

Improving Imputations of Top Incomes in the Public-Use March Current Population Survey by Using Both Cell Means and Variances

Richard V. Burkhauser
Cornell University

Shuaizhang Feng
Princeton University

Jeff Larrimore
Cornell University

The Problem

- For confidentiality reasons, the Census Bureau topcodes high incomes in the March Current Population Survey (CPS).
 - Topcoding is performed on each income source – 11 sources prior to 1987 and 24 sources since then – rather than on total personal income.
 - Topcodes vary over time and the changes are not tied to inflation.
 - In 2007, 5.7% of the population was topcoded on at least one source of household income
- Topcoding causes us to understate the income of top earners and, as a result, understate income inequality.
- Changes in topcode thresholds cause us to misstate the trends in earning inequality.

Objectives

- Using Internal March CPS data, determine the extent to which topcoding impacts measures of income inequality using different correction procedures
- Develop summary data on topcoded individuals that can be used in conjunction with Public-Use March CPS data to improve inequality estimates that are based on the public data.

Previous Topcode Correction Methods

We compare 6 methods of correcting for the topcode problem in the public use data to results using internal data. Four are previous correction methods or are based on the assumptions of earlier methods:

1. Unadjusted Public Use Data

Prior to 1996, topcoded incomes are replaced with the topcode threshold. Starting in 1996, topcoded values are assigned a cell mean – the mean source income of all individuals who are topcoded from the specified income source in the given year.

2. Public Use – No Cell Mean

In all years topcoded incomes are replaced with the topcode threshold. Identical to the Unadjusted Public Use data before 1996 but removes cell-means after that year to create a consistent series.

3. Public Use – Multiple

Replaces topcoded incomes with 1.4 times the topcode threshold. This approach has commonly been used in the literature to account for the fact that the Public No Cell Mean series must understate top incomes.

4. Public Use – Pareto

- It is commonly assumed that the income distribution fits a Pareto distribution. This series uses our calculated cell mean from the internal data and observes the inequality measure if topcoded incomes fit a Pareto distribution around this mean with the standard variance of a Pareto distribution.
- Pareto imputation procedure is the same as the Stoppa imputation procedure described – except that incomes are assumed to fit a Pareto distribution rather than a Stoppa distribution.

New Topcode Correction Methods

We also offer two new topcode correction methods, which we believe are superior to the previously available methods:

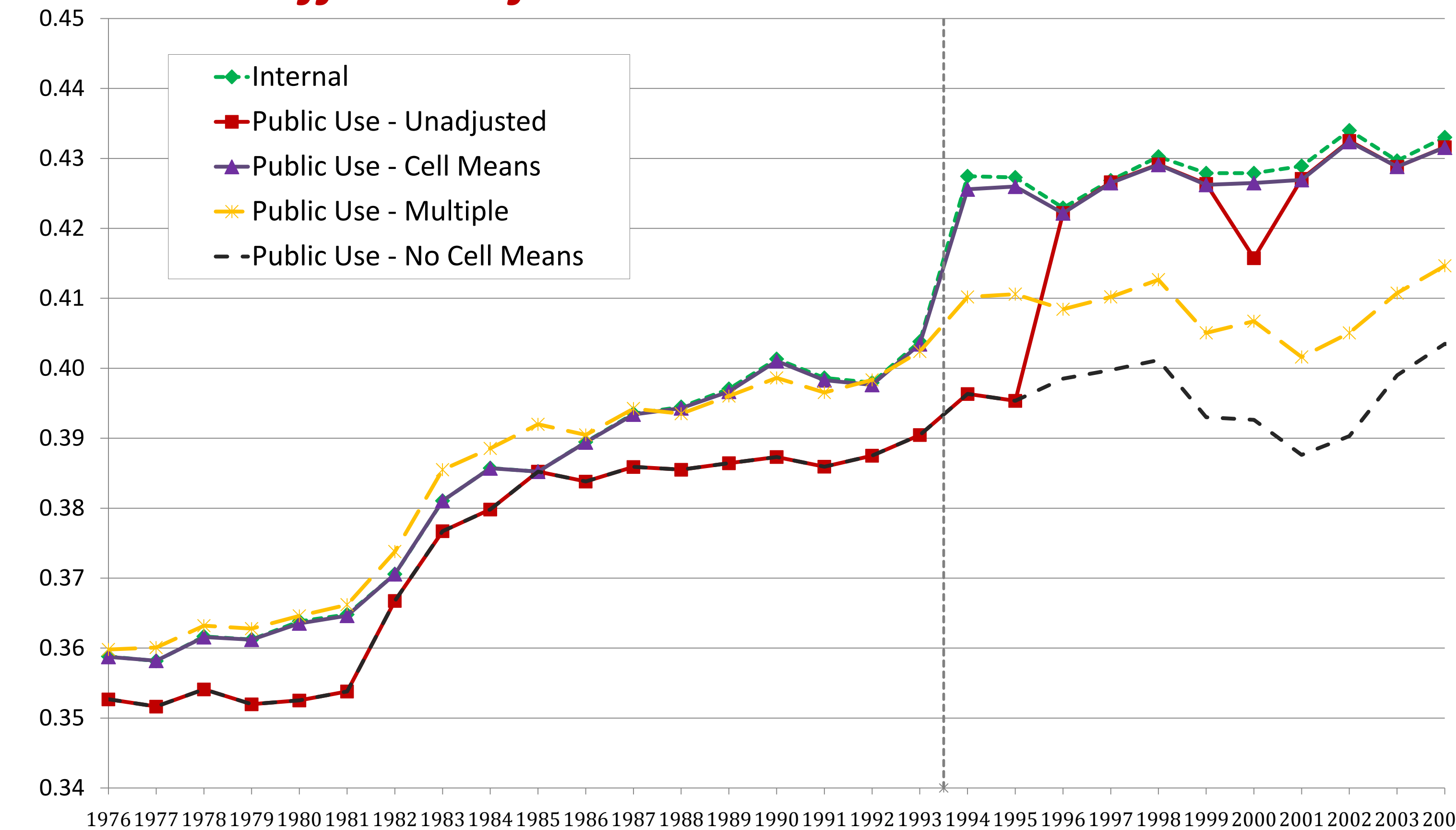
1. Public Use – Cell Means

- Using the internal March CPS data, we extended the Census provided cell means back to 1975 to create a consistent series that accurately captures the level of income held by topcoded individuals.
- Note that by design, cell means provide no information about the variance of topcoded incomes and assume all topcoded individuals receive the same income.

2. Public Use – Stoppa (Cell means and variances)

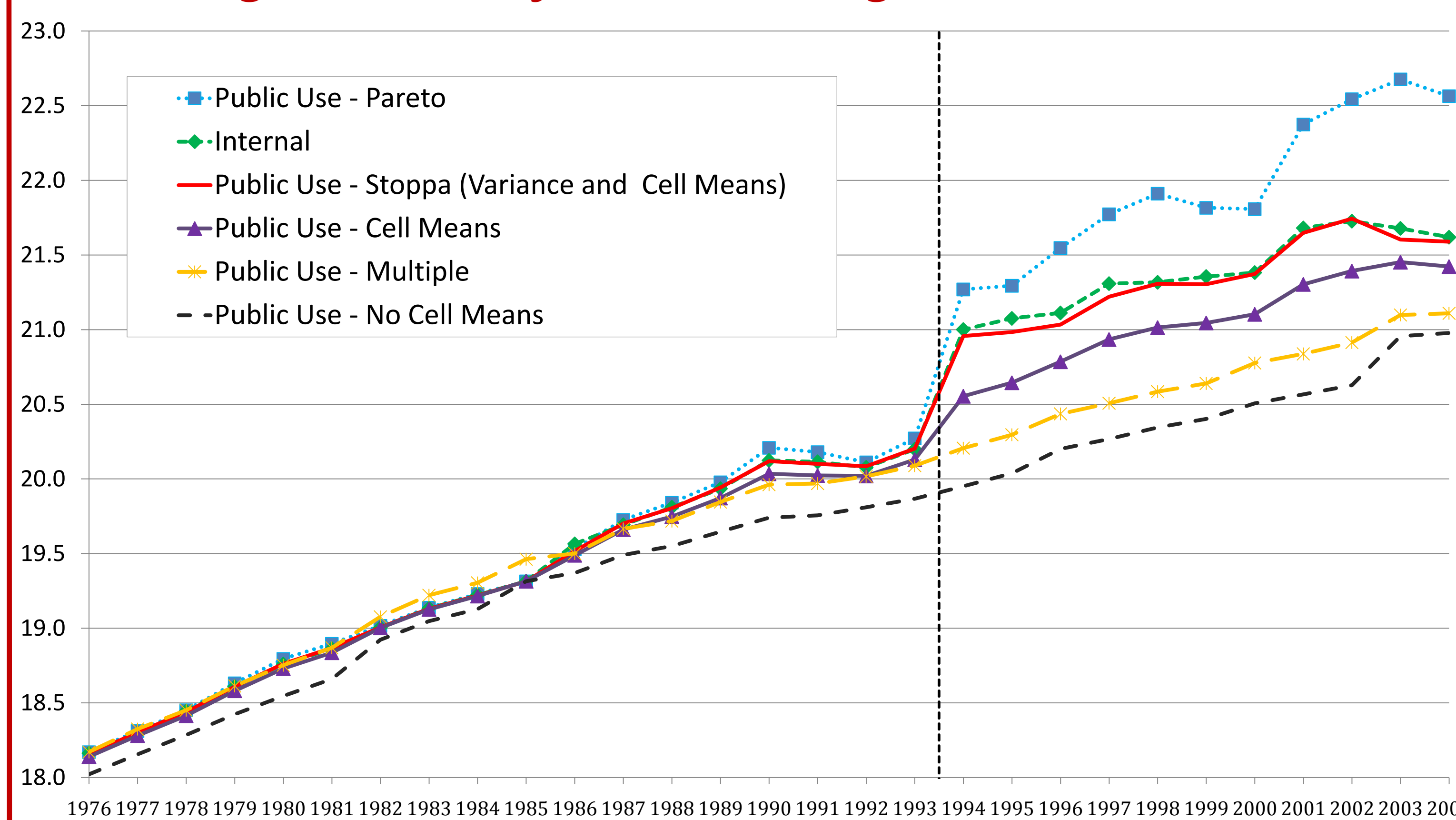
- Using the internal March CPS data, we calculated the variances of topcoded incomes from each income source in each year to complement the cell-mean series.
- This series captures both the level and distribution of income held by topcoded individuals by assuming topcoded incomes fit a Generalized Pareto (Stoppa) distribution with the cell means and variances from the internal data.

Gini Coefficient of Household Income in the March CPS



➤ The Public Use Pareto and Public Use Stoppa series are excluded given that cell means alone without additional variance information can very closely capture the results from the internal data

Log-Variance of Labor Earnings in the March CPS



➤ The Unadjusted Public Use series is excluded for the sake of clarity since it is simply a combination of the Public Use Cell Means and Public Use No Cell Means series, and the inconsistency in 1996 makes it inferior to either of these series independently

Labor Earnings Topcode Thresholds

Survey Year	Primary Labor Earnings ¹	Wage and Salaries	Self Employed Earnings	Farm Earnings
1976-1981	---	50,000	50,000	50,000
1982-1984	---	75,000	75,000	75,000
1985-1987	---	99,999	99,999	99,999
1988-1995	99,999	99,999	99,999	99,999
1996-2002	150,000	25,000	40,000	25,000
2003-2004	200,000	35,000	50,000	25,000

¹Prior to 1987, there were just 3 labor earnings categories (wages, self employment, and farm earnings). After 1987, earnings from primary employment regardless of source was separated into its own category and the other 3 sources represent earnings from secondary sources.

²In cases where under 5 individuals are topcoded from an income source, the cell-mean is combined with a similar income source to produce the cell-mean in order to protect the confidentiality of respondents. In 1995, only 3 people were topcoded on Farm income so the cell mean is combined with self employment earnings to reach the 5 person threshold for a topcode

Sample Labor Earnings Cell Means

Survey Year	Primary Labor Earnings ¹	Wage and Salaries	Self Employed Earnings	Farm Earnings
1981	---	66,367	70,528	61,356
1984	---	90,447	91,829	83,154
1987	---	140,026	135,346	122,398
1995	188,180	177,066	319,060 ²	319,060 ²
2002	325,382	55,636	99,362	167,762
2004	402,884	82,643	110,899	77,543

Disclaimer: The research in this poster was conducted while the authors were Special Sworn Status researchers of the U.S. Census Bureau at the Cornell Census RDC. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

Approach for Testing Topcode Corrections

- We consider two inequality measures:
 - The Gini coefficient of size-adjusted household income (which includes all 24 sources of income) for all individuals.
 - The Log-Variance of personal labor earnings (which includes only wage earnings, self-employment, and farm income) for working age individuals.
- Using these indices we measure US income inequality since 1975 using the 6 topcode correction methods and compare results to those in the internal data.

Explaining the Stoppa Imputation

➤ CDF of the Stoppa Distribution is:

$$F(y) = [1 - (\frac{y}{y_0})^{-\alpha}]^{\theta}$$

where y_0 is a lower bound, in this case representing the topcode threshold, and the parameters α and θ specify the shape of the distribution. The Pareto distribution is a special case of the Stoppa distribution when $\theta=1$.

➤ Use a Multiple Imputation approach to determine inequality statistics based on this distribution:

- For non-topcoded observations, use the actual value.
- For topcoded observations draw a value from the fitted Stoppa distribution with α and θ specified to fit the mean and variance from the internal data.
- Calculate inequality statistics using these estimates.
- Repeat the process 100 times for each year and average the 100 estimates to generate the reported statistics.

Results

Comparing the Gini Coefficient for Household Income

- The Unadjusted Public Use series underestimates income inequality prior to 1996 and has a clear trend-break in 1996 with the introduction of cell-means
- The Public Use No Cell Means series underestimates inequality in all years
- The Public Multiple series did relatively well approximating the internal values prior to 1993. But after 1993 when Census procedures for collecting top incomes improved it underestimated income inequality. It also picks up the artificial increases and decreases in inequality from topcode changes, as seen in the late 1990s.
- Unlike these three earlier series, the Public Cell Mean series does a good job of capturing the level of income inequality based on the Gini coefficient in all years, not in a portion of the sample period.
- Public Stoppa and Public Pareto are not necessary given that the Cell Mean series alone without the added variance data very closely matches the internal results.

Comparing the Log Variance of Labor Earnings

- As was seen for the Gini coefficient, the Public Use with No Cell Means series and the Public Multiple series) underestimate inequality, especially after 1993.
- For the log-variance, which is more sensitive to variance changes at the top of the distribution than the Gini, even the Public Cell Mean series underestimates inequality.
- The Public Use Pareto series, on the other hand, overestimates inequality because it assumes too much variance at the top of the distribution.
- The Public Use Stoppa series, however, very closely matches the results seen in the internal data as it accurately captures the variance at the top of the distribution.

Conclusions

- Previously available topcode correction methods will underestimate the level of income inequality seen in the internal data – especially after 1992 when Census collection procedures improved.
- For some income inequality statistics, such as the Gini coefficient, the Internal results can be very closely replicated by using our extended cell mean series.
- For other inequality statistics, such as the log-variance of earnings, adding the variance of topcoded incomes matters, but for such inequality statistics the internal data can be very closely replicated by using both the cell means and variance data.