exceeds the top incomes from the Larrimore et al. cell mean series in early years, neither does so after 1993 when improvements in Census Bureau collection procedures greatly improved the reporting of earnings by top earners. (See Jones & Weinberg 2000 and Ryscavage 1995 for details on this change.) Since the cell-mean series is a lower bound for top earnings, it is apparent that all previous efforts to capture the top part of the earnings distribution based soley on public-use CPS data substantially understate their level at the upper tail, as found in the internal CPS data, since 1993.

In contrast to these earlier techniques, our Maximum Likelihood Pareto estimation of internally censored observations, in conjunction with the internal data where available, produces mean earnings of the top 5 percent which exceed those of Larrimore et al. (2008). In years before 1993, when the Census Bureau increased their internal censoring thresholds, this

[3] Alternate cutoffs of the 85th, 90th, and 95th percentiles were also considered. In general increasing the income cutoff for the lower bound of the estimation lowered the mean earnings of the top 5 percent.

[4] Alternate cutoffs of the 95th, 97th and 98th percentiles were also considered and produced largely consistent results for the mean earnings of the top 5 percent.

increased the mean earnings of the top 5 percent by between 7 and 14 percent in each year. In more recent years, the gap has been smaller, ranging from a 1 to 6 percent increase over the values from Larrimore et al. [5]

Recognizing that these improved estimates are based on internal data which are not generally available, in order to allow researchers with access to just the public-use data to benefit from this approach we have created an enhanced cell-mean series which uses the actual internal data when available and these Pareto estimates for the internally censored data. This series, which is available in the data appendix, allows researchers with only public-use data to obtain the best available estimate of top earnings in the CPS data.

IV: Comparison to Social Security Administration Records

Kopczuk, Saez, and Song (2010) provide the first research using administrative records data to analyze long-run earnings inequality. Their study uses Social Security Administration (SSA) earnings data from 1937 to 2004 to examine earnings inequality of "Commerce and Industry" workers between the ages of 18 and 70 with wages over $2,575 in 2004.[6] Although their administrative earnings data may be subject to some concerns of avoidance techniques for tax reporting, this study is the current gold standard of earnings inequality trends and hence an excellent benchmark for testing the validity of our CPS-based results. If results from Kopczuk, Saez, and Song (2010) can be replicated in the CPS data, then it validates the use of CPS data for analyzing earnings trends. To this end, we limit our data sample to Commerce and Industry workers and compare Gini coefficient results across the two datasets.

[5] As a further test of the validity of the Pareto at this income level, we compare the Pareto scale parameter for the 95th, 97th, 98th, and 99th percentile. The Pareto parameters are generally stable, with the average difference between the maximum and minimum scale parameter in this range being just 16 percent apart. Pareto scale parameters are available upon request of the authors.

[6] "Commerce and Industry" workers are all non-farm, non-self-employment wage and salary workers not working in agriculture, forestry, fishing, hospitals, educational services, social services, religious organizations, private households, and public administration.

These series are shown in Figure 3, along with the results using the raw public-use data with no cell means as well as those using the cell-mean series from Larrimore *et al*. (2008) and the fixed multiple series where topcoded incomes are replaced with 1.4 times the topcode threshold. The public-use data with no cell means is clearly well below the level of earnings inequality observed by Kopczuk, Saez, and Song (2010). On the other hand, both the internal cell means series from Larrimore *et al*. (2008) and the enhanced cell-mean series are closer to the top income shares observed by Kopczuk, Saez, and Song. In particular, the enhanced cell-mean series which largely overcomes internal censoring, can largely match the trends from Kopczuk Saez, and Song back to 1967 when annual CPS data is first available. The primary exceptions occur between 1992 and 1993 when the Census Bureau improved their collection techniques, and between 1986 and 1988, when the tax code changes from the Tax Reform Act of 1986 provided some incentives for top income tax payers to switch their reported income from Subchapter-C corporation profits, which are not reported as earnings in IRS or SSA administrative records, to wage income, which are reported in the SSA data (Slemrod, 1995). This provides evidence that, with appropriate corrections to capture the top of the earnings distribution, the CPS data can be used to accurately measure and analyze United States earnings trends.

Interestingly, however, the inequality trends found here, and potentially those found by Kopczuk, Saez, and Song (2010), are sensitive to the decision to limit the sample to Commerce and Industry workers. Figure 4 compares the Gini coefficient using our preferred topcode correction of the internal Pareto series for two samples. The first is the Commerce and Industry worker sample, which matches Kopczuk, Saez, and Song's (2010) sample and was previously presented in Figure 3. The second is all workers, regardless of industry, with positive wage or salary income regardless of age. In the restricted sample, earnings inequality increased by 21.5

percent (from 0.382 to 0.464) between 1967 and 2004. However, in the full sample it increased by only 5.9 percent (from 0.463 to 0.490). Thus, earnings inequality for the full population may have increased less than Kopczuk, Saez, and Song observed when looking at just Commerce and Industry workers.

V: Conclusion

Despite the common use of CPS data for earnings inequality research, the current methods of correcting for topcoding in the data result in clear and substantial understatements of top earnings. Using a hybrid approach of internal data and Pareto imputations, this paper provides improved estimates of top earnings in the CPS data. These estimates produce earnings inequality levels that are consistent with those observed in administrative Social Security records from Kopczuk, Saez, and Song (2010). Using this hybrid approach for estimating top earnings, we have produced an enhanced cell-mean series, which more closely approximates the actual level of top earnings in the population than was previously available in CPS data. We then demonstrate that the choice to restrict the sample to just Commerce and Industry workers, as was done by Kopczuk, Saez, and Song (2010) may result in an overstatement of earnings inequality growth since the 1960s.

References

Acemoglu, Daron. 2002. "Technical Change, Inequality, and the Labor Market." *Journal of Economic Literature*, 40(1): 7-72.

Atkinson, A.B., Piketty, T., & Saez, E. (2011). Top Incomes in the Long Run of History. *Journal of Economic Literature , 49* (1), 3-71.Autor, D.H., Katz, L.F., & Kearney, M. S. (2008). Trends in U.S. Wage Inequality: Revising the Revisionists. *Review of Economics and Statistics , 90* (2), 300-323.

Bishop, J.A., Chiou, J.R., & Formby, J.P. (1994). Truncation Bias and the Ordinal Evaluation of Income Inequality. *Journal of Business and Economic Statistics , 12*, 123-127.

Card, D., & DiNardo, J.E. (2002). Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles. *Journal of Labor Economics , 20* (4), 733-782.

Fichtenbaum, R., & Shahidi, H. (1988). Truncation Bias and the Measurement of Income Inequality. *Journal of Business and Economic Statistics , 6*, 335-337.

Heathcote, J., Perri, F., & Violante, G.L. (2010). Unequal We Stand: An Empirical Analysis of Economic Inequality in the United States: 1967-2006. *Review of Economic Dynamics , 13* (1), 15-51.

Jones, A.F., & Weinberg, D.H. (2000). *The Changing Shape of the Nation's Income Distribution.* Washington, DC: U.S. Census Bureau.

Juhn, C., Murphy, K.M., & Pierce, B. (1993). Wage Inequality and the Rise in Returns to Skill. *Journal of Political Economy , 101* (3), 410-442.

Katz, L.F., & Murphy, K.M. (1992). Changes in Relative Wages, 1963-87: Supply and Demand Factors. *Quarterly Journal of Economics , 107*, 35-78.

Kopczuk, W., Saez, E., & Song, J. (2010). Earnings Inequality and Mobility in the United States: Evidence from Social Security Data since 1937. *Quarterly Journal of Economics , 125*, 91-128.

Larrimore, J., Burkhauser, R.V., Feng, S., & Zayatz, L. (2008). Consistent Cell Means for Topcoded Incomes in the Public-use March CPS (1976-2007). *Journal of Economic and Social Measurement , 33* (2-3), 89-128.

Lemieux, T. (2006). Increased Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill. *American Economic Review , 96* (2), 461-498.

Mishel, L., Bernstein, J., & Shierholz, H. (2013). *The State of Working America 12th Edition.* Ithaca, NY: Cornell University Press.

Parker, R.N. and Fenwick, R. (1983). The Pareto Curve and Its Utility for Open-Ended Income Distributions in Survey Research, *Social Forces* 61, 872-885

Piketty, T., & Saez, E. (2003). Income Inequality in the United States, 1913–1998. *Quarterly Journal of Economics , 118* (1), 1-39.

Quandt, R.E. (1966). Old and New Methods of Estimation and the Pareto Distribution, *Metrika* 10, 55-82

Polivka, A. (2000). Using Earnings Data from the Monthly Current Population Survey. *Unpublished Manuscript* .

Ryscavage, P. (1995). A Surge in Growing Income Inequality? *Monthly Labor Review , 118* (8), 51-61.

Saez, Emmanuel. (2000). Using Elasticities to Derive Optimal Income Tax Rates. *Review of Economic Studies* 68, 205-229.

Schmitt, J. (2003). *Creating a consistent hourly wage series from the Current Population Survey's Outgoing Rotation Group, 1979-2002.* Washington, DC: Center for Economic and Policy Research.

Shyrock, H. and Siegel, H. (1975). *The Methods and Materials of Demography*, Washington D.C., US Government Printing Office.

Slemrod, J. (1995). Income Creation or Income Shifting? Behavioral Responses to the Tax Reform Act of 1986. *American Economic Review , 85* (3), 175-180.

Figure 1: Topcoding and Censoring thresholds in the March CPS data (1975-2007)

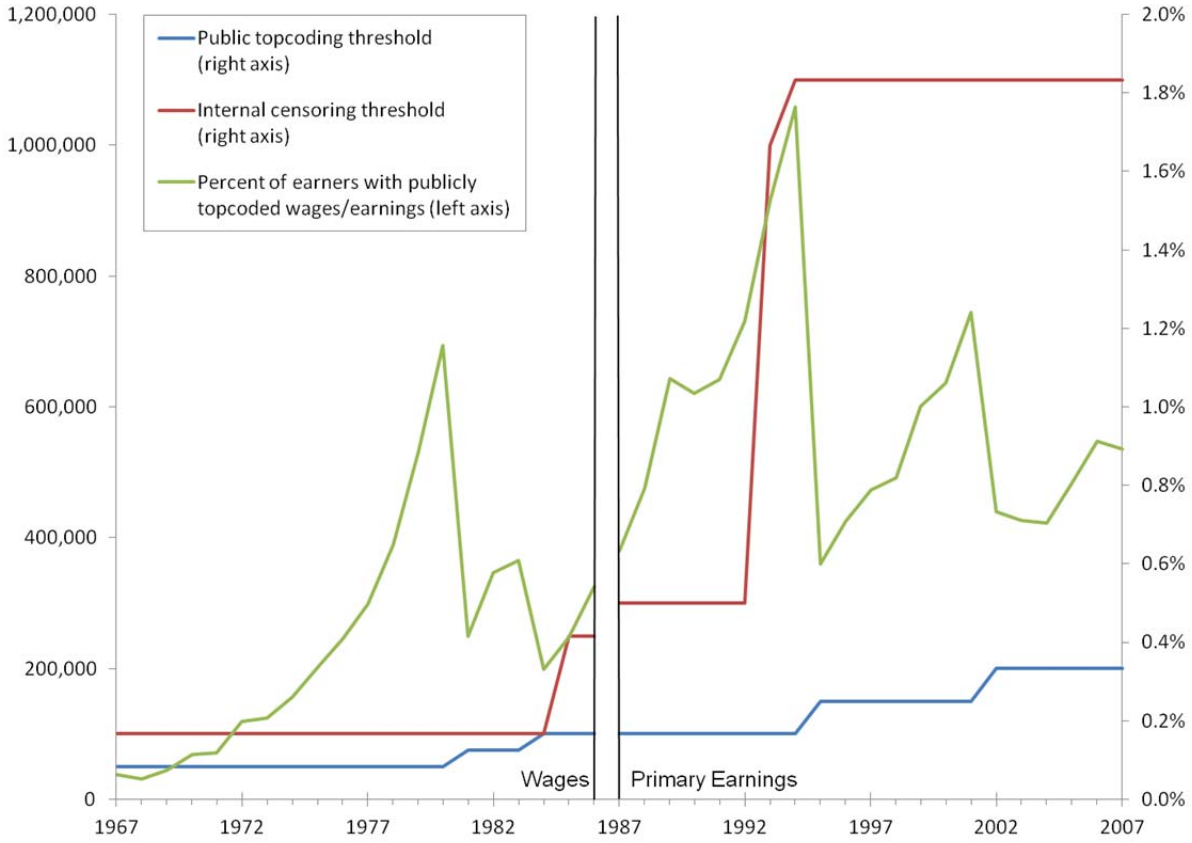


Figure 2: Mean earnings of the top 5 percent of earners by topcode correction method (in 2010 dollars)

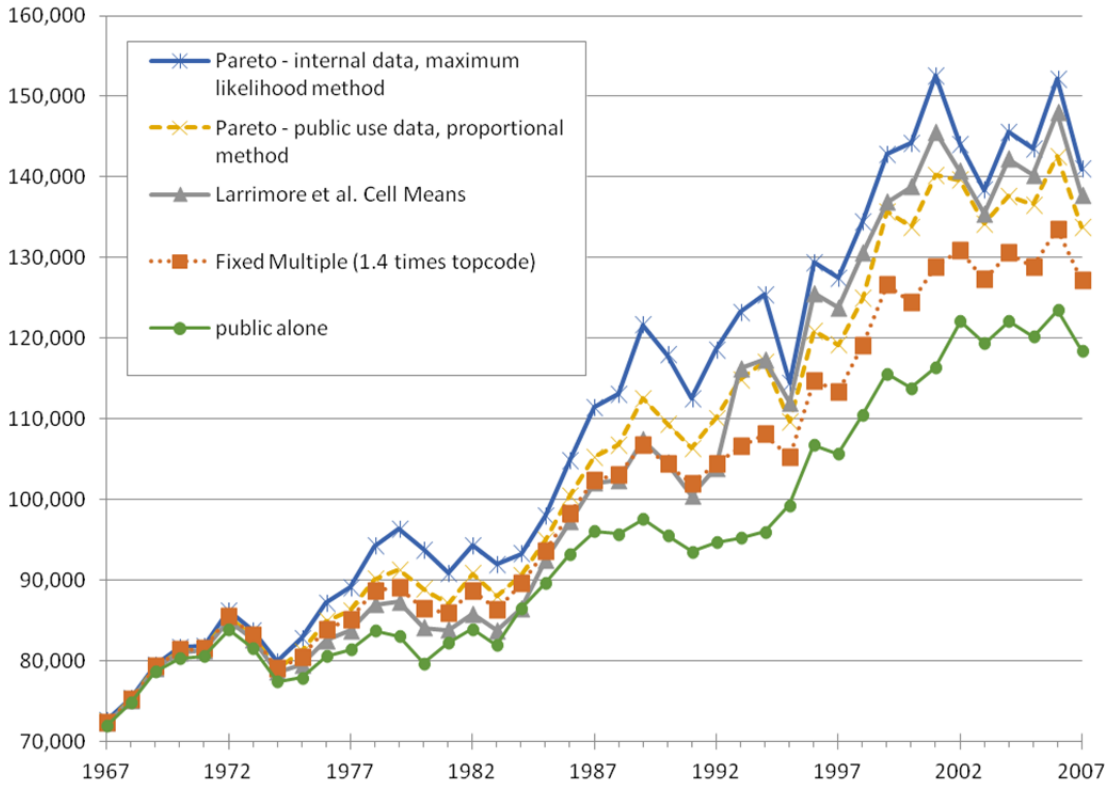


Figure 3: Gini Coefficients for Commerce and Industry workers by topcode correction method, compared to Kopczuk, Saez, and Song (2010) estimates from SSA administrative records

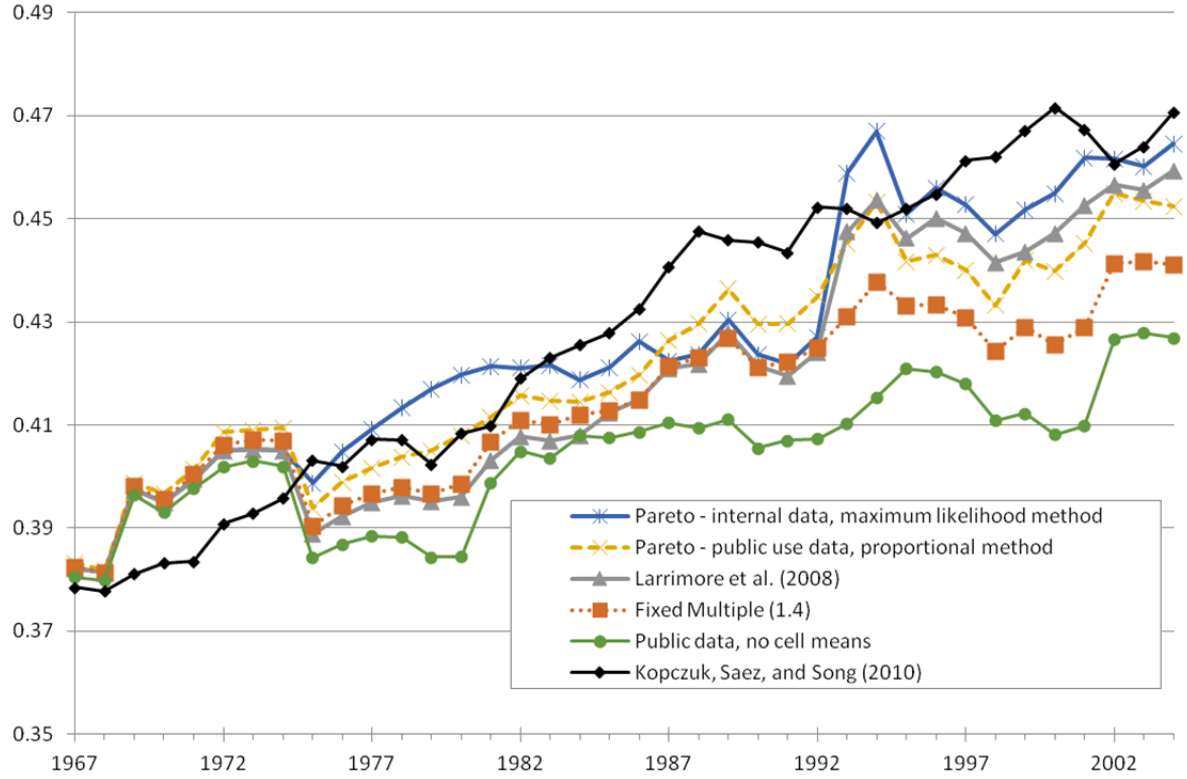
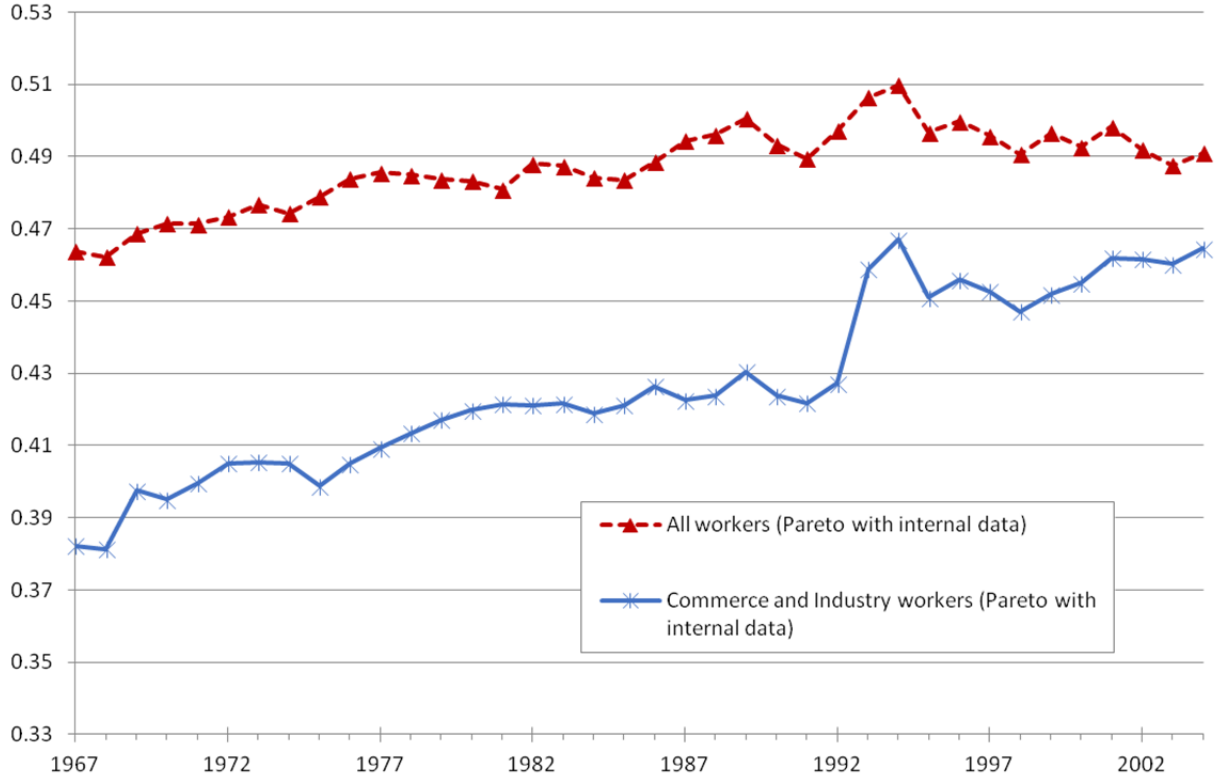


Figure 4: Gini Coefficients for all workers compared to Commerce and Industry workers, using internal CPS data with the Pareto correction method.



Enhanced Cell Mean Values for Wage (1968-1987) or Primary Earnings (1988-2008) Based on Maximum Likelihood Pareto Fit of Internal March CPS Data

Survey Year	Nominal Cell Mean
1968	66,849.15
1969	67,592.16
1970	70,508.57
1971	72,221.75
1972	70,822.31
1973	74,081.72
1974	69,862.66
1975	69,580.19
1976	104,623.10
1977	105,819.20
1978	107,545.80
1979	110,339.50
1980	112,330.50
1981	109,203.50
1982	177,926.10
1983	165,017.60
1984	168,683.20
1985	235,056.30
1986	221,293.80
1987	230,533.60
1988	236,346.60
1989	232,933.60
1990	246,791.60
1991	241,900.10
1992	225,628.70
1993	238,452.00
1994	238,452.00
1995	238,936.10
1996	357,275.70
1997	372,871.20
1998	393,602.50
1999	387,119.90
2000	343,966.10
2001	383,150.80
2002	392,626.30
2003	471,696.00
2004	449,855.00
2005	481,784.40
2006	472,174.80
2007	490,588.50
2008	459,918.10