RESEARCH REPORT SERIES
Disclosure Avoidance #2018-03

Synthesizing Housing Units for the American Community Survey

Rolando A. Rodríguez, Michael H. Freiman, Jerome P. Reiter and Amy D. Lauger

Report Issued: November 20, 2017

Center for Disclosure Avoidance Research
U.S. Census Bureau
Washington DC, 20233

Synthesizing Housing Units for the American Community Survey

Rolando A. Rodríguez

U.S. Census Bureau, Washington DC, United States

Michael H. Freiman

U.S. Census Bureau, Washington DC, United States

Jerome P. Reiter

Duke University, Durham, NC, United States & U.S. Census Bureau, Washington DC, United States

Amy D. Lauger

U.S. Census Bureau, Washington DC, United States

Abstract

The Census Bureau is charged with collecting and disseminating data while protecting the privacy of respondents. The Census Bureau must protect against several types of unauthorized disclosure of data, a task that has become more difficult in recent years. One promising line of research is the creation of synthetic data, derived from a model to mimic the original data while protecting against unauthorized disclosure. We created synthetic data for the housing variables in the American Community Survey (ACS), using standard regression methods and Classification and Regression Trees (CART). Our metrics showed that the accuracy of the synthetic data was fairly high for some variables but lower for other variables. We have not proved that our methods satisfy any formal privacy criterion, although future research does aim to have this property.

Introduction

The Census Bureau's mission is to collect and disseminate data while protecting the confidentiality of respondents. However, recent advances in technology and data availability have increased the risk of releasing data, as such data can be more easily matched to external sources than ever before. Therefore new methods are necessary to protect the data. In this paper, we will discuss one promising method, the creation of a synthetic dataset that aims to share the essential properties of the original data.

Title 13

Title 13, U.S. Code, Section 9 mandates that the Census Bureau collect and release data on United States residents and businesses. However, Title 13 also requires the Census Bureau to protect the confidentiality of individual responses, particularly directing that the Census Bureau may not "make any publication whereby the data furnished by any particular establishment or individual under this title can be identified" (Title 13, U.S. Code, Section 9). These two necessities are at odds with each other, and releasing data of maximal utility while protecting confidentiality is a topic of

ongoing study for the Center for Disclosure Avoidance Research and other areas in the Census Bureau.

The Census Bureau must protect against three main types of unauthorized disclosure:

1) **Identity disclosure** (reidentification): manipulation of released data reveals the identity of an individual or business.
2) **Attribute disclosure**: manipulation of the data reveals some feature of a respondent. An attribute disclosure may occur in addition to an identity disclosure or on its own.
3) **Inferential disclosure**: the data user can determine an identity or attribute with a high probability. An inferential disclosure occurs when the user's posterior belief regarding the particular record differs substantially from the user's prior belief.

The first two types of disclosure may be viewed as special cases of inferential disclosure where the posterior probability is 1. Hence the problem of avoiding unauthorized disclosure is equivalent to the problem of avoiding inferential disclosure. These types of disclosure are not merely theoretical concerns, as there have been several high-profile instances of reidentification of records in datasets from which personally identifiable information had been redacted (Narayanan and Shmatikov, 2006; Sweeney, 2002; Sweeney, 2013).

American Community Survey

This paper focuses on protecting the data from the American Community Survey (ACS), the Census Bureau's largest annual demographic survey, with responses from over 2.3 million households and 160,000 people in group quarters annually.

The ACS asks questions about a wide variety of housing and demographic topics, such as features of the physical housing unit, Internet access, citizenship and health insurance. As such, the Census Bureau advises respondents to set aside 40 minutes to complete the ACS questionnaire. Justification of such a respondent burden leans on the wealth of ACS-based data products released annually, the $675 billion of federal funds per year distributed directly and indirectly based on those data, and the use of the data to implement Section 203 of the Voting Rights Act. The data products take two main forms: tables and microdata (record-level responses).

The Census Bureau has used several methods to protect against disclosure in the past. Currently, data swapping is the main method of protecting ACS household data. For people in group quarters—dwellings such as dormitories, prisons, military housing, etc.—the Census Bureau currently uses partially synthetic data. Both approaches are described in Lauger *et al*. (2014).

Swapping and partial synthesis are performed to create the final weighted file, which is not public but is used to create publicly available tabulations. The Census Bureau applies additional protections, such as cell suppression and top-coding, to the public data products. One goal of the present research is to produce a final internal file from which new products can be made without the application of additional protections. Such a paradigm ideally depends on disclosure protection techniques where the concept of confidentiality can be rigorously defined and guaranteed. Equally important is the degree to which the public data products made from the protected file agree with the same products made from the non-protected file. This synthetic internal file may not be used for all purposes; it is possible that a file with less protection will remain available in the Federal

Statistical Research Data Centers (FSRDCs), relying in the FSRDCs' rules on what data may be released to ensure that the privacy of individual respondents remains protected.

Disclosure Philosophy and Synthetic Data

The data protection environment has changed dramatically in recent years, with improved knowledge and technology to attack datasets. In 2003, Dinur and Nissim (2003) proved the database reconstruction theorem, which states that any database can be reconstructed arbitrarily accurately given a sufficient but finite number of queries. A reconstruction does not directly give identities of respondents if those identities were not in the dataset being queried, but an accurate database reconstruction almost guarantees an attribute disclosure and provides much fodder for an intruder trying to create an identity disclosure.

When traditional disclosure avoidance methods were developed, the risk of disclosure was much lower for any given data release than it is today because in the past, there were relatively few external data available to correlate with any given data or statistical release and the database reconstruction theorem was unknown. The Census Bureau and other agencies curated most of the data that could compromise confidentiality, and the few commercial and private exceptions were mostly known and appropriately addressed. Database reconstruction attacks were completely unknown, and historical products were not designed to counter such attacks.

More recently, the amount of publicly available or proprietary data that can create a disclosure risk has greatly increased. "Big data" has become a crucial asset for businesses, which use the data to tailor their products. Simultaneously, research has improved the algorithms for attack, and increased computing power allows the data and algorithms to be put into practice, increasing the risk of a disclosure dramatically. The new environment requires new methods to protect the data. This paper focuses on protecting the American Community Survey (ACS) data.

Maintaining data privacy is a thorny theoretical problem, because all data publication leads to some privacy loss. This is the fundamental consequence of the database reconstruction theorem (Dinur and Nissim, 2003). No one line separates protecting privacy and not protecting privacy. Privacy loss is often incremental among several data releases, any one of which may have minimal risk (Abowd, 2017). Every piece of output leads to some actual privacy loss, if only in the form of causing slightly more accurate inference about individual records in the dataset. As more data are released, the total privacy loss can be substantial.

The Census Bureau is researching new methods to protect data, and this paper discusses the use of synthetic data (Rubin, 1993; Raghunathan *et al.*, 2003) to protect housing units. Under this method, the collected data are used to train a statistical model, which is then used to generate a new dataset that captures many of the properties of the original data. Synthetic data may be full or partial (Reiter and Raghunathan, 2007). In partially synthetic data, the data generator starts with the original dataset, blanks out some of the records or variables and replaces them with synthetic values. In fully synthetic data, we discard the entire dataset once the model has been created and generate entirely new records from the model. If created well, fully synthetic data should make reidentification attacks by linking to external files nonsensical, since no record in the synthetic dataset corresponds directly to a record in the original dataset. Depending on the method of data synthesis used, there may be some risk of reidentification if the model is sufficiently overfit that it regenerates records from the original dataset. However, there are tradeoffs: the only dataset that preserves all properties of the original dataset is the original dataset itself, so some features of the

original dataset will inevitably be lost. ACS data from group quarters are currently protected by partial synthesis.

Formal Privacy

Recently, the Census Bureau has been moving away from more traditional statistical disclosure limitation methods to newer methods that satisfy formal privacy. Under formal privacy, the privacy loss for a release or collection of releases is quantified, and the releasing agency uses methods that can be mathematically proven to limit the cumulative privacy loss. Formal privacy is not a method, but rather an umbrella term for the approach of creating a privacy loss measure and limiting the privacy loss. A formally private approach effectively creates a privacy budget and says that the cumulative privacy loss from all releases cannot exceed that budget. If much of the privacy budget is spent on creating a synthetic dataset, then any output derived from that dataset can be released without spending more of the budget. Our expectation is that when the project is complete, all public data products will likely be based on the synthetic file, but a less modified file may be available to approved researchers at the FSRDCs. Releasing data from the FSRDCs will thus result in spending more of the budget.

The choice of a privacy budget is a policy decision, reflecting a desired tradeoff between privacy and data usefulness. Statisticians can determine what level of data usefulness can be obtained for any amount of privacy loss, which can inform the decision.

The type of formal privacy that has been the subject of the most research is differential privacy. The differential privacy criterion says that if two datasets differ by only the inclusion or exclusion of a single record, the probability of observing a given set of output must differ by no more than a fixed factor depending on which dataset is used. Mathematically, a data release algorithm A has ε-differential privacy if, for any datasets D and D' differing only by the inclusion of one record, and for any set S of possible outputs of A,

$$P(A(D) \in S) \leq e^{\varepsilon} P(A(D') \in S).$$

A method satisfying differential privacy guarantees that no feature (including an identity) of an individual record in the dataset can be learned from the output of an algorithm with much higher probability than the same feature could be learned from the output of that same algorithm with the record omitted, where the choice of ε determines what "much higher" means.

Challenges Specific to the ACS

Making ACS data formally private presents a challenge due to its high dimensionality, sample size limitations and complex survey weights. The noise in formally private methods can swamp the confidential count when sample sizes are small, as they are in some of the lower-level geographies for which the ACS currently releases data. This is because individuals have proportionately much higher contribution when population counts are small.

Another challenge for the ACS, as for the Census of Population and Housing, is that not every combination of variable values could plausibly happen in the real world, and the data should reflect that. For example, a parent cannot be three years old. Combinations of variables that are not allowed to happen are called *structural zeros*. Currently, the ACS data are edited before release to ensure that no inconsistencies of this sort appear. Such inconsistencies must also not appear in any

ACS synthetic data created in the future, but forcing the model not to create them is an additional challenge.

ACS data for housing units are collected at the level of the household. Some variables (rent paid, Internet access, presence of plumbing facilities, etc.) refer specifically to features of the household, while others (sex, wage income, disability status) refer to the individual people within the household. In particular, the Relationship variable records how each person is related to the head of household, and the collection of values of this variable for all members of the household creates a household structure. The variable creates dependencies among the members of the household. Viewing the person-level data as a matrix of values each row of which may be generated separately does not consider the full complexity of the problem because of the relations between rows in the matrix. For example, parents and their biological children are likely to be of the same race and ethnicity.

The ACS uses survey weights, and to our knowledge, there is not yet a formally private method that accounts for such weights, nor has there been significant work on model-based non-formally private synthetic data with weights.

This paper examines synthesis of household data as a first step toward creating a fully synthetic dataset. Households are somewhat simpler to synthesize than the individuals within them, because there are no intra-household relationships to model. Later research will explore synthesizing people within households, respecting those relationships. Households will then be assigned to geographies in a way that is faithful to the original data.

Data and Products

The Census Bureau currently produces several products based on the ACS, most prominently the ACS official tables and the ACS Public Use Microdata Samples (PUMS).

The ACS publishes tables giving a wide range of summary statistics, with output mostly consisting of counts and their margins of error. The tables are based on one year of data or five years of data. Tables from the one-year data cover geographies with population as small as 65,000, with smaller populations allowed for a smaller set of tables. Tables from the five-year data cover all geographies down to the level of block groups, which generally have population between 600 and 3,000. The tables give less topical detail than the PUMS but more geographic detail. The ACS is the only survey with a sample size and design that can support estimates down to this geographic level of precision, although for some very small geographies, the margins of error on estimates can be quite large. The availability of data down to small geographies makes protecting the data more difficult.

The PUMS is a sample of ACS data that anyone can download and use. It includes the individual line-by-line records for a sample consisting of roughly 2/3 of the respondents to the ACS. PUMS files give geographic detail at the level of Public Use Microdata Areas (PUMAs), which have population at least 100,000.

Our approach in this paper is to protect the microdata underlying our products and then use the protected microdata to produce all external data products. This is mostly how the data are currently protected, although there are a few additional modifications made to the PUMS.

Synthetic Data Without Formal Privacy

We are researching methods for generating formally private synthetic microdata. As an initial step on that path, we are experimenting with methods for generating high quality, model-based synthetic data. These do not necessarily satisfy formal privacy as currently implemented, at least not that we are able to quantify.

Our goal is to create fully synthetic data; that is, data where every variable for every record is drawn from a model. Ideally, the synthetic number of housing units and persons should not be required to match the number in the collected sample. We begin with an empty data set and fill in variables via draws from our synthesis models. We synthesize variables in order of their appearance on the housing unit section of the ACS questionnaire.

Synthesis for household variables is performed independently for each row in the dataset. Synthetic values for the first variable, building type, come from posterior predictive draws from a multinomial distribution with a Dirichlet prior. Synthetic values for subsequent variables come from standard regressions and Classification and Regression Tree (CART) models. Our goal is to build up the joint likelihood for the set of ACS housing-unit variables as a sequence of conditional likelihoods (Reiter, 2005):

$$f_Y(y|\Theta) = f_{Y_1}(y_1|\Theta_1) f_{Y2|Y1}(y_2|y_1, \Theta_1, \Theta_2) \ldots$$

Synthetic values are nominally posterior predictive draws based on this conditional likelihood with suitable prior distributions; however, CART does not fit directly into the standard Bayesian framework.

The ACS form includes questions that are sometimes skipped based on the results of previous questions. For example, the question “How many acres is this house or mobile home on?” is only answered for houses or mobile homes, but not for apartment buildings, boats, recreational vehicles or vans. In such cases, we determine whether each question would be skipped based on the previously synthesized variables and do not synthesize it if it would be skipped.

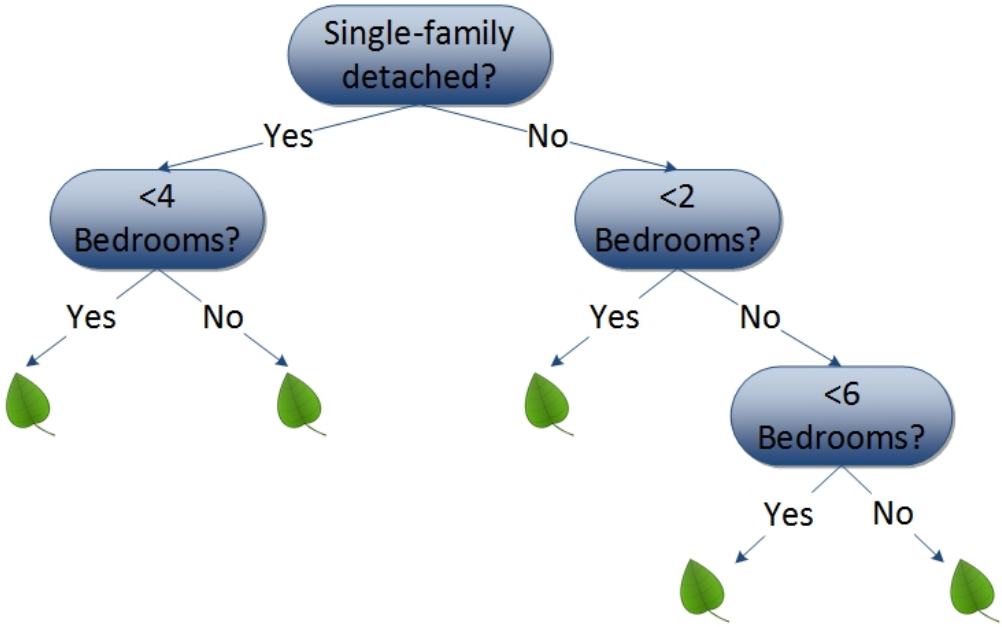
CART

To synthesize categorical variables, we use a Classification and Regression Tree (CART) approach. Classification and Regression Trees were invented by Brieman *et al.* (1984) as a non-parametric approach to predicting values of numerical or categorical variables. A tree is a recursive partitioning of the set of records based on the values of the predictor variables, with the partitions made so as to maximize the homogeneity of the response variables within the partitioned group. Each partition can be thought of as a question, with those records with the answer “yes” going into a bin in the left and those record with the answer “no” going into a bin on the right. The homogeneity resulting from a split is measured by the sum of the deviances of the two leaves if the response variable is categorical, or by the sum of the squared differences from the leaf mean if the response variable is continuous. These bins are then further split until some stopping criterion is reached. To predict the response value for a given record, we drop it down the tree until it reaches a terminal node, or leaf, a node that is not split.

For example, suppose we wish to predict how many rooms a housing unit has that are not bedrooms, based on knowing the type of housing unit, year built, year moved in and number of bedrooms. (This paragraph is based on the 2015 PUMS data from Oregon, so as not to use restricted

data when the purpose is to illustrate.) According to the deviance criterion, the best split puts single-family detached houses in one bin and all other housing units in the other. Among single-family detached houses, the optimal split puts houses with three or fewer bedrooms into one bin and houses with four or more bedrooms into another bin, and both of these bins are leaves. Among all other housing units, the optimal split is between units with one or fewer bedrooms and multi-bedroom units. Units with one or fewer bedrooms are a leaf, while multi-bedroom units are split one more time, separating units with fewer than six bedrooms from units with six or more bedrooms, and both of these are leaves. This case is shown in Figure 1 and illustrates that a tree need not have uniform depth.

Figure 1: A Classification and Regression Tree to predict the number of rooms in a housing unit that are not bedrooms.



Following Reiter (2005), we build a tree to predict the variable being synthesized from the preceding variables, using the original data as training data. We performed the tree-based synthesis in this paper using the tree package in R (R Core Team, 2013; Ripley, 2014), but we may use other languages or packages for future development. We build the tree rather deep; nodes are divided as long as there are at least five records in each child node. To allow some heterogeneity within leaves, we require that no node be split if its deviance is less than 10^{-9} times the deviance of the root node containing all of the records, but this is such a small deviance that we expect this rule will rarely be invoked. Sometimes the algorithm grows a tree deeper than the tree package's depth limit of 31 levels; when this happens, we increase the minimum allowable deviance until the package works. Once the tree is grown, each existing synthetic record is dropped down the tree, and the program notes which leaf it falls into. One of the original records in that leaf is then selected at random, and the synthetic record assumes the variable value associated with that original record.

Regression

We can consider many variables in the ACS as latently continuous or pseudo-continuous. For instance, one ACS question asks for the respondent's average commute time to work in the previous

week. The actual time in minutes is nominally a real number, but we ask the respondent to answer this question using an integer with up to three digits. While CART can certainly produce synthetic data for such variables, the privacy aspects of such data can be questionable (Freiman *et al.*, 2017). Instead, we use regression to generate synthetic data for these variables.

For variable Y_i, we fit the following regression model:

$$U_i(Y_i)|\boldsymbol{X}, \beta, \sigma \sim N(\boldsymbol{X}\beta, \sigma^2),$$

where U_i is an invertible transformation and $\boldsymbol{X}$ is the matrix of predictor variables. We use conjugate priors for the model parameters β and σ as given in Gelman *et al.* (2014). Synthesized values are drawn from the posterior predictive distribution and back-transformed. Since we are using conjugate priors, the posterior predictive distribution has a closed form and is a multivariate-t distribution.

Algorithmically, we follow the conditional method of drawing posterior predictive values given in Gelman *et al.* (2014), first drawing posterior values of the parameters and then drawing predictive values conditional on these. This kind of "proper" synthesis is typically associated with the method of multiple imputation (Rubin, 2004); however, we are currently only assessing quality based on one synthetic draw.

Ideally, the predictor matrix $\boldsymbol{X}$ would include all previously synthesized variables with all available granularity. This is not always possible, as issues such as non-invertible or ill-conditioned predictor matrices can easily and unpredictably occur. Instead, $\boldsymbol{X}$ includes a subset of previously synthesized variables, possibly at lower detail levels, possibly transformed. We do not currently include interaction terms. In the case where we cannot find any coarsening of a given predictor that will allow for a fit, we drop that variable from the current model. This results in an assumption of conditional independence of the dependent variable on the dropped variable given the remaining predictors.

As stated above, we fit each dependent variable within the data subset defined by the skip logic for the variable; we call this subset the variable's universe. Working within universes guarantees complete data for the dependent variable, but the predictor variables might not all share the same universe. In this case, we have options for including conflicting variables, depending on the predictor type. If the predictor is categorical, we can recode its being out-of-universe values as a new category. If the predictor is continuous, we can assume a reasonable value, often zero, if such a value makes sense within the model. Another consideration is the set of ACS variables defined as a mixed categorical/continuous construct. These variables require special handling during synthesis and when used as predictors in the regression models.

The regression model makes three major assumptions: linearity, normality, and homoskedasticity. Transformations can help ensure the first two of these, and we call upon them often. Given that several of our variables have a natural limit at 0 and show right skew, the logarithm is a natural choice; however, the synthetic values after the exponential back-transformation can be exceedingly large if the residuals are platykurtic or heteroskedastic. To reduce this effect we use the cube-root instead, which has the added benefit of allowing negative values in the variable.

Results and Future Work

Metrics

To be valuable, any synthetic data should preserve the accuracy of the original data and should protect against disclosure. So far, we have only measured the former, but our full analysis of the synthetic data will consider both.

Whenever data are modified to protect against disclosure, we have concerns about loss of accuracy. Our primary measure of loss of data accuracy is how much table counts change from the original output to the modified output. Our metric assumes that we start with the original table, consider the proportion of records in each table cell, and move some of the probability mass from one cell to another until we produce the modified table, also considered as a set of proportions. We measure the proportion of the total probability mass that has to be moved to get to this modified table. If the numbers in the original table are o_1, o_2, …, o_k, and the numbers in the modified table are m_1, m_2, …, m_k, then this metric is $\frac{1}{2}\sum_{i=1}^{k}\left|\frac{m_i}{\sum_{j=1}^{k} m_j} - \frac{o_i}{\sum_{j=1}^{k} o_j}\right|$, which is proportional to the L_1 distance between the two tables in the event that the original and synthetic sample sizes are the same. This metric is similar to the earth mover's distance, but does not impose a distance metric on the bins of the histogram.

Another measure of accuracy is whether one can discriminate between the original and synthetic data sets based upon some subset of synthetic variables. Propensity score measures can help us answer this question. To calculate record-level propensity scores, we stack the original and synthetic data sets and create a new variable to denote whether a row was original or synthetic. We then perform a logistic regression on this dataset indicator, using the subsets of the synthetic variables as our predictors. The fitted probabilities per record are then the estimated propensity scores. There are many ways to summarize these scores to assess dataset discrimination; we will use the following simple measure (Woo *et al.*, 2009):

$$U_p = \frac{1}{N}\sum_{i=1}^{N}[\hat{p}_i - c]^2$$

where N is the number of records in the stacked data, $\hat{p}_i$ are the estimated propensity scores, and c is the proportion of rows in the synthetic data. In the case where a synthesis model resulted in estimated propensity scores of 0 for all original data and 1 for all synthetic data, then $U_p = \frac{1}{4}$.

Results

The housing unit variables allow us to recreate 24 of the approximately 150 tables of the publicly-released ACS housing unit tables for the state of Oregon using the 2015 ACS internal data. The sample size is approximately 25,000 households. We calculate the table distance between the original and the synthetic tables, and then compare this distance to a set of distances calculated between the original tables and tables made from bootstrap samples of the original data. Table 1 shows the estimated quantile of the synthetic table distance within the bootstrap distance null distribution; subtracting this quantile from 1 gives something analogous to a p-value. Indented variable descriptions in Table 1 indicate a cross-tabulation.

Table 1: Accuracy metrics for various tables based on ACS original and synthetic data.

Table		Synthetic Table Distance	Synthetic Table Quantile
Monthly costs		2.2E-01	1.00
Units in Structure		9.3E-03	0.99
Heating Fuel		4.4E-03	0.54
Housing-unit value		9.6E-02	1.00
Housing-unit value (detail)		1.0E-01	1.00
Number of Rooms		1.1E-02	0.98
Number of Bedrooms		1.1E-02	1.00
Has a mortgage		2.3E-04	0.05
	Second loan	3.7E-02	1.00
	Monthly costs	2.3E-01	1.00
Owned/Rented		1.2E-03	0.31
	Household Size	7.3E-02	1.00
	Number of Rooms	1.3E-02	0.96
	Number of Bedrooms	1.5E-02	1.00
	Number of Vehicles	5.6E-03	0.22
	Number of Vehicles (detail)	2.9E-03	0.50
	Heating Fuel	6.0E-03	0.40
Rent (yes/no)		4.3E-03	0.93
	Rent amount	1.3E-01	1.00

Direct comparisons of the distances across different table shells are not meaningful, as tables with more granularity will tend to have greater total distances. To evaluate data quality, we create a bootstrap distribution of how much the synthetic values would differ from the original values according to our metric if the synthetic values were drawn from the same probability distribution as the original values. If the original data and the synthetic data were generated using identical mechanisms, we would expect the bootstrap quantile of the synthetic data to follow a uniform distribution on the interval [0,1]. We thus look for quantiles at or near the extremes of this interval. In most cases, this means at or near 1, meaning the synthetic data create a table deviating more from the original data than would be likely given multinomial random chance. In essence we are asking how believably the synthetic data could pose as a bootstrap of the original data.

We see a tendency for the sub-tables to have larger quantiles, which would indicate issues in multivariate relationships. We also note that several univariate tables have quantile values at or very near 1, indicating the synthetic table distance is higher than the bootstrap distance value, for all or nearly all of our 1,000 bootstrap simulations. These include variables synthesized via CART (number of bedrooms) and regression (housing-unit value). In some cases, the univariate tables have a value of the metric that is dramatically higher than even the highest of the 1,000 bootstrap simulations.

We then calculate propensity scores. We must choose predictors for the logistic regression. Given the distance-metric results for housing-unit value, we use two predictor sets: VAL alone, and the predictor set used to simulate VAL from the modeling stage. For VAL alone, we obtain the score $U_{p_{VAL}} = 5 \times 10^{-6}$. For the larger predictor set without VAL, we obtain $U_{p_{other}} = 6 \times 10^{-4}$. Neither of these scores points to easy distinguishability of the original and synthetic data. This shows an important aspect of simulation: congeniality of the synthetic data to a given data analysis depends strongly on the simulation model. We have reason to believe that the particular tables related to VAL might differ significantly between the original and synthetic data, but this broad propensity score-based measure does not indicate issues with VAL alone. A more detailed analysis of the propensity scores or a different logistic regression model (for instance, using indicators for table cells) might help us locate shortcomings in our models.

Future Work

Building up a detailed picture of where in the simulation process problems arise is difficult from the current results, as many of the synthetic variables are not represented among the dimensions of the tables we could generate, since they reference person-level characteristics which we have not synthesized. A next step would be to create additional tables not found in the public data releases to might help root out problematic model fits.

We can also avail ourselves of various model-checking procedures for CART and regression to tease out potential improvements. In particular, cursory overview of several regression fits show potential issues with regression assumptions. Such issues can be dealt with in several ways, including transformations, predictor recoding, or by first imputing broad categories via CART. Issues with multivariate outcomes, such as those noted for higher-dimensional tables in the results, might be ameliorated by including interactions.

The CART function we use limits trees to 31 levels because of limitations caused by the scheme used to label nodes. Sidestepping this restriction might improve multivariate outcomes, and other programming languages or R packages may allow us to grow deeper trees. Additionally, we aim to understand why certain variables required an increase in the minimal deviance required for making a split in the tree, as results show that outcomes for these variables might be negatively affected.

Two questions of broad import loom over the housing-unit synthesis. The first is how to produce synthetic weights or how to avoid needing them. Weights could be avoided by producing an entire synthetic population based on the collected ACS data, rather than just a sample, although this approach presents its own challenges. The second is how to assign housing units to geographies, which is necessary to maintain the ACS's value as the only ongoing Census Bureau survey whose design allows estimates at low levels of geography, a feature that can be vital for research and policy-making. Answering these questions at the housing-unit level might inform answers to the same questions at the person level, and vice versa. We must also consider how to assign persons to housing units once the former are synthesized, which is non-trivial given the gamut of within-household relationships found in the ACS.

References

Abowd, J. M. (2017). How will statistical agencies operate when all data are private? Available at http://digitalcommons.ilr.cornell.edu/cgi/viewcontent.cgi?article=1029&context=ldi. Accessed September 12, 2017.

Brieman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Belmont (CA): Wadsworth.

Dinur, I., & Nissim, K. (2003, June). Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 202-210). ACM.

Freiman, M. H., Lauger, A., & Reiter, J. P. (2017). Data synthesis and perturbation for the American Community Survey at the U.S. Census Bureau. *Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington DC.* Forthcoming.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Boca Raton, FL: CRC press.

Lauger, A., Wisniewski, B., & McKenna, L. (2014). Disclosure avoidance techniques at the U.S. Census Bureau: Current practices and research. *Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington DC*. Available at https://www.census.gov/srd/CDAR/cdar2014-02_Discl_Avoid_Techniques.pdf. Accessed September 6, 2017.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy, 2008. SP 2008* (pp. 111-125). IEEE.R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of official statistics*, *19*(1), 1.

Reiter, J. P. "Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168.1 (2005): 185-205.

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, *102*(480), 1462-1471.

Ripley, B. (2014). Tree: Classification and regression trees. R package version 1.0-35. http://CRAN.R-project.org/package=tree.

Rubin, D. B. (1993). Discussion: statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461.

Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys* (Vol. 81). John Wiley & Sons.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557-570.

Sweeney, L. (2013). Matching known patients to health records in Washington State data. Data Privacy Lab. 1089-1. Available at https://dataprivacylab.org/projects/wa/1089-1.pdf. Accessed September 6, 2017.

Title 13, U.S. Code, Section 9. Available at https://www.gpo.gov/fdsys/pkg/USCODE-2009-title13/html/USCODE-2009-title13.htm. Accessed September 6, 2017.

Woo, M. J., Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, *1*(1), 7.