Using CART to Generate Partially Synthetic, Public Use Microdata

J. P. Reiter*

Key Words: CART, Confidentiality, Disclosure, Multiple Imputation, Synthetic Data, Trees

Abstract

To limit disclosure risks, one approach is to release partially synthetic, public use microdata sets. These comprise the units originally surveyed, but some collected values, for example sensitive values at high risk of disclosure or values of key identifiers, are replaced with multiple imputations. This article presents and evaluates the use of classification and regression trees to generate partially synthetic data. Two potential applications of CART are studied via simulation: (i) generate synthetic data for sensitive variables; and, (ii) generate synthetic data for variables that are key identifiers.

1 Introduction

When releasing public use microdata, statistical agencies employ a variety of techniques to limit disclosures, including swapping data, recoding variables, and adding noise to values (see Willenborg and de Waal, 2001). Unfortunately, these techniques can distort relationships among variables in the data set and complicate estimation for the user, for example requiring non-standard, likelihood-based analyses (Little, 1993) or measurement error models (Fuller, 1993).

*Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251. E-mail: jerry@stat.duke.edu

An alternative approach with the potential to circumnavigate these problems is to release multiply-imputed, synthetic public use microdata, as proposed by Rubin (1993). In this approach, the agency (i) randomly and independently samples units from the sampling frame for each synthetic data set, and (ii) imputes unknown data values for units in the synthetic samples using models fit using the original survey data and possibly other information. This can protect confidentiality, since identification of units and their sensitive data can be difficult when the released data are not actual, collected values. And, with appropriate imputation and estimation methods developed by Raghunathan *et al.* (2003) and Reiter (2003c)–based on the concepts of multiple imputation (Rubin, 1987)–the approach can allow data users to obtain valid inferences using standard, complete-data statistical methods and software. For general discussions of fully synthetic data approaches, see also Fienberg *et al.* (1996, 1998), Dandekar *et al.* (2002a,b), and Reiter (2003b, 2002).

Although there are potentially great benefits to releasing fully synthetic data (see Raghunathan *et al.*, 2003; Reiter, 2003b), generating plausible synthetic data for all variables may be difficult in practice. Instead, agencies can release multiply-imputed, partially synthetic data sets comprising a mix of actual and imputed values, as suggested by Little (1993). This is currently being done by the U.S. Federal Reserve Board for public-use data from the U.S. Survey of Consumer Finances. They replace monetary values at high disclosure risk with multiple imputations, then release these imputed values and the unreplaced, collected values (Kennickell, 1997). A partially synthetic approach also has been used by Abowd and Woodcock (2001) to protect data in longitudinal, linked data sets. They replace all values of some sensitive variables with multiple imputations, but leave other variables at their actual values. A third example is the SMIKe algorithm of Liu and Little (2002), which simulates multiple values of key identifiers for selected units.

Partially synthetic approaches are appealing because they promise to maintain many of the benefits of fully synthetic data–protecting confidentiality while allowing users to make inferences without learning complicated statistical methods or software–with decreased sensitivity to the specification of imputation models. Still, imputation models must be specified, a task that can be daunting in surveys with hundreds of

variables, some with distributions not easily modeled with standard parametric tools. It may be advantageous to use nonparametric methods to generate imputations.

This paper presents and evaluates the use of classification and regression trees (Breiman *et al.*, 1984), typically abbreviated as CART, for generating partially synthetic data. The paper is organized as follows. Section 2 reviews the notation and methods of inference for partially synthetic data developed by Reiter (2003a). Section 3 reviews CART and suggests how it might be used for generating synthetic data. Section 4 presents results of simulation studies that use CART (i) to simulate selected units' values of potentially sensitive variables, and (ii) to simulate selected units' values of variables that are key identifiers. The simulations illustrate the validity of inferences for a variety of descriptive and model-based estimands, as well as the disclosure risks of the released data.

2 Description of partially synthetic data

To describe partially synthetic data, we use the notation of Reiter (2003a), repeated nearly verbatum here. "Let $I_j = 1$ if unit j is selected in the original survey, and $I_j = 0$ otherwise. Let $I = (I_1, \ldots, I_N)$. Let Y_{obs} be the $n \times p$ matrix of collected (real) survey data for the units with $I_j = 1$; 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 sampled units fully respond to the survey. Let X be the $N \times d$ matrix of design variables for all N units in the population, e.g. stratum or cluster indicators or size measures. We assume that such design information is known approximately for all population units. It may come, for example, from census records or the sampling frame(s).

The agency releasing synthetic data, henceforth abbreviated as the *imputer*, constructs synthetic data sets based on the observed data, $D = (X, Y_{obs}, I)$, in a two-part process. First, the imputer selects the values from the observed data that will be replaced with imputations. Second, the imputer imputes new values to replace those selected values. Let $Z_j = 1$ if unit j is selected to have any of its observed data replaced

with synthetic values, and let $Z_j = 0$ for those units with all data left unchanged. Let $Z = (Z_1, \ldots, Z_n)$. Let $Y_{rep,i}$ be all the imputed (replaced) values in the ith synthetic data set, and let Y_{nrep} be all unchanged (unreplaced) values of Y_{obs}.... The values in Y_{nrep} are the same in all synthetic data sets. Each synthetic data set, d_i, is then comprised of $(X, Y_{rep,i}, Y_{nrep}, I, Z)$. Imputations are made independently for $i = 1, \ldots, m$ times to yield m different synthetic data sets. These synthetic data sets are released to the public (Reiter, 2003a, p.)."

When using parametric imputation models, the $Y_{rep,i}$ should be generated from the Bayesian posterior predictive distribution of $(Y_{rep,i}|D, Z)$. In this article, we generate the $Y_{rep,i}$ from a series of CART models fit using the units with $Z_j = 1$. These models are described in Section 3.2.

Inferences about some scalar estimand, say Q, are obtained by combining results from the d_i. Specifically, suppose the data analyst estimates Q with some point estimator q and estimates the variance of q with some estimator v. For $i = 1, \ldots, m$, let q_i and v_i be respectively the values of q and v in synthetic data set d_i. It is assumed that the analyst determines the q_i and v_i as if d_i was in fact collected data from a random sample of (X, Y) based on the actual survey design used to generate I. The following quantities are needed for inferences for scalar Q:

$$\bar{q}_m = \sum_{i=1}^{m} q_i/m \tag{1}$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1) \tag{2}$$

$$\bar{v}_m = \sum_{i=1}^{m} v_i/m. \tag{3}$$

The analyst then can use $\bar{q}_m$ to estimate Q and

$$T_p = b_m/m + \bar{v}_m \tag{4}$$

to estimate the variance of $\bar{q}_m$. When n is large, inferences for scalar Q can be based on t-distributions with degrees of freedom $\nu_p = (m-1)(1+r_m^{-1})^2$, where $r_m = (m^{-1}b_m/\bar{v}_m)$. In many cases, a normal distribution provides an adequate approximation to the t-distribution because r_m is small. Derivations of these methods are presented in Reiter (2003a). Extensions for multivariate Q are presented in Reiter (2003c).

3 CART models for generating partially synthetic data

In this section, we propose the use of CART models to generate the $Y_{rep,i}$. We first provide some background on CART and existing proposals for using CART models to impute missing data.

3.1 Background on CART

CART models (Breiman *et al.*, 1984) are a flexible tool for estimating the conditional distribution of a univariate outcome given multivariate predictors. Essentially, the CART model partitions the predictor space so that subsets of units formed by the partitions have relatively homogeneous outcomes (Chipman *et al.*, 1998). 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. An example of a tree structure for a univariate outcome Y and two predictors, X_1 and X_2, is presented in Figure 1. Units with $X_1 \geq 2$ fall in the leaf labeled L_1, regardless of their value of X_2. Units with $X_1 < 2$ and $X_2 \geq 0$ fall in the leaf labeled L_2, and units with $X_1 < 2$ and $X_2 < 0$ fall in the leaf labeled L_3. Such trees can be grown using algorithms like the one in the software package S-Plus (Clark and Pregibon, 1992).

A common strategy for finding trees is to fit one with a large number of leaves, and then prune the tree according to some optimality or complexity criteria. For example, if the tree in Figure 1 is deemed too large or too complex, the branch to the leaves L_2 and L_3 can be cut, so that the resulting tree has only two leaves, L_1 and what was formerly the root of L_2 and L_3. Pruned trees typically do not predict the values in the observed data as well as larger ones, but they may be more robust to overfitting than larger ones.

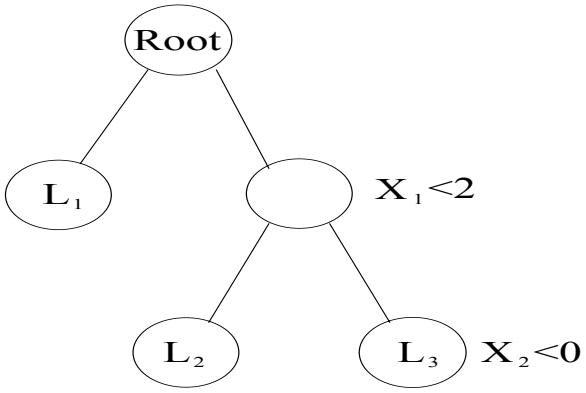

Figure 1: Example of a tree structure

As a method of estimating conditional distributions, CART models have some potential advantages over parametric models. First, CART modeling may be more easily applied than paramteric modeling, particularly for continuous data that are truncated or not smooth. Second, CART models can capture non-linear relationships and interaction effects that may not be easily revealed in the process of fitting parametric models. Third, CART provides a semi-automatic way to fit the most important relationships in the data, which can be a substantial advantage when there are many potential predictors. Primary disadvantages of CART models relative to parametric models include difficulty of interpretation, discontinuity at partition boundaries, and decreased effectiveness when the data follow relationships easily captured by parametric models (Friedman, 1991).

Because of their nonparametric nature, CART models have been proposed to impute missing data (Barcena and Tussel, 2000; Piela and Laaksonen, 2001; Conversano and Siciliano, 2002). These proposals primarily use the leaves of trees as imputation classes, assuming the data are missing at random (Rubin, 1976). As an example, suppose a single variable Y has data missing at random. A tree is grown using the observed outcomes, Y_{obs}, and all other variables as predictors, then pruned to some desired size. Units with missing Y are placed in appropriate leaves of the tree according to their predictor values, and imputed values of Y are then drawn randomly from the Y_{obs} in the corresponding leaves.

It is not clear how to implement the CART approach when data are missing for multiple variables. Imputations from single-variable trees can fail to reflect relationships among the imputed variables. For example, imputation of missing Y_a and missing Y_b from trees approximating $f(Y_a|X)$ and $f(Y_b|X)$ assumes, possibly incorrectly, conditional independence between Y_a and Y_b. One approach is to impute from chains of single-variable trees conditional on previous imputations (Conversano and Siciliano, 2002). For example, first impute missing values of Y_a using its single-variable tree fit on X, then impute missing values of Y_b using its single-variable tree fit on X and the filled in Y_a, then impute missing values of Y_c after filling in missing values of Y_a and Y_b, etc. Such conditional approaches are related to the sequential imputation algorithms of Van Buuren and Oudshoorn (1999) and Raghunathan *et al.* (2001) for parametric modeling. To this author's knowledege, there have not been published evaluations of the repeated-sampling properties of inferences from multiply-imputed data sets generated from such chained CART models.

Single variable trees can be employed for missing multivariate categorical data. All levels of the r missing categorical variables are combined into one variable with $K = \Pi_i^r n_i$ levels, where n_i is the number of levels for categorical variable i (Barcena and Tussel, 2000). Unfortunately, this can be computationally infeasible when K is large.

With any of these approaches, and regardless of the number of variables with missing data, a key issue is how to prune the tree. Pruning the tree too much may result in non-homogeneous imputation donors, so that the imputations are not drawn from plausible conditional distributions; essentially, the imputation classes are too broad. Insufficiently pruning the tree may lead to over-fitting the observed data, resulting possibly in inferences with larger variances. Given the usual advice for multiple imputation of accepting variance to avoid bias (Rubin, 1987), it may be preferable to use larger trees for imputation purposes. To this author's knowledge, this conjecture has not been substantiated with research.

3.2 Generation of $Y_{rep,i}$ from CART models

We now turn to considering CART models for generating partially synthetic data sets, $d_i = (X, Y_{rep,i}, Y_{nrep}, I, Z)$, using values of the observed data, $D = (X, Y_{obs}, I)$. The proposed CART algorithm for imputing $Y_{rep,i}$ is layed out in Section 3.1, and motivation for its specification is presented in Section 3.2.

3.2.1 Algorithm for imputations

Let $Y_{(1)}$ be the variable in Y that has the largest number of values to be replaced, and let $Y_{(k)}$ be the variable in Y that has the kth largest number of values to be replaced. Let $Z_{(k)} = 1$ for all units having $Y_{(k)}$ replaced. For each $Y_{(k)}$, we fit the tree of $Y_{(k)}$ on $(X, Y_{-(k)})$, where $Y_{-(k)}$ is all variables in Y except $Y_{(k)}$, using the values in D. Label these trees $\mathcal{Y}_{(k)}$. Whenever practical, only units with $Z_{(k)} = 1$ are used to grow $\mathcal{Y}_{(k)}$. For example, when $Z_{(k)} = 1$ only for units with $Y_{(k)} > 100,000$, the imputation model should be fit using only those units in D with $Y_{(k)} > 100,000$.

When two or more variables have the same number of values to be replaced, the order of the variables is selected as follows. First, just to avoid introducing additional notation, assume the variables are assigned a random ordering. The $\mathcal{Y}_{(k)}$ are fit for each of these variables. Let $P_{(k)}$ be the depth in $\mathcal{Y}_{(k)}$ of the first split on one of these other variables. If none of these other variables appear in $\mathcal{Y}_{(k)}$, define $P_{(k)} = \infty$. Now, re-order the variables in decreasing order of the $P_{(k)}$ to obtain the order of imputations. Figure 2 illustrates this procedure for two variables, Y_a and Y_b. Because Y_b appears higher up in $\mathcal{Y}_{(a)}$ than Y_a appears in $\mathcal{Y}_{(b)}$, the $P_{(b)} > P_{(a)}$, and we impute Y_b before Y_a.

At its largest, each $\mathcal{Y}_{(k)}$ can have exactly one leaf for every unit with $Z_{(k)} = 1$. Imputing data by sampling from leaves of maximal trees results in $d_i = D$ for all i, which obviously fails to protect confidentiality if D is not releasable. The maximal trees must be pruned so as to preserve as much as possible the relationships in D, while limiting disclosure risks. For continuous $Y_{(k)}$, we propose pruning until the observed values in all leaves have variance larger than some imputer-defined threshold. For categorical $Y_{(k)}$, we propose pruning

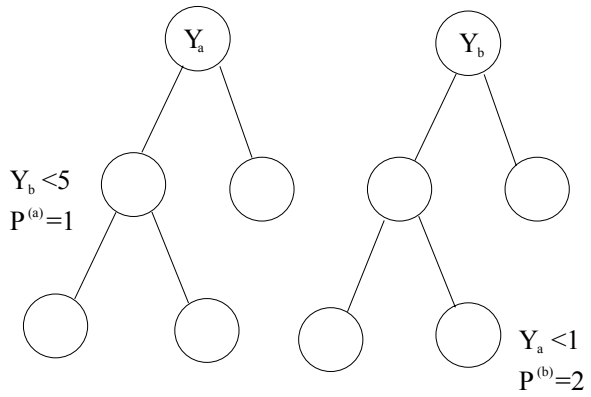

Figure 2: Example of ordering of imputations when two variables have equal numbers of replaced values. Here, Y_b is imputed before Y_a.

trees so that no one value of $Y_{(k)}$ appears in any leaf more than an imputer-specified percentage of the time. Additional protection can be obtained by requiring a minimum number of units in each leaf of the tree.

Once trees are pruned to satisfy disclosure criteria, imputations are generated sequentially using the pruned trees, beginning with $Y_{(1)}$. Let L_{1w} be the wth leaf in the pruned $\mathcal{Y}_{(1)}$, and let $Y_{(1)}^{L_{1w}}$ be the $n_{L_{1w}}$ values of $Y_{(1)}$ in leaf L_{1w}. In each L_{1w} in the tree, we generate a new set of values by drawing from $Y_{(1)}^{L_{1w}}$ using the Bayesian bootstrap (Rubin, 1981, and described in Section 3.2). When $Y_{(1)}$ is categorical, these sampled values are the replacement imputations, $Y_{(1)rep,i}$, for the $n_{L_{1w}}$ units that belong to L_{1w}. When $Y_{(1)}$ is continuous, we take an additional step to avoid purposefully releasing real values of $Y_{(1)}$. In each leaf, we estimate the density of the bootstrapped values, for example by using a Gaussian kernel density estimator (Wegman, 1972). Then, for each unit, we sample randomly from the estimated density in that unit's leaf using an inverse-cdf method. We allow the support of the estimated density to stretch from the smallest to the largest value of $Y_{(1)}$. The sampled values are the $Y_{(1)rep,i}$.

Imputations are next made for $Y_{(2)}$ using the same procedure. To maintain consistency with the $Y_{(1)rep,i}$, units' leaves in $\mathcal{Y}_{(2)}$ are located using $Y_{(1)rep,i}$ in place of $Y_{(1)}$. Occasionally, some units may have combinations of $(X, Y_{-(1,2)}, Y_{(1)nrep}, Y_{(1)rep,i})$ that do not belong to one of the leaves of $\mathcal{Y}_{(2)}$. For these units, we search up the tree until we find a node that contains the combination, then treat that node as if it were the

unit's leaf. Once each unit's leaf is located, values of $Y_{(2)rep,i}$ are generated using the Bayesian bootstrap and kernel density procedure used to impute $Y_{(1)}$. Imputing any $Y_{(k)}$ follows the same process: we place each unit in the leaves of $\mathcal{Y}_{(k)}$ based on their values in $(X, Y_{-(1,2,\ldots,k-1)}, Y_{(1,2,\ldots,k-1)nrep}, Y_{(1,2,\ldots,k-1)rep,i})$, then impute using the Bayesian bootstrap and kernel density procedure.

Each released, partially synthetic data set $d_i = (X, Y_{nrep}, Y_{rep,i}, I, Z)$. The process is repeated independently m times, and these m data sets are released to the public.

3.2.2 Motivation for algorithm

When fitting the $\mathcal{Y}_{(k)}$, only units with $Z_{(k)} = 1$ are used to grow the tree. This helps ensure the estimated conditional distributions for the $Y_{(k)}$ are in the space of $Y_{(k)}$ where data need to be replaced. For example, when replacing incomes only when they are greater than \$100,000, all imputed incomes must be at least \$100,000 if inferences for the population mean income are to be potentially valid. Using trees grown from observed data that include units with incomes below \$100,000 may result in imputed incomes below \$100,000, which may lead to biased estimates. As another example, when replacing some outcome only for certain small subpopulations (e.g., replace incomes for single Native American males), the imputations should be drawn from that sub-population's outcome distribution. A tree grown using units outside the sub-population may not accurately capture the outcome distribution in the small sub-population. As a result, the imputations for the sub-population would not be consistent with the corresponding distribution of outcomes in the observed data.

It may be necessary for practical reasons or disclosure limitation purposes to use units with $Z_{(k)} = 0$ when growing some $\mathcal{Y}_{(k)}$. For example, there may be insufficient number of units with $Z_{(k)} = 1$ to fit an accurate tree model from only those units. Or, the values of $Y_{(k)}$ for the units with $Z_{(k)} = 1$ may not be sufficiently varied, so that disclosure criteria for pruning the trees cannot be satisfied.

Imputations are made from sequential CART models. Essentially, the $\mathcal{Y}_{(k)}$ estimate $f(Y_{(k)}|X, Y_{-(k)}, Z_{(k)})$.

All $Y_{-(k)}$ are predictors so that as much information as possible is used for imputations, which helps to maintain consistency in relationships. For example, suppose there are two strongly related variables to be replaced, $Y_{(a)}$ and $Y_{(b)}$, , and $Y_{(a)}$ has many more values to be replaced than does $Y_{(b)}$. Including $Y_{(b)}$ as a predictor when fitting $\mathcal{Y}_{(a)}$, and vice-versa, appropriately results in imputations that reflect dependencies between $Y_{(a)}$ and $Y_{(b)}$ (assuming $\mathcal{Y}_{(a)}$ splits on $Y_{(b)}$ and $\mathcal{Y}_{(b)}$ splits on $Y_{(a)}$). On the other hand, fitting $\mathcal{Y}_{(a)}$ without including $Y_{(b)}$, or vice-versa, inappropriately produces imputations that reflect conditional independece of $Y_{(a)}$ and $Y_{(b)}$, at least for some units.

Variables are ordered for sequential imputation by the number of values to be replaced, going from largest to smallest. This helps preserve relationships for variables with smaller numbers of values to be replaced. To illustrate, consider two variables, $Y_{(a)}$ and $Y_{(b)}$, where $a < b$, and $Z_{(a)} = 1$ for all units with $Z_{(b)} = 1$. Suppose $Y_{(a)}$ is a strong predictor of $Y_{(b)}$ for the units with $Z_{(b)} = 1$, so that $\mathcal{Y}_{(b)}$ contains splits on $Y_{(a)}$. Further, suppose that there are many units with $Z_{(a)} = 1$ and $Z_{(b)} = 0$, and that $Y_{(b)}$ is not a strong predictor of $Y_{(a)}$ for these units. The $\mathcal{Y}_{(a)}$, dominated by the units with $Z_{(a)} = 1$ and $Z_{(b)} = 0$, may not not contain splits on $Y_{(b)}$. If so, when Y_b is imputed before Y_a, the imputations for units with $Z_b = 1$ will reflect conditional independence between Y_a and Y_b implied in $\mathcal{Y}_{(a)}$. On the other hand, imputing Y_a before Y_b avoids this problem.

When two or more variables have equal values of $Z_{(k)}$, the trees are fit in decreasing order of $P_{(k)}$, as illustrated in Figure 2. Essentially, this aims to impute the variables in decreasing order of dependency on each other. This helps preserve the strongest relationships among the $Y_{(k)}$ in the imputations. To illustrate, consider the example in Figure 2, in which $Y_{(b)}$ appears in $\mathcal{Y}_{(a)}$ before $Y_{(a)}$ appears in $\mathcal{Y}_{(b)}$, so that $b < a$. The trees indicate that $Y_{(b)}$ is a stronger predictor of $Y_{(a)}$ than $Y_{(a)}$ is of $Y_{(b)}$. Setting $b < a$ passes this relationship on to the imputations, whereas setting $a < b$ results in imputations that reflect a weaker relationship between $Y_{(a)}$ and $Y_{(b)}$ than those implied by the trees.

The above examples illustrate some drawbacks to using sequential, single-tree CART models. First,

sequential CART models cannot guarantee relationships will be preserved, even when conditioning on all variables and choosing the order judiciously. Second, the imputations may not and probably will not come from legitimate joint probability density functions. The examples of the previous paragraphs illustrate this problem. Third, the imputations can be sensitive to the ordering of the sequence.

Once the trees are fit and pruned, Bayesian bootstraps and, for continuous data, kernel density estimators are used to generate imputations. The Bayesian bootstrap draws values of some variable Y from a donor pool of selected values of the observed data. Let Y_{elig} be the $n_0 \times 1$ vector of values of that make up the donor pool. Let n_{rep} be the number of values to be drawn. The Bayesian bootstrap proceeds as follows.

1. Draw $(n_0 - 1)$ uniform random numbers. Sort these numbers in ascending order. Label these ordered numbers as $a_0 = 0, a_1, a_2, \ldots, a_{n_0-1}, a_{n_0} = 1$.

2. Draw n_{rep} uniform random numbers, $u_1, u_2, \ldots, u_j, \ldots, u_{n_{rep}}$. For each of these u, impute $Y_{elig,j}$ when $a_{j-1} < u \leq a_j$.

The Bayesian bootstrap incorporates the additional uncertainty in the conditional distributions in each leaf due to having only a sample of values in each leaf. Sampling values from Y_{elig} directly, i.e. the usual bootstrap, underestimates this uncertainty. Arguments for preferring the Bayesian bootstrap can be found in Rubin (1987, Chapter 4).

For continuous data, we take the additional step of drawing values from an estimated density, fit using the bootstrapped values and a kernel density estimator. Values are drawn from the estimated density by the inverse cdf method. As stated previously, the primary reason for drawing from the density estimator rather than releasing the bootstrapped values is to avoid releasing real data values. The support of the density in each leaf L_{kw} stretches from the largest to the smallest value of $Y_{(k)}$ to allow for a wider range of possible values, which can help protect confidentiality. Additionally, this avoids assigning zero probability mass to values outside the range of observed values in each leaf. To ensure the density can be reasonably estimated,

Table 1: Description of variables used in the empirical studies

Variable	Label	Range
Sex	X	male, female
Race	R	white, black, American Indian, Asian
Marital status	M	7 categories, coded 1–7
Highest attained education level	E	16 categories, coded 31–46
Age (years)	G	0 – 90
Child support payments ($)	C	0, 1 – 23,917
Social security payments ($)	S	0, 1 – 50,000
Household alimony payments ($)	A	0, 1 – 54,008
Household property taxes ($)	P	0, 1 – 99,997
Household income ($)	I	-21,011 – 768,742

the imputation algorithm must require that the bootstrapped values within any L_{kw} are not all identical. To reduce variance, the algorithm should require a minimum number of observed units in each L_{kw}. In the simulations in Section 4, ten observations are the minimum. Imputers can evaluate the sensitivity of their imputations to alternative minimum values.

4 Simulation studies

This section illustrates the performance of these sequential CART models using genuine data. All CART models are fit in S-Plus using the algorithm of Clark and Pregibon (1992). The first set of simulations mimics replacing sensitive variables and the second set mimics replacing key identifiers. Both simulations are based on a subset of public release data from the March 2000 U.S. Current Population Survey. The data comprise ten variables measured on 51,016 heads of households. The variables, displayed in Table 1, were selected and provided by statisticians at the U.S. Bureau of the Census. Similar data are used by Reiter (2003b) to illustrate and evaluate releasing fully synthetic data.

Marital status, M, has seven types: $M = 1$ for married civilians with both spouses present at the home; $M = 2$ for married people in the armed forces with both spouses present at the home; $M = 3$ for married people with one spouse not present at the home; $M = 4$ for widowers; $M = 5$ for divorced people; $M = 6$ for

separated people; and, $M = 7$ for people who never have been married. Highest attained education level, E, increases from 31 to 46 in correspondence with years of schooling. As examples, $E = 31$ represents highest educational attainments of less than first grade; $E = 35$ represents highest educational attainments of ninth grade; $E = 39$ represents a high school degree; $E = 43$ represents a bachelor's degree; $E = 44$ represents a master's degree; $E = 45$ represents a professional school degree; and, $E = 46$ represents a doctoral degree.

Marginally, there are ample numbers of people in each sex, race, marital status, and education category. Many cross-classifications have few or zero people, especially those involving minorities with $M \notin \{1, 7\}$. Out of the 51,106 people, there are 33,076 who have positive property taxes, 12,021 who receive social security payments, 1,677 who receive child support payments, and 206 who receive alimony payments. There are 132 households with negative income, 582 with zero income, 5371 households with incomes at least $100,000, and the remainder with incomes between 0 and $100,000. The negative incomes are legitimate values: some households actually report paying out more money than they took in over the year. The distributions of positive values for all monetary variables are right-skewed.

4.1 Simulating sensitive variables

Imputers may decide to replace selected units' values of sensitive variables with multiple imputations, then release the imputed and unreplaced values. This typically does not reduce the risks of re-identifications, but it can limit the risks of attribute disclosures. We mimic this strategy by considering S, I, C, and A to be sensitive, replacing S for all people with $S > 0$, I for all people with $I > 100,000$, C for all people with $C > 0$, and A for all people with $A > 0$. Other values are not replaced and are released in all d_i.

Each observed dataset, D, comprises $n = 10,000$ randomly sampled households from the 51,106 households. There are $m = 5$ synthetic data sets generated for each D. Each d_i is generated using the CART models outlined in Section 3, with sequential order of imputation $S - I - C - A$. The trees are pruned so that in each leaf the values have variance greater than 0.1. This essentially requires the ten (or more)

Table 2: Simulation results when imputing sensitive variables: Simple estimands

Estimand	Q	Avg. $\bar{q}_5$	95% CI Coverage Observed	Synthetic
Average income	52632	53351	94.6	67.4
Average social security	2229	2229	94.6	94.5
Average child support	139	136	93.7	94.0
Average alimony	41	42	91.3	91.1
% of households with income > 200,000	2.10	2.34	94.8	69.1
% of households with social security > 10,000	10.53	10.27	94.2	87.6
Coefficient in regression of alimony on:				
Intercept	4315	6537	89.2	84.4
Income	.14	.074	64.1	68.8
Coefficient in regression of alimony on:				
Intercept	9846	10157	90.8	92.4
Child support	.078	.054	95.7	97.4
Coefficient in regression of social security on:				
Intercept	2999	2988	93.1	93.0
Income	-.015	-.014	94.2	92.4

observed values in each leaf not to be identical. In these data, the trees are almost never pruned, allowing us to illustrate, for all practical purposes, the highest attainable utility for this population when using the CART algorithm to simulate these variables.

Table 2 and Table 3 summarize the results of 1000 runs of the simulation for a variety of estimands. Inferences are made using the methods of Section 2. In Table 2, the regressions include A on I and A on S for units with $A > 0$, and S on I for all units. In Table 3, the regression involving $\sqrt{S}$ uses only people with $S > 0$ and $G > 54$; the regression involving $\sqrt{C}$ uses only people with $C > 0$; and, the regression involving $\log(I)$ uses only people with $I > 0$. For all estimands, the finite population correction factor is used when determining the variances v. Reported statistics include the population values Q, the averages of the $\bar{q}_5$ across the 1,000 simulations, and the percentages of observed data 95% confidence intervals ($q_{obs} \pm 1.96\sqrt{v_{obs}}$) and synthetic data 95% confidence intervals that cover their corresponding Q.

For most estimands, the averages of the synthetic point estimates are close to their corresponding Q, and the coverages of synthetic 95% confidence intervals are reasonably close to the coverages for the cor-

Table 3: Simulation results when imputing sensitive variables: Model estimands

Estimand	Q	Avg. $\bar{q}_5$	95% CI Coverage Observed	Synthetic
Coefficient in regression of $\sqrt{C}$ on:				
Intercept	-93.28	-59.09	93.3	77.3
Indicator for sex=female	13.30	1.74	96.8	42.9
Indicator for race=black	-9.69	-6.72	96.6	94.9
Education	3.37	2.84	93.2	89.2
Number of youths in house	2.95	1.50	93.4	82.3
Coefficient in regression of $\sqrt{S}$ on:				
Intercept	79.87	82.87	94.7	92.4
Indicator for sex=female	-13.30	-12.84	95.2	92.5
Indicator for race=black	-5.85	-4.62	94.9	89.0
Indicator for race=American Indian	-7.00	-5.28	94.2	97.9
Indicator for race=Asian	-3.27	-2.09	90.1	98.2
Indicator for marital status=married in armed forces	2.08	-0.51	92.7	89.9
Indicator for marital status=widowed	7.30	6.47	94.1	89.7
Indicator for marital status=divorced	-0.88	-1.07	94.7	93.4
Indicator for marital status=separated	-5.44	-4.64	96.0	97.9
Indicator for marital status=single	-1.54	-0.92	92.4	92.6
Indicator for education=high school	5.49	5.52	95.9	95.9
Indicator for education=some college	6.77	7.01	94.0	95.1
Indicator for education=college degree	8.28	8.85	92.8	93.6
Indicator for education=advanced degree	10.67	10.71	90.8	94.0
Age	0.21	0.17	94.7	89.5
Coefficient in regression of $\log(I)$ on				
Intercept	4.92	4.88	92.5	90.8
Indicator for race=black	-0.17	-0.17	95.5	95.0
Indicator for race=American Indian	-0.25	-0.25	87.5	87.8
Indicator for race=Asian	-0.0064	-0.0080	92.6	93.4
Indicator for sex=female	0.0035	-0.00090	97.6	96.9
Indicator for marital status=married in armed forces	-0.028	-0.030	95.4	95.7
Indicator for marital status=widowed	-0.015	-0.017	95.4	95.7
Indicator for marital status=divorced	-0.16	-0.17	93.4	93.5
Indicator for marital status=separated	-0.24	-0.24	88.1	88.9
Indicator for marital status=single	-0.17	-0.18	92.5	92.5
Education	0.11	0.11	92.9	90.8
Indicator for household size > 1	0.50	0.50	91.1	91.0
Interaction for females married in armed forces	-0.52	-0.52	91.9	92.1
Interaction for widowed females	-0.31	-0.30	96.9	97.0
Interaction for divorced females	-0.31	-0.30	94.1	93.3
Interaction for separated females	-0.52	-0.52	90.8	90.6
Interaction for single females	-0.32	-0.31	93.9	93.7
Age	0.044	0.044	93.5	93.5
Age^2	-0.00044	-0.00044	93.9	93.7
Property tax	0.000037	0.000040	56.5	57.7

Table 4: Attribute disclosure limitation in simulation of imputing sensitive variables

Variable	Min.	1st Quartile	Median
RMSE			
S	175	1445	2317
I	2138	20453	37664
C	200	1179	1915
A	1020	3040	5627
RelRMSE			
S	.02	.16	.25
I	.02	.15	.26
C	.08	.37	.57
A	.17	.41	.65

responding observed data intervals. Several estimands–in particular average income, percentage of incomes above \$200,000, and the coefficient of sex in the regression involving $\sqrt{C}$–have poor synthetic data confidence interval coverages even though the observed data intervals have near 95% coverage. This results from biases in the $\bar{q}_5$, stemming from imputation models that do not perfectly reflect relationships in the data.

To assess attribute disclosure risks for each $Y_{(k)}$, we assume the intruder would estimate unit j's outcome $Y_{(k),j}$ by averaging the unit's replaced values, $\hat{Y}_{(k),j} = \sum_{i=1}^{m} Y_{(k)rep,ij}$. We then calculate the root mean squared error ($RMSE$) and relative root mean squared error ($RelRMSE$) of this estimator for each unit:

$$RMSE_{(k),j} = \sqrt{(Y_{(k),j} - \hat{Y}_{(k),j})^2 + \sum_{i=1}^{m}(Y_{(k)rep,ij} - \hat{Y}_{(k),j})^2/((m-1)m)} \quad (5)$$

$$RelRMSE_{(k),j} = RMSE_{(k),j}/Y_{(k),j} \quad (6)$$

For any data set, the distributions of the $RMSE_{(k),j}$ and $RelRMSE_{(k),j}$ across all units with replaced values can be examined to ensure sufficient variability in the imputations. Table 4 displays averages across the 1000 simulation runs of various summaries of the distributions of these quantities. Median $RelMSEs$ are typically 25% or more, suggesting imputations for most units have a wide range of uncertainty. When imputers require larger errors, stricter disclosure criteria can be used to prune the trees.

4.2 Simulating key identifiers

Imputers may decide to replace selected units' values of key identifiers with multiple imputations. This approach aims to reduce the risks of re-identifications. We mimic it by considering G, M, X, and R to be key identifiers, and replace their values for the same set of units specified in Section 4.1. Other values are not replaced and are released in all d_i.

As before, each D comprises $n = 10,000$ randomly sampled households, and there are $m = 5$ synthetic data sets generated for each D. The sequential order of imputation is $G - M - X - R$, which is decreasing in the $P_{(k)}$. The tree for G, treated as a continuous variable, is pruned so that the observed ages in each leaf have variance greater than 0.1. Imputed ages are rounded to the nearest integer. The trees for M, X, and R are pruned so that the observed values in each leaf are not identical. This guarantees a non-zero chance for each unit that its imputed values will differ from the ones in D. There is hardly any pruning of the age tree, but the sex and race trees are pruned by typically about 33% to meet the disclosure criteria. The marital status tree typically is pruned only slightly.

Table 5 summarizes the results of 1000 runs of the simulation for estimands like those in Table 3. A few of the indicator variables from Table 3 are collapsed to speed up the simulations. Inferences for the averages of S, I, C, and A are not reported because they are identical to the observed data inferences. As a replacement, the average education level of black females is reported. For most estimands, the averages of the synthetic point estimates are close to their corresponding Q, and the coverages of the synthetic and observed data 95% confidence intervals are reasonably similar. The coefficient of sex in the regression involving $\sqrt{C}$ again has poor synthetic data confidence interval coverage, indicating that the relationship between sex and C is not easily captured using CART models.

To assess re-identification risks when releasing these partially synthetic data sets, we assume the intruder follows a simple strategy for guessing true values of the simulated key identifiers. For marital status, sex, and race, the intruder uses the most frequently occurring value among that unit's imputations. When all

Table 5: Simulation results when imputing key variables

Estimand	Q	Avg. $\bar{q}_5$	95% CI Coverage Observed	95% CI Coverage Synthetic
Avg. education for married black females	39.44	39.49	95.0	94.3
Coefficient in regression of $\sqrt{C}$ on:				
Intercept	-93.28	-84.77	94.1	92.7
Indicator for sex=female	13.30	4.84	95.7	68.5
Indicator for race=black	-9.69	-4.39	95.2	84.0
Education	3.37	3.35	93.8	93.6
Number of youths in house	2.95	2.59	92.5	91.5
Coefficient in regression of $\sqrt{S}$ on:				
Intercept	79.50	83.11	93.7	88.0
Indicator for sex=female	-13.34	-12.37	93.6	82.0
Indicator for race=black	-6.04	-5.70	94.6	94.1
Indicator for race=American Indian	-7.12	-4.27	93.3	96.1
Indicator for race=Asian	-3.22	-2.00	90.5	96.7
Indicator for marital status=widowed	7.37	6.47	94.0	87.7
Indicator for marital status=divorced	-0.79	-0.83	93.4	95.8
Indicator for marital status=single	-1.46	-0.07	94.0	93.8
Indicator for education=high school	5.51	5.36	96.4	96.2
Indicator for education=some college	6.78	6.56	94.5	93.9
Indicator for education=college degree	8.31	8.12	92.9	93.4
Indicator for education=advanced degree	10.72	11.14	91.2	90.7
Age	0.22	0.16	93.9	87.5
Coefficient in regression of $\log(I)$ on				
Intercept	4.92	4.93	92.4	92.7
Indicator for race=black	-0.17	-0.17	93.5	92.5
Indicator for race=American Indian	-0.25	-0.24	88.6	90.7
Indicator for race=Asian	-0.0064	-0.0015	91.5	90.7
Indicator for sex=female	0.0035	-0.0030	96.0	94.0
Indicator for marital status=married in armed forces	-0.028	-0.11	94.4	88.6
Indicator for marital status=widowed	-0.015	-0.083	95.8	80.0
Indicator for marital status=divorced	-0.16	-0.16	92.1	92.3
Indicator for marital status=separated	-0.24	-0.23	87.6	89.9
Indicator for marital status=single	-0.17	-0.17	92.1	94.3
Education	0.11	0.11	94.3	93.7
Indicator for household size > 1	0.50	0.50	92.7	92.7
Interaction for females married in armed forces	-0.52	-0.40	91.0	84.6
Interaction for widowed females	-0.31	-0.25	97.0	86.3
Interaction for divorced females	-0.31	-0.29	93.0	94.0
Interaction for separated females	-0.52	-0.48	89.2	89.8
Interaction for single females	-0.32	-0.30	92.0	91.4
Age	0.044	0.043	94.4	93.6
Age^2	-0.00044	-0.00043	94.8	95.1
Property tax	0.000037	0.000040	51.7	52.0

five of a unit's imputations are unique, the intruder picks one at random. Using this strategy, typically an intruder matches exactly the martial status, sex, and race in 53% of the units with replaced data. For age, we consider two intruder strategies: (i) use the most frequently occurring value among the unit's imputed ages, and (ii) use the average of the unit's imputed ages. Using the first strategy, typically .002% of the intruder's guesses match exactly on all four key identifiers. Using the second strategy, typically 2.6% of the guesses match on all four key identifiers. With either strategy, about 12.5% of the guesses have, simultaneously, exact matches on marital status, sex, and race, and ages within two years of the age in the observed data. Clearly, simulating age accounts for most of the disclosure protection.

CART imputations for key identifiers can be sensitive to the disclosure criteria. For the categorical variables, requiring leaves not to have more than 90% of any one value typically results in pruned sex and race trees with just a handful of splits, producing conditional independences in the imputations. Using the 90% criterion to generate synthetic data, four of the synthetic 95% confidence intervals have less than 1% coverage, and five have between 1% and 50% coverage. The gains in disclosure protection are not large: 2.2% of units match on all four characteristics as compared to 2.6%, and 46% match on all characteristics but age as compared to 53%. These reductions in disclosure risk are not worth the large sacrifices in utility.

5 Concluding Remarks

The simulations in this article suggest that CART models have promise as a method for generating partially synthetic data sets. The primary drawback of the approach is the sequential nature of the imputations, which can introduce conditional independence structures into the released data. This issue also affects the use of CART models, or any sequential imputation scheme, for imputation of missing data. Further research would help quantify the sensitivity of inferences from multiply-imputed data sets to the ordering of the variables used in the sequential imputations.

This article represents some initial results on the use of CART models for imputations, and it suggests

topics for future research. The potential advantages of CART models over parametric models could be quantified, at least somewhat, by simulation studies comparing CART-generated and parametric imputations in a realistically complex partially synthetic setting. It also would be informative to investigate the payoffs to using Bayesian approaches to generating trees (Denison *et al.*, 1998a; Chipman *et al.*, 1998, 2000). Additionally, there may be some advantage to using multivariate adaptive regression splines (Friedman, 1991; Denison *et al.*, 1998b) to build trees instead of CART algorithms.

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 Applications for Statistical Agencies*, 215–277. Amsterdam: North-Holland.

Barcena, M. J. and Tussel, F. (2000). Multivariate data imputation using trees. In *COMPSTAT–Proceedings in Computational Statistics, 14th Symposium*, 193–204.

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

Chipman, H., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search (with discussion). *Journal of the Statistical Association* **93**, 935–960.

Chipman, H., George, E. I., and McCulloch, R. E. (2000). Hierarchical priors for Bayesian CART shrinkage. *Statistics and Computing* **10**, 17–24.

Clark, L. and Pregibon, D. (1992). Tree-based models. In J. Chambers and T. Hastie, eds., *Statistical Models in S*. Belmont, CA: Wadsworth, Inc.

Conversano, C. and Siciliano, R. (2002). Tree based classifiers for conditional incremental missing data imputation. In *Proceedings of Data Clean 2002 Conference*.

Dandekar, R. A., Cohen, M., and Kirkendall, N. (2002a). Sensitive micro data protection using Latin hypercube sampling technique. In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases*, 117–125. Berlin: Springer-Verlag.

Dandekar, R. A., Domingo-Ferrer, J., and Sebe, F. (2002b). LHS-based hybrid microdata versus rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases*, 153–162. Berlin: Springer-Verlag.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998a). A Bayesian CART algorithm. *Biometrika* **85**, 363–377.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998b). Bayesian MARS. *Statistics and Computing* **8**, 337–346.

Fienberg, S. E., Makov, U. E., and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics* **14**, 485–502.

Fienberg, S. E., Steele, R. J., and Makov, U. E. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: Data swapping and log-linear models. In *Proceedings of Bureau of Census 1996 Annual Research Conference*, 87–105.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics* **19**, 1–141.

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. Washington, D.C.: National Academy Press.

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

Liu, F. and Little, R. J. A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. In *ASA Proceedings of the Joint Statistical Meetings*, 2133–2138.

Piela, P. and Laaksonen, S. (2001). Automatic interaction detection for imputation–Tests with the WAID software package. In *Proceedings of Federal Committee on Statistical Methodology 2001 Conference*.

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

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. (2003a). Inference for partially synthetic, public use microdata sets. *Survey Methodology* forthcoming.

Reiter, J. P. (2003b). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Tech. rep., Institute of Statistics and Decision Sciences, Duke University.

Reiter, J. P. (2003c). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. Tech. rep., Institute of Statistics and Decision Sciences, Duke University.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63**, 581–592.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

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

Van Buuren, S. and Oudshoorn, C. G. M. (1999). *Flexible multivariate imputation by MICE*. Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054.

Wegman, E. J. (1972). Nonparametric probability density estimation. *Technometrics* **14**, 533–546.

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