Journal of Privacy and Confidentiality
Vol. 8 (1) 2018 Issue in honor of S. Fienberg
Submitted Mar. 20, 2017
Published Nov, 2018

PUBLIC-USE VS. RESTRICTED-USE: AN ANALYSIS USING THE AMERICAN COMMUNITY SURVEY

SATKARTAR K. KINNEY AND ALAN F. KARR

RTI International, 3040 East Cornwallis Road, Research Triangle Park, NC 27709 USA
e-mail address: skinney@rti.org

RTI International, 3040 East Cornwallis Road, Research Triangle Park, NC 27709 USA
e-mail address: karr@rti.org

Abstract. Statistical agencies frequently publish microdata that have been altered to protect confidentiality. Such data retain utility for many types of broad analyses but can yield biased or insufficiently precise results in others. Research access to de-identified versions of the restricted-use data with little or no alteration is often possible, albeit costly and time-consuming. We investigate the advantages and disadvantages of public-use and restricted-use data from the American Community Survey (ACS) in constructing a wage index. The public-use data used were Public Use Microdata Samples, while the restricted-use data were accessed via a Federal Statistical Research Data Center. We discuss the advantages and disadvantages of each data source and compare estimated CWIs and standard errors at the state and labor market levels. We find the results from the publicly available data are generally good relative to the restricted-use data, with greater similarity for larger areas and less similarity for smaller areas. Standard errors are higher in the public-used data but may still be underestimated.

1. Introduction

American Community Survey (ACS) microdata are released publicly in the form of Public Use Microdata Sample (PUMS) files [U.S. Census Bureau, 2015b] published annually by the Census Bureau and made available for download on their website.[1] We refer to these data as ACS/PUMS. In order to allow public release without violating confidentiality pledges to respondents, ACS/PUMS data are subsetted and altered to protect against disclosure of personally identifiable information (PII). Modifications include a small amount of data swapping—the exact amount is not disclosed, top-coding, rounding, coarsening, and limited geographic detail. See Crimi and Eddy [2014] for a discussion of top-coding and other disclosure protection steps applied to ACS/PUMS data. The most detailed geographic information available in ACS/PUMS data is the Public Use Microdata Area (PUMA), a statistical geographic area designed to have at least 100,000 residents.

Key words and phrases: Data swapping; Top-coding; Data quality.

[1] https://www.census.gov/programs-surveys/acs/data/pums.html.

www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.661

The ACS/PUMS data are used widely by researchers and policy makers, sometimes with potentially biased results resulting from modifications to the data. In particular, because of top-coding and data swapping, the relationship between age and wages can be distorted [Alexander et al., 2010]. Wages in ACS/PUMS are top-coded: values exceeding a threshold are reported as equal to a replacement value. In fact, state-dependent thresholds are employed, which range from $209,000 for Montana to $607,000 for the District of Columbia. For the ACS/PUMS, the marginal distribution of wages is preserved approximately at the state population level by using replacement values for each state that differ from the truncation value and are chosen to preserve weighted population means [Crimi and Eddy, 2014]. Age is also top-coded in ACS/PUMS data.

Alternatively, research access to confidential, including ACS, data collected by the Census Bureau and other agencies is possible at approximately 30 Federal Statistical Research Data Centers (FSRDCs). Agencies frequently work with Census to produce estimates and special tabulations, but the FSRDCs are specifically for research. Data access via a FSRDC requires a proposal and approval process, including background checks on researchers, in addition to travel to a FSRDC, as well as user or institutional fees. The approval process, while straightforward, can take several months.[2]

Data in the FSRDCs are de-identified, and are altered for disclosure protection, but significantly less so than public-use data. For example, wages in ACS data available in FSRDCs are top-coded at a much higher threshold than in PUMS. Detailed geography is available, down to the Census block level.

My professional relationship with Steve began in the early 1990s, when I came to NISS as Associate Director and he was a member of the Board of Trustees. We sometimes disagreed, or perhaps more accurately, I failed to grasp his wisdom. Something must have worked, though, because Steve also chaired the committee that selected me to be Director of NISS.

Our scientific collaboration arose in late 1990s, when I was PI, and he co-PI, on two grants from NSF's Digital Government initiative. These grants, as did the entire collaboration, stemmed from Steve's fervent belief that deep mathematics can be brought to bear on pressing personal and societal problems. The first had to do with web-based query systems now known as restricted data access systems (RDAS), and specifically with table servers. We were frontiersmen together in formulating and applying risk-utility frontiers, released table frontiers and unreleasable table frontiers.

With his usual prescience, Steve knew before data breaches were daily news that privacy and confidentiality are major concerns. We wrote only a few papers together, but we exchanged sometimes wildly complementary ideas for more than twenty years. I still remember a meeting with a number of federal statistical agencies at which what I proposed as a risk measure was exactly what Steve construed as a utility measure.

From the science grew a multi-year, multi-continent friendship that drew in Joyce and Senora as well. It mattered not whether the last encounter was three weeks or three years ago. Sadly, only one of the four of us now remains, but in keeping with the advice of Dr. Seuss, instead of crying because what ended, I smile because what happened.

Alan Karr

DOI: 10.29012/jpc.691

This paper investigates the difference between using the restricted-use ACS data [U.S. Census Bureau, 2015a], which we will refer to as the ACS/FSRDC, and the ACS/PUMS data. The context of comparison is statistical modeling, used to construct a Comparable Wage Index (CWI), that predicts wages from age, gender and other variables. The two data

[2]See `http://www.census.gov/fsrdc` for details.

sources produce results that are more similar for large areas, but less similar for smaller areas. Standard errors are higher in ACS/PUMS, although they may still be underestimated.

A description of the data sources and the CWI appear in §2. The public-use and restricted-use wage indices are contrasted in §3, and §4 summarizes the relative advantages and disadvantages of using public-use and restricted-use data.

2. The Wage Index

The CWI is an index calculated and published by the National Center for Education Statistics (NCES) for the years 1999 to 2005. It is intended primarily to facilitate comparison across states and labor market areas of educational expenditures, typically on a per-pupil basis, at the elementary and secondary (K-12) levels, as well as in allocations [Taylor and Fowler, Jr., 2006]. In part, the CWI functions by resolving controllable and uncontrollable expenditures on instructional salaries, which constitute the largest category of expenditures at the K-12 level. The CWI was developed using PUMS data from the 2000 long-form Census for the base year and then updated using Occupational Expenditure Survey (OES) data from the Bureau of Labor Statistics.

Using a slightly different approach from that in Taylor and Fowler, Jr. [2006], which we describe in §2.3, we calculated CWIs using both ACS/PUMS data and ACS/FSRDC data. These two versions are compared in §3.

2.1. **Data Preparation.**

Starting with the ACS/FSRDC and ACS/PUMS data files for 2010, the samples were refined to include only employed college graduates. Specifically, the samples are restricted to persons with Bachelors degrees (or higher) with a minimum salary of $5,000 who worked at least 20 hours a week for at least 27 weeks in the year prior to responding to the ACS questionnaire. Place of work was restriced to the fifty states and the District of Columbia. Table 1 lists the ACS/PUMS and ACS/FSRDC variables used in constructing CWIs. For details about the ACS/PUMS variables, see U.S. Census Bureau [2011].

Log transforms were applied to annual wages and to the number of hours worked per week. Based on Carrillo and Karr [2013], both age and its square appear in the model. Several categorical predictors, including Industry, Occupation, and Field of Degree were coarsened in order to simplify the analysis. All Hispanic categories were combined into a single indicator for Hispanic ethnicity.

The variables Labor Market and Place-of-Work PUMA were not on the ACS files, and so were merged into the sample using Census place-of-work variables for State, County, and Place. The ACS is released in single-year and multi-year formats.[3] In this paper, we focus on the single-year 2010 data files. Results examined from other years and multi-year files were similar.

[3]See https://www.census.gov/programs-surveys/acs/guidance.html for more details.

Table 1: Variables used in the CWI model

Variable	**ACS/PUMS Name**	**Remarks**
Wages/Salary	WAGP	Must be at least $5,000; top-coded
Age	AGEP	Top-coded
Sex	SEX	
Race	RACE1P	Recoded to 9 values
Ethnicity	HISP	
Education	SCHL	Must have Bachelors or higher
Field of Degree	FOD1P	2009 and later; coarsened
Occupation	SOCP	No exclusions; coarsened
Industry	INDP	No exclusions; coarsened
Place of Work PUMA	POWPUMA	Defined by Census Bureau
Place of Work State	POWSP	
Hours Worked per Week	WKHP	Must be at least 20
Weeks Worked per Year	WKW	Must be at least 27
Base Weight	PWGTP	
Replicate Weights	PWGTP1, ..., PWGTP80	

2.2. **Labor Markets.** A "labor market" is an area in which employers compete for labor. Under perfect competition and perfect worker mobility, all workers in the same labor market with the same skills and doing the same work would earn the same salary. In this case, salary variation *among* labor markets would reflect such factors as desirability of particular locations and differences in the cost of living. In reality of course, various frictional forces, such as certification and residency requirements for teachers, as well as differing student populations and workforce composition, lead to salary variation across Local Education Agencies (LEAs) even within labor markets. The CWI is meant to capture *inter*-labor market variations, and not intra-labor market variations.

In urban areas, labor markets correspond approximately to metropolitan areas. In our analysis, labor markets are made up of one or more PUMAs, which are the smallest geographic units available in ACS/PUMS. While PUMAs are not designed for this purpose, labor markets can be approximated by aggregating contiguous PUMAs. The labor markets used in this analysis are based on either Place-of-work PUMA or the Place of Work-Metropolitan Area variable POWMETRO as defined by the Integrated Public Use Microdata Series projects (IPUMS) [Ruggles et al., 2010]. Even though the ACS/FSRDC files contain detailed place-of-work geography, down to the Census block level, we use the PUMA-based labor markets for both datasets in order to allow a direct comparison between them.

An advantage of PUMAs is that they represent areas with minimum population areas of 100,000, and large populations are necessary to achieve sufficient accuracy. A disadvantage is that they are numerous—there are more than 1200 POWPUMAs—and there is no official labor market definition for place-of-work metro that depends on PUMAs or POWPUMAs. We borrowed the definition from IPUMS (POWMETRO); however, these metropolitan area definitions do not appear to be regularly updated. A further disadvantage is that PUMAs are not in widespread use. Users may be unfamiliar with them, and they complicate tasks such as merging PUMS data with datasets that use different geographical aggregation.

While some further aggregations of POWPUMAs or POWMETROs may be warranted, we made no adjustments to the IPUMS definition of POWMETRO. Thus, the PUMA-based labor market is constructed as follows: If a POWPUMA is in an IPUMS POWMETRO,

then it is assigned to that metro. If not, it becomes its own labor market. This process yields 789 labor markets.

Many POWPUMAs are aggregations of PUMAs, and roughly correspond to Core Based Statistical Areas (CBSAs), though neither PUMAs nor POWPUMAs strictly respect county boundaries. They do respect state boundaries. However, POWMETROs in urban areas may overlap multiple states.

2.3. **Wage Model.** While our modeling approach differs from Taylor and Fowler, Jr. [2006], the overall construction and interpretation are similar. The CWI for a given labor market (or state) is calculated as the ratio of its predicted wage level to a predicted national wage level. The predicted wage level for a labor market is the average wage one would expect *if everyone in the nation, as represented by the weighted ACS sample, worked in that market.* Therefore, differences in CWIs are not confounded with differences in workforce characteristics.

For simplicity, we ignore sample weights in our notation. For labor markets ℓ that lie entirely with one state S, the predicted labor market wage level is obtained using a linear mixed effects model:

$$\overline{\log\left(\mathrm{Sal}(\ell)\right)} = \frac{1}{N}\sum_i \left[X(i)\cdot\hat{\beta} + \widehat{\gamma_\ell} + \widehat{Y_S}\right] = \overline{X}\cdot\hat{\beta} + \widehat{\gamma_\ell} + \widehat{Y_S}. \tag{2.1}$$

For labor markets ℓ intersecting more than one state, we employ:

$$\overline{\log\left(\mathrm{Sal}(\ell)\right)} = \frac{1}{N}\sum_i \left[X(i)\cdot\hat{\beta} + \widehat{\gamma_\ell} + \sum_{S:\cap\ell\neq\emptyset} w_S\widehat{Y_S}\right] = \overline{X}\cdot\hat{\beta} + \widehat{\gamma_\ell} + \sum_{S:\cap\ell\neq\emptyset} w_S\widehat{Y_S}, \tag{2.2}$$

where the w_S are weighting factors defined by population fractions. Essentially, when we assume everyone in the nation lives in a given labor market, they must necessarily live in the state(s) corresponding to that labor market, so the labor market and state coefficients are changed correspondingly and predictions made. Where the labor market overlaps states, the state effects are averaged according to the worker population-weighted fractions of the states covered by the labor market. The variables used in the model are those listed in Table 1.

For state-level CWIs, we employ a similar approach; however, if state is the unit of analysis it makes sense for it to be a fixed effect. It may also make sense for labor market to be a random effect in this case; however, for this analysis a fixed effects model was used. Because most states overlap or contain multiple labor markets, we use an equation analogous to (2.2):

$$\overline{\log\left(\mathrm{Sal(s)}\right)} = \frac{1}{N}\sum_i \left[X_i\cdot\hat{\beta} + \widehat{Y_s} + \sum_{l:\cap s\neq\emptyset} w_l\widehat{\gamma_l}\right] = \bar{X}\cdot\hat{\beta} + \widehat{Y_s} + \sum_{l:\cap s\neq\emptyset} w_l\widehat{\gamma_l} \tag{2.3}$$

Finally, the CWI for region A (labor market or state) is computed as:

$$CWI_A = \frac{\overline{\exp[\log(\mathrm{Sal(A)})]}}{\overline{\exp[\log(\mathrm{Sal}(*))]}} \tag{2.4}$$

where $\overline{\exp(\log(\mathrm{Sal(A)}))}$ is the weighted average predicted regional wage level if every one in the sample lived in region A and $\overline{\exp(\log(\mathrm{Sal}(*)))}$ is the predicted national wage level, computed as a weighted average of predicted wages over the sample.

Standard errors for CWIs are estimated using sets of 80 replicate weights that appear in both the ACS/FSRDC and ACS/PUMS files. The procedure is straightforward, but

extremely demanding computationally. The methodology used by the Census Bureau to construct the replicate weights is a version of successive difference replication, and is described in U.S. Census Bureau [2009].

3. The Comparisons

In this section we compare CWIs produced using ACS/FSRDC and ACS/PUMS data. The CWIs and their associated standard errors were calculated for 51 (including the District of Columbia) states, and separately for the 789 labor markets. As noted, for simplicity, we focus on data from 2010; however, similar comparisons were done for other years and using multi-year ACS data, with similar results.

We found only modest differences between state-level CWIs produced from ACS/PUMS datasets and those produced from ACS/FSRDC datasets. Larger differences were observed for labor markets. Relationships among standard errors are more complex, in part because of the more severe statistical disclosure limitation (SDL) applied to the ACS/PUMS datasets. Figure 1 contains maps showing ACS/FSRDC and ACS/PUMS state-level CWIs for 2010. States that have higher labor costs are shown to have higher CWIs. The District of Columbia, which has the highest state CWI, is too small to appear on the maps. Differences between ACS/FSRDC and ACS/PUMS estimates are small enough that they are nearly indistinguishable on these maps.

Figure 2 shows similarities and differences between ACS/FSRDC and ACS/PUMS CWIs more clearly. Panel (a) illustrates the high correlation between the two versions of the state-level CWI estimates, with most points clustering along the diagonal line, which represents equality, and a few modest differences at the lower end of the range. There is also a high correlation for labor market-level CWIs, as shown in Panel (b) of Figure 2. At the labor market level, however, there is a great deal more spread around the diagonal.

Figure 3 shows a high correlation between the standard error estimates, though it is evident from these scatterplots that ACS/PUMS standard errors are generally larger. Moreover, Figure 4 shows that the discrepancy increases as standard errors increase and sample sizes decrease. This is expected given the disclosure control measures applied. In Figure 3 we see that some ACS/PUMS standard errors are smaller than ACS/FSRDC standard errors. This may be due to a differing impact of disclosure control in these regions: top-coding of salaries has artificially decreased the range of salary values.

As an illustration of CWI usage in practice, Figure 5 compares state per-pupil expenditures (SPPE) when adjusted by the CWI from ACS/PUMS and when adjusted by the ACS/FSRDC-calculated CWI. Expenditures are adjusted by dividing by the appropriate CWI. Again, we see a close correspondence between ACS/FSRDC and ACS/PUMS results for state-level data. Lastly, we include a comparison for a five-year data file, 2005–09, in Figure 6, which illustrates that the issues persist in the larger ACS/PUMS data files.

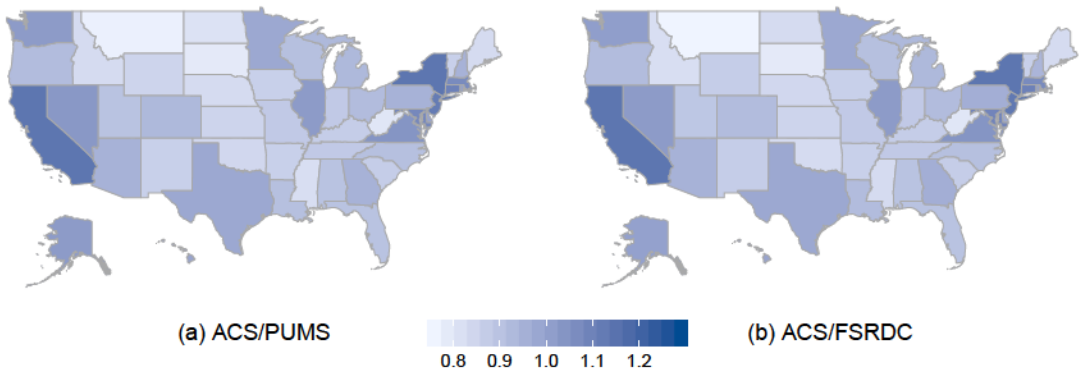

Figure 1: 2010 state-level CWIs calculated using ACS/PUMS data (a) and ACS/FSRDC data (b).

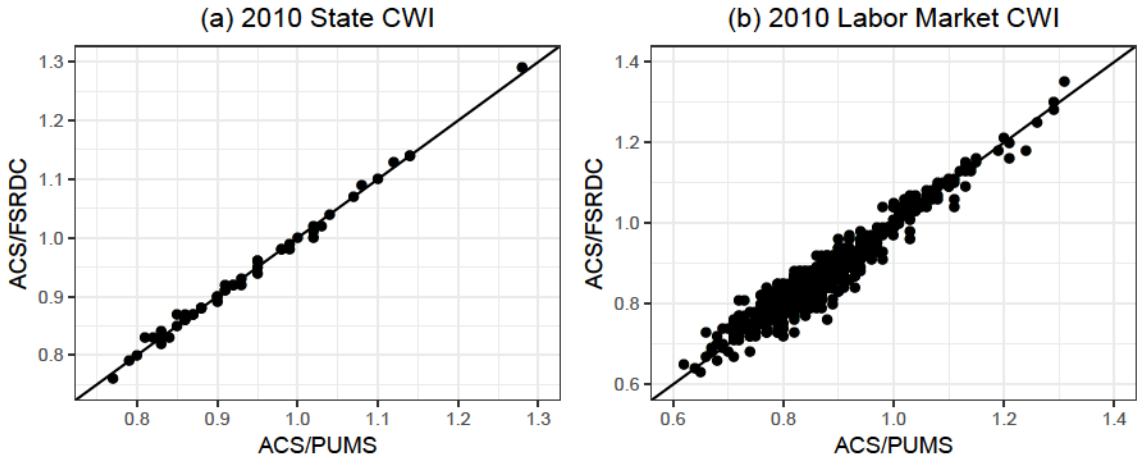

Figure 2: Comparison of ACS/FSRDC and ACS/PUMS CWIs, by states (a) and labor markets (b).

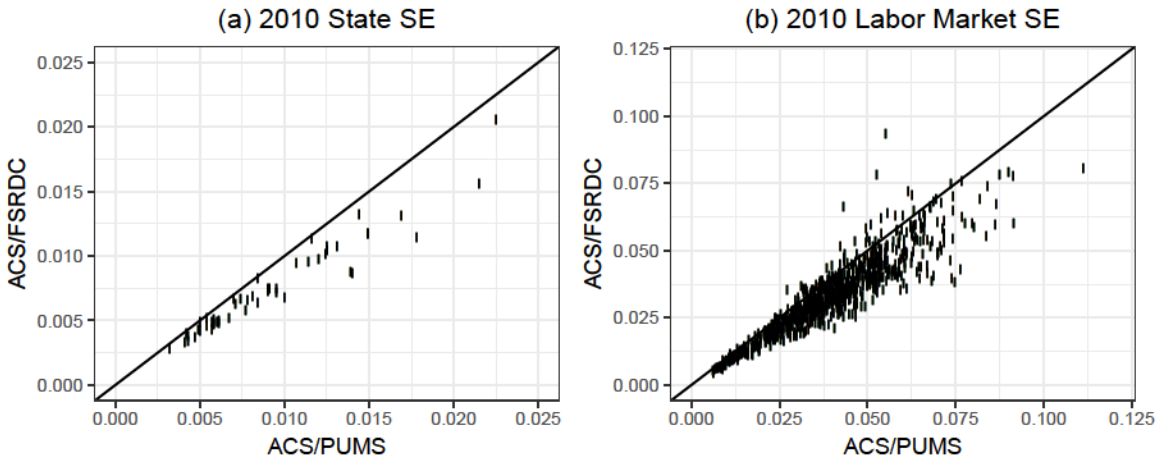

Figure 3: Comparison of ACS/FSRDC and ACS/PUMS standard errors, by states (a) and labor markets (b).

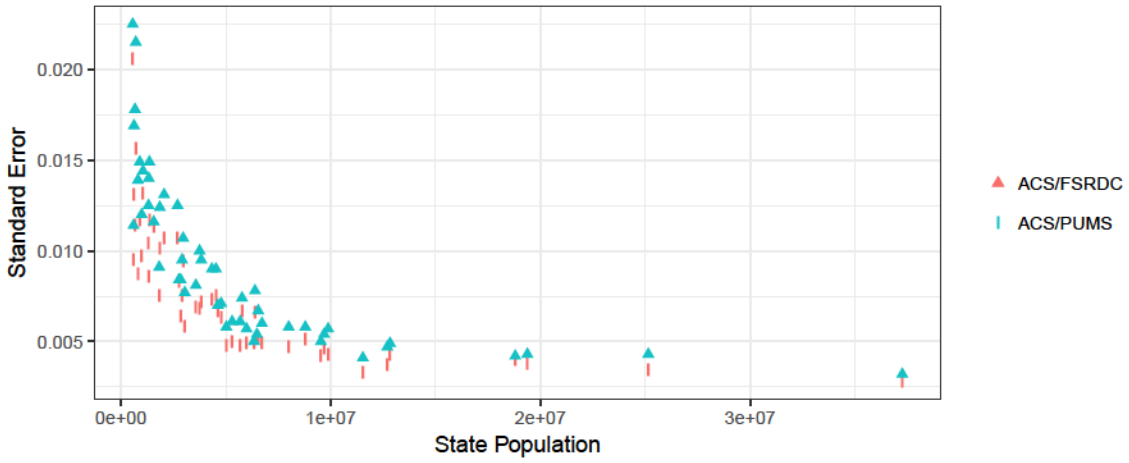

Figure 4: State-level standard error vs. population: ACS/PUMS data and ACS/FSRDC data.

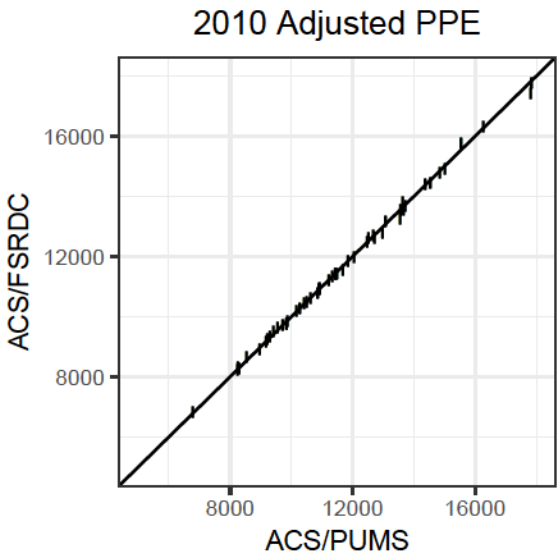

Figure 5: Adjusted per-pupil expenditure, ACS/PUMS data vs. ACS/FSRDC data.

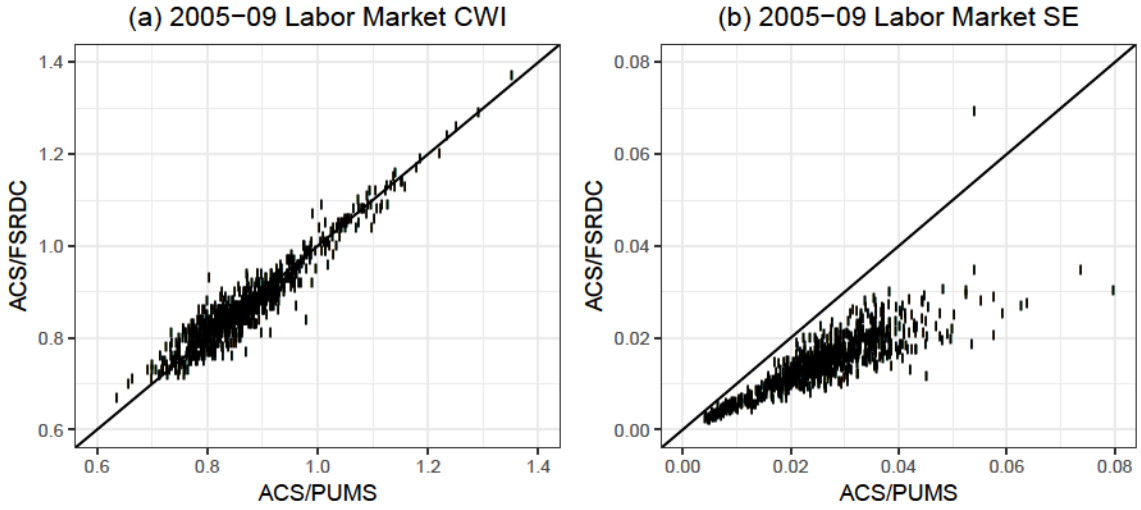

Figure 6: Comparison of 2005-09 ACS/FSRDC and ACS/PUMS CWI estimates (a) and standard errors (b), by labor market.

4. Discussion

The advantages of the restricted-use data are several. First, the sample sizes are larger, by about one-third. While there is a small degree of top-coding in the ACS/FSRDC data, the threshold is much higher than for ACS/PUMS. To our knowledge, there are not other disclosure control procedures applied. Additional variables are available in ACS/FSRDC data that are not provided in ACS/PUMS data. Those relevant to our analysis include CPI-adjusted wages, source of health insurance (e.g., employer or union), urbanicity, and crucially, fine-level geography. While the finest level geography available on ACS/PUMS is PUMA, ACS/FSRDC data contain workplace locations at the Census block level.

Comparisons between public-use and restricted-use estimates were quite close for states, likely due to the top-coding methodology which was designed to preserve state-level means. Comparisons for smaller areas were much noisier with larger standard errors, or in some cases, underestimated when public-use data are employed. Standard errors tend to be underestimated when top-coding and data swapping are applied, particularly for small domains, as analysts do not have methods to properly account for bias or uncertainty due to the disclosure treatment applied. Analyses restricted to wages below the top-coding thresholds, which can be determined by inspection of the data, will not be impacted by top-coding but may still be impacting by data swapping.

Top-coding and data swapping are said to be *nonignorable* disclosure treatment methods because the estimates obtained from the treated data differ from what would be obtained from the confidential data; however, ACS/PUMS, and similar data, are usually analyzed as if the the disclosure treatment is ignorable. If the parameters of the treatment are available, as is the case for top-coding, then it is possible to conduct an *SDL-aware* analysis. The precise details of the swapping rate and algorithm are secret, confounding SDL-aware analyses and evaluation of the uncertainty introduced by the swapping. The uncertainty introduced by swapping for ACS, however, is reported to be less than that introduced by data editing and imputation [Abowd and Schmutte, 2015].

These comparisons presented were done using the same model for both public-use and restricted-use data; however, the restricted-use data has additional information available that can allow for construction of better models. As an example, we compare results using the 'limited model' used for comparison with ACS/PUMS, and an improved model utilizing features available in ACS/FSRDC that were unavailable in PUMS. As seen in Figure 7, the results are highly correlated; however, researchers will prefer to utilize the best data and model available to them, and will need to consider the tradeoffs in deciding which data source to use.

Other issues can arise when using ACS/PUMS data. The statistical disclosure control methods applied to ACS/PUMS data have in the past resulted in errors that took considerable time to be detected and corrected [Alexander et al., 2010, Crimi and Eddy, 2014]. We were unable to use the five-year ACS/PUMS for 2006–10 because at the time of our work, there was not a consistent industry code for the whole file due to the switch from the Standard Industrial Classification (SIC) system to the North American Industry Classification System (NAICS). This problem had been addressed in the ACS/FSRDC file.

The key advantage of public use data, of course, is their ready availability. This is no trivial matter, and as we have shown, it is possible to get good results from them for large samples, particularly for states, which seem likely to be sufficient for many purposes. Of

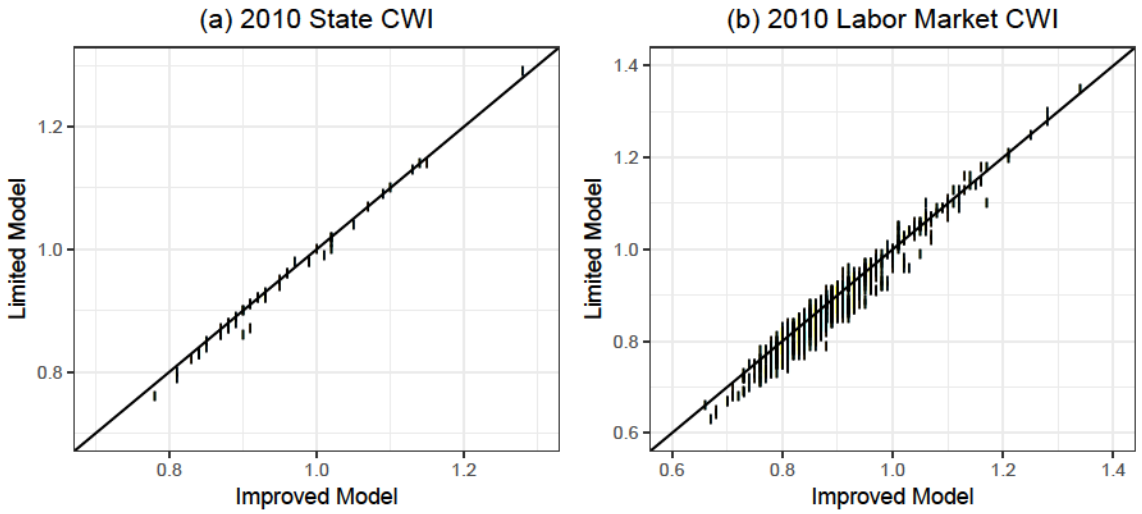

Figure 7: Comparison of improved model and limited model by state (a) and labor market (b), for 2010 CWI, using ACS/FSRDC data.

course, we could not have determined that our estimates derived from the ACS/FSRDC and ACS/PUMS datasets were substantially similar without access to both.

The high correlation between the ACS/PUMS and ACS/FSRDC state-level CWI suggests that using ACS/PUMS data to generate state-level estimates may be appropriate, while greater caution may be warranted for smaller areas. There is some indication that CWIs developed from ACS/PUMS for small areas, using similar methodology to Taylor and Fowler, Jr. [2006], have been used to inform policy decisions. (See for example, Taylor [2012, 2015], TXSmartSchools [2016].) In cases where small differences in an ACS estimate translate into appreciable financial impacts, it might be prudent in such cases to make sure the quality and uncertainty associated with such estimates is better understood, and to use the best data, models, and geography, available, subject to cost considerations and availability.

Disclaimer and Acknowledgments

This work was supported by the National Center for Education Statistics and the National Institute of Statistical Sciences. We thank Marie Marcum for helpful discussions and Ivan Carillo-Garcia for assistance with the ACS/PUMS computations.

A portion of this work was conducted by Special Sworn Status researchers of the U.S. Census Bureau at the Triangle Federal Statistical Research Data Center. Research results and conclusions expressed are those of the authors and do not necessarily reflect the views of the Census Bureau, the National Center for Education Statistics, or the National Institute of Statistical Sciences. Results have been screened to ensure that no confidential data are revealed.

References

J. M. Abowd and I. M. Schmutte. Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*, 2015.

J. T. Alexander, M. Davern, and B. Stevenson. Inaccurate age and sex data in the Census PUMS files: Evidence and implications. *Public Opinion Quarterly*, 74(3):551–569, 2010.

I. Carrillo and A. F. Karr. Combining cohorts in longitudinal surveys. *Survey Methodology*, 39(1), 2013.

N. Crimi and W. Eddy. Top-coding and Public Use Microdata Samples from the U.S. Census Bureau. *Journal of Privacy and Confidentiality*, 6(2), 2014.

S. Ruggles, J. T. Alexander, K. Genadek, R. Goeken, M. B. Schroeder, and M. Sobek. *Integrated Public Use Microdata Series: Version 5.0, [Machine-readable database]*. University of Minnesota, Minneapolis, 2010.

L. Taylor. An ACS-based regional cost adjustment for the state of washington, 2012. Available on-line at http://www.k12.wa.us/Compensation/Meetings/2012/WashingtonACSCostIndex.pdf.

L. Taylor. External cost adjustments for the Wyoming school funding model: 2015, 2015. Retrieved on-line at http://legisweb.state.wy.us/InterimCommittee/2015/SSRRpt1001AppendixD-1.pdf.

L. L. Taylor and W. J. Fowler, Jr. A comparable wage approach to geographic cost adjustment, 2006. NCES Research and Development Report, NCES 2006-321. Available online at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2006321.

TXSmartSchools. TXSMartSchools methodology: Technical description, 2016. Available on-line at http://www.txsmartschools.org/pdf/TXSmartSchoolsMethodologyTechnicalDescription.pdf.

U.S. Census Bureau. Design and methodology: American Community Survey, 2009. ACS–DM1, available on-line at https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html.

U.S. Census Bureau. 2010 ACS PUMS Data Dictionary, 2011. Available on-line at https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMSDataDict10.pdf.

U.S. Census Bureau. American Community Survey - restricted access, [machine-readable database], 2015a. URL https://www.census.gov/ces/dataproducts/demographicdata.html.

U.S. Census Bureau. American Community Survey - Public-Use Microdata Sample (PUMS), [machine-readable database], 2015b. URL https://www.census.gov/programs-surveys/acs/data/pums.html.