Data Synthesis and Perturbation for the American Community Survey at the U.S. Census Bureau

Michael H. Freiman, U.S. Census Bureau

U.S. Census Bureau, Washington DC

Amy D. Lauger, U.S. Census Bureau

U.S. Census Bureau, Washington DC

Jerome P. Reiter, Duke University and U.S. Census Bureau

Duke University, Durham, NC; U.S. Census Bureau, Washington DC

Summary. This paper assesses an empirical measure of disclosure risk of synthetic demographic data generated using classification and regression trees. We synthesized a dataset with 50 implicates and tried to infer from the synthetic data the maximum income in the original dataset. If synthetic values were determined by drawing without noise from a leaf of the regression tree, then the maximum value across implicates was a very good estimate of the maximum value in the original dataset. If synthetic values were determined by drawing from the leaf with noise, then skewness in the incomes within the leaves led to substantial bias in the mean wage for the synthetic dataset. Furthermore, the maximum income could still be determined with unreasonable accuracy, estimable by the median of the maxima of the implicates, or in some cases by rescaling the maximum across all of the implicates. We conclude that this method of generating synthetic data does not adequately protect continuous variables such as income from reconstruction, at least not when many implicates are created.

Keywords: Classification and regression trees; Confidentiality; Data quality; Disclosure; Kernel density estimator; Synthetic data

1. Introduction

The U.S. Census Bureau's mission is "to serve as the leading source of quality data about the nation's people and economy. We honor privacy, protect confidentiality, share our expertise globally, and conduct our work openly" (U.S. Census Bureau, 2017b). To further the mission, the Census Bureau is undergoing a transformation in the way it disseminates data by making data easier to access, more flexible and more customizable. To this end, we are combining multiple data sources into integrated data products. An example is the Census Bureau's APIs, which allow programmers to access several Census Bureau datasets for custom use in apps (U.S. Census Bureau, 2017a).

Several types of unauthorized disclosure of data pose potential concerns. The most dramatic is re-identification, where a Census record is attached to an external file containing personally identifiable information, revealing the identity of the person, household or business in the Census Bureau file. Another concern is attribute disclosure, where a person may not be re-identified, but some attribute of that

person is revealed, either exactly or within a range. Finally, an inferential disclosure occurs when a data user can determine an identity or attribute with high confidence but not certainty. A person's presence within a survey is itself confidential, so any attack that could reveal this information must also be protected against. These disclosure concerns are discussed in Reiter (2012).

Preventing attacks while maintaining data quality is a particular concern with outliers, especially if they represent well-known records. For example, a public figure (athlete, entertainer, etc.) may be known to have wages of $31 million in a given year, so if a data user discovers someone with those wages in the appropriate geographic region for that year, the user can hypothesize with some confidence that the person discovered was the aforementioned public figure. Since it is possible that no one else in that county has wages in that specific range, even adding noise to the value may not do much to mask it, unless the noise is so extreme that the resulting impact on accuracy is unacceptable for most intended data uses. Topcoding (truncating variables if they are above a certain value) can protect against some disclosure risk but limits the legitimate information a data user can derive about the tail of the distribution.

Protecting confidentiality is both a legal requirement and an ethical obligation, so the data dissemination transformation necessitates a corresponding transformation in the Census Bureau's disclosure avoidance procedures. This transformation is also driven by the increasing amount of publicly available data, as such external data can now more readily be used to identify a person, household or business in a Census Bureau data release. Increasing computing power and sophisticated data mining techniques make it easier to leverage an external dataset to cause disclosure, compounding these concerns.

Broadly speaking, there are two ways to protect confidentiality—suppression and perturbation. Suppression has traditionally been used when providing data would make individual records or their attributes easy to identify, but its ability to protect privacy is limited if the suppressed records are also used in the computation of marginals or summary statistics. Perturbation, wherein the data are modified, may allow something closer to the full dataset to be released (perhaps still with some loss of detail) but gives data that deviate from the original data and are therefore less accurate.

A recent Census Bureau research project aligned with the data dissemination transformation was the Microdata Analysis System (MAS), which would allow users to request a custom table of Census Bureau data and have the table delivered to them quickly online. Tables would be based on the official Census Bureau microdata, with relatively minimal perturbation. A primary means of disclosure avoidance for the MAS was suppression of tables when the cell counts were too small, based on a set of rules. Freiman *et al.* (2016) showed that for queries on census tracts, and to a lesser but substantial extent counties, tables were often suppressed. However, the rules were often insufficient to prevent the attacks we simulated. Hence we could not loosen the rules without creating unacceptable risk. Tightening the rules, or even leaving them as originally planned, would lead to frustratingly many tables being denied. We concluded that a table suppression approach could not be successful and discontinued research on the MAS.

Subsequent research has focused on perturbation by creation of synthetic data (Reiter and Raghunathan, 2007)—where the unprotected data are used to train a model, which is used to generate data sharing many of the properties of the original dataset. The synthetic data may then be released, either as microdata or as the basis for some other data product. Synthetic datasets come in two varieties:

- Fully synthetic: All of the records are completely generated from the model. This can be done by treating the data fields to be filled in as missing records to be imputed.
- Partially synthetic: Only some records or some variables are generated, with the rest of the synthetic dataset being identical to the original data.

Fully synthetic data were proposed by Rubin (1993) and expanded upon by Raghunathan *et al.* (2003), and partially synthetic data were introduced by Little (1993), with Reiter (2003) proposing the name "partially synthetic." See Reiter and Raghunathan (2007) for a review of the differences between full and partial synthesis.

The Census Bureau currently releases several synthetic data products, including the Survey of Income and Program Participation (SIPP) Synthetic Beta (Benedetto *et al.* 2013), the Synthetic Longitudinal Business Database (SynLDB) (Kinney *et al.*, 2011) and the OnTheMap application (Machanavajjhala *et al.*, 2008). In addition, the official microdata for group quarters in the Census of Population and Housing and American Community Survey (ACS) are partially synthetic (Hawala, 2008).

The Census Bureau increasingly aims for its products to satisfy formal privacy, a broad class of criteria where the loss of privacy from the release of output is formally defined and quantified, so that the tradeoff between privacy loss and data accuracy can be set by policy-makers. For example, a disclosure protection algorithm satisfies differential privacy, one such criterion, if the probability of observing a given collection of output from the algorithm changes by no more than a certain amount depending on whether a single record is included in or excluded from any dataset (Dwork, 2006). This criterion protects against any arbitrary set of information potential intruders may already have about the records in the dataset. In this case, and in most formal privacy mechanisms, the "certain amount" is a tuning parameter that is set according to the determination of policy-makers. A higher value means less protection but usually better data quality. Formal privacy is part of the Census Bureau's response to the proliferation of external data sources and the availability of massive computation power, which makes it possible to reconstruct databases of confidential data based on the release of too many statistics drawn from the same confidential dataset (see Dinur and Nissim, 2003).

Work is ongoing at the Census Bureau to create products that are formally private, but this paper does not consider such methods. The synthetic data approach we investigate here does not satisfy formal privacy, but it might be modified to do so in the future.

2. Creating Synthetic Data Using Trees

A wide variety of methods may be used to synthesize data, notably draws from posterior predictive distributions as given in Raghunathan *et al.* (2003) and Reiter (2005a). This research applies classification and regression trees to generating synthetic data.

Proposed by Breiman *et al.* (1984), classification and regression trees (CART) are a method of prediction based on a series of binary splits of the predictor variables to predict a response variable. Figure 1 gives an example of a tree to predict wages. The tree generation algorithm considers all possible binary splits on a single variable, using a deviance criterion to determine what split creates two bins (or nodes) that are the most homogeneous. In this case, the algorithm determines that the optimal split is to put people with a

graduate or professional degree in the right group and people without such a degree in the left group. The algorithm then splits each of these two groups into subgroups. The non-graduate/professional degree group can best be split again on education, putting those who have at least a Bachelor's degree in one group and those who do not in the other group. The graduate and professional degree group can best be split on age, with those under age 70 going into one group and those age 70 and older going into the other group. We now have four groups, which we continue splitting. In the example, the optimal split for graduate/professional degree holders under 70 is on sex, illustrating that even after three splits, the formula can consider three-way interactions between the predictor variables.

We continue splitting the bins to maximize homogeneity of wages until we reach some stopping criterion, creating a set of paths that looks like a flow chart or an upside-down tree. In our research, we used a simple stopping criterion that said that no bin could have fewer than five observations in the training dataset. We also required that no node could be split if its pre-split deviance was less than 10^{-9} times the deviance of the root node of the tree. The deviance criterion ensures that if a bin is extremely homogeneous, then the algorithm will not try to find more signal within the bin, providing some minimal protection against overfitting. In practice, the deviance used is very small, so we would expect this deviance requirement to affect the growing of a tree only very rarely. When trees are used for prediction rather than for synthesis, the criteria often place more emphasis on not splitting leaves once they reach a certain level of homogeneity.

In most trees, the number of splits before one reaches a leaf depends on the path taken; the depth of the tree need not be uniform.

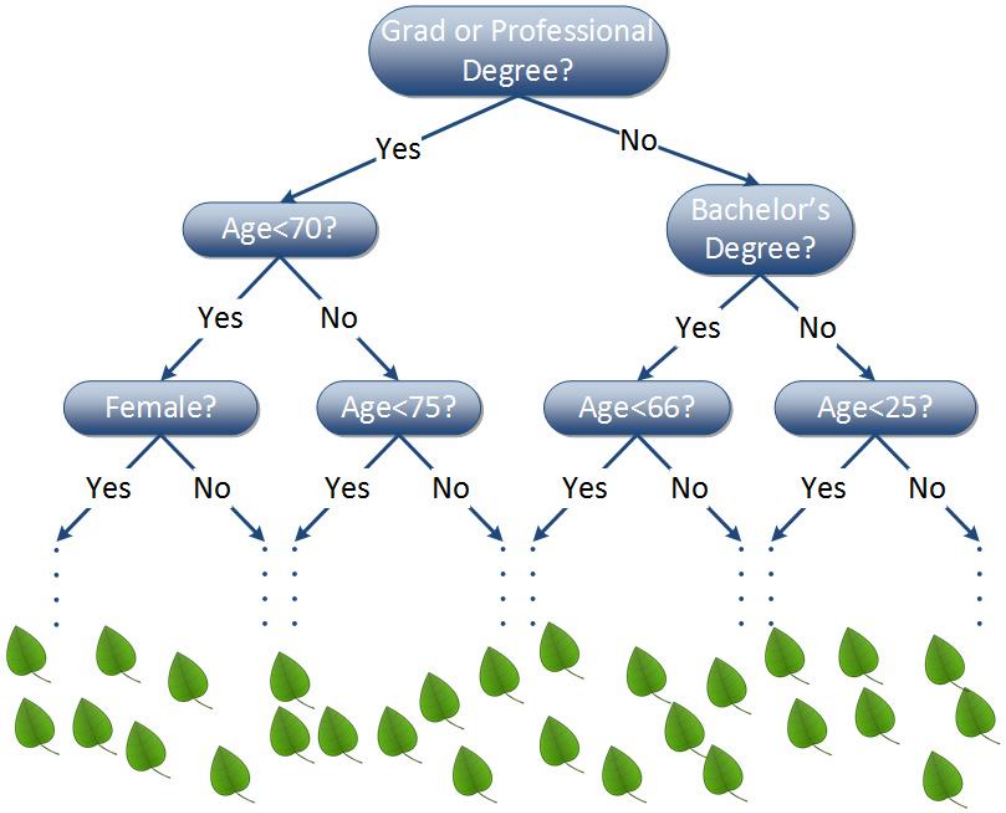


Figure 1: Example of a tree to predict wages. The figure illustrates how a tree works and is not based on fitting a tree to actual data.

To synthesize the value of wages using a tree, we take a new observation for which the other variable values have already been synthesized and run it through the series of splits, until we reach a node that is not split, which is called a terminal node or leaf. We then draw as the synthetic wage one of the values of wages from the original records that fell into that leaf.

Synthesis using trees was proposed by Reiter (2005b). Reiter notes that this synthesis approach:

- is easily applied and makes no parametric assumptions;
- can capture non-linear relationships and interaction effects that may not be easily revealed; and
- provides a semi-automatic way to fit the most important relationships in the data.

However, Reiter also notes drawbacks as compared to parametric models:

- Trees partition the dataset into pieces (represented by the leaves), creating discontinuities at the partition boundaries.
- When relationships between variables can be accurately described by parametric models, a tree is not as effective as using the appropriate parametric model.

Since synthetic data are not based on the collected data as directly as other types of data, it can be useful to have multiple synthetic datasets, known as implicates. The researcher can evaluate the variation in

results across the datasets to allow for proper estimation of variances (Reiter and Raghunathan, 2007). In our research, we used the same tree to create all of the implicates.

Our approach to creating synthetic data is as follows. List the variables in sequence, with the sequence determined roughly so that later variables in the sequence are likely to be response variables in an analysis where the other variables are predictors. For example, sex and age may frequently be used as predictors to model educational attainment, but using educational attainment to predict sex and age seems less natural in most cases. Hence, we synthesize sex and age before educational attainment. This ordering of variables is somewhat subjective, and we did not test different orderings of variables to determine whether the results were sensitive to the ordering. We make the first variable categorical, allowing us to draw probabilities for the first variable from a Dirichlet distribution whose parameters are the observed counts of the first variable. The synthetic values for the first variable are then drawn from a multinomial distribution with these probabilities. For each subsequent variable, we build a tree based on the original data, using variables earlier in the list to predict the current variable. Depending on the tuning parameters, the synthetic value may be generated in two ways:

1) No noise. The synthetic value is the value of the variable from a randomly selected record in the leaf in the original data.
2) Noise. The values from the leaf are run through a kernel density estimator and a value is chosen from this distribution. A tuning parameter controls the bandwidth of the kernel density smoother. Using a kernel smoother to synthesize data was originally proposed by Abowd and Woodcock (2004), and the methodology was adapted by Woodcock and Benedetto (2006), although both papers do so in contexts other than trees.

If no noise is used, then the synthesis method will output actual values of variables such as income, albeit not necessarily in such a way that they can easily be tied back to records in the original data. Hence noise is desirable for variables that can be treated as continuous (Reiter, 2005b). In this research, we consider no noise as one possibility, although most of our synthesis includes noise. We use a smoother bandwidth equal to a multiplier times the value given by Silverman's (1986) rule of thumb, using the nrd0 option in R. The multiplier varies in different iterations between 1, 1.5, 2 and 2.5.

3. Pilot Study

In our research, we synthesize the variables using tree methods described above, which are similar to tree methods available in the well-known synthpop package in R (Nowok *et al.*, 2015). We use the five-year ACS public use microdata sample (PUMS) from 2010-2014. Although subsequent work uses the internal microdata, using the PUMS for this research allows us to share the results quickly and publicly. We use data from 2012 to 2014 in Public Use Microdata Area (PUMA) 00101 in Washington DC, considering only people age 18 or older. This dataset gives us a sample size of about 2,500. We use the inflation adjustment factor in the file to convert all wages to 2014 dollars.

We synthesize variables in this order:

- Sex
- Age
- Race

- Hispanic Origin
- Educational Attainment
- Marital Status
- Wages

We also make some adjustments to counteract some of the disclosure avoidance methods applied to public use microdata, giving data whose properties should more closely match the originally collected data. Since the wage data in the PUMS are rounded, we add noise to "undo" the rounding, essentially replacing every wage in the file with a wage uniformly distributed among the unrounded wages that could have led to the observed wage. The public use data are also top-coded; we approximate the distribution of actual wages by replacing every top-coded wage with a value drawn from a shifted exponential distribution whose minimum value is the top-coding threshold and whose mean is the mean of the top-coded values. We then use these values of wages as the "ground truth" for creating our synthetic dataset. The results in this paper could potentially be somewhat sensitive to the way in which the "true" dataset was generated.

We recode race into four categories: white alone, black alone, Asian alone and other, which includes multiple races. We recode Hispanic origin into two categories: Hispanic and not Hispanic. We consider education as continuous rather than categorical, since for the most part, the categories of education are ordered. Converting ordered data to magnitude data introduces some judgment (is the difference between an associate's degree and a bachelor's degree the same as the difference between a bachelor's and a master's?), but given a fixed ordering of the values, trees are relatively insensitive to the choices made about the distances between values.

Generating the data for all variables other than wage is a direct application of the tree methodology we have described. We use a minimum leaf size of 5.

When synthetic wages are generated, there are a number of tuning parameters:

- We can generate data with and without noise.
- If we use noise, the bandwidth multiplier can vary.
- If we use noise, the kernel density function can have two different types of support:
    - Method 1: Support only from the lowest value to the highest value in the leaf.
    - Method 2: Support from the lowest value to the highest value in the leaf, except if the highest value exceeds a certain threshold, in which case the upper bound of the support is 1.5 times the maximum value in the leaf. The threshold we use is the same as the topcoding threshold in the public use data.

Method 2 ensures that there is some possibility of the largest values in the synthetic dataset being larger than the largest values in the original dataset.

For each collection of settings, we create 50 implicates, i.e., 50 synthetic copies of the dataset. Creating multiple synthetic implicates to correspond to one set of collected data is a common though not universal practice. In practice, the number of implicates released to the public is rarely that large; we use 50 here to lessen effects on our analysis from having a small number of implicates.

4. Results

4.1 Data Quality

We found that the mean value of wages increased as the amount of noise increased, indicating that noise degraded overall data quality. Table 1 shows the means under the two methods and under varying amounts of noise. Asterisked means are significantly different from the original mean of the data.

Table 1. Mean wages as a function of noise factor.

	Method 1		**Method 2**	
	Mean	**SE**	**Mean**	**SE**
Original Data	69,311		69,311	
Noise Factor = 1	69,210	2,088	73,521	2,405
Noise Factor = 2	71,482	2,147	77,186*	2,579
Noise Factor = 3	73,647	2,236	80,461*	2,753
Noise Factor = 4	74,895*	2,273	83,280*	2,892

Asterisks denote means statistically significantly different from the original mean.

We believe that the upward bias in wages is due to the leaves of the tree usually being right-skewed. Drawing from a normal kernel density estimator is equivalent to the following procedure:

- Choose at random one of the points used to train the estimator.
- Draw from a normal distribution centered at that point.
- If the value selected is within the allowable range, choose this value. If the value is outside the range, throw the point out and start again.

The last bullet point leads to the bias. If no points were thrown out and redrawn, the kernel density estimator should have the same mean as the original points. Since there are usually more points near the bottom of the leaf than near the top, more values will be thrown out for being too low than for being too high, creating a positive bias. Extending the bounds downward for the allowable values of the kernel density might lessen or eliminate this bias but might also lead to implausible synthetic values, so for this research the bound remains in place. We did not consider transformations of the wages, such as log or cube root, but it is possible that these would alleviate the skewness.

Figure 2 is a histogram of wages in the dataset, which are dramatically right-skewed (sample skewness≈3.43). Figure 3 is a histogram of the skewness of the individual leaves, which shows that most leaves are also right-skewed, reinforcing the point above about leaves having more values near the bottom than near the top. The average skewness is .33, and 68% of leaves are at least slightly right-skewed. About 14% of leaves have skewness greater than 1 (substantial right-skewness), while no leaves have skewness less than -1. When Method 2 is used, few if any points are rejected and redrawn because they are too large, but some points are rejected and redrawn because they are too small, further adding to the bias.

Figure 2: Histogram of wages.

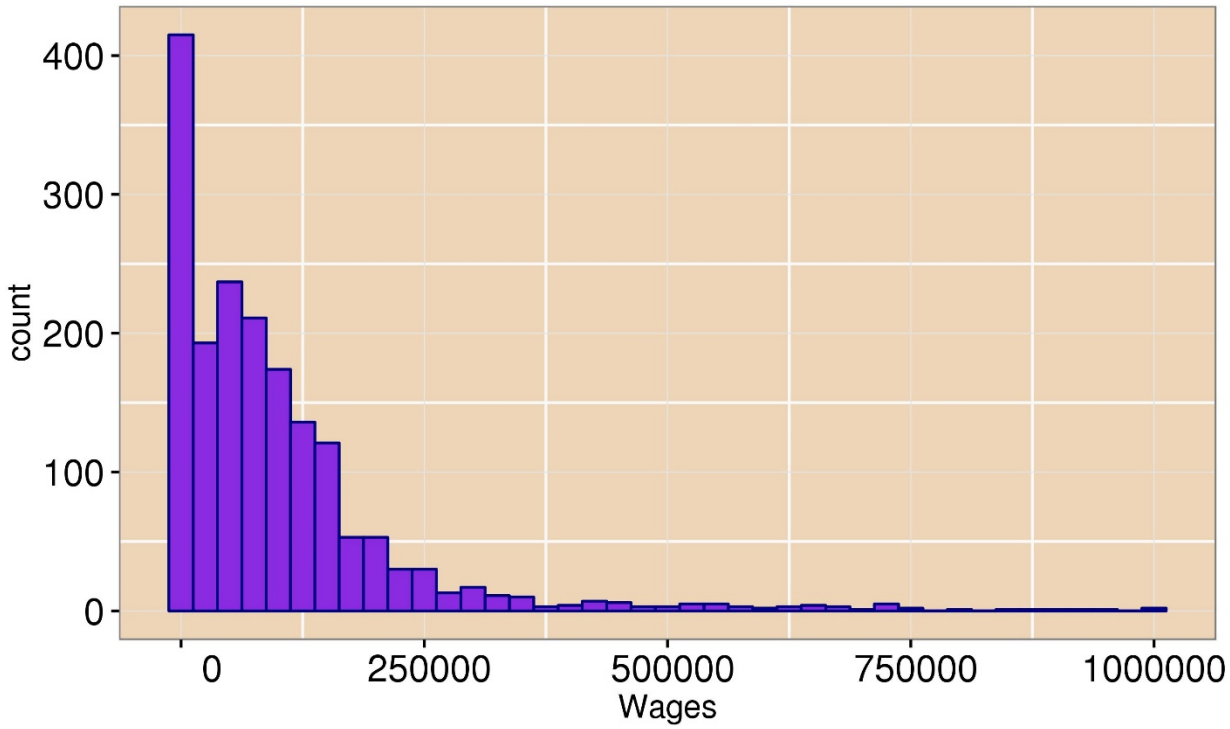
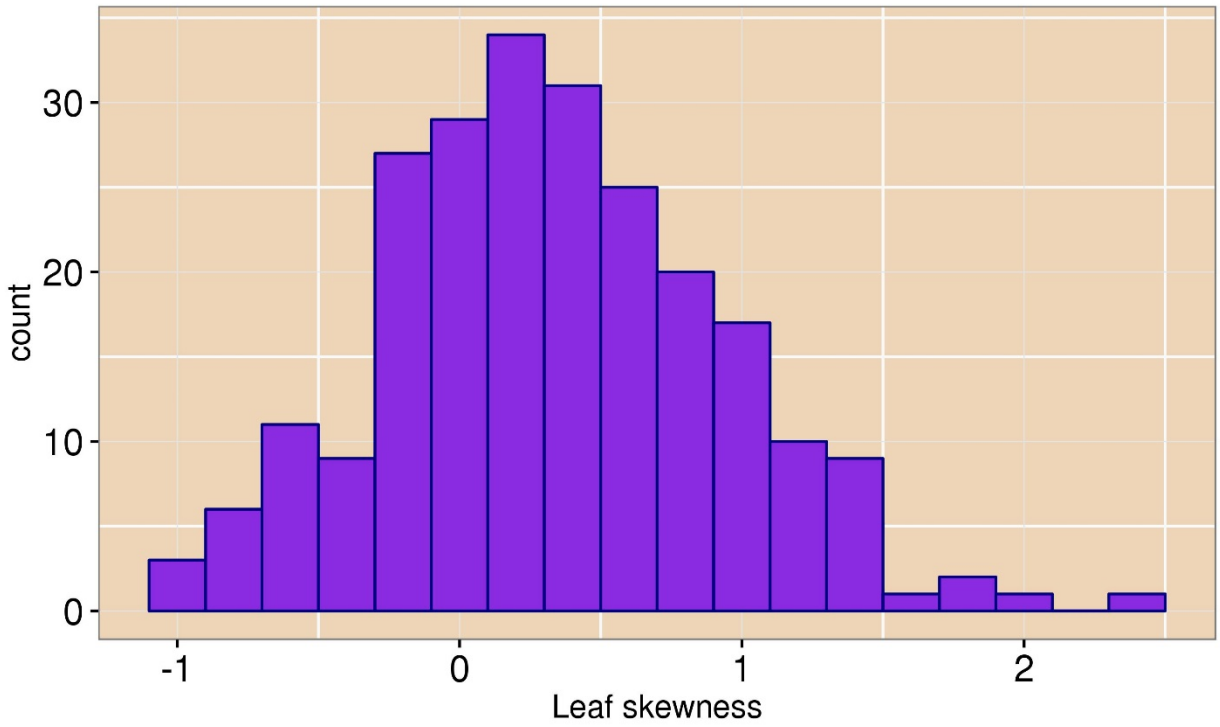


Figure 3: Histogram of leaf skewness.



4.2 Disclosure Protection

We examine whether the maximum income can be reconstructed, since the maximum is likely to be the riskiest income. For our initial dataset, the maximum income is just over $1 million. The risk to the maximum is likely to depend on whether it substantially exceeds the other values in the dataset. Hence, we make alternative starting datasets to use as the "ground truth" for synthesis. Each one is the same as the original dataset we described, except that the maximum value is replaced by values from $1 million to $2 million in increments of $200,000. The initial dataset includes a second record where the wages are just above $1 million—in this case, $1,001,399. We do not change this record. Hence the first alternative starting dataset includes the recoded version of the original maximum record, with wages of $1 million, and a record with wages slightly above $1 million, which is the new maximum for this dataset. Since this record is so close to $1 million, we do not believe synthesis using this value to be appreciably different from synthesis if there were two records with a wage of $1 million.

Figure 4 shows the synthetic maxima for each of the 50 implicates for different noise multipliers and different "true" maxima, using Method 1. Since the range of a leaf cannot go beyond the range of the data, an intruder estimating the maximum wage this way knows he cannot overestimate the maximum. Even with noise, we would expect that at least one of the 50 implicates would have a maximum that is very close to the original maximum. Figure 4 shows this is the case: although some implicates have maxima dramatically below the original maximum, almost invariably at least a few synthetic maxima are very nearly equal to the original maximum. Hence, the maximum of all synthetic implicates is a good estimate of the maximum of the original data.

Figure 4: Implicate maxima for different synthetic dataset maxima and bandwidth factors. (Method 1)

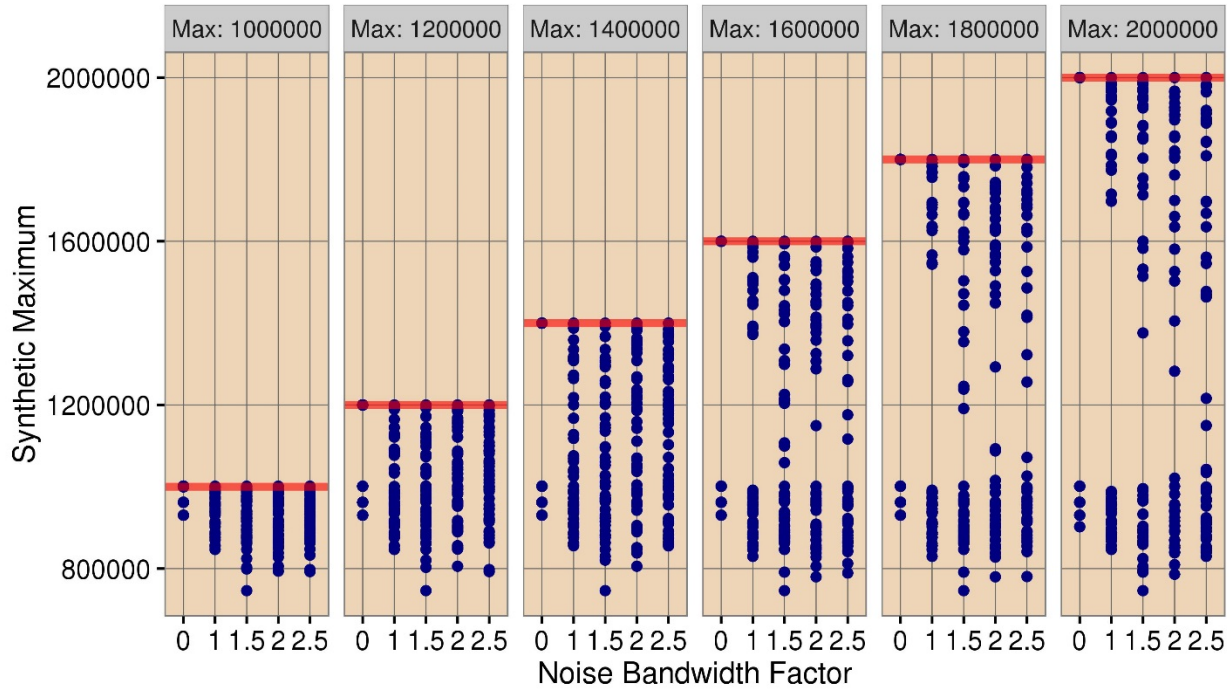


Figure 5 indicates that for a given original dataset using Method 2, more noise is associated with a higher overall maximum across all implicates. The highest possible synthetic maximum under Method 2 is 1.5

times the original data maximum, and this value is generally almost reached for some implicates for a noise factor of 2.5, although less so for the lower noise factors. Indeed, with a high level of noise but with synthetic values capped at 1.5 times the maximum value in the leaf, the maximum can be estimated well by dividing the maximum across all implicates by 1.5. With less noise—particularly a noise multiplier of 1 or 1.5—the maximum is not as close to 1.5 times the original maximum and using this estimation method for the original maximum may somewhat underestimate. However, if the factor used to determine the cap is known to the intruder, the intruder can divide by the factor to find a lower bound on the maximum wage, and depending on the synthesis method, the intruder may be able to infer that the maximum wage probably only slightly exceeds this bound. Table 2 shows that for the higher noise factors and 50 implicates, a user is likely to be able to get a very close estimate of the maximum wage.

Figure 5: Implicate maxima for different original maxima and bandwidth factors. (Method 2)

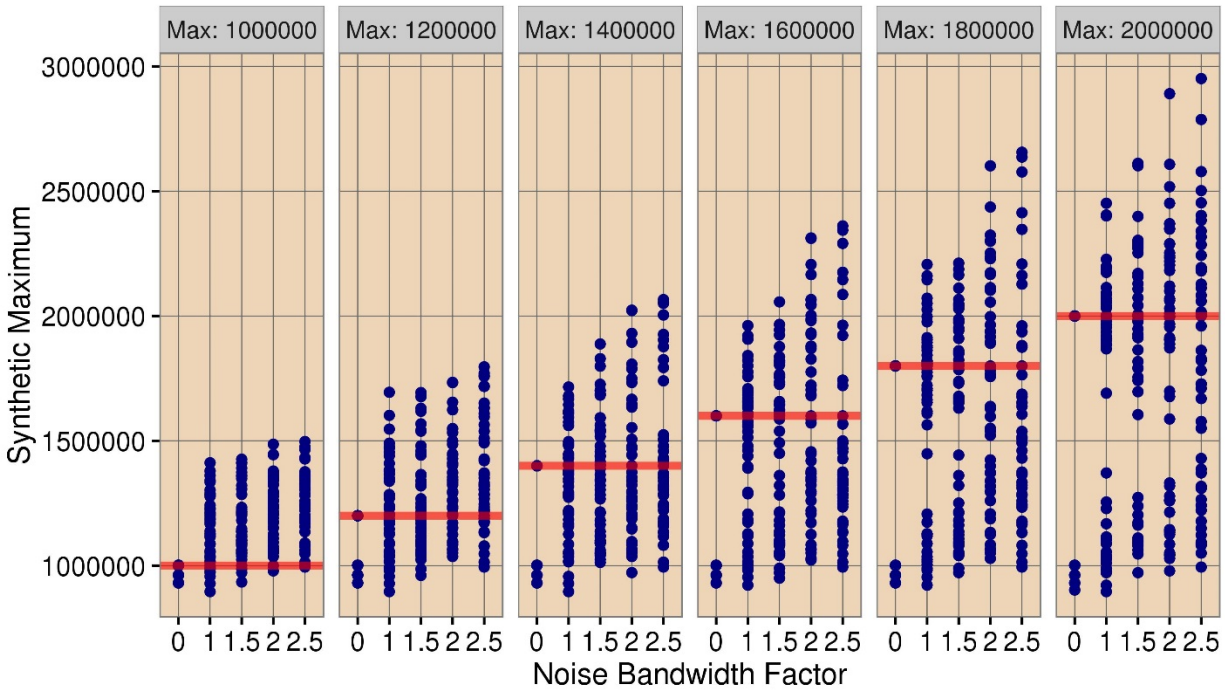


Table 2. 1/1.5 times Maximum of implicate maxima (in thousands of dollars) and percent error by noise factor.

Noise Factor	1,001	1,200	1,400	1,600	1,800	2,000
2	991 (-1.1%)	1,156 (-3.6%)	1,349 (-0.4%)	1,542 (-3.6%)	1,734 (-3.6%)	1,927 (-3.6%)
2.5	998 (-0.3%)	1,198 (-0.2%)	1,377 (-0.2%)	1,574 (-1.6%)	1,771 (-1.6%)	1,967 (-1.6%)

More generally, the maximum can be estimated by taking the median of the implicate maxima. For smaller noise multipliers (1 and 1.5), the median implicate maximum generally gives a more accurate estimate of the maximum income than the mean, while for larger noise multipliers (2 and 2.5), the mean is usually but not always more accurate. The difference between the mean and the median may also be a

result of skewness in the data. A higher noise multiplier seems to protect the maximum from being derived very closely using the median, but an intruder can still get a general range for the maximum this way and can get a closer estimate of the maximum by the methods in Table 2. Also, Table 1 indicates that noise factors this large lead to noticeable positive bias in the mean.

Table 3. Median of implicate maxima (in thousands of dollars) and percent error by noise factor.

Noise Factor	1,001	1,200	1,400	1,600	1,800	2,000
0	1,000 (-0.1%)	1,200	1,400	1,600	1,800	2,000
1	1,123 (+12.1%)	1,195 (-0.4%)	1,357 (-3.1%)	1,561 (-2.4%)	1,740 (-3.4%)	1,906 (-4.7%)
1.5	1,120 (+11.8%)	1,207 (+0.6%)	1,342 (-4.1%)	1,471 (-8.1%)	1,674 (-7.0%)	1,913 (-4.3%)
2	1,259 (+25.7%)	1,296 (+8.0%)	1,296 (-7.4%)	1,513 (-5.4%)	1,633 (-9.3%)	1,935 (-3.3%)
2.5	1,295 (+29.3%)	1,311 (+9.3%)	1,331 (-4.9%)	1,334 (-16.6%)	1,446 (-19.7%)	1,755 (-12.2%)

5. Conclusion

Our results suggest that the risk of disclosure of the maximum wage increases as more implicates are released. The wide range of implicate maxima in the plots indicates that for a single implicate, the kernel density smoother may provide enough protection for the maximum if the smoothing bandwidth is large enough and the support of the kernel density smoother extends beyond the range of the leaf. For multiple implicates, the tree-based method of creating synthetic data may allow close estimation of the maximum wage, particularly when the maximum is an outlier. If Method 1 is used, the maximum of the implicate maxima is a good estimate of the original maximum. If Method 2 is used, the original maximum can be estimated from the median of the implicate maxima or in some cases from the maximum of the implicate maxima. If the approximate original population maximum is already known, the synthetic data may lead to an inference about whether the highest-wage person in the population is in the sample.

Some modifications could improve the protection of the maximum. In this research, the upper bound of synthetic values was at most 1.5 times the top value in the leaf. Increasing this multiplier would make an attack based on the maximum of the implicate maxima less effective but would probably further bias the mean wage and could lead to synthetic wages that were not anywhere near the range of the original wage. Such a change would also do little to prevent a median-based attack. We could gain some protection by giving very little information about the data synthesis mechanism, including the bound parameter, but such secrecy limits the utility of the data to researchers, who may need to know the nature of the disclosure avoidance to make proper inferences. The formal privacy approach dictates that the data should be safe even if the user knows the exact synthesis algorithm, with the exception of random seeds. This includes the assumption that all parameter values, such as the factor 1.5, are publicly known.

Regardless of what method is used, smoothing has substantial negative effects on data quality, positively biasing the mean. Higher bandwidths increase the bias, as does increasing the upper bound of allowable

smoothed values. This problem could be alleviated by decreasing the lower bound, but extending the lower bound too far would lead to the possibility of negative wages, which are not allowed. The problem may also be due to the skewness of the variable we examined; in future work, we may consider transformations that may lessen the skewness.

We conclude that other methods may be necessary to protect continuous variables such as wages where outliers are possible. Current research continues to use a tree-based approach for categorical variables and for continuous variables that have a limited number of possible values, but we are now looking into regression-based approaches for synthesizing continuous variables such as wages.

References

Benedetto, G., Stinson, M., & Abowd, J. M. (2013). The creation and use of the SIPP Synthetic Beta. https://ecommons.cornell.edu/bitstream/handle/1813/43924/SSBdescribe_nontechnical.pdf?sequence=3

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Cynthia, D. (2006). Differential privacy. *Automata, languages and programming*, 1-12.

Freiman, M. H., Schar, B., Hasenstab, K., & Lauger, A. (2016). Evaluating a remote access system. Working Paper CDAR2016-04, U.S. Census Bureau.

Hawala, S. (2008). Producing partially synthetic data to avoid disclosure. In *Proceedings of the Joint Statistical Meetings. Alexandria, VA: American Statistical Association*.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., & Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *International Statistical Review*, *79*(3), 362-384.

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official statistics*, *9*(2), 407.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008, April). Privacy: Theory meets practice On The Map. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on* (pp. 277-286). IEEE.

Nowok, B., Raab, G. M., & Dibben, C. (2015). synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*.

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. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, *29*(2), 181-188.

Reiter, J. P. (2005a). 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), 185-205.

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

Reiter, J. P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public opinion quarterly*, *76*(1), 163-181.

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

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of official Statistics*, *9*(2), 461-468.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). CRC press.

U.S. Census Bureau (2017a). Available APIs. Available at https://www.census.gov/data/developers/data-sets.html.

U.S. Census Bureau (2017b). What we do. Available at https://www.census.gov/about/what.html.

Woodcock, S. D., & Benedetto, G. (2009). Distribution-preserving statistical disclosure limitation. *Computational Statistics & Data Analysis*, *53*(12), 4228-4242.