Journal of Official Statistics, Vol. 21, No. 3, 2005, pp. 441–462

Using CART to Generate Partially Synthetic Public Use Microdata

J.P. Reiter[1]

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.

Key words: CART; confidentiality; disclosure; multiple imputation; synthetic data; trees.

1. Introduction

When releasing public use microdata, statistical agencies employ a variety of techniques to limit disclosures, including recoding variables, swapping data, and adding noise to values (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 nonstandard, likelihood-based analyses (Little 1993) or measurement error models (Fuller 1993). These are unfamiliar to many users, who are comfortable with and therefore likely to use standard statistical techniques and software.

An alternative approach with the potential to circumnavigate these problems is to release multiply-imputed, fully synthetic public use microdata, as proposed by Rubin (1993). In this approach, the agency (i) randomly and independently samples units from the sampling frame to comprise each synthetic data set, (ii) imputes unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) releases multiple versions of these synthetic datasets to the public. This limits disclosure risk, 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 (2005b) – based on the concepts of multiple imputation (Rubin 1987) – the approach allows data users to obtain valid inferences using standard statistical methods and software. For discussions of synthetic approaches, see Fienberg et al. (1996; 1998), Dandekar (2002a; b), and Reiter (2002; 2005a).

[1] Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251, U.S.A. Email: jerry@stat.duke.edu

Acknowledgments: This work was supported by the U.S. Census Bureau through a contract with Datametrics Research. The author thanks four referees and an associate editor for their valuable comments and suggestions.

Although there are potentially great benefits to releasing fully synthetic data (Raghunathan et al. 2003; Reiter 2005a), 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). For example, agencies seeking to prevent identifications of certain records can simulate values of key variables – like age, race, sex, and marital status – for those records, but leave all other values unchanged. Agencies seeking to protect certain records' values of sensitive variables, like income or disease status, can replace those values with imputed values. The nature of the partial synthesis depends on the degree of confidentiality protection and level of data utility deemed acceptable by the agency.

Partial synthesis maintains many of the potential benefits of full synthesis – limiting disclosure risk while allowing users to obtain valid inferences using standard statistical software and methods – with decreased sensitivity to the specification of the imputation models, since only some values are imputed. Hence, inferences from partially synthetic data are generally less affected by artifacts of inaccurate imputation models than those from fully synthetic data. However, partially synthetic data carry greater disclosure risk than fully synthetic data, because the original units and some genuine values are released.

Several agencies and statistical researchers have adopted partially synthetic approaches to protecting public use data. In the Survey of Consumer Finances, the U.S. Federal Reserve Board replaces monetary values at high disclosure risk with multiple imputations, then releases these imputed values and the unreplaced, collected values (Kennickell 1997). The U.S. Bureau of the Census has adopted a partially synthetic approach to protect data in longitudinal, linked data sets (Abowd and Woodcock 2001). 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.

Even when simulating only a few variables, specification of imputation models can be daunting in surveys with hundreds of variables, some with distributions not easily modeled with standard parametric tools. It may therefore be advantageous to use nonparametric methods to generate imputations.

This article presents and evaluates a nonparametric approach for generating partially synthetic data: the use of classification and regression trees, typically abbreviated as CART (Breiman et al. 1984). The article is organized as follows. Section 2 reviews the notation and methods of inference for partially synthetic data developed by Reiter (2003). 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 types of disclosure risks and level of data utility that can be expected when using CART models to generate partially synthetic data. Section 5 concludes with a general discussion of partially synthetic data approaches.

2. Description of Partially Synthetic Data

To describe partially synthetic data, we use the notation of Reiter (2003). 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. Methods for handling simultaneously missing data and synthetic data are described in Reiter (2004). Let X be the N x 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, then comprises $(X, Y_{rep,i}, Y_{nrep}, I, Z)$. 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.

To protect confidentiality it may be necessary to simulate values of the design variables X. If so, these values are considered part of $Y_{rep,i}$. To keep notation simple, we assume that no values of X are simulated.

The values in Z can and frequently will depend on the values in D. For example, the imputer may choose to simulate sensitive variables or identifiers only for units in the sample with unusual combinations of identifiers; or, the imputer may replace only those incomes above $100,000 with imputed values. To avoid bias, imputers should account for such selections by imputing from the distribution of Y for those units with $Z_j = 1$. In practice, this can be done by using only the units with $Z_j = 1$ as the data when finding the distributions for imputations. Using all units with $I_j = 1$ can result in biased estimates or wider confidence intervals with overly conservative coverage rates, as illustrated in the simulations of Reiter (2003).

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 $Z_j = 1$. This approach is 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 $\upsilon_p = (m-1)(1 + r_m^{-1})^2$, 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 (2003). Extensions for multivariate Q are presented in Reiter (2005b).

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. Many pruning criteria remove leaves that do not add much to the explanatory power of the tree. For example, the branch for L_2 and L_3 might be cut if the distribution of Y in L_2 is very similar to the distribution of Y in L_3. Pruning unimportant leaves of a tree is analogous to

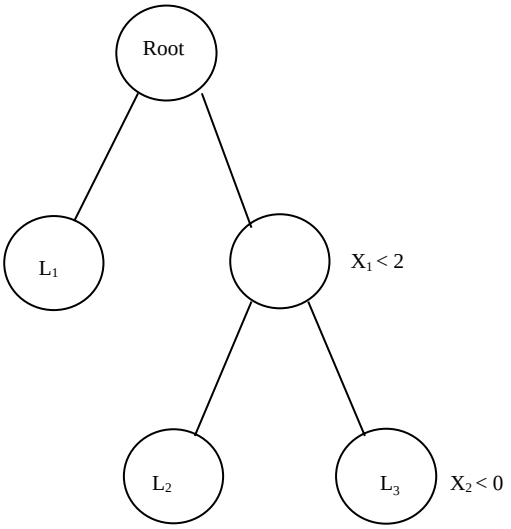

Fig. 1. Example of a tree structure

removing unimportant variables from standard regression models, and hence can reduce the variances of out-of-sample predictions. Of course, pruning important branches can introduce substantial bias, just like removing important predictors in standard regression.

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 parametric modeling, particularly for data with irregular distributions. Second, CART models can capture nonlinear 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 relationships are accurately described 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 less straightforward 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 knowledge, there have been no evaluations published of the repeated-sampling properties of inferences from multiply-imputed data sets generated from such chained CART models. A related approach to multivariate CART imputation has been used by the Euredit project (results available at http://www.cs.york.ac.uk/euredit).

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 $K = \prod_i^r n_i$ levels, where n_i is the number of levels for categorical variable i (Barcena and Tussel 2000). This can be computationally infeasible for large K.

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 nonhomogeneous 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 of missing data.

3.2. *Generation of $\mathbf{Y_{rep,i}}$ 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 laid out in Section 3.2.1, and motivation for its specification is presented in Section 3.2.2.

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 $TREE_{(k)}$. Whenever practical, only units with $Z_{(k)} = 1$ are used to grow $TREE_{(k)}$. 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 $TREE_{(k)}$ are fit for each of these variables. Let $P_{(k)}$ be the depth in $TREE_{(k)}$ of the first split on one of these other variables. If none of these other variables appear in $TREE_{(k)}$, define $P_{(k)} = \infty$. Now, reorder 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 $TREE_{(a)}$ than Y_a appears in $TREE_{(b)}$, the $P_{(b)} > P_{(a)}$, and we impute Y_b before Y_a.

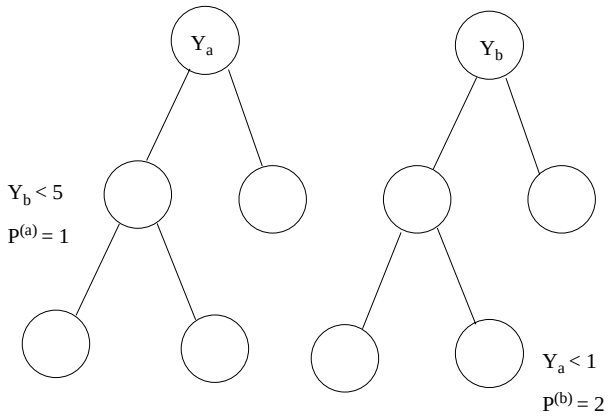

Fig. 2. Example of ordering of imputations when two variables have equal numbers of replaced values. Here, Y_b is imputed before Y_a

At its largest, each $TREE_{(k)}$ 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 far as possible the relationships in D, while limiting disclosure risks. For continuous $Y_{(k)}$, one approach is pruning until the observed values in all leaves have variance larger than some imputer-defined threshold, thereby ensuring that replacement values are imputed from distributions with some minimum variance. For categorical $Y_{(k)}$, trees can be pruned so that no one value of $Y_{(k)}$ appears in any leaf more than an imputer-specified percentage of the time. Another approach is to require a minimum number of units, say ten, in each leaf of the tree. It is also possible to use pruning criteria based on formal measures of disclosure risk, for example pruning until probabilities of identification, or mean squared errors of imputed values, for the records with $Z_j = 1$ are deemed sufficiently small.

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 $TREE_{(1)}$, 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 3.2.2). When it is safe to release real values of $Y_{(1)}$, these sampled values are the replacement imputations, $Y_{(1)rep,i}$, for the $n_{L_{1w}}$ units that belong to L_{1w}. When it is not safe to release real values of $Y_{(1)}$, we take an additional step. In each leaf, we estimate a density by fitting a Gaussian kernel density estimator (Wegman 1972) to the bootstrapped values. Then, for each unit, we sample randomly from the estimated density in that unit's leaf. The support of the estimated density stretches from the smallest to the largest value of $Y_{(1)}^{L_{1w}}$. 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 $TREE_{(2)}$ are located using $Y_{(1)rep,i}$ in place of $Y_{(1)}$. Occasionally, some units may have combinations of $(X, Y_{-(1,2)}, Y_{nrep}, Y_{(1)rep,i})$, that do not belong to one of the leaves of $TREE_{(2)}$. For these units, we search up the tree until we find a node that contains the combination, and 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 the kernel density procedure is used to impute $Y_{(1)}$. Imputing any $Y_{(k)}$ follows the same process: we place each unit in the leaves of $TREE_{(k)}$ based on their values in $(X, Y_{-(1,2,\ldots,k-1)}, Y_{nrep}, Y_{(1,2,\ldots,k-1)rep,i})$, and 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 each $TREE_{(k)}$, 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 above \$100,000 only, 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 subpopulations (e.g., replace incomes for single native American males), the imputations should be drawn from that subpopulation's outcome distribution. A tree grown using units outside the subpopulation may not accurately capture the outcome distribution in the subpopulation. As a result, the imputations for the subpopulation 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 $TREE_{(k)}$. 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. The $TREE_{(k)}$ 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 $TREE_{(a)}$, and vice versa, appropriately results in imputations that reflect dependencies between $Y_{(a)}$ and $Y_{(b)}$ (assuming $TREE_{(a)}$ splits on $Y_{(b)}$ and $TREE_{(b)}$ splits on $Y_{(a)}$). On the other hand, fitting $TREE_{(a)}$ without including $Y_{(b)}$, or vice versa, inappropriately produces imputations that reflect conditional independence of $Y_{(a)}$ and $Y_{(b)}$.

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 $TREE_{(b)}$ 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 $TREE_{(a)}$, dominated by the units with $Z_{(a)} = 1$ and $Z_{(b)} = 0$, may 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 $TREE_{(a)}$. 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. This aims to impute the variables in decreasing order of dependency on each other, which 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 $TREE_{(a)}$ before $Y_{(a)}$ appears in $TREE_{(b)}$, 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. When two or more variables have equal values of $Z_{(k)}$ and $P_{(k)}$, imputers can permute the orderings to determine which reproduces the joint distribution of the replaced values most closely.

Different sequencing and tree specification methods may be effective in certain partially synthetic settings. For example, suppose the imputer simulates all values of three key variables for certain records. Then, $f(Y_a, Y_b, Y_c|rest) = f(Y_a|rest)\ f(Y_b|Y_a, rest)f(Y_c|Y_a, Y_b, rest)$. The imputer can approximate draws from this distribution by drawing from three univariate trees that approximate the three conditional distributions. Research is needed to compare this specification with the specification used here.

A Bayesian bootstrap (Rubin 1981) is employed in each leaf as part of the imputation process. Let Y^L be the n_L values of the dependent variable in leaf L. The Bayesian bootstrap in leaf L proceeds as follows:

1. Draw $(n_L - 1)$ uniform random numbers. Sort these numbers in ascending order. Label these ordered numbers as $a_0 = 0, a_1, a_2, \ldots, a_{n_L-1}, a_{n_L} = 1$.
2. Draw n_L uniform random numbers, $u_1, u_2, \ldots, u_j, \ldots, u_{n_L}$. For each of these u, impute Y_j^L when $a_{j-1} < u \leq a_j$.

The Bayesian bootstrap draws values of Y from the leaf L, but it differs from the standard bootstrap. In Step 1, the resampling probabilities for each Y_j^L are randomly drawn; they do not all equal $1/n_L$. Varying the selection probabilities accounts for the additional uncertainty in the conditional distributions in each leaf due to having small samples of values in each leaf. Sampling values from Y^L directly, i.e., the standard bootstrap, does not incorporate this uncertainty. Justification for the Bayesian bootstrap over the standard one can be found in Rubin (1987, Chapter 4).

For some data, we take the additional step of drawing values from an estimated density, fit using the bootstrapped values and a kernel density estimator. 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 $Y_{(k)}$ in that leaf. If the range of values in any leaf is too narrow to protect confidentiality, the range can be extended beyond the range of values $Y_{(k)}$ in the leaf. Such extensions can compromise data utility if they result in many implausible imputations. To ensure the density can be reasonably estimated, the bootstrapped values within any L_{kw} cannot be all identical; they are redrawn if this is the case.

It is extremely difficult to prove analytically that sequential CART models are proper in the sense of Rubin (1987, Chapter 4). Indeed, it is difficult to prove that even simple regression imputation models are proper in complex samples (Binder and Sun 1996). The performance of imputation methods is therefore best evaluated in real data settings:

does the method provide inferences with reasonable frequentist properties? To assess this crucial question, it is helpful to simulate applications of the approach.

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. Census Bureau. Similar data are used by Reiter (2005a) to illustrate and evaluate releasing fully synthetic data.

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. There are negative incomes in the data: 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 may not reduce the risks of reidentifications, 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. The notes in Table 1 indicate the percentages of values of S, I, C, and A typically replaced in any D. Approximately 37.3% of the households have at least one value replaced, and about 1.5% have two or more values replaced. 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 for each variable are grown using only the units satisfying the conditions for that variable, e.g., the trees for C are fit using only the roughly 3.3% of households with $C > 0$. Trees are pruned so that each leaf has a minimum of ten values with at least two distinct values in each leaf. Requiring two distinct values per leaf ensures that the same value is not forced to be always imputed for units within that leaf.

Table 2 and Table 3 summarize the results of 1,000 runs of the simulation for a variety of estimands. Inferences are made using the methods of Section 2. 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. The median ratio of the mean squared error for $\bar{q}_5$ over the mean squared

Table 1. Description of variables used in the empirical studies

Variable	Label	Range	Notes
Sex	X	male, female	
Race	R	white, black, Amer. Indian, Asian	
Marital status	M	7 categories	
Highest attained education level	E	16 categories	
Age (years)	G	15 – 90	integers
Household alimony payments ($)	A	0 – 54,008	0.4% have $A > 0$
Child support payments ($)	C	0 – 23,917	3.3% have $C > 0$
Social security payments ($)	S	0 – 50,000	23.6% have $S > 0$
Household property taxes ($)	P	0 – 99,997	64.8% have $P > 0$
Household income ($)	I	$-21{,}011$ – 768,742	11.7% have $I > 100,000$

error for q_{obs} equals 1.06, indicating that most synthetic and observed point estimators yield similar estimates. The coverages of synthetic 95% confidence intervals are reasonably close to the coverages for the corresponding observed data intervals, with the exception of the coefficient of sex in the regression involving $\sqrt{C}$. This results because the tree for C rarely splits on sex. This forces a conditional independence between C and X in the imputation of positive C, which is reflected in the near-zero coefficient of X in Table 2. The ratio of mean squared errors for this coefficient is a very large 5.64, again reflecting the bias due to the implied conditional independence.

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, $\bar{Y}_{(k),j} = \Sigma_{\iota=1}^{m} Y_{(k)rep,ij}$. We then calculate the root mean squared error (*RMSE*) and relative root mean squared error (*RelRMSE*) of this estimator for each unit:

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

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

Table 2. Simulation results when imputing sensitive variables: Simple estimands and a multiple regression involving child support payments

			95% CI Coverage	
Estimand	Q	Avg. $\bar{q}_5$	Observed	Synthetic
Average income	52,632	52,893	96.4	92.6
Average social security	2,229	2,225	94.9	94.8
Average child support	139	137	93.9	92.6
Average alimony	41	42	92.5	92.4
% of households with income $> 200,000$	2.10	2.10	95.3	95.9
% of households with social security $> 10,000$	10.53	10.25	96.5	85.4
Coefficient in regression of A on:				
Intercept	4,315	6,087	89.6	88.6
Income	.14	.08	67.7	73.8
Coefficient in regression of A on:				
Intercept	9,846	10,046	92.2	92.9
Child support	.078	.065	97.2	96.4
Coefficient in regression of S on:				
Intercept	2,999	3,017	93.7	92.0
Income	$-.015$	$-.015$	93.0	91.0
Coefficient in regression of $\sqrt{C}$ on:				
Intercept	-93.28	-64.91	94.7	79.8
Indicator for sex = female	13.30	1.57	96.0	38.1
Indicator for race = black	-9.69	-6.49	96.9	93.4
Education	3.37	3.01	95.2	89.8
Number of youths in house	2.95	1.69	93.1	82.5

Population means and percentages calculated using all records. See Table 1 for percentages of imputed values.
Alimony regressions fit using records with $A > 0$. 100% of these records have imputed A.
Social security regression fit using all records. 33% of these records have imputed S or I.

Table 3. Simulation results when imputing sensitive variables: Multiple regressions involving incomes and social security payments

Estimand	Q	Avg. $\bar{q}_5$	95% CI Coverage	
			Observed	Synthetic
Coefficient in regression of $\sqrt{S}$ on:				
Intercept	79.87	82.97	93.7	84.6
Indicator for sex = female	−13.30	−12.94	94.2	89.5
Indicator for race = black	−5.85	−4.68	95.5	84.7
Indicator for race = American Indian	−7.00	−5.01	94.3	96.7
Indicator for race = Asian	−3.27	−2.11	90.2	96.2
Indicator for marital status = married in armed forces	2.08	−0.71	92.6	84.2
Indicator for marital status = widowed	7.30	6.47	95.2	88.4
Indicator for marital status = divorced	−0.88	−1.12	95.1	91.3
Indicator for marital status = separated	−5.44	−4.67	96.6	97.0
Indicator for marital status = single	−1.54	−1.05	93.9	91.2
Indicator for education = high school	5.49	5.60	95.3	92.3
Indicator for education = some college	6.77	7.13	96.3	93.9
Indicator for education = college degree	8.28	9.10	93.7	88.3
Indicator for education = advanced degree	10.67	11.90	89.2	90.6
Age	0.21	0.17	94.1	85.1
Coefficient in regression of log(I) on				
Intercept	4.92	4.90	92.9	93.2
Indicator for race = black	−0.17	−0.17	94.5	94.4
Indicator for race = American Indian	−0.25	−0.25	89.5	89.0
Indicator for race = Asian	−0.0064	−0.010	92.5	92.8
Indicator for sex = female	0.0035	−0.0011	96.9	96.4
Indicator for marital status = married in armed forces	−0.52	−0.52	94.5	95.5
Indicator for marital status = widowed	−0.31	−0.30	96.5	96.6
Indicator for marital status = divorced	−0.31	−0.30	94.1	93.8
Indicator for marital status = separated	−0.52	−0.52	88.8	89.0
Indicator for marital status = single	−0.32	−0.31	92.7	92.7
Education	0.11	0.11	93.0	92.9

Table 3. Continued

Estimand	Q	Avg. $\bar{q}_5$	95% CI Coverage	
			Observed	Synthetic
Indicator for household size > 1	0.50	0.50	93.0	93.2
Interaction for females married in armed forces	−0.52	−0.52	92.5	92.4
Interaction for widowed females	−0.31	−0.30	95.6	95.8
Interaction for divorced females	−0.31	−0.30	94.6	94.5
Interaction for separated females	−0.52	−0.52	91.1	91.0
Interaction for single females	−0.32	−0.31	90.8	91.0
Age	0.044	0.044	93.1	93.2
Age^2	−0.00044	−0.00044	93.4	93.3
Property tax	0.000037	0.000040	52.3	53.1

Social security regression fit using records with $S > 0$ and $G > 54$. 100% of these records have imputed S.

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 1,000 simulation runs of various summaries of the distributions of these quantities. Median *RelMSEs* are typically around 24% 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 reidentifications. We mimic it by considering G, M, X, and R to be key identifiers, and replace their values for the households in the union of households with $S > 0$, $A > 0$, $C > 0$, or $I > 100,000$. Other values are not replaced and are released in all d_i. Typically, about 37.3% of sampled households have G, M, X, and R replaced.

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 $P_{(k)}$. Each tree is grown using the union of households with $S > 0, A > 0, C > 0$, or $I > 100,000$. The trees are pruned to have a minimum of ten observations in each leaf. It is assumed that observed data values of G, M, X, and R are safe to release, so that imputations are drawn using only the Bayesian bootstrap in each leaf.

Table 5 summarizes the results of 1,000 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. Instead, the table reports the average education level of married black females, about 1.1% of the population.

For most estimands, the averages of the synthetic point estimates are close to their corresponding Q. The median ratio of the mean squared error for $\bar{q}_5$ over the mean squared error for q_{obs} equals 1.10, indicating that most synthetic and observed point estimators yield similar estimates. The maximum ratio is 2.71, belonging to the coefficient of age in the regression for $\sqrt{S}$. The coverages of the synthetic and observed data 95% confidence intervals are reasonably similar. The synthetic and observed intervals differ most for the

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

Variable	Min.	1st Quartile	Median
RMSE			
S	168	1,365	2,194
I	2,078	18,832	34,336
C	190	1,038	1,806
A	931	2,833	5,201
RelRMSE			
S	.02	.15	.24
I	.02	.14	.24
C	.08	.34	.55
A	.16	.39	.62

Table 5. Simulation results when imputing key variables

Estimand	Q	Avg. $\bar{q}_5$	95% CI Coverage	
			Observed	Synthetic
Avg. education for married black females	39.44	39.46	94.4	94.1
Coefficient in regression of $\sqrt{C}$ on:				
Intercept	−93.28	−88.11	94.5	93.8
Indicator for sex = female	13.30	7.46	96.2	81.3
Indicator for race = black	−9.69	−5.26	94.3	88.2
Education	3.37	3.38	94.2	94.5
Number of youths in house	2.95	2.67	93.9	93.6
Coefficient in regression of $\sqrt{S}$ on:				
Intercept	79.50	83.79	94.6	81.3
Indicator for sex = female	−13.34	−12.94	93.8	91.3
Indicator for race = black	−6.04	−6.12	94.5	94.2
Indicator for race = American Indian	−7.12	−4.48	94.7	95.0
Indicator for race = Asian	−3.22	−2.19	89.3	94.7
Indicator for marital status = widowed	7.37	7.20	94.5	94.2
Indicator for marital status = divorced	−0.79	−0.96	93.7	96.4
Indicator for marital status = single	−1.46	0.18	93.8	92.3
Indicator for education = high school	5.51	5.53	94.8	95.8
Indicator for education = some college	6.78	6.77	94.5	94.8
Indicator for education = college degree	8.31	8.12	92.7	92.4
Indicator for education = advanced degree	10.72	10.99	89.1	90.6
Age	0.22	0.16	93.8	80.6
Coefficient in regression of log(I) on:				
Intercept	4.92	4.95	91.2	90.2
Indicator for race = black	−0.17	−0.17	94.9	94.3
Indicator for race = American Indian	−0.25	−0.25	88.6	91.0
Indicator for race = Asian	−0.0064	−0.0045	92.5	92.0
Indicator for sex = female	0.0035	−0.0018	96.2	95.5
Indicator for marital status = married in armed forces	−0.028	−0.091	94.9	90.4

Table 5. Continued

Estimand	Q	Avg. $\bar{q}_5$	95% CI Coverage	
			Observed	Synthetic
Indicator for marital status = widowed	−0.015	−0.057	96.6	89.4
Indicator for marital status = divorced	−0.16	−0.16	93.5	93.9
Indicator for marital status = separated	−0.24	−0.23	87.3	88.5
Indicator for marital status = single	−0.17	−0.17	93.3	94.1
Education	0.11	0.11	93.0	92.2
Indicator for household size > 1	0.50	0.50	93.5	92.1
Interaction for females married in armed forces	−0.52	−0.43	92.2	88.9
Interaction for widowed females	−0.31	−0.27	96.8	90.0
Interaction for divorced females	−0.31	−0.30	92.8	93.1
Interaction for separated females	−0.52	−0.48	89.0	89.1
Interaction for single females	−0.32	−0.31	92.2	92.7
Age	0.044	0.043	94.1	91.3
Age^2	−0.00044	−0.00043	94.4	92.8
Property tax	0.000037	0.000040	51.8	51.8

Average education calculated using all black females. 29.2% of these records have imputed G, M, X, and R.
Child support regression fit using records with $C > 0$. 100% of these have imputed G, M, X, and R.
Social security regression fit using records with $S > 0$ and $G > 54$. 100% of these have imputed G, M, X, and R.

aforementioned coefficient of age and the coefficient of sex in the regression for $\sqrt{C}$. These differences result because the CART models inadequately capture these relationships.

To assess reidentification 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 five of a unit's imputations are unique, the intruder picks one at random. Using this strategy, typically an intruder matches exactly the marital status, sex, and race in 54% 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 3.0% of the intruder's guesses match exactly on all four key identifiers. Using the second strategy, typically 2.7% 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 especially sensitive to the pruning criteria. 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 54%. 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 are a promising approach for generating partially synthetic data sets. Most synthetic confidence intervals in the simulation have coverage properties like those of the corresponding real-data intervals, even for regression models that have 100% imputed dependent variables. The approach can be used to limit attribute or identification disclosures, or even both simultaneously, depending on the variables synthesized. The degree of confidentiality protection can be assessed by modeling intruder behavior and attempting attribute or identification disclosures. When the partially synthetic data are found not to provide sufficient protection, imputers can prune branches off the trees used to generate synthetic data.

One notable exception to the overall reasonableness of the synthetic data inferences is the coefficient of sex in Table 2. It is attenuated to zero because the tree for child support payments typically fails to split on sex. These conditional independences are most likely to occur when trees are built from only a small number of units. Imputers may not want to use CART models in such cases, opting instead for parametric imputation models.

Imputers should provide information that helps users decide what inferences can be supported by the synthetic data. For example, imputers can include the imputation models as attachments to public releases of data. When using CART models, this could raise

confidentiality risks if the branching points are sensitive. Alternatively, imputers can include generic statements that describe the imputation models, such as "The tree for child support payments is built from 40 records with positive payment values. It splits on age and income only." Analysts who desire finer detail than afforded by the imputations may have to apply for special access to the observed data.

As argued by Reiter (2005a), releasing or describing the imputation models is necessary, but it is not sufficient: imputers also should release synthetic data generated from the models. Some analysts are not able to generate synthetic data given the models; they need imputers to do it for them. Even when analysts can do so, it is a cumbersome burden to place on them. Additionally, when analysts want to compare competing analyses, it is advantageous if these analyses are performed on the same data sets, thereby eliminating simulation variance from comparisons. Finally, analysts may desire some function of the synthetic data that is hard to estimate from the model parameters, but easy to determine from the synthetic data.

CART models can produce implausible imputations if built carelessly. For example, suppose imputers replace both a detailed race code and an indicator for whether or not the person is a minority. Imputing both separately might produce a nonminority race with a minority indicator. To avoid such inconsistencies, imputers should simulate the most detailed variables before creating any derived variables from the imputed values. Imputers can check manually for inconsistent data before releasing the synthetic data and, as an overall check on the accuracy of the imputation models, can compare the distributions of the synthetic data to those of the observed data being replaced. If the synthetic distributions are too dissimilar from the observed ones, the imputation models should be altered.

Synthetic data methods, or any other disclosure limitation method, cannot be guaranteed to reproduce closely the results for all analyses of the observed data. Nonetheless, synthetic data can serve effectively for wide classes of simple analyses, as illustrated by the simulations in this article. Synthetic data also could play the role of training data: users build their models based on the publicly available, synthetic data, and submit requests to the imputers for results based on the original data.

As with all disclosure limitation strategies, partially synthetic data do not eliminate the risk of disclosures. Users can utilize the released, unaltered values to facilitate disclosure attacks. Additionally, users may be able to estimate actual values of Y from the synthetic data with reasonable accuracy. For example, if all people in a certain demographic group have the same value of an outcome variable, the CART models likely will generate that value for imputations. Imputers may need to prune the trees or otherwise coarsen the imputations for these people. As another example, if users know that a certain record has the largest value of some Y in the database, they can obtain a lower bound for Y by taking the maximum value of the synthetic Y. Such outlying records are difficult to protect using synthetic data approaches.

Is releasing partially synthetic data generated from CART models an effective approach to disclosure limitation? This question cannot be fully answered from this article, although the simulation results are encouraging. Within the partially synthetic data context, research is needed to compare the merits of CART models with those of parametric models. It also would be informative to investigate the use of Bayesian CART approaches

(Denison et al. 1998a; Chipman et al. 1998; 2000) or multivariate adaptive regression splines (Friedman 1991; Denison et al. 1998b) as imputation models. More broadly, research is needed to compare synthetic data methods with standard disclosure limitation methods. These comparisons should focus on measures of disclosure risks, obtained by simulating intruder behavior, and on measures of data utility for estimands of interest to users, including properties of point and interval estimates. Simulation studies in genuine, realistically complex settings would provide valuable information for gauging the risk-utility trade-offs for the various approaches to disclosure limitation.

6. References

Abowd, J.M. and Woodcock, S.D. (2001). Disclosure Limitation in Longitudinal Linked Data. In Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds). North-Holland: Amsterdam, 215–277.

Barcena, M.J. and Tussel, F. (2000). Multivariate Data Imputation Using Trees. COMPSTAT–Proceedings in Computational Statistics, 14th Symposium, 193–204.

Binder, D.A. and Sun, W. (1996). Frequency Valid Multiple Imputation for Surveys with a Complex Design. Proceedings of the American Statistical Association, Section on Survey Research Methods, 281–286.

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 American 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 Statistical Models in S, J. Chambers and T. Hastie (eds). Belmont, CA: Wadsworth, Inc., 377–420.

Conversano, C. and Siciliano, R. (2002). Tree Based Classifiers for Conditional Incremental Missing Data Imputation. Proceedings of Data Clean 2002 Conference, 30–34.

Dandekar, R.A., Cohen, M., and Kirkendall, N. (2002a). Sensitive Micro Data Protection using Latin Hypercube Sampling Technique. In Inference Control in Statistical Databases J. Domingo-Ferrer (ed.). Berlin: Springer-Verlag, 117–125.

Dandekar, R.A., Domingo-Ferrer, J., and Sebe, F. (2002b). LHS-based Hybrid Microdata versus Rank Swapping and Microaggregation for Numeric Microdata Protection. In Inference Control in Statistical Databases, J. Domingo-Ferrer (ed.). Berlin: Springer-Verlag, 153–162.

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. (1998b). 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. Proceedings of the U.S. Bureau of the Census 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 Record Linkage Techniques, W. Alvey and B. Jamerson (eds). Washington, D.C. National Academy Press, 248–267.

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. Proceedings of the Joint Statistical Meetings of the American Statistical Association, 2133–2138.

Piela, P. and Laaksonen, S. (2001). Automatic Interaction Detection for Imputation – Tests with the WAID Software Package. Proceedings of Federal Committee on Statistical Methodology 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. (2003). Inference for Partially Synthetic, Public Use Microdata Sets. Survey Methodology, 29, 181–189.

Reiter, J.P. (2004). Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. Survey Methodology, 30, 235–242.

Reiter, J.P. (2005a). 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. (2005b). Significance Tests for Multi-component Estimands from Multiply-imputed, Synthetic Microdata. Journal of Statistical Planning and Inference, 131, 365–377.

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 and 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.

Received June 2003
Revised September 2004