Preserving Privacy in Person-Level Data for the American Community Survey

Rolando A. Rodríguez[1], Michael H. Freiman[1],
Jerome P. Reiter[1,2], Amy D. Lauger[1]
[1]U.S. Census Bureau
[2]Duke University

Abstract

The Census Bureau is researching model-based synthetic person-level data that maintain many of the properties of the original American Community Survey (ACS) data while protecting individual privacy. Protecting the ACS while maintaining data quality presents particular challenges because of the ACS's sample weighting, the survey's large number of variables and the small geographies for which ACS data are desired. This paper discusses the reasons adapting existing differential privacy methods is difficult and describes the approaches we are investigating to protect the data, including tree-based methods for categorical or discrete variables and regression for continuous variables.

Key Words: privacy, confidentiality, synthetic data, survey data

1. Privacy in The American Community Survey

Federal statistical agencies often release data from official surveys and censuses under legal strictures requiring the protection of the identities and personal attributes of respondents. The means to this end often differ greatly among data products, due both to the nature of the particular information released and to the philosophy of privacy of the agency. The Census Bureau uses the term *disclosure avoidance* to describe methods and actions used to provide privacy to respondents. Disclosure avoidance methods historically and currently used at the Census Bureau include: removal of direct identifiers such as name and address, swapping of households with similar characteristics, and suppression of tables with small counts (Lauger, Wisniewski, & McKenna, 2014).

The American Community Survey (ACS) is the largest demographic survey conducted by the Census Bureau. It collects information on individuals including demographics, relationships, schooling, occupation, and income, and on domiciles, including location, utility services, and unit value. The ACS is the basis for the distribution of approximately 675 billion dollars of federal funding annually.

The ACS has several types of data releases, including:

- Profiles
- Tables
- Public-Use Microdata Samples (PUMS)
- Summary files

Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Disclosure avoidance procedures affect both the internal data underlying these releases and the releases themselves: tables generated from swapped households may still fail to meet count thresholds for release, for instance. The Census Bureau is researching solutions for disclosure avoidance that involve the creation of underlying privatized data that can serve as a source for all public releases from the ACS, without the need for per-release protections. These methods have additional theoretical advantages over the current methods. One such method is model-based synthetic data.

2. Model-Based Synthetic Data

Model-based synthetic data is data generated from a model and a dataset such that certain statistical properties of the dataset remain while the record-level data itself changes (Raghunathan, Reiter, & Rubin, 2003). Often the model is a Bayesian model on the data and the synthetic data are predictions (draws) from the posterior predictive distribution of the model. In that instance, one may compute Bayesian or asymptotic frequentist properties of a given statistical output from the data. This is an improvement over methods such as swapping, where one cannot calculate such theoretical properties.

In our research, we plan to generate what is called fully synthetic data for unweighted individuals in the ACS. To create fully synthetic data, we replace every variable for every respondent in the data with a prediction from a model (Drechsler, 2011). The extent to which that model captures the multivariate properties of the data determines which statistics will have the most quality after synthesis. Our goals are to generate a synthetic data set from which all other ACS data products may stem and to maximize the number of those products that maintain useful analytic properties.

Ideally we would generate synthetic data from a multivariate density across all the ACS columns, but specifying such a distribution across the scores of ACS variables is untenable due to numerous issues, including dimensionality. An alternative approach is to specify a sequence of conditional distributions to generate the synthetic data, which offers a certain degree of flexibility over a fully multivariate specification.

The theoretical workflow for generating fully synthetic data from conditional models is as follows:

1. Define an unconditional model for an initial set of variables:

   $f(X_{init}|\Theta_{init})$

2. Define a conditional model for the next variable or set of variables:

   $f(X_i|X_{init}, \Theta_{init}, \Theta_i)$

3. Define conditional models for the remaining variables (or sets of variables) in turn:

   $f\left(X_j\middle|X_{init}, \Theta_{init}, X_i, \cdots, X_{j-1}, \Theta_i, \cdots, \Theta_{j-1}, \Theta_j\right),\ j \in \{i+1, \cdots, q\}$,

   where q is the number of variables.

4. Generate synthetic data sequentially via conditional posterior predictive distributions.

The product of the set of distributions yields a joint distribution on all the variables, and the set of conditional predictions thus constitutes a posterior prediction from this joint distribution.

Using conditional models allows for easier adaptations to certain survey realities, such as question skip patterns and logical constraints, but specifying a full set of models for a survey with as many variables as the ACS is still not trivial. In regressions, for instance, issues such as multicollinearity can easily effect model fitting towards the end of the conditional chain. Given that most of the variables in the ACS are discrete, we use classification trees as our main discrete structural model to help avoid some of the inherent issues in conditional modeling.

Classification trees perform a series of binary splits on a discrete data vector, choosing the split based upon a set of predictors (also called factors) and a homogeneity criterion (Breiman, Friedman, Olshen, & Stone, 1984). Typical algorithms for the fit are greedy, in that they always pick the split with the highest resultant homogeneity. As the algorithm continues to split the vector based on the predictors, a tree-like structure appears, with certain branches having more splits, ultimately reaching an end in nodes called leaves. The leaves then contain values of the vector that are more homogenous, in terms of the criterion, than the initial vector.

Classification trees are useful in our research as they can cull a large set of possible predictors into a set with the strongest relationships to the variable under synthesis. This is particularly useful when synthesizing data within geographies, as the strength of the relationship among a given set of variables can vary widely geographically. The sword is two-sided, though, in that the tree may never split on variables with weaker relationships but that nonetheless impose logical restrictions on the variable under synthesis.

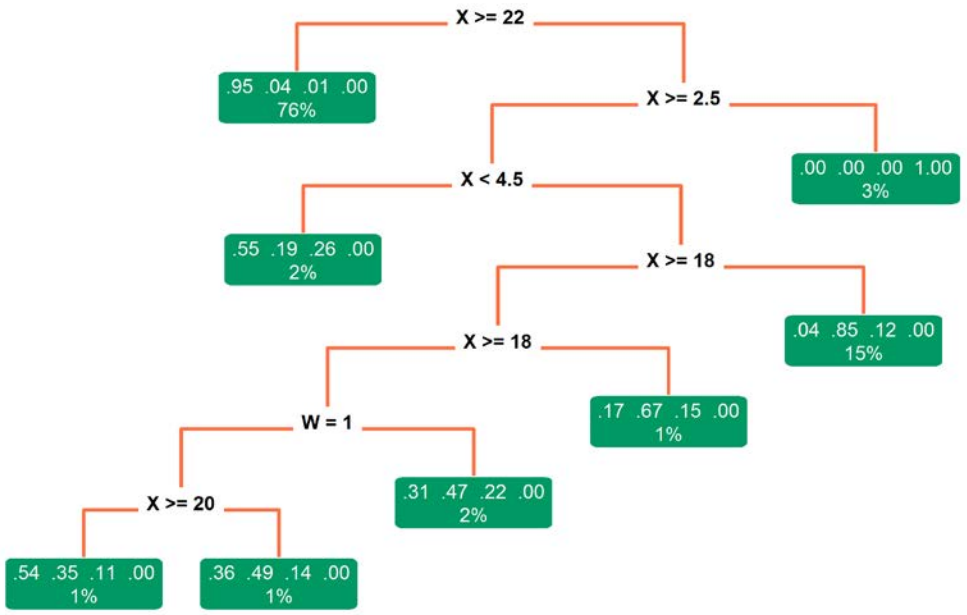

Figure 1: Structure of a Classification Tree

The typical method of prediction for a tree is to pick the modal category found in a leaf. Since we wish to use trees within the context of synthetic data, we must have some stochastic method to obtain a prediction, ideally one that we can reasonably couch as simulating a posterior prediction. The method we use is to take a single draw from a Bayesian bootstrap within the leaf (Reiter J. , 2005), which functions as a multinomial model with a Dirichlet prior where the multinomial outcomes are the values within the leaf.

5. ACS Public-Use Microdata Test Case

As an example of using classification trees for synthetic data, we synthesize unweighted variables related to education from the 2014 ACS 1-year Public Use Microdata Sample (PUMS) for the state of Oregon. In particular, we synthesize:

- School enrollment (SCH)
- Grade level attending (SCHG)
- Educational attainment (SCHL)

These variables are useful for initial study for a number of reasons: they have strong relationships with one another, the strength of those relationships can vary substantially across other variables, they have strict logical requirements with other variables, and they are often coarsened for analysis.

We use these three variables as initial predictors/factors:

- Sex (SEX)
- Age (AGE)

- Relationship (RELP)

Age in particular is strongly tied with education, and forms a set of logical requirements for school-age children. For our example, we will keep these initial predictors as observed so that we can isolate the effects of the classification tree synthesis of SCH, SCHG, and SCHL. Thus in this example we create column-partially synthetic data (Drechsler, 2011).

To assess the accuracy of the synthesis, we generate 896 synthetic data sets and produce the following unweighted tables from each data set:

- Sex by Age by Educational Attainment for the Population 18 Years and Over (T1)
- School Enrollment by Level of School for the Population 3 Years and Over (T2)

We compare each table to the original unweighted table from PUMS, using as a metric the proportion of total variation:

$$\tau = 1 - \frac{L^1}{2n}$$

Here, L^1 is the sum of the absolute cell deviations and n is the total count in the table. If we consider a table cell count as a number of stones, and if we consider two tables having the same number of cells and same total number of stones, then τ is the proportion of stones that remain in place when changing one table into the other. A value of zero indicates that all the mass of the table has moved to a new cell. For the sake of analytics (but not necessarily privacy), we seek values of τ near 1.

We then produce the same tables and metric for 4480 bootstrap samples of the PUMS and compare the distribution of τ between the synthetic and bootstrap tables:

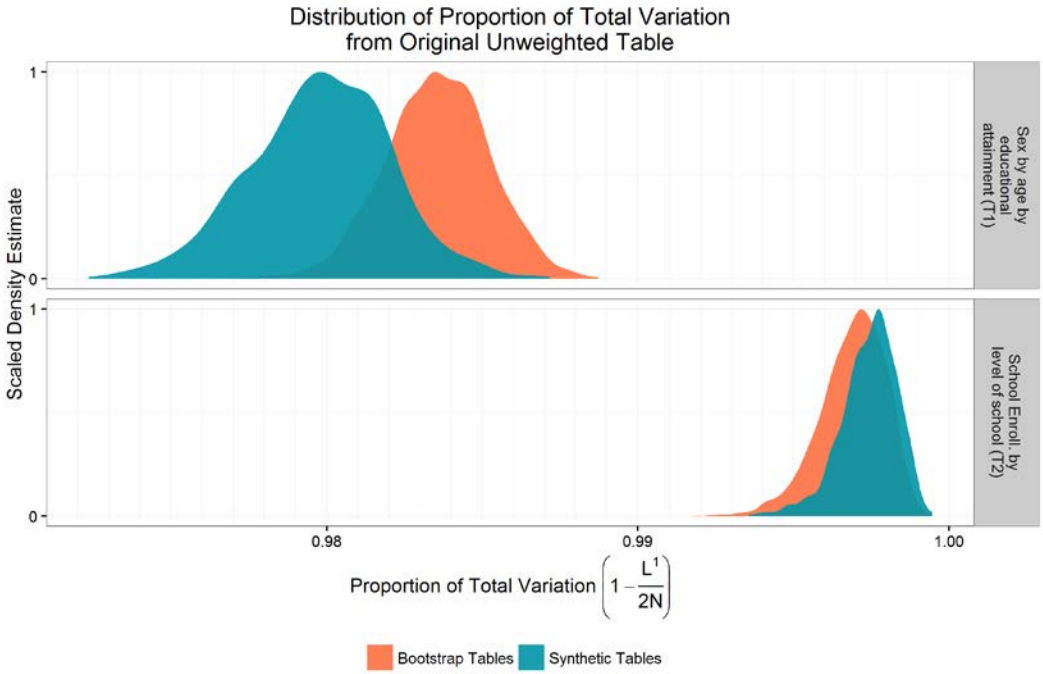

Figure 2: Distribution of Proportion of Total Variation

For both base tables, we can see that the synthetic tables have values of τ near one, indicating good data quality. For T1, the synthetic tables have a median τ slightly lower than the bootstrap tables, but whether the difference in meaningful is debatable. The results for T2 are even closer, and we see in fact that the synthetic tables have slightly higher quality than the bootstrap tables. This is wholly possible if we consider the nature of the data and the methods: the bootstrap data are simple random samples with replacement, while the synthetic data are essentially stratified on SEX, AGE and RELP, and the stratification may improve data quality enough to counteract the degradation from the synthesis. Again, how much these differences matter with such high values of τ is not clear.

In addition to τ, we also considered the relative change in counts for the values of each variable across the synthetic tables. For educational attainment we found the following result for adults:

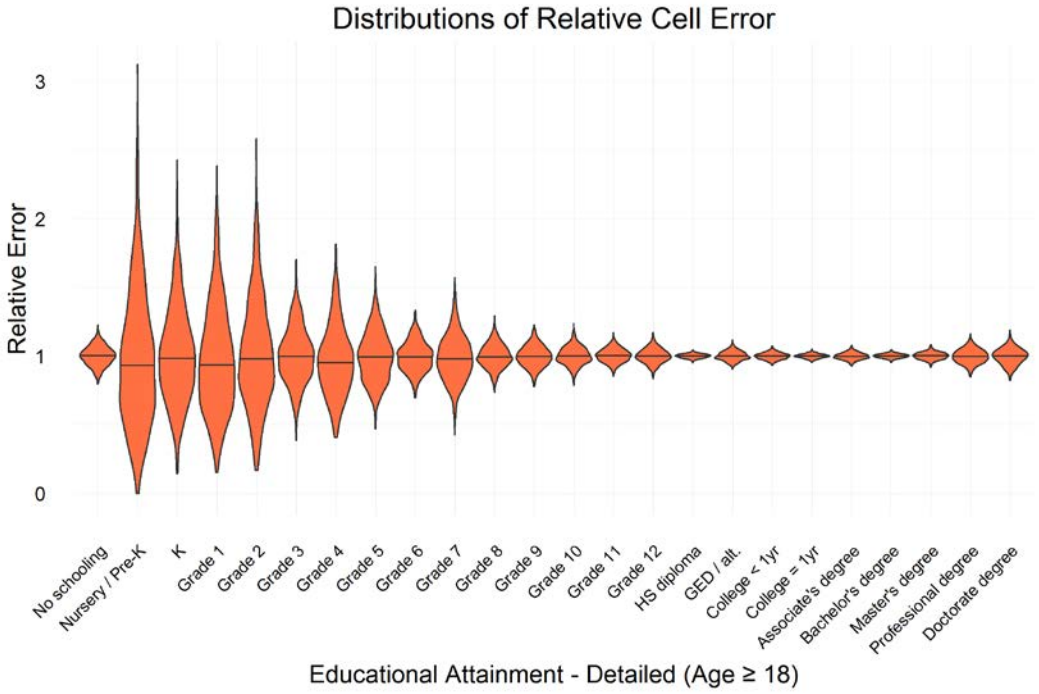

Figure 3: Distribution of Relative Error of Synthetic Cell Counts for Educational Attainment

We see that the largest errors in adult educational attainment occur for grades below high school, where the table is sparse. In such a case, the relationship between the variance added due to synthesis versus the variance due to sampling becomes important, especially for data users analyzing this population.

For grade attending, we assessed the relative count changes for the population in school:

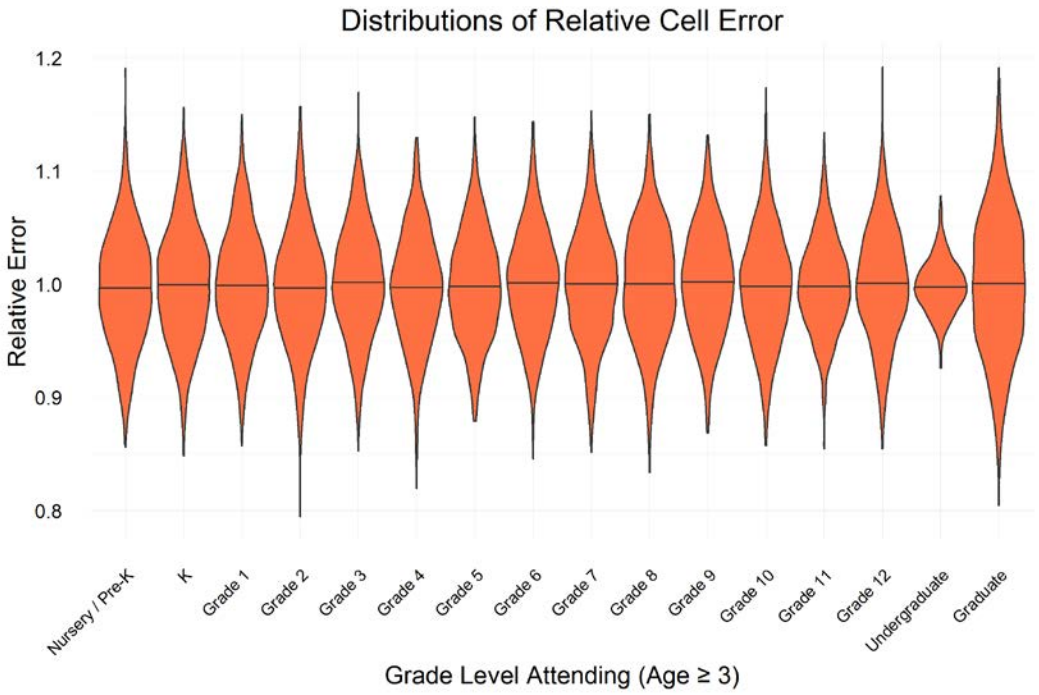

Figure 4: Distribution of Relative Error of Synthetic Cell Counts for Grade Level Attending

We see that the most represented category, people in undergraduate studies, has the least variation in cell totals. This table has no small cells and so we do not see the larger relative errors that we saw for adult educational attainment.

6. Formal Privacy

Applying any disclosure avoidance method to a data release raises two immediate concerns: the effect on data quality and the effect on respondent privacy. For many disclosure avoidance methods, these effects are rarely quantified theoretically, especially the effect on privacy. Rather, heuristic arguments or ad-hoc quantification often form the justification for using these methods for public data releases. For instance, one can emulate an attempt by an attacker to link a particular external data set to a given data product in order to expose respondent participation or attributes, but a successful or unsuccessful link says nothing about the privacy of the release in relation to any other data the attacker might have.

To overcome such limitations, the Census Bureau is researching and implementing definitions of privacy that are mathematically rigorous. A disclosure avoidance method that meets such a privacy definition can call upon the mathematical foundations of that definition to quantify the privacy (or privacy loss) of a data release made with the method. This quantification allows the expression of a meaningful, mathematical tradeoff between privacy and accuracy for any given accuracy metric. Such *formal privacy* methods, most notably *differential privacy* (Dwork & Roth, 2014), have other benefits, such as expressing

the privacy loss associated with multiple public data releases from the same underlying data, allowing the calculation of a privacy-loss 'budget' for those releases. Quantification of the tradeoff of the budget versus data accuracy also allows for meaningful graphical representations:

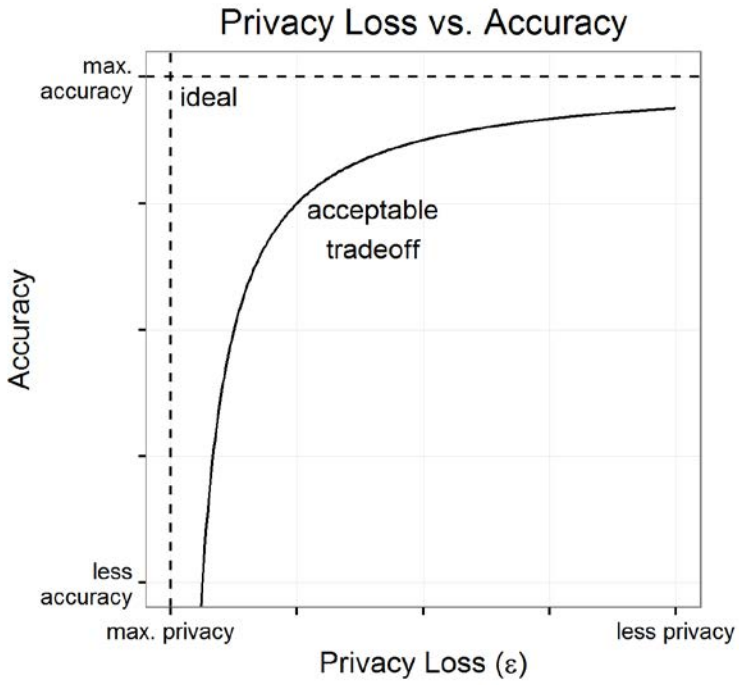

Figure 5: Typical form of the tradeoff between privacy loss and accuracy

Another tangible benefit of formal privacy methods is that many of them allow the public release of the method implementation, including algorithms, software, and parameters. This benefits not only data users in terms of transparency, but also the releasing agency by allowing public testing and verification. Ad-hoc methods often restrict the release of code and parameters due to uncertainty about the privacy implications.

The Census Bureau is currently researching formally private methods to supersede current disclosure avoidance strategies for the American Community Survey, with the aim of producing an underlying privatized data set from which all other releases spring. Such a system takes advantage of another underlying feature of certain formal privacy methods: that summaries made solely from privatized data remain private (Dwork & Roth, 2014).

A formally private ACS is an enticing concept, but the ACS poses significant challenges even for non-formally-private disclosure avoidance methods. Two of the largest hurdles are dimensionality and weighting.

The Census Bureau is developing a formally private disclosure avoidance method for the 2020 Census of Housing and Population (Leclerc, Clark, & Sexton, 2017). This method, like several common formally or differentially private methods, views the data as a frequency table rather than record-level microdata: we define a cell for each possible combination of the variables in the data and note the total number of respondents in each cell. We then add noise to the cell counts in such a way as to ensure privacy. Currently, the

frequency table under consideration has approximately 230,000 cells excluding geography. For the ACS, that number could increase by several orders of magnitude, which means that the method ultimately used for the 2020 Census may be computationally infeasible for the ACS. Even if computation were possible, dimensionality complicates privacy in other ways. While the ACS has a large sample, the total number of cells in a tabular representation of the data would certainly dwarf the sample size, and indeed the size of the sampling frame, resulting in a sparsely-filled table. Adding noise to a sparse table can easily result in large variances. In addition, if we require that all final cell counts be positive, which is the case if we wish to transform the table back into respondent-level records, the required rounding can introduce bias.

The ACS releases data each year for the current collection year and for the five most recent collection years combined. Tabular estimates for the 5-year data contain geographies down to the block-group level. There are approximately 200,000 block groups, each containing around 600 to 3,000 individuals. Data users regularly combine results from various block groups to obtain estimates for coarser geographies not directly available in the ACS data releases. This use-case conflicts with the need for privacy, as the addition of such detailed geography can further serve as a source for dimensionality issues when viewing the data as a frequency table.

Weighted surveys pose a number of issues for several disclosure avoidance methods, and the ACS is no exception. For formal privacy in particular, weighting interplays with the concept of sensitivity, which broadly means the amount of change one sees in a given summary of the data due to the addition or deletion of one respondent. The higher the sensitivity, the more privacy loss one can expect for a given accuracy. For weighted surveys, that sensitivity can be high if we consider the case of a person with maximal weight entering or leaving the survey, leading to a choice between too much privacy loss or too little accuracy.

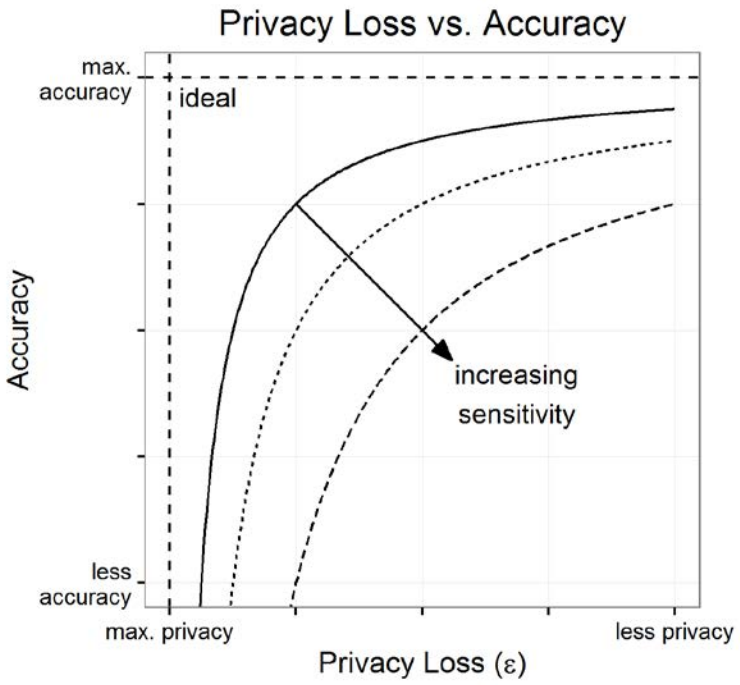

Figure 6: Effect of sensitivity on the tradeoff between privacy loss and accuracy

Other obstacles to a formally private ACS exist—household structure, data editing/cleanup, variance estimation—and can interact with dimensionality and weighting. The Census Bureau is researching methods to overcome these challenges.

7. Conclusion

Our example uses classification trees to produce synthetic data for three variables in the American Community Survey Public-Use Microdata Sample. We see promising results for accuracy on these variables in certain tables, but results for single variables make an important point: that the suitability of a given synthetic data method depends greatly on the analyses data users expect to perform (Reiter J. P., 2005). Our research will test classification trees and other synthetic data models on internal ACS data as a means to replace certain disclosure avoidance methods currently in use. Many obstacles to a final production synthetic data method remain unconquered: synthesis of geography, the effect of weighting, maintenance of household-level relationships. But research into these and other issues can only help in the effort to bring better privacy methods to surveys and censuses.

Governments, academic institutions, businesses, and individuals depend upon surveys such as the ACS for timely and accurate data for making decisions. Disclosure avoidance methods have the potential to change such decisions. But those same groups of data users are also respondents, and statistical agencies have not only an ethical but sometimes a legal obligation to protect respondents from unwanted identification and attribution. Disclosure avoidance methods have to potential to provide the necessary protections. Negotiating the tradeoff between accuracy and privacy should be of foremost concern to any agency releasing information, not just data, about respondents. Formal privacy methods provide a

mathematical means of quantifying the tradeoff, but as research into their use for large sample surveys continues, other methods can help bridge the gap between formal privacy and older methods lacking meaningful guarantees for privacy and analytics.

References

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.* Belmont, CA: Wadsworth, Inc.

Drechsler, J. (2011). Synthetic datasets for statistical disclosure control: theory and implementation. In *Lecture Notes in Statistics, Vol. 201.* New York: Springer.

Dwork, C., & Roth, A. (2014). *The Algorithmic Foundations of Differential Privacy.* Now Publishers.

Lauger, A., Wisniewski, B., & McKenna, L. (2014). *Disclosure avoidance techniques at the US Census Bureau: Current practices and research.* Center for Disclosure Avoidance Research, US Census Bureau.

Leclerc, P., Clark, S., & Sexton, W. (2017, October 24). 2020 Decennial Census: Formal privacy implementation update. Piscataway, NJ: Presented at DIMACS/Northeast Big Data Hub Workshop on Overcoming Barriers to Data Sharing including Privacy and Fairness.

Raghunathan, T., Reiter, J., & Rubin, D. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics, 19(1)*, 1-16.

Reiter, J. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics, 21(3)*, 441.

Reiter, J. P. (2005). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A, 168*, 185-205.