Improving Disclosure Avoidance Procedures for the Current Population Survey Public Use File[1]

Gary Benedetto, Assistant Center Chief, Center for Enterprise Dissemination – Disclosure Avoidance

Kyra Linse, Survey Director for the Current Population Survey, Demographic Programs Directorate

Eloise Parker, Assistant Director for Demographic Programs

U.S. Census Bureau

Federal Economic Statistics Advisory Committee (FESAC)

June 10, 2022

[1] Pre-decisional draft

Introduction

In accordance with the Census Act (13 U.S. Code §§ 8(b) & 9) and the Confidential Information Protection and Statistical Efficiency Act of 2018 (44 U.S. Code § 3563), the U.S. Census Bureau is required by law to ensure that data are released in a manner that protects the confidentiality of its survey respondents. As large stores of data are increasingly available in combination with advanced methods and technologies for matching data, the Census Bureau has grappled with the degree to which traditional disclosure avoidance practices continue to offer the protections they once did. Importantly, the agency's legal obligation to protect confidentiality can operate in juxtaposition with the Census Bureau's mission to release high-quality data that meets the needs of its data users.

Data from the Current Population Survey Public Use File (CPS PUF) are not immune to these challenges. In 2020, the Census Bureau conducted a re-identification study on the CPS PUF that revealed vulnerabilities in select geographies. In January 2022, the Census Bureau issued an announcement that it would mitigate these vulnerabilities by increasing the threshold for suppressing geographic areas from populations less than 100,000 to populations less than 250,000. In addition, the Census Bureau announced that it would round wages and earnings. Feedback from a broad constituency of CPS PUF data users about the way in which these additional protections would adversely impact the utility of the PUF prompted the Census Bureau to revisit its approach.

In anticipation of preparing the 2023 CPS PUF, this paper sets forth a revised proposal designed to meet disclosure avoidance requirements while preserving the PUF's value as a critical source of data for understanding the labor economy. Further, this plan would be phased in order to enable data users to conduct year-over-year analysis and other key research components.

Risk Mitigation

Current State for Geographical Detail

The CPS microdata files currently identify numerous substate areas within our confidentiality restrictions. These definitions are based on the results of the 2010 Decennial Census and were defined in Office of Management and Budget Bulletin No. 13-01, dated February 28, 2013. Within our confidentiality restrictions, indicators are provided for 260 selected core-based statistical areas (CBSA), 42 selected combined statistical areas (CSA), 277 counties, and 97 principal cities in multi-principal city core-based statistical areas or combined statistical areas. Within each of those areas, the metro status is also defined.

Historically, the Census Bureau has required that the minimum geography as defined by the intersection of CBSA and metropolitan status contain a population of at least 100,000. With the amount of external data continuously increasing and the computational means to use these data to attack Census Bureau data publications also expanding, the vulnerabilities in public data are becoming more evident. The Census Bureau must continuously evaluate disclosure risk and statistical disclosure limitation methods in public-use products. An internal reidentification study using national simulated attacker data revealed a risk of reidentification of respondents for the smallest geographies identified by the CPS. This study also revealed that a minimum population threshold of 250,000 was currently sufficient to adequately mitigate this risk. For the sake of communicating the problem and proposed solutions clearly, we provide an example with fictional data in a fictional state. Table 1 shows the geographies in this fictional example where columns 2 and 3 represent the full internal geographical detail which has never been

provided in the public files, columns 5 and 6 represent the current practices in the CPS respecting the 100,000 threshold, and columns 7 and 8 represent what geographical detail would be provided in the hypothetical situation where the traditional suppression approach was taken.

Table 1. Fictional example of geographies in a single state

State	Internal Value		Population Size Category for Geo	Old PUF Value		Hypothetical Suppressed Value	
	Geocode	Met Status		CBSA	Met Status	CBSA	Met Status
AA	1	1 (Metro)	4 (500,000+)	1	1 (Metro)	1	1 (Metro)
AA	2	1 (Metro)	4 (500,000+)	2	1 (Metro)	2	1 (Metro)
AA	3	1 (Metro)	3 (250,000-499,999)	3	1 (Metro)	3	1 (Metro)
AA	4	1 (Metro)	1 (<100,000)	Other	1 (Metro)	Other	3 (Not identified)
AA	5	1 (Metro)	2 (100,000-249,999)	5	1 (Metro)	Other	3 (Not identified)
AA	6	1 (Metro)	2 (100,000-249,999)	6	1 (Metro)	Other	3 (Not identified)
AA	7	1 (Metro)	1 (<100,000)	Other	1 (Metro)	Other	3 (Not identified)
AA	8	2 (Non-met)	1 (<100,000)	Other	2 (Non-met)	Other	3 (Not identified)
AA	9	2 (Non-met)	1 (<100,000)	Other	2 (Non-met)	Other	3 (Not identified)

Hypothetical Suppression of Risky Geographies

The simplest, and traditional, method for mitigating the confidentiality risk of low-population geographies in microdata is to coarsen (or suppress) the information contained in the geography fields, usually to meet a population threshold from which other protections, including sampling, usually provide sufficient confidentiality protection. In light of the recent analysis regarding risk of reidentification in the CPS, the Census Bureau would normally have coarsened all geographies identified by the intersection of the CBSA field and the metropolitan status field that fell beneath the 250,000 population threshold. As demonstrated in Table 1, the current practices would already combine geographies associated with the internal codes, 4 and 7, into one metropolitan category, and internal codes, 8 and 9, into one non-metropolitan category. With the new confidentiality standard, geographies with codes 5 and 6 would also need to be coarsened because they fall under the 250,000 threshold. This has a very large impact. Because these are both metropolitan areas, they would have to be combined with areas 4 and 7. However, while the combination of {Other, Metro} is now large enough, the combination of {Other, Non-metro} is now also below the 250,000 threshold. As a result, the detail on metropolitan status must be removed for all geographies in areas 4 through 9. This coarsening with suppression approach can have a very large impact on the utility of the data.

Proposed Partial Synthesis of Risky Geographies

Since the traditional approach of coarsening and suppression reduces the number of geographic areas available for some important use cases for the public use data, Census Bureau staff are exploring the use of partial synthesis of the geography fields to accomplish the same confidentiality protection goal of coarsening to the 250,000 population threshold while providing the same level of geographical detail (albeit perturbed for areas with populations below the 250,000 threshold) for data users. Partial synthesis (Little, 1993) is a powerful method in statistical disclosure limitation because, unlike suppression, it allows for many analyses to be preserved while meeting the same confidentiality

standards. With synthesis, a model of the sensitive variables conditional on all the other variables is estimated using the observed, internal data. Then, samples are drawn from the posterior predictive distribution implied by the estimation step, and these sampled values replace the values of the sensitive variable. As a result, the variables on the public use file have the same level of detail as before, and many of the underlying relationships in the data are preserved in the new, synthetic values.

We start by selecting only the respondents in the risky geographies for the partial synthesis. For these records, we model to which low-population geography within a state a household belongs. The model we use is a multinomial logistic regression conditioning on as many features of the household as the sample size permits. Going back to Table 1, we want to provide the same level of geographical detail as provided by the current CPS practices (columns 5 and 6) but with a lower disclosure risk to the respondents. Therefore, in the fictional example, for a sample defined by households in geographical areas 4 through 9, the values for the dependent variable will be defined by the {CBSA, Met Status}-tuples:

1. {Other, Metro}
2. {5, Metro}
3. {6, Metro}
4. {Other, Non-metro}

To add clarity to the mechanics of the proposed method, Table 2 shows a few records from fictional microdata associated with our fictional geographies discussed earlier.

Table 2: Fictional example continued with a small set of microdata records

		Internal	Old PUF Value					Hypothetical Suppressed Value		Proposed Synthetic Value	
Household	ST	Geo-code	CBSA	Metro Status	Geo Pop Size Category	House-holder Age	House-holder Earnings	CBSA	Metro Status	CBSA	Metro Status
1	AA	6	6	1	2	52	2200	Other	3	5	1
2	AA	6	6	1	2	35	600	Other	3	6	1
3	AA	4	Other	1	1	64	1900	Other	3	Other	1
4	AA	5	5	1	2	43	1700	Other	3	6	1
5	AA	7	Other	2	1	72	0	Other	3	Other	2
6	AA	3	3	1	3	38	800	3	1	3	1
7	AA	1	1	1	4	49	1200	1	1	1	1
8	AA	2	2	1	4	57	1000	2	1	2	1
9	AA	2	2	1	4	61	1500	2	1	2	1

Only households 1 through 5 fall into geographies that would need to be coarsened to address the reidentification risk, so only these records would have their geography variables replaced by a synthetic value. Moreover, as Table 2 shows, no other variable is affected by the partial synthesis; in this limited example, age and earnings remain exactly the same for all records. Also, one can see that even for the modeled records, the model is not prevented from predicting the original value. In this example, we see that the model predicts the value a data user would see if the current practices could be continued for households 2, 3, and 5, while households 1 and 4 have their geographies replaced with different values.

Impacts to the Data User

The Census Bureau's obligation to mitigate the reidentification risk must be addressed; the decision now is how best to do so. With both of the solutions – 1) applying traditional coarsening and suppression for geographies with populations less than 250,000; and 2) applying traditional suppression for geographies with populations less than 100,000 and partial synthesis for geographies with populations between 100,000 and 250,000 – all analyses at state or higher level are unchanged. For analyses that rely on geographies falling below the 250,000 threshold, the partial synthesis solution will result in a little more uncertainty relative to the uncertainty in the model as estimated on the sample. The Census Bureau will quantify this additional uncertainty for users. However, the alternative of suppression would remove the possibility of such analyses altogether. For partial synthesis, analyses that condition on metropolitan status will have a small amount of noise added based on the modeling of the relatively small number of households falling into risky geographies that would have had metropolitan status defined in the previous public use files. Suppression does not get around this problem. Coarsening geographies would also introduce more uncertainty, and systematic bias (because the suppressions create nonignorable missing data (Little and Rubin 2002), from dropping cases where metropolitan status is no longer identified. As we develop these alternatives further, we will closely examine the nature of the added uncertainty and bias of both approaches.

Precision in Wage and Earnings Data

A possible identifier can be a wage or earning. The goal is to remove the full precision reporting that can act as an identifier. The Census Bureau has sought to apply rounding to the minimal extent possible to eliminate specific reporting. For more common wages and earnings, less rounding will be required. As reported wages and earnings increase, the distribution becomes sparser, requiring a graduated rounding scheme.

A dynamic topcode will also now be implemented monthly, with the top 3 percent of earning reported being topcoded. Moving forward, this will result in fewer cases being topcoded. For individuals whose earnings have been topcoded, the earnings variable will contain the weighted average of the top 3% and not the cutoff value.

PTERNHLY, PTERNH1O, PTERNH2, and PTERNH1C are the hourly reporting variables. The current proposal for rounding is:

Usual Hourly Earning as Reported	**Rounded value on Public Use File**
$00.01 - $00.07	$00.05
$00.08 - $19.99	Nearest $00.05
$20.00 - $39.99	Nearest $00.25
$40.00 +	Nearest $00.50 until topcode

The weekly earnings variables (PTERNWA, PTERN, and PTERN2) will also be rounded minimally following the same distribution suppression.

Usual Hourly Earning as Reported	Rounded value on Public Use File
$0	$0
$1 - $7	$5
$8 - $1000	Nearest $5
$1001 +	Nearest $25 until topcoded

Because of the interest in workers earning at or below the prevailing Federal minimum wage of $7.25 per hour, the variable PRERNMIN will be added to the public use files that will identify workers who earn $7.25 per hour or less. To maintain confidentiality with the flag, the value of $7.25 will be rounded down.

Conclusion

The proposal set forth here represents the Census Bureau's recommendation for meeting its legal requirements for protecting the CPS PUF from disclosure while minimizing impacts to the utility of the data. Based on feedback from both the CPS PUF user community and disclosure experts, the Census Bureau plans to refine and release its final approach in September 2022. The additional protections are currently slated to be introduced with the release of the January 2023 CPS PUF.

References

Little RJA. 1993. Statistical analysis of masked data. *J. Off. Stat.* 9(2):407-407. https: //www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-analysis-of- masked-data.pdf cited on June 1, 2022.

Little RJA, Rubin D. 2002 Statistical analysis with missing data 2nd Edition. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781119013563