

# Making public use, synthetic files of the Longitudinal Business Database

Satkartar K. Kinney and Jerome P. Reiter  
Duke University

## Abstract

Longitudinal business data are widely desired by researchers, but difficult to make available to the public because of confidentiality constraints. In this paper, we discuss the generation of synthetic public use datasets for establishment data. The basic idea is to release simulated values of sensitive variables, generated from probability distributions fit using genuine data. This can protect confidentiality, since attributes are synthetic rather than real. And, when the models describe the data well, broad-scale inferences from the synthetic datasets will be inferentially valid. We discuss the approaches used for generating synthetic public-use files for the U. S. Census Longitudinal Business Database.

KEY WORDS: Multiple imputation; Partially synthetic data; Statistical Disclosure Limitation

## 1. Introduction

Many statistical agencies disseminate microdata, i.e., data on individual units, in public use files. These agencies strive to release files that are (i) safe from attacks by ill-intentioned data users seeking to learn respondents' identities or attributes, (ii) informative for a wide range of statistical analyses, and (iii) easy for users to analyze with standard statistical methods. Doing this well is a difficult task. The proliferation of publicly available databases, and improvements in record linkage technologies, have made disclosures a serious threat, to the point where most statistical agencies alter microdata before release. Some of the methods employed, such as data swapping (Dalenius and Reiss, 1982) or adding random noise (Fuller, 1993) reduce the utility of the released data and require complex procedures to analyze properly (Reiter, 2004).

One approach that does allow for valid inferences to be made using standard statistical methods is the use of multiple imputation to generate synthetic datasets. Rubin (1993) first proposed using multiple imputation to generate fully synthetic datasets for the purpose of statistical disclosure limitation. In this approach, the actual units and collected values from the confidential microdata are not released, only multiply-imputed datasets generated from models fit with the original survey data. Simple combining rules developed by Raghunathan *et al.* (2003) and Reiter (2005) allow users to make valid inferences us-

ing standard statistical methods and software. Fully synthetic data are further described in Rubin (1993), Raghunathan *et al.* (2003), and Reiter (2002).

While fully synthetic data represents a promising disclosure limitation approach, it is difficult to implement, and no agencies have yet used it to generate public use files. A variant, partially synthetic datasets, proposed by Little (1993), has been used successfully (Kennickell, 1997; Abowd and Woodcock, 2001). Partially synthetic datasets retain the originally sampled units and some of their observed values while other values, such as key identifiers or sensitive values, are replaced with multiple imputations. Hence confidentiality can be protected while allowing valid inferences to be made. An advantage over fully synthetic datasets is that the inferences are less sensitive to the imputation models; however, some risks of disclosure remain. Combining rules for inferences with univariate and multivariate estimands from partially synthetic data have been developed by Reiter (2003, 2005). It is not expected that every possible analysis will be valid in the synthetic data. Users desiring complex or detailed inferences at the individual unit level may need to seek special access to the confidential microdata.

This paper describes current efforts to create partially synthetic datasets for longitudinal establishment data from the Longitudinal Business Database (LBD). In particular, entry and exit information, and annual payroll and employment data are desired. Real geographic and industry classification information are also included. The synthetic data generated should ideally preserve relationships among the different variables. Section 2 describes the methods used to generate synthetic data. Section 3 describes how these are being applied to LBD. Section 4 discusses further work to be done. This work represents preliminary efforts to synthesize a longitudinal establishment dataset.

## 2. Synthesis Methods

We use the notation of Reiter (2003). Let  $I_j = 1, j = 1, \dots, N$  indicate unit  $j$  was selected in the original survey. Let  $Y_{obs}$  be the  $n \times p$  matrix of observed survey data for  $I_j = 1$ ; let  $Y_{nobs}$  be the  $(N - n) \times p$  matrix of unobserved survey data for the units with  $I_j = 0$ ; and let  $Y = (Y_{obs}, Y_{nobs})$ . For simplicity we assume that all units respond fully to the survey, i.e., there are no missing values. We let  $X$  be the  $N \times p$  matrix of design vari-

ables for all units in the population and assume that this is known. Synthetic datasets will be constructed based on the observed data  $D = (X, Y_{obs}, I)$ .

Partially synthetic datasets are constructed by replacing selected values in the observed data with  $r$  independent draws from posterior predictive distributions. Let  $Z_j = 1$  indicate that unit  $j$  has been selected to have any observed values replaced with imputations. Let  $Y_{rep}^{(i)}$  be the imputed values in the  $i$ th synthetic dataset,  $i = 1, \dots, r$ , assumed to be drawn from  $(Y_{rep}|D, Z)$ , and  $Y_{nrep}$  be the unchanged values, which are the same in each dataset. Each of  $r$  synthetic datasets,  $D^{(i)}$ , is comprised of  $(X, Y_{rep}^{(i)}, Y_{nrep}, I, Z)$ . Imputations should only be made from the posterior predictive distribution of those units with  $Z_j = 1$ . We assume here that all units will have their values of confidential variables replaced, i.e.,  $Z_j = 1$  for all units.

When several variables are considered confidential, i.e.,  $Y$  has dimension  $N \times d$ , specification of the joint posterior density  $Y|X$  may be difficult. We write the joint distribution as a product of conditional densities. For  $Y = (y_1, \dots, y_d)$ , sampling from  $Y|X$  is thus achieved by sampling from  $f(y_1|X)$ ,  $f(y_2|y_1, X)$ ,  $\dots$ ,  $f(y_d|y_1, \dots, y_{d-1}, X)$ . This allows complex relationships to be modeled in a computationally feasible fashion. Some predictors can be omitted from the imputation model if an independence relationship is reasonable. For example, with longitudinal variables, values in year  $t$  may be assumed to be dependent on values in year  $t - 1$ , but not on values in previous years.

A similar approach was taken by Abowd and Woodcock (2004) for generation of longitudinal linked data, where observations are taken from multiple sampling frames. They approximated the joint density using a sequence of conditional densities defined by generalized linear models. A key difference is that each confidential variable  $y_k$  is drawn from the conditional distribution  $f(y_k|X, Y_{-k})$  where  $Y_{-k}$  represents all confidential variables excluding  $y_k$ . While this may produce satisfactory results, it is not guaranteed to converge to a joint distribution.

Typical confidential variables for establishment data include highly skewed variables such as income, in which case normal linear models do a poor job of modelling the data. In addition, categorical variables common in establishment data such as geographic or industry indicators can have numerous categories which can make model fitting difficult. Below we describe some approaches used in the generation of synthetic data for the LBD.

## 2.1 Normal Method

A common approach is to model the posterior distribution using a normal linear regression model, possibly on trans-

formed data. Initially we experimented with more flexible approaches utilizing a Generalized Additive Model (GAM), but ultimately were unsatisfied with the preservation of the correlation structure. Using appropriate transformations, we are able to impute these using the simpler and faster linear approach.

We would like to generate a synthetic variable  $\tilde{y}_1$  for confidential variable  $y_1$  by drawing from an appropriate conditional distribution  $f(y_1|X)$ . Using the normal approach,  $\tilde{y}_1$  is generated simply by taking draws from the posterior predictive distribution:

1. Apply appropriate transformations, if needed, to satisfy linear regression assumptions
2. Fit a linear model to the observed data, obtain estimates of  $\beta$  and  $\sigma^2$ .
3. For each imputation, draw new values  $\tilde{\sigma}_{(i)}^2$  and  $\tilde{\beta}_{(i)}$  from the posterior distributions  $f(\sigma^2|y, X)$  and  $f(\beta|\sigma^2, y, X)$ .
4. Draw  $\tilde{y}_1^{(i)}$  from  $N(X\tilde{\beta}_{(i)}, \tilde{\sigma}_{(i)}^2)$

If we would also like to generate  $\tilde{y}_2$  from  $f(y_2|X, y_1)$ , we can follow the same procedure but in Step 4 we need to draw  $\tilde{y}_2^{(i)}$  from  $N(X^{(i)}\tilde{\beta}^{(i)}, \tilde{\sigma}^{2(i)})$ , where  $X^{(i)} = (X, \tilde{y}_1^{(i)})$ . Similarly, we can take draws from  $f(y_k|X, Y_{-k})$ .

## 2.2 Nonnormal Models

The normal approach can be modified for nonlinear models; however, adaptations will be needed for nonparametric or semiparametric models such as the GAM to model the error distribution. For binary and categorical responses without very many categories, one can sample from binomial and multinomial distributions, using appropriate generalized linear models to obtain the sampling probabilities.

To generate a synthetic variable  $\tilde{y}_1$  for binary response  $y_1$ , we first fit a logistic model using the observed data to obtain predicted probabilities  $\hat{p}_l(X_l)$ ,  $l = 1, \dots, n$ . The synthetic  $\tilde{y}_1^{(i)}$  is obtained by sampling from  $Bin(1, \hat{p}_l(X_l))$ . To approximate draws from  $f(y_2|y_1, X)$  when  $y_2$  is binary, we use the observed data to fit a logistic model to obtain  $\hat{p}_l(X_l, y_{1l})$ , and save the coefficients so that we may determine  $\hat{p}_l(X_l, \tilde{y}_{1l}^{(i)})$ . The synthetic  $\tilde{y}_2^{(i)}$  are then obtained by sampling from  $Bin(1, \hat{p}_l(X_l, \tilde{y}_{1l}^{(i)}))$ . For categorical responses, the same approach is used, but a generalized logit model is used in place of a logistic model to obtain posterior probabilities  $\hat{p}_{lj}(x_l)$ ,  $l = 1, \dots, n$ ;  $j = 1, \dots, c$ , where  $c$  is the number of categories in the response. A multinomial distribution is used in place the binomial.

### 2.3 Dirichlet-multinomial method

When there are several categories in the response, or several categorical predictors, the generalized logit model can become impossible to fit. The multinomial-Dirichlet model provides a convenient framework for sampling from the posterior predictive distribution for a categorical  $y$  when  $X$  are categorical, and can also be used to impute missing data.

In the disclosure limitation setting, problems may arise when categories and categorical variables are too numerous. Let  $C$  be a unique category determined by categorical predictors in  $X$  and let  $y_C$  be the observed values of a categorical response variable corresponding to the  $n_C$  units in  $C$ . If  $n_C = 1$ , or  $y_{Ci}, i = 1, \dots, n_C$  all have the same value, then the above procedures will impute synthetic values  $\tilde{y}_C$  for  $y_C$  such that  $\tilde{y}_C = y_C$ . This creates a high risk of re-identification of  $y_C$ .

Hence in our approach we add a positive probability that for a given category  $C$ , the  $\tilde{y}_C$  generated may contain values not present in  $y_C$ . We add the positive probability by putting positive prior probability in the prior. The “prior” is estimated by replacing one of the categorical predictors with a coarsened version, i.e., collapsing categories to increase the sample size and, likely, the variability. The prior counts are adjusted downward so that the larger sample size does not overwhelm the likelihood. This is a data-driven prior, and, in fact the sample used to compute a prior contains the data at hand; however, it serves the purpose of adding noise in a controlled fashion to serve the goals of reducing disclosure risks with minimal loss of utility. This approach is most easily implemented when there are no synthetic variables to condition on, i.e., for  $f(y_1|X)$ .

### 2.4 Simple multinomial method

When synthesizing  $(y_2|y_1, X)$  with categorical variables, an alternative to the generalized logit is a simple multinomial model. When the number of unique categories determined by the predictors becomes too numerous, the Dirichlet-multinomial approach just described can be cumbersome. In the simple multinomial approach, the cell probabilities are estimated from the observed data. For a given cell in  $(y_1, X)$ , the corresponding cell in the observed data is found and the corresponding cell probabilities used to sample from the multinomial distribution to obtain  $\tilde{y}_2$ . If an exact cell match is not found in the observed data, a possibility depending on the disclosure control applied to  $y_1$ , then the cell is collapsed until a match is found. Similarly, one can collapse cells until a threshold such as a maximum allowable cell probability is met to ensure that sufficient variability exists to provide disclosure limitation. This method can also be used to synthesize  $(y_k|y_1, \dots, y_{k-1}, X)$ .

This approach may provide insufficient disclosure control in some cases, for example, if this approach is used to predict  $(y_1|X)$ , and there are subgroups in  $X$  that have little or no variability in  $y_1$ . With sufficient variability in the observed data, disclosure control is provided by the sampling from the multinomial distribution, and by the disclosure control methods applied to any predictors. This approach is very fast computationally and produced good results when applied to the LBD. In other settings, one would need to evaluate whether this method provided adequate disclosure protection.

## 3. Imputation of the LBD

The development of the methods described in Section 2 was motivated in part by the desire to generate public use files for longitudinal establishment data from the U. S. Census Longitudinal Business Database. Currently controlled access to this data is only granted to researchers by special agreement with the U. S. Census Bureau. The version of the LBD that we are synthesizing is based on the most recent release, made available to authorized users by the Center for Economic Studies in May 2007. A small fraction of the data values are missing. These are imputed during the synthetic data generation. Some additional data cleaning is also performed on the observed data prior to generating the synthetic data.

The variables that we are working with are described in Table 1. County and industry codes are not synthesized but all other variables must be synthesized for the data to be considered for public release. Establishments started after 2001 will not be included in the synthetic public release files until methodology is developed to align the NAICS and SIC systems for industry coding. As this project is ongoing, the procedures described here may not correspond to any eventual public release files. We describe our current approach and provide a few summary tables to show some features of the observed data that are preserved in the synthetic versions.

Our strategy is to build up the joint distribution as described in Section 2. Here we describe our approach and present preliminary results for a subgroup defined by a 3-digit SIC, with approximately 130,000 establishments. We generate the synthetic datasets as follows:

1. Impute Firstyear using the Dirichlet-multinomial method to approximate draws from  $f(y_1|x_1, x_2)$ .
2. Impute Lastyear using the simple multinomial approach to approximate draws from  $f(y_2|y_1, x_1, x_2)$ .
3. Impute Multiunit status using the simple multinomial approach to draw from  $f(y_3|y_2, y_1, x_1, x_2)$ .

Table 1: Variable Descriptions

Variable	Name	Type	Description
$x1$	County	categorical	Geographic Location
$x2$	SIC	categorical	Industry Code
$y1$	Firstyear	categorical	First Year Establishment is Observed
$y2$	Lastyear	categorical	Last Year Establishment is Observed
$y3$	Multiunit	categorical	Multiunit Status
$y4$	Employment	continuous	March 12 Employment (26 years)
$y5$	Payroll	continuous	Annual Payroll (26 years)

4. Impute Employment and Payroll variables using the normal method, with transformations, to draw from  $f(y_4^{(t)} | y_4^{(t-1)}, y_5^{(t-1)}, y_3, y_2, y_1, x_1, x_2)$  and  $f(y_5^{(t)} | y_4^{(t)}, y_5^{(t-1)}, y_3, y_2, y_1, x_1, x_2)$ , where  $t$  indicates a year between 1976 and 2001.

Table 2: Observed and Synthetic Distributions of Multiunit Status

Value	Observed Percent	Synthetic Percent
1	.8408	.8386
2-4	.0096	0.0090
5	.1497	.1525

### 3.1 Firstyear

The variable Firstyear contains 27 categories, namely the years 1975 through 2001. We predict this conditional on 4-digit SIC and County. There are over 3000 counties in the United States and numerous SIC groups. This results a large number of unique county-SIC groups, that with 27 categories in the response, it is not always possible to use a generalized logit or similar model to predict the response. Furthermore, there are many county-SIC groups for which the observed unit(s) has only one observed value of Firstyear. Hence the multinomial approach of Section 2.3 was developed to handle this case. The marginal distribution is well-preserved using this approach as seen in Figure 1. Initial exploratory analysis suggests that conditional relationships are also preserved.

### 3.2 Lastyear

The variable Lastyear contains 30 categories, the years 1976 through 2005, representing the last year an establishment is observed. We predict Lastyear using a simple multinomial approach. Frequencies of Lastyear values are determined for each category determined by a combination of Firstyear and 4-digit SIC using the observed data. Dependencies on geographic variables are not accounted for. For logical consistency, the probability that the value of Lastyear for a given unit can be less than the imputed value of Firstyear is set to zero and the remaining cell probabilities are normalized. Figure 2 shows the close correspondence between the observed frequencies of the variable Lastyear and one synthetic implicate.

### 3.3 Multiunit status

The variable Multiunit indicates whether or not an establishment was ever part of a multi-unit firm, i.e., whether an establishment was ever part of a parent enterprise conducting business at multiple locations. In the real data, multiunit status is a longitudinal binary indicator of multiunit status for a given year. To facilitate synthesis, a categorical variable was defined such that a value of 1 indicates an establishment was never part of a multi-unit firm; values of 2-4 indicate a change in multi-unit status at some point in the lifetime of the establishment; and a value of 5 indicates the establishment was always part of a multi-unit firm. The synthesis of this categorical variable using the simple multinomial approach was straightforward. The predictors used include Firstyear, Life (Lastyear - Firstyear), 4-digit SIC, and State. For units that change their multiunit status over the course of their lifetime, the year when the change occurs is also of interest and is planned for synthesis at a later time. Firm structure and linkages between establishments in the same firm are not planned for synthesis. Table 2 shows the observed frequency of Multiunit compared to one synthetic implicate for one 4-digit SIC group.

### 3.4 Payroll and employment

Payroll and employment data are collected for each active establishment in every year between 1976 and 2005. If the synthetic values of Firstyear and Lastyear indicate an establishment was inactive in a given year, then no payroll or employment value is generated. For establishments



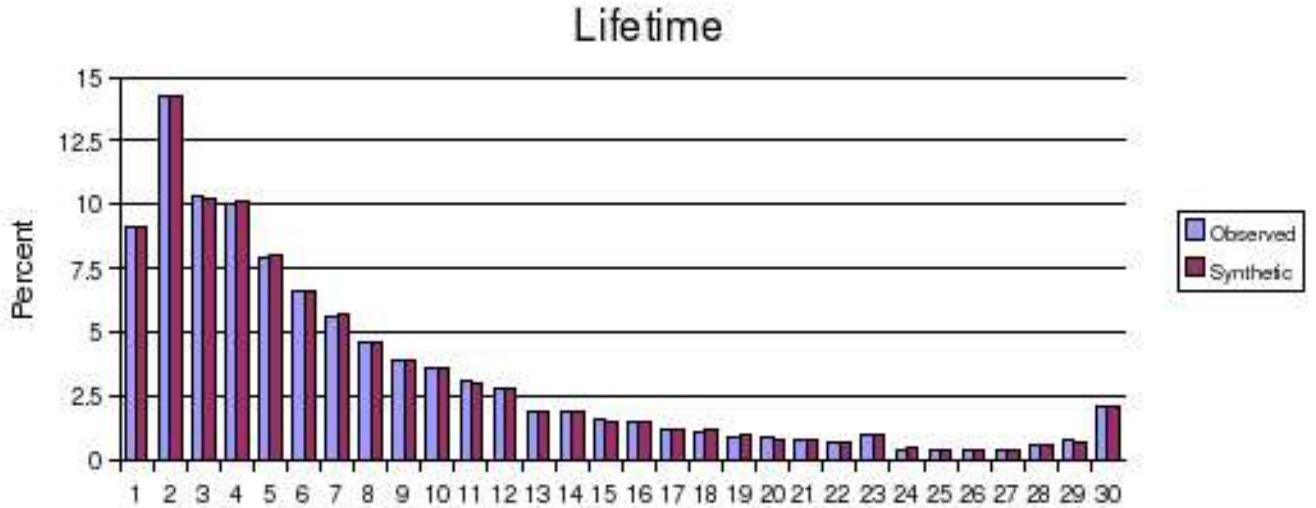


Figure 3: Observed and Synthetic Distributions of Lifetime

in their first year, first-year employment and payroll are predicted from observed data corresponding to units in their first year, using Multiunit, Life, and SIC as predictors. Payroll for continuers is predicted from the previous year’s payroll, previous year’s employment, as well as Multiunit, Life, SIC, and Age. Employment for continuers is predicted from previous year’s employment, current year’s payroll, and other predictors. For these variables, normal models are used, with transformations applied to the continuous variables to satisfy the normal linear model assumptions.

Given the highly skewed nature of payroll and employment data, a kernel density estimation procedure is used to transform the response variable so that marginally it has a standard normal distribution as in Abowd and Woodcock (2004). For each observed value  $y_k$ , the transformed values  $y_k^*$  are computed as  $\Phi^{-1}(\hat{k}(y_k))$ , where  $\Phi$  denotes the standard normal CDF and  $\hat{k}(y_k)$  is a kernel density estimate of  $y_k$ . The normal method is applied to obtain synthetic  $\tilde{y}_k^*$ , and then an inverse transformation is applied to obtain  $\tilde{y}_k = \hat{k}^{-1}(\Phi(\tilde{y}_k^*))$ . Log transforms are applied to the continuous predictors. Applying the KDE transform to the predictors can affect the preservation of correlations between the back-transformed variables in the synthetic data.

#### 4. Discussion

This paper has described ongoing work to develop synthetic data for longitudinal establishment data from the U. S. Census Bureau’s Longitudinal Business Database for public release. The methodology is flexible and can be

adapted for other datasets. Preliminary results illustrate the feasibility of using synthetic data for releasing microdata for public use while protecting confidentiality and allowing valid inferences to be made. We have not addressed the risks associated with releasing partially synthetic data in this paper. While preliminary results are promising, once the synthesis of the data has been completed further analysis of utility and risk will be completed before the data are released. Further evaluation of the randomization-validity of the proposed imputation methods is also needed.

#### Acknowledgments

The research in this paper was conducted while the authors were Special Sworn Status researchers of the U.S. Census Bureau at the Triangle Census Research Data Center. Research results and conclusions expressed are those of the authors and do not necessarily reflect the views of the Census Bureau. This paper has been screened to insure that no confidential data are revealed. This work was supported by NSF grant ITR-0427889. The authors are also grateful for the assistance of many project participants and Census Bureau staff.

#### References

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds., *Confidentiality, Disclosure, and Data Access: Theory and Practical*

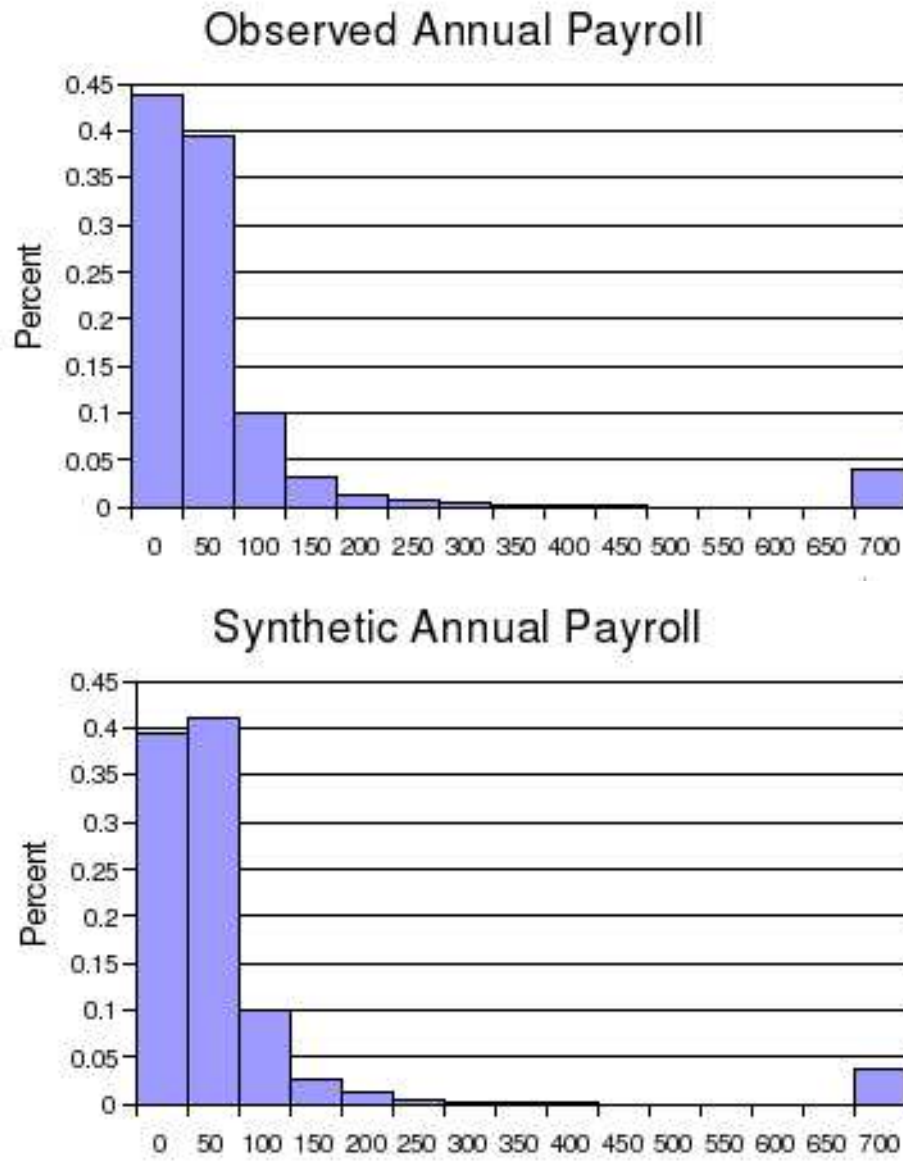


Figure 4: Observed and Synthetic Distributions of Annual Payroll (in \$1000)

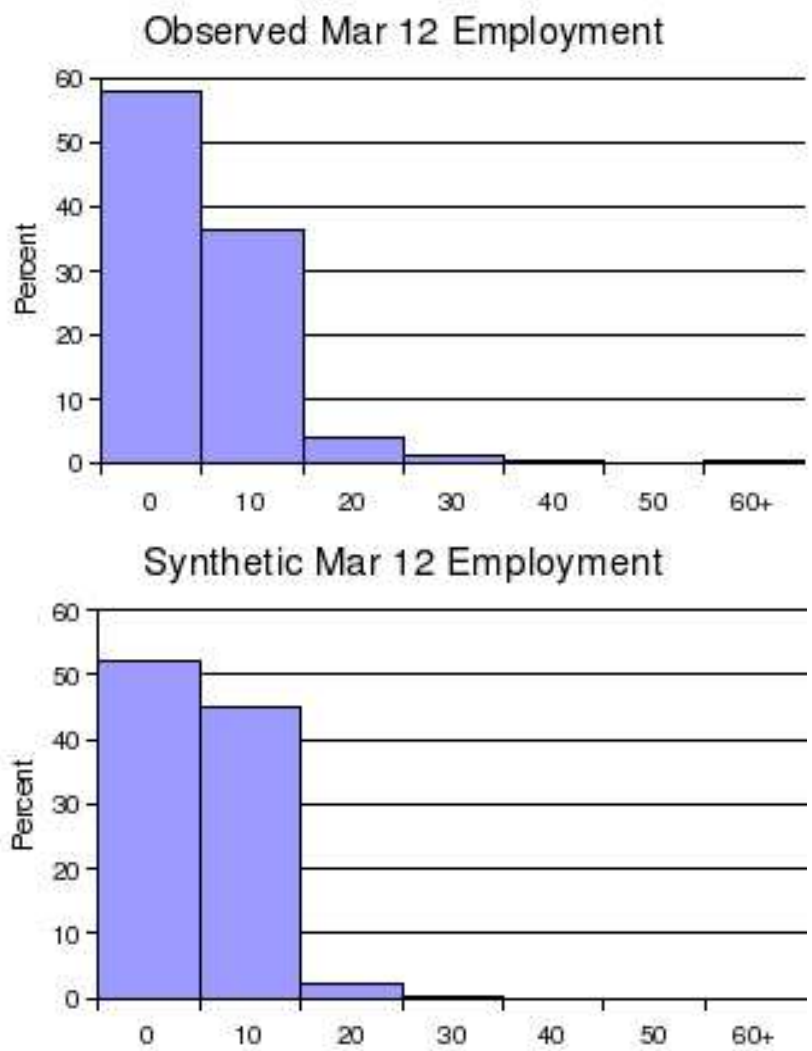


Figure 5: Observed and Synthetic Distributions of March 12 Employment

- Applications for Statistical Agencies*, 215–277. Amsterdam: North-Holland.
- Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*. New York: Springer-Verlag.
- Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference* **6**, 73–85.
- Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9**, 383–406.
- Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, eds., *Record Linkage Techniques, 1997*, 248–267. National Academy Press, Washington, D.C.
- Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* **9**, 407–426.
- Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**, 1–16.
- Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531–544.
- Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**, 181–189.
- Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235–242.
- Reiter, J. P. (2005). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* **131**, 365–377.
- Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.