An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets

Jörg Drechsler*

Institute for Employment Research, Department for Statistical Methods, Regensburger Straße 104, 90478 Nürnberg, Germany.

Jerome P. Reiter

Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708-0251.

Abstract

When intense redaction is needed to protect the confidentiality of data subjects' identities and sensitive attributes, statistical agencies can use synthetic data approaches. To create synthetic data, the agency replaces identifying or sensitive values with draws from statistical models estimated from the confidential data. Many data producers are reluctant to implement this idea because (i) the quality of the generated data depends strongly on the quality of the underlying models, and (ii) developing effective synthesis models can be a labor intensive and difficult task. Recently there have been suggetions that agencies use nonparametric methods from the machine learning literature to generate synthetic data. These methods can detect non-linear relationships that might otherwise be missed and run with minimal tuning, thus considerably reducing burdens on the agency. Four synthesizers based on machine learning algorithms—classification and regression trees, bagging, random forests, and support vector machines—are evaluated in terms of their potential to preserve analytical validity while reducing disclosure risks. The procedures are run with minimal tuning in a repeated sampling simulation

*Corresponding author. Tel: (+49)911-179-4021; fax: (+49)911-179-1728

Email addresses: joerg.drechsler@iab.de (Jörg Drechsler), jerry@stat.duke.edu (Jerome P. Reiter)

based on a subset of the 2002 Uganda census public use sample data. The simulation suggests that synthesizers based on regression trees can result in synthetic dataset that provide reliable estimates and low disclosure risks, and that these synthesizers can be implemented easily by statistical agencies.

Keywords: Census, Confidentiality, Disclosure, Imputation, Microdata, Synthetic

---

1. Introduction

One of the primary missions of most national statistical agencies is to disseminate data to the public. Wide access to data has great benefits, leading to advances in research, improvements in policy making, opportunities for students to learn data analysis skills, and resources for individuals to better understand and participate in their society. Additionally, citizens who pay for data collection via taxes arguably should have a right to have access to that data.

However, government agencies are under increasing pressure to limit access to data because of growing threats to data confidentiality. Even stripping obvious identifiers like names, addresses, and identification numbers may not be sufficient to protect confidentiality. Ill-intentioned data users, henceforth called intruders, may be able to link records in released data to records in other files by matching on common key variables, such as demographic variables when data subjects are individuals or employee size when data subjects are business establishments. For example, Sweeney (1997) showed that 97% of the records in publicly available voter registration lists in Cambridge, Massachusetts, could be uniquely identified using birth date and nine digit zip code. By matching on the information in these lists, she was able to identify the governor of Massachusetts in a supposedly anonymized medical database.

To protect confidentiality in public use datasets, many statistical agencies release data that have been altered to protect confidentiality. Common strategies include aggregating geography, top-coding variables, swapping data values across records, and adding random noise to values (Willenborg and de Waal, 2001). As the threats to confidentiality grow, these techniques may have to be applied with high intensity to ensure adequate protection. However, applying these methods with high intensity can have serious consequences for secondary statistical analysis. For example, aggregation of geography to high levels disables small area estimation and hides spatial variation;

top-coding eliminates learning about tails of distributions—which are often most interesting—and degrades analyses reliant on entire distributions (Kennickell and Lane, 2006); swapping at high rates destroys correlations among swapped and not swapped variables (Winkler, 2007); and, adding random noise introduces measurement error that distorts distributions and attenuates correlations (Fuller, 1993). In fact, Elliott and Purdam (2007) use the public use files from the UK census to show empirically that the quality of statistical analyses can be degraded even when using recoding, swapping, or stochastic perturbation at modest intensity levels. These problems would only get worse with high intensity applications.

Motivated by the shortcomings of standard disclosure limitation at high intensity, Rubin (1993) and Little (1993) suggested that agencies release partially synthetic data, which comprise the original units surveyed with some collected values replaced with multiple imputations. The imputations are drawn from distributions designed to preserve important relationships in the confidential data. With partially synthetic data, analysts can obtain frequency-valid inferences for wide classes of estimands by combining standard likelihood-based or survey-weighted estimates with simple formulas; the analyst need not learn new statistical methods or software to adjust for the effects of the disclosure limitation. This is true even for high fractions of replacement, whereas swapping high percentages of values or adding noise with large variance produces worthless data. The released data can include simulated values in the tails of distributions (no top-coding) and avoid category collapsing. Finally, because many quasi-identifiers can be simulated, finer details of geography can be released.

Several national statistical agencies have started to release partially synthetic data products. For example, the U.S. Federal Reserve Board in the Survey of Consumer Finances replaces monetary values at high disclosure risk with multiple imputations, releasing a mixture of these imputed values and the unreplaced, collected values (Kennickell, 1997). The U.S. Bureau of the Census has released a partially synthetic, public use file for the Survey of Income and Program Participation that includes imputed values of Social Security benefits information and dozens of other highly sensitive variables (Abowd et al., 2006). The Census Bureau protects the identities of people in group quarters (e.g., prisons, shelters) in the American Community Survey by replacing quasi-identifiers for records at high disclosure risk with imputations (Hawala, 2008). The Census Bureau also has developed synthesized origin-destination matrices, i.e. where people live and work, available to the

public as maps via the web (On The Map, http://lehdmap.did.census.gov/). In the U.S., partially synthetic, public use datasets are in the development stage for the Longitudinal Business Database (Kinney and Reiter, 2007), the Longitudinal Employer-Household Dynamics database, and the American Community Survey full sample data. Statistical agencies in Germany (Drechsler et al., 2008b,a; Drechsler, 2011) and New Zealand (Graham and Penny, 2005) also are developing synthetic data products. Other examples of synthetic data are in Abowd and Woodcock (2001, 2004), Reiter (2005b), Little et al. (2004), and Woodcock and Benedetto (2009).

The key to the success of synthetic data approaches, especially when replacing many values, is the data generation model. Current practice for generating synthetic data typically employs sequential modeling strategies based on parametric or semi-parametric models similar to those for imputation of missing data in Raghunathan et al. (2001). The basic idea is to impute Y_1 from a regression of Y_1 on $(Y_2, Y_3, etc.)$, impute Y_2 from a regression of Y_2 on $(Y_1, Y_3, etc.)$, impute Y_3 from a regression of Y_3 on $(Y_1, Y_2, etc.)$, and so on. Specifying these conditional imputation models can be daunting in surveys with many variables. Many datasets include numerical, categorical, and mixed variables, some of which may not be easy to model with standard tools. The relationships among these variables may be non-linear and interactive. Finally, specifying models for many variables is a resource-intensive task, and many statistical agencies simply do not have the time to invest in careful specification of these conditional models for many variables.

Given these issues, it can be advantageous for agencies to adapt nonparametric regression methods to generate synthetic data. These approaches can handle diverse data types in high dimensions. They can capture non-linear relationships and interaction effects that may not be easily revealed in the process of fitting standard models. Finally, they can be implemented with comparatively far less tuning, and hence much faster, than approaches in current practice by statistical agencies planning timely data releases. The use of nonparametric methods has also been recently suggested in the context of missing data imputation (Iacus and Porro, 2007)

In this article, we empirically evaluate and compare four synthesizers based on nonparametric regression algorithms from the machine learning literature, namely classification and regression trees (Breiman et al., 1984), bagging (Breiman, 1996), random forests (Breiman, 2001), and support vector machines (Boser et al., 1992). We do so by means of a repeated sampling simulation using a subset of data from the 2002 Uganda census public use

files. The results suggest a clear risk-utility tradeoff among the procedures, with regression trees and support vector machines at the relatively high end of data utility and disclosure risk, and random forests and bagging at the relatively low end of data utility and disclosure risk. The results also suggest that synthesizers based on regression trees are a particularly attractive option for statistical agencies seeking to release datasets with intense synthesis without intense labor.

The remainder of the article is organized as follows. In Section 2, we briefly review partially synthetic data. In Section 3, we describe the four nonparametric data synthesizers. In Section 4, we present results of the empirical evaluation. In Section 5, we conclude with a summary of our findings.

2. Review of partially synthetic data

To provide context for the empirical evaluations of the nonparametric synthesizers, we briefly review methods for assessing disclosure risks and methods for obtaining inferences from partially synthetic datasets. See Drechsler and Reiter (2008) and Reiter and Mitra (2009) for further details on the former, and see Reiter (2003) and Reiter and Raghunathan (2007) for further details on the latter.

Let Y_{rep} be the values of the collected data that are replaced with synthetic values, and let Y_{nrep} be the values that remain unchanged. Let $D^{(l)} = (Y_{rep}^{(l)}, Y_{nrep})$, where $Y_{rep}^{(l)}$ is a set of synthetic values of Y_{rep}. The agency releases $D = \{D^{(1)}, \ldots, D^{(m)}\}$, i.e., a set of m partially synthetic datasets each with independently simulated $Y_{rep}^{(l)}$.

2.1. Identification disclosure risk measures

Suppose that the intruder has a vector of information, $\mathbf{t}$, on a particular target unit in the population. The target may not be in D. Let t_0 be the unique identifier (e.g., person or establishment name) of the target, and let d_{j0} be the (not released) unique identifier for record j in D, where $j = 1, \ldots, n$.

The intruder seeks to match unit j in D to the target when $d_{j0} = t_0$, and not to match when $d_{j0} \neq t_0$ for any $j \in D$. Let J be a random variable that equals j when $d_{j0} = t_0$ for $j \in D$ and equals $n + 1$ when $d_{j0} = t_0$ for some $j \notin D$. The intruder thus seeks to calculate the $Pr(J = j|\mathbf{t}, D)$ for $j = 1, \ldots, n + 1$. She then would decide whether or not any of the identification probabilities for $j = 1, \ldots, n$ are large enough to declare an

identification. Because the intruder does not know Y_{rep}, for each record in D she computes

$$Pr(J = j|\mathbf{t}, D) = \int Pr(J = j|\mathbf{t}, D, Y_{rep})Pr(Y_{rep}|\mathbf{t}, D)dY_{rep}. \quad (1)$$

Reiter and Mitra (2009) discuss estimation of (1) for different degrees of knowledge possessed by the intruder, for example knowledge about the conditional distribution of Y_{rep}. For purposes of comparing the risks of the nonparametric synthesizers, we assume that the intruder approximates (1) by treating the simulated values in the released datasets as plausible draws of Y_{rep}. This also represents what intruders might do absent strong beliefs about the conditional distribution of Y_{rep}. The matching probability for any record j is then $Pr(J = j|\mathbf{t}, D) = (1/m)\sum_l (1/F_{\mathbf{t}})(D_j^{(l)} = \mathbf{t})$, i.e., one over the total number of like-valued units in the population. Here, the logical expression $(D_j^{(l)} = \mathbf{t})$ equals one when the values of variables used for matching by intruders—which are specified by the agency using its best judgments—for record j are in accord with the corresponding values in $\mathbf{t}$; the expression equals zero otherwise. Also, $F_{\mathbf{t}}$ is the number of records in the population that satisfy the matching criteria. Using $F_{\mathbf{t}}$ instead of $N_{\mathbf{t}}^{(l)}$—the number of records in $D^{(l)}$ that satisfy the matching criteria—accounts for the fact that the original data comprise only a sample from the population, so that the intruder generally may not know if the target is included in D. If the intruder knows who is in D, $F_{\mathbf{t}}$ is replaced with $N_{\mathbf{t}}^{(l)}$ (Reiter and Mitra, 2009). When the agency does not know $F_{\mathbf{t}}$, it can be estimated using a log-linear modeling approach (Skinner and Shlomo, 2008; Drechsler and Reiter, 2008). For some target records, $N_{\mathbf{t}}^{(l)}$ might exceed $F_{\mathbf{t}}$. For such cases, we presume that the intruder sets $Pr(J = n + 1|\mathbf{t}, \mathbf{D}) = 0$ and picks one of the matching records at random.

Following Reiter (2005a), we quantify disclosure risk with summaries of these identification probabilities. It is reasonable to assume that the intruder selects as a match for $\mathbf{t}$ the record j with the highest value of $Pr(J = j|\mathbf{t}, D)$, if a unique maximum exists. Furthermore, we assume that the intruder will never declare any record in the dataset to be a match if $Pr(J = n+1|\mathbf{t}, D) > Pr(J = j|\mathbf{t}, D)$ for $j = 1, ..., n$. Let c_j be the number of records in the dataset with the highest match probability for the target t_j for $j = 1, ..., n$; let $I_j = 1$ if the true match is among the c_j units and $I_j = 0$ otherwise. Let $K_j = 1$ when $c_j I_j = 1$ and $K_j = 0$ otherwise. The *true match rate* equals $\sum_j K_j/n$.

Finally, let $G_j = 1$ when $c_j(1-I_j) = 1$ and $G_j = 0$ otherwise; and, let s equal the number of records with $c_j = 1$. The *false match rate* equals $\sum G_j/s$.

2.2. Inferences from partial synthesis

Let Q be the secondary analyst's estimand of interest, such as a regression coefficient or population average. For $l = 1, \ldots, m$, let q_l and u_l be respectively the estimate of Q and the estimate of the variance of q_l in synthetic dataset $D^{(l)}$. Here, each q_l and u_l are computed acting as if each $D_{(l)}$ was collected with the original sampling design. See Mitra and Reiter (2006) for a discussion of how agencies might alter survey weights for design-based analysis. Secondary analysts use $\bar{q}_m = \sum_{l=1}^{m} q_l/m$ to estimate Q and $T_m = \bar{u}_m + b_m/m$ to estimate $var(\bar{q}_m)$, where $b_m = \sum_{l=1}^{m}(q_l - \bar{q}_m)^2/(m-1)$ and $\bar{u}_m = \sum_{l=1}^{m} u_l/m$. For large samples, inferences for Q are obtained from the t-distribution, $(\bar{q}_m - Q) \sim t_{\nu_m}(0, T_m)$, where the degrees of freedom $\nu_m = (m-1)\left[1 + m\bar{u}_m/b_m\right]^2$. Reiter (2005c) describes methods for multivariate significance tests.

3. Algorithmic data synthesizers

To describe the synthesizer for each algorithmic method, we first presume that the agency seeks to replace values of only one variable, Y_i, given values of all other variables, Y_{-i}. We extend to multiple variables in Section 3.4. For $j = 1, \ldots, n$, let $Z_j = 1$ when record j has its value of Y_i replaced, and let $Z_j = 0$ otherwise. Let $Z = (Z_1, \ldots, Z_n)$.

3.1. CART

Classification and regression trees (CART) seek to approximate the conditional distribution of a univariate outcome from multiple predictors. The CART algorithm partitions the predictor space so that subsets of units formed by the partitions have relatively homogeneous outcomes. The partitions are found by recursive binary splits of the predictors. The series of splits can be effectively represented by a tree structure, with leaves corresponding to the subsets of units. The values in each leaf represent the conditional distribution of the outcome for units in the data with predictors that satisfy the partitioning criteria that define the leaf.

CART has been adapted for generating partially synthetic data (Reiter, 2005d). First, using only records with $Z_j = 1$, the agency fits the tree of Y_i on Y_{-i} so that each leaf contains at least k records; call this tree $\mathcal{Y}^{(i)}$.

In general, we have found that using $k = 5$, which is a default specification in many applications of CART, provides sufficient accuracy and reasonably fast running time. For categorical variables, we grow $\mathcal{Y}^{(i)}$ by finding the splits that successively minimize the Gini index; for numerical variables, we successively minimize the deviance of Y_i in the leaves. We cease splitting any particular leaf when the deviance in that leaf is less than some agency-specified threshold d—which we vary in the empirical evaluations—or when we cannot ensure at least k records in each child leaf. We use only records with $Z_j = 1$ to ensure that the tree is tailored to the data that will be replaced.

For any record with $Z_j = 1$, we trace down the branches of $\mathcal{Y}^{(i)}$ until we find that record's terminal leaf. Let L_w be the wth terminal leaf in $\mathcal{Y}^{(i)}$, and let $Y_{L_w}^{(i)}$ be the n_{L_w} values of Y_i in leaf L_w. For all records whose terminal leaf is L_w, we generate replacement values of Y_{ij} by drawing from $Y_{L_w}^{(i)}$ using the Bayesian bootstrap (Rubin, 1981). Repeating the Bayesian bootstrap for each leaf of $\mathcal{Y}^{(i)}$ results in the lth set of synthetic values. We repeat this process m times to generate m datasets with synthetic values of Y_i.

Reiter (2005d) describes two further steps that agencies can take to protect confidentiality. First, the agency can prune $\mathcal{Y}^{(i)}$ to satisfy confidentiality criteria, e.g., values in leaves must be sufficiently diverse, before using the Bayesian bootstrap. Second, if it is desired to avoid releasing genuine values from some leaf, as may be the case for sensitive numerical data, the agency can approximate a smooth density to the bootstrapped values using a Gaussian kernel density estimator with support over the smallest to the largest value of the outcome in the leaf. Then, for each unit, the agency samples randomly from the estimated density in that unit's leaf using an inverse-cdf method.

3.2. Random forests and bagging

As described in Caiola and Reiter (2010), random forests are collections of CARTs, e.g., 500 or more trees grown on the same data. Each tree is based on a different random subset of the original data; usually, the subsets include around 2/3 of the full sample. Each branch of the tree is grown using a randomly selected subset of the predictor variables to determine the binary splits; usually, the subset includes roughly $\sqrt{p}$ variables, where p is the total number of predictors. Typically each tree is grown to the maximum size, so that each terminal leaf contains one observation. There is no pruning of

leaves for random forests, although it is possible to force leaves to contain more than one observation to speed up the algorithm.

To generate synthetic data, we fit a random forest of Y_i on Y_{-i} using only those records with $Z_j = 1$. For any record j with values of predictors $Y_{-i,j}$, we run it down each tree in the forest to obtain a predicted value of Y_i. That is, we follow the sequence of partitioning for record j until we reach the terminal leaf in the tree. For categorical Y_i, we tabulate the predictions for Y_{ij} to form the data for a multinomial distribution. For example, if the forest generates 500 predictions for a particular Y_{ij} such that 300 predict a race of white, 100 predict a race of black, 75 predict a race of Asian, and 25 predict a race of American Indian, we form a multinomial distribution with $p(white) = .6$, $p(black) = .2$, $p(Asian) = .15$, and $p(Amer.Ind.) = .05$. To generate the synthetic Y_{ij}, we randomly sample one value from the implied multinomial distribution. For continuous data, we randomly sample one value from the predictions, possibly after approximating with a smooth density estimate.

For any tree that includes the jth record in the training sample, the terminal leaf associated with the jth record will contain Y_{ij}, since the tree is fit with actual values of Y_{-i}. Hence, there will be a large probability of sampling the actual Y_{ij}, which might not lead to adequate protection. One alternative approach, called out of bag prediction, is to use the predictions only for trees where j does not appear in the training sample. We investigate this approach in the empirical evaluations. We note that this is not problematic when synthetic values of Y_{-i} are used to run leaves down the trees, as may be the case in multivariate synthesis.

In the empirical evaluations, we also investigate bagging as a data synthesizer. Bagging is a predecessor of random forests. It grows a large number of trees based on different subsets of the original data. The decisions for the splits in each tree are always based on all predictors rather than a random sample of predictors. In that sense, bagging is like running random forests without any sampling of predictors in the trees.

3.3. Support vector machines

Support vector machines (SVM) are used to predict the outcome of some categorical variable Y_i from some set of predictors Y_{-i}. The algorithm finds hyperplanes from Y_{-i} that separate the different classes of Y_i to satisfy some optimality criterion, e.g., find the hyperplanes resulting in the largest gaps among the separated classes. To find the hyperplanes, it is often beneficial to map Y_{-i} into a higher dimensional space using kernel functions.

Support vector machines depend on a tuning parameter C, which controls the number of misclassifications allowed by the supporting hyperplanes; for example, see Moguerza and Muñoz (2006). The tuning is performed by cross-validation, in which the algorithm repeatedly splits the data into two random subsets, a training and test dataset. The SVM is run on the training set with different values of C, and each time the performance of the algorithm is evaluated by measuring how well it predicts the known classes of Y_i in the test dataset. The level of C that provides the best results on the test data across all cross validations is used when the SVM is applied to new data for which Y_i is not observed.

Support vector machines also can be tuned for continuous variables, which is called support vector regression; see Smola and Schölkopf (1998) for a review. The approach is comparable to the SVM for classification. However, the optimization is based on a different loss function, and the mean squared error of predictions is used as an evaluation criterion for tuning. Support Vector Machines are especially useful if a very large number of potential classifiers is available for example if the classification is based on gene expression data (Choi et al., 2011; Shim et al., 2009).

Support vector machines and support vector regression can be used to generate synthetic data (Drechsler, 2010). This involves several adaptations to the standard implementations of these algorithms. First, we need to draw values from the conditional distribution $p(Y_i|Y_{-i})$ rather than simply classify or predict Y_i. For categorical Y_i, the loss function suggested by Wu et al. (2004) results in an approximation of $p(Y_i|Y_{-i})$; see Drechsler (2010) for details. Using this distribution, we compute the vector of probabilities for each record j by running its Y_{-ij} through the support vector machine. We then sample randomly from a multinomial distribution with those probabilities. For continuous variables, we assume $p(Y_i|Y_{-i})$ follows a Laplace distribution; thus, we add zero-mean Laplace noise to the predicted values. We obtain the scale of the Laplace distribution from the variance of the prediction errors in the test datasets used in the tuning stage. As with CART and random forests, we find the support vectors for Y_i on Y_{-i} using only those records with $Z_j = 1$.

Second, in the tuning to select C, we train the support vector machine on a subset of the data as in the standard implementation, but we evaluate the performance on the complete dataset rather than only on the remaining test data. In partially synthetic data settings, we are not concerned with overfitting, since the goal is to preserve the features in the original data as

closely as possible (without violating confidentiality).

Third, for categorical variables we adopt an alternative evaluation criterion for tuning suggested in Drechsler (2010) that basically searches for the solution that places the highest confidence in the predicted classification. This differs from the usual performance evaluation for categorical variables that is based on how often the SVM predicts the correct class for Y_i in the test dataset.

3.4. Multivariate versions of each synthesizer

In this empirical evaluation, we consider the case of synthesizing all values of r variables identified as sensitive by the agency. Let $Y_{(0)}$ be all variables with no values replaced. In such cases, for an arbitrary ordering of the variables and any of the synthesizers considered here, the agencies can proceed as follows. Let $Y_{(i)}$ represent the ith variable in the synthesis order.

1. Run the algorithm to regress $Y_{(1)}$ on $Y_{(0)}$ only. Replace $Y_{(1)}$ by synthetic values using the corresponding synthesizer for $Y_{(1)}$. Let $Y_{(1)rep}$ be the replaced values of $Y_{(1)}$.
2. Run the algorithm to regress $Y_{(2)}$ on $(Y_{(0)}, Y_{(1)})$ only. Replace $Y_{(2)}$ with synthetic values using the corresponding synthesizer for $Y_{(2)}$. Use the values of $Y_{(1)rep}$ and $Y_{(0)}$ for predicting new values for $Y_{(2)}$. Let $Y_{(2)rep}$ be the replaced values of $Y_{(2)}$.
3. For each i where $i = 3, \ldots, r$, run the algorithm to regress $Y_{(i)}$ on $(Y_{(0)}, Y_{(1)}, \ldots, Y_{(i-1)})$. Replace each $Y_{(i)}$ using the appropriate synthesizer based on the values in $(Y_0, Y_{(1)rep}, Y_{(2)rep}, \ldots, Y_{(i-1)rep})$.

The result is one synthetic dataset. These three steps are repeated for each of the m synthetic datasets, and these datasets are released to the public.

When replacing only parts of variables rather than all values, the process is adapted by running the algorithms at each step using only records with $Z_{ij} = 1$ for that Y_i, and by conditioning on $Y_{(-i)}$, i.e., all variables except Y_i, when fitting the algorithms.

As noted in Caiola and Reiter (2010), there is no mathematical theory underpinning the ordering of variables for synthesis. Different orderings could produce different risk and utility profiles. One approach is to order the variables by decreasing amount of synthesis. This bases the largest number of synthetic imputations on the most genuine predictor values. This should afford the highest data quality for the variable with the most synthesis. Alternatively, one could order the variables by increasing amount of synthesis,

which could result in lower disclosure risks since the protection from synthesizing propagates down the chain.

When two or more variables have the same amount of synthesis, as happens when entire variables are synthesized, one approach is to select orderings to ease computation; for example, impute categorical variables with small numbers of categories early in the sequence and those with large numbers of categories later in the sequence. Saving the variables with many levels until the end can speed up computation for tree-based methods, since splitting a categorical variable with many levels is time consuming. Another approach is to experiment with several orderings to determine which produces datasets with the most desirable risk-utility profile. When practical, this is the optimal approach.

4. Empirical evaluation

It is challenging to evaluate the relative merits of these nonparametric synthesizers from analytical perspectives. Instead, we compare them using simulation studies based on genuine data. Specifically, we identify a subset of public use census microdata to treat as a population, and repeatedly take random samples from it. For each sample, we generate partially synthetic datasets from each nonparametric synthesizer, and we use the inferential methods in Section 2.2 to compute point estimates and 95% confidence intervals for 162 estimands spanning representative analyses, including one linear and two logistic regressions with non-linear predictor functions and a variety of marginal and conditional population percentages. These estimands are described in the appendix. Using the repeated samples, we compare the empirical biases in the point estimates and the empirical coverage rates of the intervals for the synthesizers. We also compute disclosure risks for ten of these replications using the methods in Section 2.1. In this way, we can assess the risk-utility tradeoff for the different methods.

We do not claim that the synthesis strategy used in the empirical evaluations is ideal for creating public use files for data with these characteristics. Further protection may be required, and better analytic validity can be obtained by additional tuning. Nonetheless, these simulations provide measurements of the relative effectiveness of the different nonparametric synthesizers using genuine data.

4.1. Simulation design

The empirical evaluations are based on the public use microdata sample from the 2002 Uganda Population and Housing Census provided by IPUMS International (Minnesota Population Center, 2010). The public use microdata file is a 10% systematic sample of the population living in Uganda, comprising 2,497,449 questionnaire records (households and institutions). The file contains more than 100 variables at the household and personal level for each respondent. Parts of the questionnaire are not answered by all respondents, e.g., only females aged 12 to 54 who ever had a child are asked questions about children. To avoid dealing with such skip patterns in the repeated sampling study, we focus only on male heads of households, and we drop the variables on migration that are answered only by persons who migrated previously. The final dataset comprises 394,307 male respondents and 54 variables, including detailed information on the living conditions, demographics, education, employment status, and more.

We treat these 394,307 records as the population, and we repeatedly draw 1% simple random samples from it. We consider these samples as the original data from which synthetic datasets should be generated for public release. We synthesize the following variables: number of persons in the household, which ranges from 1 to 30; age, which ranges from 10 to 95; marital status, which has five categories; literacy, which has two categories; and employment status, which has three categories. These variables represent a mix of nominal and numerical data that are used for both descriptive statistics and analytical models. We generate synthetic data by replacing all records' values for these variables; all other variables are left unchanged. In practice, it is not always necessary to synthesize entire values. It can be sufficient to synthesize values only for records deemed at high risk of re-identification; see Drechsler and Reiter (2010) for further discussion. We chose to synthesize all values rather than a fraction as a more stringent evaluation of the analytic properties of the nonparametric synthesizers.

For each drawn observed dataset, we synthesize the five variables using the steps of Section 3.4 for each synthesizer in Section 3. We synthesize in the order: persons, age, marital status, literacy, and employment status. We did not consider other orderings for computational expediency. In actuality, there has been little empirical research on the impacts of synthesis order on data usefulness and disclosure risks; see Reiter (2005d) and Caiola and Reiter (2010) for further discussions of this issue. We stratify each sample in four geographic regions, and run separate synthesizers in each region. This

helps to preserve variation in relationships across these geographies, which improves the quality of the synthetic data. For each synthesizer, we generate $m = 5$ partially synthetic datasets.

For the CART synthesizer, we use two different parameter settings for the minimum deviance in each node, namely $d = .01$ and $d = .0000001$. For the former value, the CART algorithm stops splitting a branch of the tree when the deviance of Y_i in the leaf under consideration is less than 1% of the deviance of all values of Y_i in the sample. Thus, using $d = .01$ tends to grow comparatively small trees compared to using $d = .0000001$.

For the random forest synthesizer, we create 500 trees such that each terminal leaf contains only one value of Y_i. We use the standard defaults for the tuning parameters in random forests: random samples of roughly (2/3)3943 records and random selection of roughly $\sqrt{54}$ predictors. We use the Gini index (for categorical Y_i) and deviance (for numerical Y_i) as criteria to determine the binary splits, which is a default criteria for many applications of random forests. For bagging, we use the same tuning parameters without sampling predictors. We use both the in-bag and out-of-bag methods in Section 3.2 for selecting trees from the forests.

For the support vector machine synthesizer, we use the fitting method described by Hsu et al. (2010), which uses a radial basis function kernel indexed by two tuning parameters, γ and C. Because tuning support vector machines is time consuming, an exhaustive search to find optimal values of these tuning parameters for each variable in each replication is computationally prohibitive in a repeated sampling evaluation. Instead, we tuned the SVMs on ten independent samples of observed data using ten-fold cross validation, each time searching over an exhaustive grid of possible tuning parameter values for all five variables. In the repeated sampling simulation, we limit the SVM tunings to searches over the much smaller spaces of parameter combinations identified as near optimal in the ten independent samples.

4.2. Analytical validity

To evaluate the quality of the data generated with the different approaches, we repeat the process of sampling from the population and generating synthetic datasets 1,000 times. Figure 1 displays scatter plots of the population values versus the simulated expected values of the 162 point estimates for the original sample and for each synthesizer. Not surprisingly, the best results are obtained with the original unaltered data, but the estimates for the CART synthesizer allowing for large trees (CART BIG) and

the support vector machine synthesizer (SVM) also are close to the population values for most of the estimates. The point estimate for the CART synthesizer based on smaller trees (CART SML) are somewhat less accurate, suggesting that some important relationships in the data are missed because the growing of the trees stops too early. The point estimates for the bagging and for the random forest approaches are very similar, as are the point estimates when using all trees (BAG and RF) or only the out-of-bag trees (BAG.ooB and RF.ooB) for the predictions.

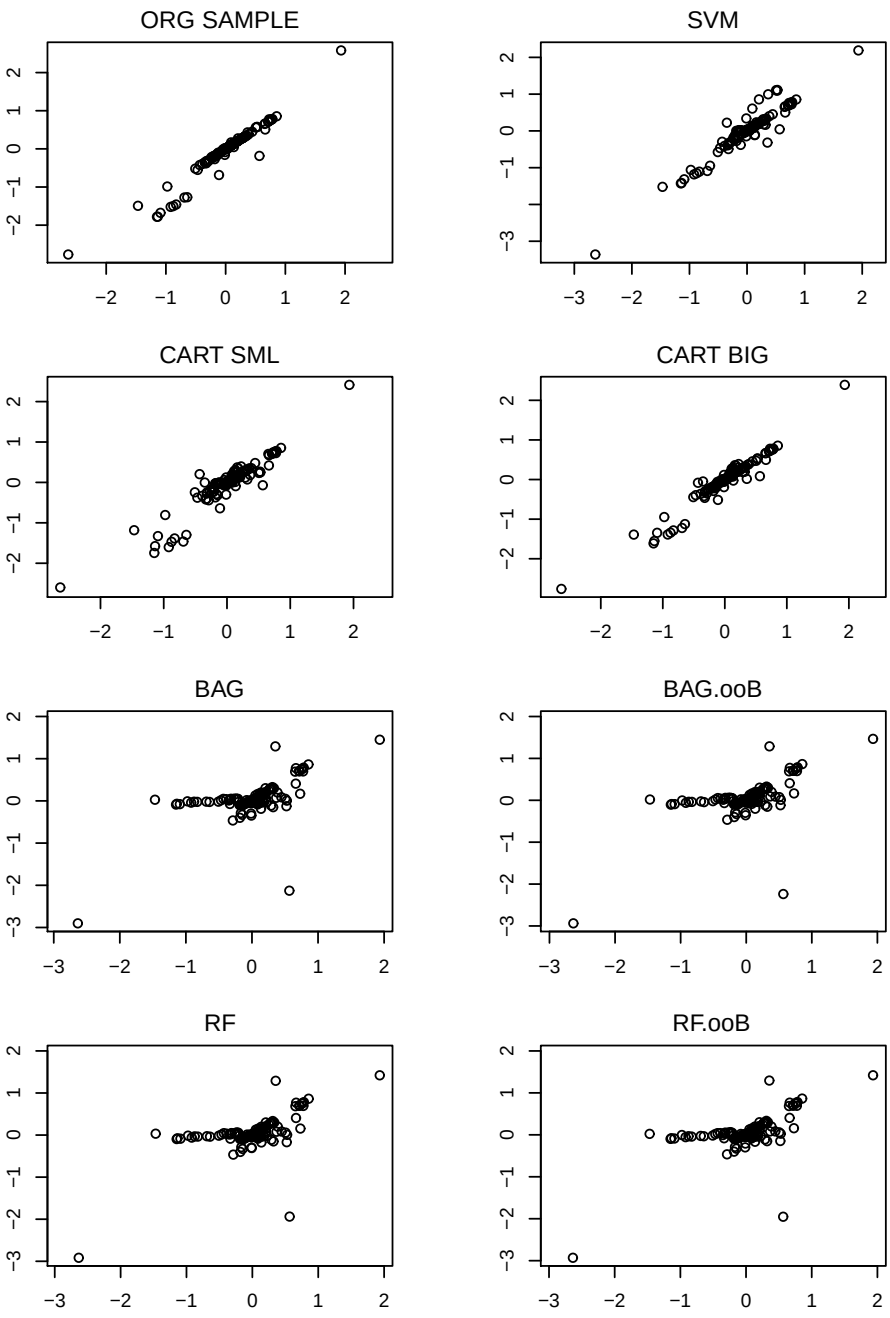

Figure 1: Comparison of the true population quantities with the average of the point estimates across the 1,000 simulation runs for the original and synthetic datasets.

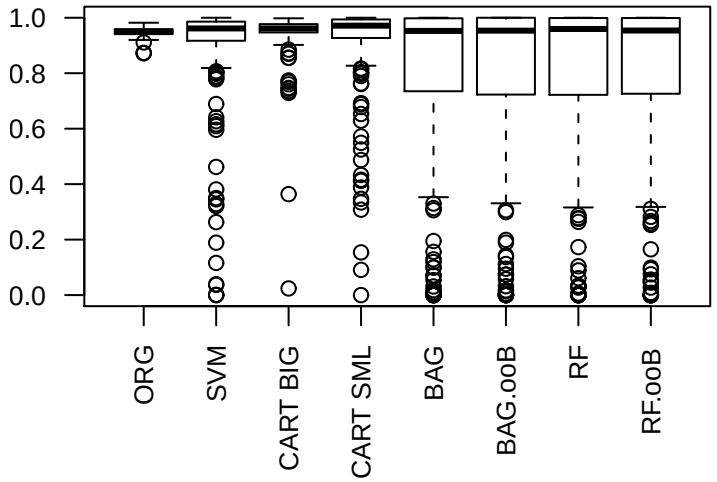

Figure 2: Boxplots of the percentage of 95% confidence intervals that cover the true population quantity across the 1,000 simulation runs for all estimands obtained for the original and synthetic datasets.

Figure 2 displays boxplots of the simulated coverage rates for 95% confidence intervals obtained using the original data and synthetic data. The results are in accord with the results in Figure 1. For the original samples, the average coverage rate is 94.8% and the minimum rate is 87.2%. CART BIG provides the highest data quality among the synthesizers: its average coverage rate is 94.2%, and only two estimands have coverage rates below 50%. CART SML results in a lower average coverage rate of 89.8% with 11 rates below 50%. The SVM synthesizer performs slightly worse than CART SML, resulting in an average coverage rates of 87.7% and 14 rates below 50%. The random forest and bagging synthesizers do not perform as well: average coverage rates for the random forest and bagging synthesizers are between 78.5% and 78.7%, and 31 to 35 rates fall below 50%.

For CART BIG, the two estimands with very poor coverage rates are the percentages of people in the age interval 10 - 20 and in the age interval 80 - 90. These arise because of modest biases produced by the synthesizers; the population percentages in these two categories are 4.35% and 3.42%, respectively, whereas the estimated expected values resulting from CART BIG are 3.13% and 2.77%. For CART SML and SVM, the estimands with less than 50% coverage rates also are primarily marginal probabilities. For ran-

dom forests and bagging, the low coverage rates tend to be for estimands involving relationships among variables, especially the interaction effects between marital status and age (see Appendix A.2) and between literacy and education (see Appendix A.1) in the regressions, as well as the conditional probabilities involving marital status and religion (Table 7 in the appendix) and education and age (Table 6 in the appendix). Thus, it appears that the random forest and bagging synthesizers are not as effective at preserving relationships as the CART and SVM synthesizers.

We also separately examine the average coverage rates for the 29 percentages associated with the marginal distributions of the five synthesized variables. For CART BIG, this average coverage rate is 86.9%. For CART SML and SVM, these average coverage rates are 78.7% and 63.0%, respectively, suggesting that CART outperforms SVM. For random forests and bagging, the average coverage rates for the 29 univariate percentages are between 82.7% and 83.4%, which actually are closer to 95% than either the CART SML or the SVM rates. This emphasizes the comparatively poor performance of random forests and bagging for the estimands involving relationships among variables.

Taking all the evidence together, it appears that CART BIG has the best overall performance in terms of data utility.

4.3. Disclosure risk

We assume that the intruder has perfect information on all five variables that have been synthesized and also on the district in which the respondent lives. We evaluate disclosure risks under two scenarios. For both scenarios, we evaluate two matching strategies. With the first strategy the intruder declares a match only if the target record matches exactly on all six variables. With the second strategy, the intruder allows a deviation of ± 2 years between the age in the target record and the age for the declared match. For either strategy, if no record in the synthetic samples fulfills all matching criteria, we assume that the intruder matches on the district alone, since she knows that this variable is not synthesized. We repeat the process of sampling from the population and synthesizing the datasets ten times, and report average risk measures across the ten simulation runs. For comparison, we also report the risk measures computed on the samples before replacing the five variables with synthetic values.

We begin the evaluation by assuming that the intruder does not know who participated in the survey. This is a realistic assumption for many

Table 1: Disclosure risk assuming the intruder does not know if the target participated in the survey.

		Org	SVM	CART BIG	CART SML	RF ooB	RF	Bagg. ooB	Bagg.
Age	true mr	21.83	0.03	0.13	0	0	0	0	0
	false mr	0	0	0	0	0	0	0	0
Age ±2	true mr	7.62	0.53	0.32	0	0	0	0	0
	false mr	0	0.32	5.80	0	0	0	0	0

surveys, especially for household surveys where sampling rates are low. It is particularly appropriate in public use samples from census files, since no one but the agency knows who was in the sample. We assume that the intruder always picks the record j with the highest match probability, including the possibility that the target record is not in D. For this scenario, for each target $\mathbf{t}$ we use the value of $F_{\mathbf{t}}$, i.e., the number of potential matches in the population for the target, when computing the match probabilities.

Table 1 displays the results based on the risk measures described in Section 2.1. The results in the second column indicate that, under the given assumptions about the intruder knowledge, disclosure risks are high for the unaltered data. On average 21.8% of the records in the sample are population uniques if the intruder matches on the exact age. This number decreases to 7.2% if the intruder allows for some deviation for the age variable. We note that there are no false matches with the original data.

For the synthetic datasets, the intruder's match probabilities select $j = n + 1$ for the great majority of records, i.e., the target most likely is not in D. The intruder is able to find some true matches only with synthetic data from the SVM and CART BIG synthesizers. However, the true match rate is under 1% in all simulations. When the intruder matches on the exact age, there are no false matches whenever she finds a single match among D. The true match rate increases when the intruder allows for a two year difference in ages, but now the intruder incorrectly declares a match more often than she declares a match correctly. The disclosure risk measures for all other synthesizers equal zero.

Overall, the disclosure risks are small for all synthesizers, suggesting that

Table 2: Disclosure risk assuming the intruder knows who participated in the survey.

		Org	SVM	CART BIG	CART SML	RF ooB	RF	Bagg. ooB	Bagg.
Age	true mr	90.46	10.50	16.81	2.29	0.21	0.20	0.20	0.16
	false mr	0	61.15	50.23	88.28	98.66	98.74	98.69	98.95
Age ± 2	true mr	71.70	46.89	30.53	5.36	0.30	0.34	0.32	0.23
	false mr	0	25.84	43.56	81.76	98.47	98.34	98.42	98.86

the methods provide reasonable protection for this type of risk scenario. However, it is difficult to compare the procedures based on the results of Table 1. Therefore, we also examine a scenario for which the intruder knows all the targets in the sample and is trying to identify them; that is, we set $Pr(J = n + 1|\mathbf{t}, D) = 0$ and replace $F_{\mathbf{t}}$ with $N_{\mathbf{t},i}$ when estimating $Pr(J = j|\mathbf{t}, D)$. This is a conservative assumption, but it enables comparisons of the relative risks attached to the synthesizers.

Table 2 displays the results for this scenario. More than 90% of the records are sample uniques in the original samples; even when we allow for some uncertainty in age, 71.7% of the records are uniquely identified. Two of the methods that provide the highest data utility, SVM and CART BIG, also lead to the highest disclosure risk. The risks are substantially lower for all other approaches. Interestingly, the risk for the SVM approach is smaller than the risk for CART BIG when we force exact matching on age, and the order changes when we allow for deviations in age. This arises because the CART approach leads to a higher probability that the unit's original age is selected as the synthetic value than the SVM approach does.

When the intruder knows which records are in the sample, the risks for CART SML are substantially lower than for CART BIG. This suggests that agencies can tune the minimum deviance parameter d to release data with adequate protection.

5. Concluding remarks

The empirical evaluations in this paper illustrate that it is possible to obtain synthetic datasets that provide reliable estimates paired with low disclosure risk by using nonparametric synthesizers. It is important to note

that we achieved these results without any tuning of the different synthesizers. Better results are obtainable if the methods are tailored to the data at hand. For example, values of the tuning parameter $d < .01$ might provide more accurate inferences with still acceptably low disclosure risks than seen in CART SML. Nevertheless, the evaluations demonstrate that good results can be obtained with minimal effort, even with intense synthesis.

The results indicate that the SVM and CART synthesizers outperform the bagging and random forests synthesizers in terms of analytical validity, albeit for the price of an increased risk of identification disclosure (when intruders know who is in the sample). The results for the SVM synthesizer are promising, but implementing the approach can be difficult. SVMs are sensitive to tuning, and it is not obvious which variation of SVM should be used for synthesis. This complexity may make the SVM synthesizer less attractive for agencies seeking automated methods of generating synthetic data. In contrast, the CART synthesizer offers a straightforward way to balance analytical validity and disclosure risks. With appropriate d, it can provide a high level of data utility for potentially acceptable disclosure risks. Thus, among these four nonparametric synthesizers, we believe that the CART synthesizer is best suited as a general-purpose, low-cost approach to generating partially synthetic datasets with good utility and acceptable risks.

Acknowledgements

This research was supported by U.S. National Science Foundation grant SES-0751671 and grants from the German Science Foundation, the German Federal Ministry of Education and Research, and the European Commission.

References

Abowd, J.M., Stinson, M., Benedetto, G., 2006. Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Technical Report. Longitudinal Employer–Household Dynamics Program, U.S. Bureau of the Census, Washington, DC.

Abowd, J.M., Woodcock, S.D., 2001. Disclosure limitation in longitudinal linked data, in: Doyle, P., Lane, J., Zayatz, L., Theeuwes, J. (Eds.), Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. Amsterdam: North-Holland, pp. 215–277.

Abowd, J.M., Woodcock, S.D., 2004. Multiply-imputing confidential characteristics and file links in longitudinal linked data, in: Domingo-Ferrer, J., Torra, V. (Eds.), Privacy in Statistical Databases. New York: Springer, pp. 290–297.

Boser, B., Guyon, I., Vapnik, V., 1992. A training algorithm for optimal marign classifiers, in: Proceedings of the Fifth ACM Workshop on Computation Learning Theory (COLT), New York: ACM Press. pp. 144–152.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Belmont, CA: Wadsworh, Inc.

Caiola, G., Reiter, J.P., 2010. Random forests for generating partially synthetic, categorical data. Transactions on Data Privacy 3, 27–42.

Choi, H., Yeo, D., Kwon, S., Kim, Y., 2011. Gene selection and prediction for cancer classification using support vector machines with a reject option. Computational Statistics and Data Analysis 55, 1897–1908.

Drechsler, J., 2010. Using support vector machines for generating synthetic datasets, in: Domingo-Ferrer, J., Magkos, E. (Eds.), Privacy in Statistical Databases. New York: Springer, pp. 148–161.

Drechsler, J., 2011. New data dissemination approaches in old europe – synthetic datasets for a german establishment survey. Journal of Applied Statistics , (forthcoming).

Drechsler, J., Bender, S., Rässler, S., 2008a. Comparing fully and partially synthetic data sets for statistical disclosure control in the German IAB Establishment Panel. Transactions on Data Privacy 1, 105–130.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., Zwick, T., 2008b. A new approach for disclosure control in the IAB Establishment Panel – multiple imputation for a better data access. Advances in Statistical Analysis 92, 439–458.

Drechsler, J., Reiter, J.P., 2008. Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data, in: Domingo-Ferrer, J., Saygin, Y. (Eds.), Privacy in Statistical Databases. New York: Springer, pp. 227–238.

Drechsler, J., Reiter, J.P., 2010. Sampling with synthesis: A new approach for releasing public use census microdata. Journal of the American Statistical Association 105, 1347–1357.

Elliott, M., Purdam, K., 2007. A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymized Records. Environment and Planning A 39, 1101–1118.

Fuller, W.A., 1993. Masking procedures for microdata disclosure limitation. Journal of Official Statistics 9, 383–406.

Graham, P., Penny, R., 2005. Multiply Imputed Synthetic Data Files. Technical Report. University of Otago. `http://www.uoc.otago.ac.nz/departments/pubhealth/pgrahpub.htm`.

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

Hsu, C., Chang, C., Lin, C., 2010. A Practical Guide to Support Vector Classification. Technical Report. Department of Computer Science, National Taiwan University.

Iacus, S.M., Porro, G., 2007. Missing data imputation, matching and other applications of random recursive partitioning. Computational Statistics and Data Analysis 52, 773–789.

Kennickell, A., Lane, J., 2006. Measuring the impact of data protection techniques on data utility: Evidence fromthe Survey of Consumer Finances, in: Domingo-Ferrar, J. (Ed.), Privacy in Statistical Databases 2006 (Lecture Notes in Computer Science). New York: Springer-Verlag, pp. 291–303.

Kennickell, A.B., 1997. Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances, in: Alvey, W., Jamerson, B. (Eds.), Record Linkage Techniques, 1997. Washington, DC: National Academy Press, pp. 248–267.

Kinney, S.K., Reiter, J.P., 2007. Making public use, synthetic files of the Longitudinal Business Database, in: Proceedings of the Joint Statistical Meetings, Alexandria, VA: American Statistical Association.

Little, R.J.A., 1993. Statistical analysis of masked data. Journal of Official Statistics 9, 407–426.

Little, R.J.A., Liu, F., Raghunathan, T.E., 2004. Statistical disclosure techniques based on multiple imputation, in: Gelman, A., Meng, X.L. (Eds.), Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. New York: John Wiley and Sons, pp. 141–152.

Minnesota Population Center, 2010. Integrated Public Use Microdata Series, International: Version 6.0 [Machine-readable database]. Technical Report. Minneapolis: University of Minnesota.

Mitra, R., Reiter, J.P., 2006. Adjusting survey weights when altering identifying design variables via synthetic data, in: Domingo-Ferrer, J., Franconi, L. (Eds.), Privacy in Statistical Databases. New York: Springer-Verlag, pp. 177–188.

Moguerza, J., Muñoz, A., 2006. Support vector machines with applications (with discussion). Statistical Science 21, 322–362.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., Solenberger, P., 2001. A multivariate technique for multiply imputing missing values using a series of regression models. Survey Methodology 27, 85–96.

Reiter, J.P., 2003. Inference for partially synthetic, public use microdata sets. Survey Methodology 29, 181–189.

Reiter, J.P., 2005a. Estimating identification risks in microdata. Journal of the American Statistical Association 100, 1103–1113.

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

Reiter, J.P., 2005c. Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. Journal of Statistical Planning and Inference 131, 365–377.

Reiter, J.P., 2005d. Using CART to generate partially synthetic, public use microdata. Journal of Official Statistics 21, 441–462.

Reiter, J.P., Mitra, R., 2009. Estimating risks of identification disclosure in partially synthetic data. Journal of Privacy and Confidentiality 1, 99–110.

Reiter, J.P., Raghunathan, T.E., 2007. The multiple adaptations of multiple imputation. Journal of the American Statistical Association 102, 1462–1471.

Rubin, D.B., 1981. The Bayesian bootstrap. The Annals of Statistics 9, 130–134.

Rubin, D.B., 1993. Discussion: Statistical disclosure limitation. Journal of Official Statistics 9, 462–468.

Shim, J., Sohn, I., Kim, S., Lee, J., Green, P., Hwang, C., 2009. Selecting marker genes for cancer classification using supervised weighted kernel clustering and the support vector machine. Computational Statistics and Data Analysis 53, 1736–1742.

Skinner, C.J., Shlomo, N., 2008. Assessing identification risk in survey microdata using log-linear models. Journal of the American Statistical Association 103, 989–1001.

Smola, A., Schölkopf, B., 1998. A Tutorial on Support Vector Regression. Technical Report. NeuroCOLT2 Technical Report Series, NC2-TR-1998-030.

Sweeney, L., 1997. Computational disclosure control for medical microdata: the Datafly system, in: Proceedings of an International Workshop and Exposition, pp. 442–453.

Willenborg, L., de Waal, T., 2001. Elements of Statistical Disclosure Control. New York: Springer-Verlag.

Winkler, W.E., 2007. Examples of Easy-to-implement, Widely Used Methods of Masking for which Analytic Properties are not Justified. Technical Report. Statistical Research Division, U.S. Bureau of the Census.

Woodcock, S.D., Benedetto, G., 2009. Distribution-preserving statistical disclosure limitation. Computational Statistics and Data Analysis 53, 4228–4242.

Wu, T., Lin, C., Weng, R., 2004. Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005.

Appendix

Here we describe the variables and different analyses we use to evaluate the analytical validity associated with the nonparametric data synthesizers. Table 3 contains descriptive information on all the variables included in the evaluation. The validity evaluations are based on multivariate regressions and descriptive statistics for these variables.

Table 3: Description of the variables included in the evaluation.

label	description	type	range
age	age in years	continuous	10–95
age.cat10	age in 10 year categories (10–90)	categorical	8 categories
citizen	citizen of Uganda	binary	2 categories
edattand	highest level of education attained	categorical	9 categories
empstat	employment status	categorical	3 categories
lit	literacy	binary	2 categories
marstd	marital status	categorical	5 categories
migration	Previous residence outside Uganda	binary	2 categories
nchild	number of children	categorical	10 categories
ownrshp	ownership of dwelling	binary	2 categories
persons	number of persons in the household	continuous	1–30
relig	religion	categorical	11 categories
tv	own a television set	binary	2 categories
urban	urban-rural status	binary	2 categories

A. Multivariate Analysis

We use two logistic and one linear regression in the evaluation. Details about the different models are provided below.

A.1. Regression involving employment status

We fit a logistic regression of employment status on 41 predictors. The model is

$$empstat \sim age.cat10 + urban + marstd + lit + persons + edattand + tv + ownrshp + nchild + lit^{*}edattand$$

Some interactions between literacy and education are not included due to collinearity. The dataset used in fitting the model is restricted to individuals age 25 or older who are not inactive in the labor force, i.e., they are either employed or unemployed. All predictors are included as a series of dummy variables. All household sizes greater than 10 are included in one category.

A.2. Regression involving household size

We fit a linear regression of the logarithm of household size on 24 predictors. The model is

$$\log(persons) \sim marstd + age + age^2 + marstd^{*}age + marstd^{*}age^2 + lit + empstat + edattand + citizen$$

The model is fit with all individuals in the data (there is only one record per household). Age is treated as a continuous variable; all other predictors are included as a series of dummy variables.

A.3. Regression involving migration status

We fit a logistic regression of migration status on 32 predictors. The model is

$$migration \sim age.cat10 + edattand + urban + marstd + lit + empstat + tv + persons$$

The model is fit with all individuals in the data. All predictors are included as a series of dummy variables. All household sizes greater than 10 are included in one category.

B. Demographics

We also compare several descriptive statistics. Here we present the statistics with their population values to illustrate that some of the quantities are for small subgroups of the population. Where convenient, we present results in tabular format.

- Single/never married: 7.78%
- Married, monogamous: 71.14%
- Married, polygamous: 14.62%
- Separated or divorced: 4.38%
- Widowed: 2.08%
- Illiterate: 25.03%
- Literate: 74.97%
- Employed: 76.91%
- Unemployed: 2.81%
- Inactive: 20.28%
- Own household, given employed and in urban area: 28.13%
- Own household, given employed and in rural area: 85.43%
- Employed, given literate and in one person household: 78.23%
- Employed, given illiterate and in one person household: 66.46%

Table 4: Distribution of number of people in household.

persons	1	2	3	4	5	6	7	8	9	10	11+
%	11.4	9.9	12.2	13.3	13.1	11.7	9.5	7.0	4.7	3.6	3.6

Table 5: Distribution of age in ten year categories.

age	(10,20]	(20,30]	(30,40]	(40,50]	(50,60]	(60,70]	(70,80]	(80,90]
%	4.35	30.94	27.85	15.68	9.56	6.89	3.42	1.04

Table 6: Distribution of education for different age groups.

	< 35	$35 - 55$	$>= 55$
No schooling	13.38	19.03	38.57
Some primary completed	28.68	26.92	31.60
Primary (6 yrs) completed	33.24	30.63	20.42
Lower secondary general completed	15.82	12.87	5.12
Secondary, general track completed	3.50	2.66	0.67
Post-secondary technical education	4.34	5.88	2.76
University completed	1.04	2.02	0.86

Table 7: Probability of living in a monogamous/polygamous marriage for different religions.

	monogamous	polygamous
Muslim	65.34	20.48
Catholic (Roman or unspecified)	71.90	13.23
Pentecostal	77.32	8.69
Anglican	72.03	14.31