Microdata Simulation for Confidentiality Protection Using Regression Quantiles and Hot Deck

Jennifer C. Huckett, Michael D. Larsen
Iowa State University, Department of Statistics, Ames Iowa 50011, jhuckett@iastate.edu

Abstract

Government agencies must simultaneously maintain confidentiality of individual records and disseminate useful microdata. We study options for creating full synthetic data files for public release. Specifically, we study combining quantile regression, hot deck imputation, and additional confidentiality-preserving methods to produce releasable, usable data. The result of the implementation of our ideas is a releasable data set containing original values for a few key variables, synthetic values for several variables, and perturbed values for remaining variables. The procedure should simulataneously provide quality data to the user and protect the confidentiality of the respondents. In this paper we describe quantile regression, hot deck imputation, and rank swapping and present results from an application of generating synthetic values using quantile regression for veterans data in the American Community Survey at the U.S. Census Bureau.

KEY WORDS: Hot deck imputation; quantile regression; rank swapping; statistical disclosure limitation; synthetic data.

1. Introduction

Federal statistical agencies exist in the United States and other countries to inform the public on matters that affect the welfare of the people, both individually and collectively. Each of over 70 statistical agencies in the United States was founded in response to specific needs for data about critical areas in public policy (Duncan, et al. 2001). How do federal agencies fill these needs? They collect and disseminate quality data and information to users such as policymakers and researchers. There are several challenges faced by an agency to collect quality data; these topics are not covered in this paper. Here we focus on the challenges associated with disseminating quality data once the data have been collected.

From a user's perspective, an ideal data product might be the actual data collected by the agency, perhaps cleaned, audited, or with missing values filled-in. However, agencies are bound by legal and internal obligations to protect the identities of individuals and organizations from whom they collect data, making it impossible for them to provide the user's ideal. This challenge is reflected by many agencies in their published mission and policy statements. These documents include explicitly stated goals to

- *collect and disseminate quality information (data)* , and
- *uphold privacy and protect confidentiality.*

In the field of statistical disclosure limitation (SDL), addressing these competing goals is explored. One option is to restrict access to data through licensing agreements that specify severe penalties for inappropriate use of data and security protocols to prevent unintended access and use. A second restriction option is to use remote access data servers that allow users to request summaries and analyses of data, but do not actually provide unit-level data to users. In such an arrangement the requests for data summaries are monitored, possibly in an automated fashion, to prevent disclosure of sensitive information. A third option is to perturb or aggregate information on sensitive variables so that individual confidentiality is not compromised. Several techniques studied and implemented are based on preparing aggregated tabular data, identifying sensitive cells, perturbing the original microdata in specific ways to address the disclosure risk, and recomputing the aggregated tabular data. Other procedures for tables involve further aggregation of tabular cells, especially cells with small counts, or blanking out enough cells in a table, but releasing margins and the other non-sensitive cell counts, so that disclosure risk is sufficiently small. A fourth option consists of methods for perturbing the microdata before release. These techniques include noise addition and data swapping or rank swapping. Noise addition simply adds random errors generated from a distribution (parametric or empirical) to the observed values before release. Data swapping randomly switches values on some variables for some records. Rank swapping is data swapping for quantitative or multi-valued ordinal variables with some control for the degree of alteration in the records that is allowed. The issue of confidentiality protection is also referred to in the literature as disclosure control, disclosure prevention or avoidance, and inference control. Publications on the topic include Willenborg and de Waal (1996, 2001) Domingo-Ferrer (1999, 2002), Domingo-Ferrer and Franconi (2006), Domingo-Ferrer and Torra (2004), and Doyle et al. (2001), a special issue of the *Journal of Official Statistics* (1998), and several technical reports at the National Institute of Statistical Sciences (NISS).

The approach pursued in this paper is to simulate data for variables in a data set to produce synthetic microdata. A synthetic microdata data set is comprised of values that are simulated, hence artificial, but similar in important ways to the original unit-level data. The similarities might not be at the level of individual units. Rather, the similarities between the simulated and the original data sets should be apparent in marginal and conditional distributions of the values in the data. Simulation could be accomplished in numerous ways and involve various degrees of modeling assumptions. Rubin (1993) originally proposed creating a full artificial set of data for public release to satisfy confidentiality constraints. Reiter (2002, 2003, 2005), Abowd and Woodcock (2004), Raghunathan, Reiter and Rubin (2003), Little and Liu (2002, 2003) have considered methods for implementing the proposal to create such artificial data sets, applications of some methods are presented in Kinney and Reiter (2007) and Hawala (2003).

We propose an approach to SDL that combines traditional (perturbing data using hot deck imputation and rank swapping) and modern approaches (creating synthetic data using conditional quantile

regression models). We create a releasable data set containing original values for a few key variables, synthetic values for several variables, and perturbed values for remaining variables. The procedure should produce data that has high utility for inferential purposes and low disclosure risk.

In this paper we describe quantile regression, hot deck imputation, and rank swapping in detail and present results from an application of generating synthetic values using quantile regression for data on veterans in the American Community Survey at the U.S. Census Bureau. The proposed synthetic data method is described in Section 2. An application of simulating data using conditional quantile regression models is presented in Section 3. Some conclusions and planned future work are described in Section 4.

2. Proposed Synthetic Data Method

As the economy becomes more complex, and interactions among household, businesses, governments become more entangled, more detailed data is required for researchers to attempt to develop a full understanding of the economy and society (Doyle et al. 2001). We propose using quantile regression models to provide an accurate model for variables that have complex marginal and conditional distributions such as those found in large data sets collected and maintained by government statistical agencies and to simulate values from these models. Further, we propose to implement hot deck imputation and random rank swapping to fill-in values for other variables. This combination of methods can be used to produce a synthetic data set containing many variables for release. In principle, either simulations or a series of hot-deck imputations with rank swapping could be used to generate a large synthetic data set. The use of both corresponds to our involvement in applications for which there are a few key variables (modeled using quantile regression) and many other variables (imputed using a hot deck procedure) related to each other. Details on the proposed method are described in the following three sections.

2.1 Quantile Regression

Quantile regression uses a function of predictors X to model the distribution of random variable Y as at distinct quantiles in its distribution. Quantile regression can provide insight beyond what is learned in least squares regression if the relationship between Y and X differs depending on the portion of the distribution of Y being examined. For example, the effects describing the relationship between income and age may differ depending on whether the individuals have high income or low income. If this is the case, performing quantile regression at various quantiles can provide a better understanding of the relationship between age and income than mean regression might. As is often the case in large databases representing a wide range of respondents, variables such as age and wages have skewed and non-standard distributions whose relationships to other variables do indeed vary depending on where one looks in the distribution. In order to fully represent these (conditional) relationships, one can perform quantile regression at a set of quantiles in the interval [0, 1]. In order to create synthetic or artificial data that mimic these complex relationships, one can use the resulting quantile regression model estimates to simulate values conditional on predictors. In Section 2.1.1 we describe quantile regression as a general estimation procedure. In Section 2.1.2 we describe the details of using quantile regression to simulate data values. The use of the simulation procedure to create synthetic data for statistical disclosure limitation is outlined in Section 2.1.3.

2.1.1 Quantile Regression

Quantile regression is explored and described in Koenker (2005), Bassett and Koenker (1978), and Koenker and Hallock (2001), among articles by these authors. These, and additional papers by the authors, offer technical details and examples to illustrate quantile regression. Here, we summarize quantile regression as a general estimation procedure. Consider random variable Y to have right-continuous distribution function $F_Y(y) = P(Y \leq y)$. Use the distribution function to denote the τ^{th} quantile of Y as $F_Y^{-1}(\tau) = inf\{y : F_Y(y) \geq \tau\}$. Express the τ^{th} quantile of Y, conditional on predictor variables X, using a linear function $\xi(x, \beta_\tau) = x^T\beta_\tau$ and regression equation $Q_y(\tau|x) = \xi(x, \beta_\tau) + F^{-1}(\tau)$, where $F^{-1}(\tau)$ are independent and identically distributed (iid) errors. The coefficients β_τ can be estimated by minimizing $\sum \rho_\tau(y_i - \xi(x_i, \beta_\tau))$ over choices of β_τ.

The function ρ_τ is called the *tilted absolute value function* and has the form $\rho_\tau(y_i - \xi(x_i, \beta_\tau)) = (y_i - \xi(x_i, \beta_\tau))(\tau - I_{[y_i - \xi(x_i, \beta_\tau) < 0]})$, where $I_{[y_i - \xi(x_i, \beta_\tau) < 0]}$ is an indicator function with a value of 1 if $y_i - \xi(x_i, \beta_\tau) < 0$ and 0 otherwise. The minimization problem $min_{\beta_\tau \in \Re^p} \sum \rho_\tau(y_i - \xi(x_i, \beta_\tau))$ can be solved by reformulating it into the linear programming problem

$$min\{\tau 1_n^T u + (1-\tau)1_n^T v | 1_n \xi(x, \beta_\tau) + u - v = y\},$$

where $\{u_i, v_i\}$ are artificial variables corresponding to the positive and negative parts of the vector of residuals. If $e_i = y_i - \xi(x_i, \beta_\tau) < 0$, then $u_i = |e_i|$ and $v_i = 0$; if $e_i \geq 0$, then $u_i = 0$ and $v_i = e_i$. Estimation using this formulation can be performed in R using the *rq* function in the *quantreg* package (Koenker 2005).

2.1.2 Simulation From Quantile Regression Models

In this section, we describe simulating values of random variable Y at a particular quantile of its distribution. We can characterize the distribution of Y with respect to X using a linear model $\xi(X, \beta_\tau)$ for quantile $\tau \in (0, 1)$. We obtain estimates for β_τ using the methods described in Section 2.1.1 and use them to compute predicted values of the τ^{th} quantile of Y given the values of X in each record. For example, if $\tau = 0.5$, we estimate $\hat{\beta}_{\tau=0.5}$ in the quantile regression model to simulate Y. The resulting values are denoted $\hat{Y}_{\tau=0.5}$, the predictions of Y at the 0.5 quantile, or median, of its distribution, conditional on predictors X. Predicted values can be computed for each record, resulting in a set of values around the (conditional) median of Y.

Suppose now that Y and X represent the variables income and age, respectively, and that the relationship between income and age differs depending on whether individuals have high or low income. We choose to model income (or some transformation of it) as a linear function of age at the $\tau = 0.05$ and $\tau = 0.95$ quantiles in order to examine the low and the high levels of income with respect to age. We obtain regression estimates $\hat{\beta}_{\tau=0.05}$ and $\hat{\beta}_{\tau=0.95}$ which are then

used to compute predicted values of income at both quantiles. The result is two predicted values of income, $\hat{y}_{\tau=0.05}$ and $\hat{y}_{\tau=0.95}$, per record. Imagine extending this example to obtain predicted values at several quantiles for each record. Or, randomly selecting quantiles for which to generate predicted values. (There are limitations to the number of quantiles that make sense to estimate regression equations for. For example, with fifty data points it would hardly make sense to select more that fifty quantiles; see Koenker (2005) for a discussion of nonuniqueness.) We propose using randomly selected quantiles to perform SDL in confidential data sets.

2.1.3 Quantile Regression Simulation for SDL

The motivation for this work is a consideration of the challenges faced by statistical agencies to simultaneously disseminate quality data and protect confidentiality. In this section we describe the proposed method to address both challenges using quantile regression predictions for variables in a data set. Suppose variable X can be released to the public without concern about confidentiality, but that variables $Y_1, Y_2, ..., Y_s$ are sensitive. Quantile regression, as described in Sections 2.1.1 and 2.1.2, can be used to generate synthetic values according to the conditional distributions

$$\begin{aligned} &Y_1|X \\ &Y_2|X, Y_1 \\ &Y_3|X, Y_1, Y_2 \\ &\vdots \\ &Y_s|X, Y_1, Y_2, ..., Y_{s-1}. \end{aligned}$$

That is, given values for variable X, one can randomly generate quantiles for each record, compute the quantile regression of Y_1 on X at the chosen quantiles, and predict Y_1 based on the estimated quantile regression function. Given values for variables X and Y_1, one can randomly generate a second set of quantiles for each record, compute the quantile regression of Y_2 on X and Y_1 at the chosen quantiles, and predict Y_2 based on the estimated quantile regression function. This procedure continues through the prediction/simulation for variable Y_s.

A simplification could occur if not all Y values are needed in later predictions. For example, $Y_3|X, Y_2$ might be a sufficient model to predict Y_3. Such simplifications would correspond to assumptions about conditional independence between sets of variables. In other cases, it might be quite difficult to fit large models to predict a variable Y_j using all previous Y variables and X, especially in small to moderate size data sets with many variables. General considerations for statistical modeling and prediction will need to be considered when variables are selected for prediction. Note also that the sequential procedure above is designed to be expedient. An alternative would be to build a full model for the joint distribution of all variables and simultaneously generate a vector of values for all Y variables. In most large-scale surveys this will be prohibitively difficult. Future extensions could examine intermediate options between the sequential procedure adopted here and something closer to sampling from the full joint distribution.

For each sensitive variable Y_j we run through the procedure described in 2.1.1 and 2.1.2 to generate synthetic values for each record in the data set. Note that the regression coefficients depend on both the chosen τ and the predictor variables in the model. A different τ_j is chosen for each sensitive variable $j = 1, 2, ..., s$ and each record. Consider a single record. We randomly select τ_j from a $Uniform(0, 1)$ distribution. Using the regression estimates at quantile $\tau = \tau_j$ we compute the predicted value $\hat{Y}_{j,\tau_j}$, conditional on X values in that record. Suppose we do this for each of the s sensitive variables in the record. The result is one predicted value for each sensitive variable at its distinct randomly selected quantile, $\hat{Y}_{1,\tau_1^*}, \hat{Y}_{2,\tau_2^*}, \ldots, \hat{Y}_{s,\tau_s^*}$.

This process is repeated for each record in the data set to obtain synthetic values for each sensitive variable in every record. By generating the synthetic values in this way, we

- *maintain the distributions of each variable both marginally and conditionally with respect to the predictors in the quantile regression models* , and
- *create synthetic values on records, reducing the risk of identification and protecting the confidentiality of respondents.*

2.2 Hot Deck Imputation

Typically, hot deck imputation is used to handle missing-data problems in data sets with several variables. An inventory of various methods is presented in Little and Rubin (2002); we present a summary here. Broadly, two approaches are taken to impute values: explicit and implicit modeling. Explicit approaches include mean modeling, regression modeling, and stochastic regression modeling methods. Implicit approaches include hot deck, substitution, and cold deck imputation methods. It is also common to consider a combination of these methods to approach missing-data problems. We focus on hot deck imputation because it is a flexible imputation methodology that imputes actual values observed in the data set.

Hot deck imputation is a method in which individual values from complete records (donors) are drawn to fill in missing values of incomplete records where the complete and incomplete records are similar with respect to some variables with recorded values in both. For each incomplete record, potential donors can be identified based on their similarity. Picking the donor from the set of potential donors is done using a selection procedure. The value of the missing variable is imputed to the incomplete record from the selected donor. The term *hot deck* literally refers to computer cards that match on some characteristic in the complete and incomplete records due to the fact that the cards are sorted according to the characteristic (Little and Rubin 2002). In early applications at the U.S. Census, cards corresponded to households and were sorted by sequential address listing. Donors were determined according to this address order and a few other matching variables. The nearest eligible donor was selected. In Section 2.2.1 we list several options for implementing hot deck imputation. In Section 2.2.2 we consider using hot deck imputation for SDL.

2.2.1 Hot Deck Options

A hot deck imputation procedure involves identifying potential donors from which to impute values to incomplete records and selecting the donor record using one of several selection procedures. In Little and Rubin (2002), the authors present several approaches to determine which records match: exact matching with respect to some key fields, matching based on calipers (or ranges) of observed covariates, sequential matching ordered by covariates, and matching based on distance to nearest neighbors. Exact matching occurs when the potential donors have exactly the same values of the key variables as the unit with missing values on other variables. In many cases there will not be exact matches in a data set that contains a large collection of variables or variables with many distinct values, such as quantitative variables or multi-valued discrete variables such as county or race. Matching within calipers or ranges often is necessary for finding matches and accomplishes the goal of making donors and recipients very similar. For example, age is often matched within a range of ages. Sequential methods require exact matches on some variables and then attempt to match as closely as possible on others. For example, one could require matching on sex, county, and broad age ranges. Within this set of initially acceptable matches, additional matching requirements could be specified. If an exact match for all requirements is not possible, then matching criteria can be removed or relaxed one at a time until acceptable matches are found.

In nearest neighbor matching, a metric is defined to measure the distance between respondents in the data set, usually based on some covariate values. Values to impute are chosen from the respondents' records closest to the respondent with the missing value. Possible metrics include maximum deviation, predictive mean, and Mahalanobis distance. Mahalanobis distance is $d(i,j) = \sqrt{(y_i - y_j)^T S_{yy}^{-1} (y_i - y_j)}$, where y_i are values for variable Y in the complete records (potential donors) and y_j are values in the incomplete records. S_{yy} is the estimated variance of Y, or the variance-covariance matrix of variables in Y if matching is done on more than one variable. For each incomplete record j, record i with the smallest $d(i,j)$ is declared a match (the donor). The value of the variable missing from incomplete record j is imputed from complete donor record i (original data).

Nearest neighbor matching can be combined with other matching requirements. For example, nearest neighbor matching can be implemented on various quantitative variables among record residing in the same county and having the same gender or marital status. The distance metric alternatively could be used to define a 'neighborhood' of potential donors around the intended recipient. One could choose the closest M (say, $M = 5$) potential donors to define the neighborhood, or one could specify a distance threshold such that all potential donors within the specified distance comprise the neighborhood of the intended recipient. Instead of picking the closest record to donate values, some schemes call for randomly picking a donor from among the potential donors in the neighborhood.

In Section 2.2.2 we discuss the use of hot deck imputation via nearest neighbor matching within categories, using the Mahalanobis distance as the metric to measure closeness between complete and incomplete records, in order to implement SDL in confidential data sets. Other versions of hot deck imputation could similarly be used for disclosure limitation and could be examined in future work.

2.2.2 Hot Deck for SDL

To implement hot deck imputation for SDL, we consider the original data as the set of complete records (potential donors). No original records have missing values. The synthetic data (generated using quantile regression) are the set of incomplete records. All synthetic records have missing values for variables that are neither the set of variables that can be released to the public (X) nor the synthetic variables (the first few Y variables), but those remaining variables that cannot be released to the public without some type of SDL. We fill in the missing values on incomplete records using hot deck imputation via nearest neighbor matching within categories (defined by releasable variables x), using the Mahalanobis distance as the metric to measure closeness. Mahalanobis distance is used to compare the original and synthetic values within records with exactly matching categories. In the formulation of Mahalanobis distance in Section 2.2.1, the original data values, or complete records, are the y_i and the synthetic data values, or incomplete records are the y_j. The estimated S_{yy} and $d(i,j)$ are computed within categories of the releasable variables.

To use hot deck imputation for SDL, we implement the procedure in two stages. First, we match on one variable to determine a set of close original records. Second, we match on two or more variables among that set to determine the closest record to be declared a match. The result is the data set containing all of the original data that cannot be released and a second data set containing the original releasable variables, several synthetic variables, and several variables with imputed values. This choice of procedures reflects the large size of our intended applications (tax records for an entire state or a large survey for the U. S. Census). The initial categorization and first stage matching greatly reduce the number of distances that must be computed for each record.

To further decrease disclosure risk in our application, we go one step further and perturb the hotdeck imputed values by performing random rank swapping on the imputed values. This would not be necessary in some applications, but is contemplated here as further protection for extremely sensitive databases. A general description of rank swapping, or data swapping, is presented in Section 2.3.

As was mentioned previously, one could contemplate simply using hot deck methods sequentially on variables (or on a set of variables) one at a time in order to generate synthetic data. Due to particular interest in some variables in our applications and their extreme sensitivity, we initially generate values of some variables from the regression quantile models, and then apply hot deck imputation for remaining variables.

2.3 Rank Swapping

Data swapping is a procedure in which values from individual records are exchanged. It has been used as an SDL technique for both tabular data and microdata. The basic idea is that if a user looking at the released data set cannot know for sure which records are perturbed through swapping and which are in their original form, then

the user cannot make a sure identification of individuals in the data file. Rank swapping is a version of data swapping useful for quantitative variables or ordinal variables with many levels that seeks to improve the preservation of conditional relationships between variables in the released data set while still protecting confidentiality of respondents. Rank swapping swaps values that are close in rank to one another. Using ranks avoids assumptions about distributional forms for variables. References include Dalenius and Reiss (1982), Moore (1996), Dandekar et al. (2002). Rank swapping for SDL is summarized in Section 2.3.1.

2.3.1 Rank Swapping for SDL

Data swapping involves exchanging values between records for one or more variables. For example, the recorded value of gender (female or male) could be exchanged between two randomly selected records. Sometimes the two randomly selected records would have the same gender and no effective change would take place. Others would produce actual changes. Categorical variables such as county or race and quantitative values such as age also could be swapped.

Sufficiently many variables and values need to be swapped so that a user looking at the released data set cannot know for sure which records are perturbed through swapping and which are in their original form. Swapping takes place at some rate (the swap rate); the higher the swap rate, the more perturbed the data are. Since values are swapped between records, but no values are omitted or changed, the marginal distribution of the variable being swapped remains unchanged. The conditional distribution of the variable being swapped and other variables, however, is affected by the swap rate and complexity of the swapping procedure (Moore 1996). Rank swapping is a limited version of data swapping that aims to limit the distortion of conditional relationships in the released data.

2.3.2 Rank Swapping with Hot Deck

We use rank swapping to further perturb imputed values obtained using the hot deck imputation procedure described in Section 2.2. We propose this combination of SDL techniques to decrease disclosure risk in a data set for release. This could be important in extremely sensitive databases. In summary, our procedure holds some variables as they are (such as county and gender). It then uses regression quantile simulation to generate totally artificial values for some variables conditional on those that are held constant. This is done sequentially so that some dependencies among variables are preserved. Then we implement hot deck imputation for SDL as described in Section 2.2.2 using the Mahalanobis distance to determine the closest match in the real data set to an artificial data record based on the unchanged and previously imputed variables. The remaining variables are temporarily imputed based on the values of these variables for the nearest neighbor. One option would be to simply use those imputed values together with the other values as a releasable data set.

In order to avoid identification of a record through the release of actual information (the hot deck imputations insert actual data values from a survey respondent into the data set), we propose to perturb the data further through rank swapping applied to the hot deck imputations. For the closest match in the original data, we compute its sample rank, r, for the variable to be imputed via hot deck. We then randomly draw rank r^* from a discrete $Uniform(r-\delta, r+\delta)$ distribution, and impute the value from the record in the original data set with sample rank r^*. The value of δ would be determined based on the size of the data set and the amount of confidentiality protection we wish to impose. Future work will study sensitivity to the choice of δ. The resulting data set for release has several variables with synthetic values (generated using quantile regression) and several variables with imputed and perturbed values (obtained using hot deck imputation and rank swapping procedures).

In Section 3 we apply the proposed SDL methods to an American Community Survey data set and present some initial results and a discussion.

3. American Community Survey Application

The U.S. Census Bureau collects and maintains data collected in surveys. To achieve its goal of simultaneously disseminating information while protecting confidentiality, the Census Bureau takes several approaches. Published statistics and summaries, tables, and subsamples with limited number of variables and geographic information are among them. Users who wish to compute other statistics and perform their own analyses can apply for access to microdata through a Research Data Center or at the site of the Census Bureau itself. The process requires a proposal of research, oaths and contracts to protect confidentiality, and restriction to physical location where research can be performed if proposals are accepted and access is granted. We suggest that the SDL methods described in this paper could be implemented on a number of Census data sets to produce releasable data to users, lessening the burden on users and on the Bureau itself.

Results from an application of generating synthetic values using quantile regression for veterans data in the American Community Survey at the U.S. Census Bureau are presented. In Section 3.1 we provide a description of the American Community Survey. In Section 3.2 we describe the quantile regression models used to generate synthetic data and some initial results. We discuss the results and some concerns in Section 3.3.

3.1 The American Community Survey

The U.S. Census Bureau administers a decennial census to provide population counts consistent with a Constitutional mandate to apportion seats in the House of Representatives. The long form that has historically accompanied the decennial census to collect data on the social, economic, housing, and demographic characteristics of the population. With a growing population and increased needs for current and more frequent information about these characteristics, the American Community Survey (ACS) was designed. It is administered yearly and will replace the long form starting in 2010, thereby enabling the Census Bureau to provide pertinent and timely data products every year about communities with larger populations and every 3 and 5 years about communities with smaller populations. More detailed information is available in the ACS Handbook and a document describing the design and implementation of the ACS (www.census.gov/acs/www/).

3.2 Application of SDL Method to ACS Veterans Data

We apply the methods presented in Section 2.1 to ACS data on vet-

erans. Specifically, we simulate synthetic values for age and wages using conditional quantile regression models. The values of age and wages in the data have distinct distributions for male and female respondents, so we consider separate models for the two groups. Based on discussions with members of the Statistical Research Division at the Census Bureau, some variables are included in the models to maintain important conditional distributions. Others are included based on empirical plots and correlations that indicate they will help to characterize the distributions of age and wages well.

3.2.1 Models and Procedure for Simulation

We use a conditional model containing variables that reflect education level (several) (*educ*), current employment in the military (*mil*), social security income (*ss*), and fertility (*fer*) for female respondents. Define $x = \{educ, mil, ss, fer\}$ and $x = \{educ, mil, ss\}$ for female and male respondents, respectively. The quantile regression model is $Q_{age}(\tau|x) = \xi_{age}(x, \beta_{age,\tau}) + F^{-1}(\tau)$, where ξ_{age} is a linear function of x and $\beta_{age,\tau}$ and $F^{-1}(\tau)$ represents iid errors.

Values of wages are simulated using a conditional model containing age, commute time (*com*), race group (*race*), retirement income (*retire*), social security income (*ss*), and two variables reflecting the amount of time spent at work (*work*). With $x = \{age, com, race, retire, ss, work\}$ for both female and male respondents, the regression model is $Q_{l.wages}(\tau|x) = \xi_{wages}(x, \beta_{wages,\tau}) + F^{-1}(\tau)$, where ξ_{wages} is a linear function of x and $\beta_{wages,\tau}$ and $F^{-1}(\tau)$ are iid errors. Two additional considerations are made for wages. A large number of records have recorded wages of zero, so rather than including them in the estimation procedure, we first perform logistic regression to predict whether $wages > 0$ or $wages = 0$. In records with predicted wages of 0 we consider the synthetic value to be 0 and proceed with estimating the quantile regression models using only records predicted to have positive wages. We also notice that many records contain very large values for wages. We perform this modified log-transformation to lessen the effect of the highest values in the estimation:

$$l.wages = \begin{cases} log(wages), & wages \neq 0 \\ 0, & wages = 0. \end{cases}$$

For both age and wages, we simulate values using the method described in Section 2.1.3. Specific methods are presented here. First we fit the models for all quantiles in the set $\tau = \{0.001, 0.01, 0.02, ..., 0.98, 0.99, 0.999\}$. Next randomly select τ_{age} and τ_{wages} for each record from a *Uniform*(0, 1) distribution. For age, using $\hat{\beta}_{age,\tau}$ at quantiles $\tau_{age,a}$ and $\tau_{age,b}$, directly above and below the randomly selected τ_{age} from the set τ, we compute predicted values $\hat{y}_{\tau_{age,a}}, \hat{y}_{\tau_{age,b}}$. Finally, we interpolate to obtain synthetic values $\hat{y}_{\tau_{age}}$ for each record. For wages, using $\hat{\beta}_{wages,\tau}$ at quantiles $\tau_{wage,c}$ and $\tau_{wages,d}$, directly above and below the randomly selected τ_{wages} from the set τ, we compute predicted values $\hat{y}_{\tau_{wages,c}}, \hat{y}_{\tau_{wages,d}}$. Finally, we interpolate to obtain synthetic values $\hat{y}_{\tau_{wages}}$ for each record. Results are presented in Section 3.3.

3.2.2 Initial Results

Recall that the procedure was implemented separately on records for female and male respondents. We compare marginal distributions

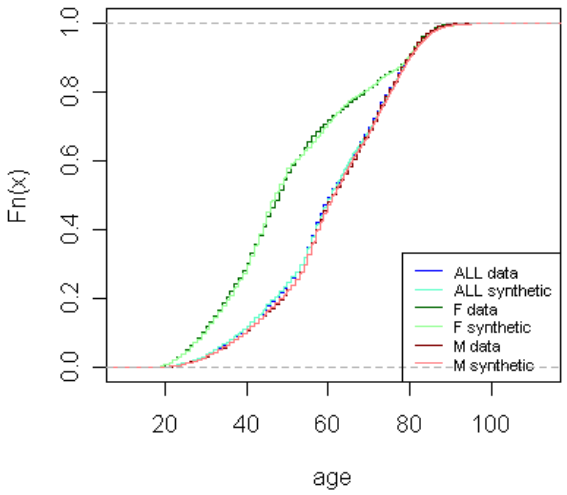

Figure 1: Empirical cumulative distribution functions of age in original and synthetic data. Simulated using quantile regression equation $Q_{age}(\tau|x)$.

of age and wages in the original and synthetic data in plots of their empirical distributions in Figures 1 and 2. We see that marginally, the distributions of age and l.wages in the data are fairly well preserved with the synthetic values, within female and male records as well as across all records.

We examine the distributions of l.wages with respect to age, commute time (*commute*), and two variables reflecting the amount of time spent at work, *work*$_1$ and *work*$_2$. To do so, we compare regression estimates, standard errors, and R^2 values presented in Table 1. We see that estimates are quite close. Standard errors for the synthetic set are higher than for the original data and the R^2 value in the synthetic set is lower than for the original data. The distributions of age values in the original and synthetic data with respect to Veteran Period of Service (VPS) are also compared. We use box plots

Table 1: *l.wages = f(age, work*$_1$*, work*$_2$*, commute,* γ*)* + ϵ

	coefficient	estimate	$s.e.(\hat{\gamma})$
Original data	$\hat{\gamma}_{age}$	0.0027	0.00036
	$\hat{\gamma}_{work_1}$	0.0304	0.00027
	$\hat{\gamma}_{work_1}$	0.0432	0.00027
	$\hat{\gamma}_{commute}$	0.0033	0.00014
	R^2	0.46	
Synthetic data	$\hat{\gamma}_{age}$	-0.0007	0.00033
	$\hat{\gamma}_{work_1}$	0.0289	0.00037
	$\hat{\gamma}_{work_1}$	0.0411	0.00036
	$\hat{\gamma}_{commute}$	0.0033	0.00018
	R^2	0.30	

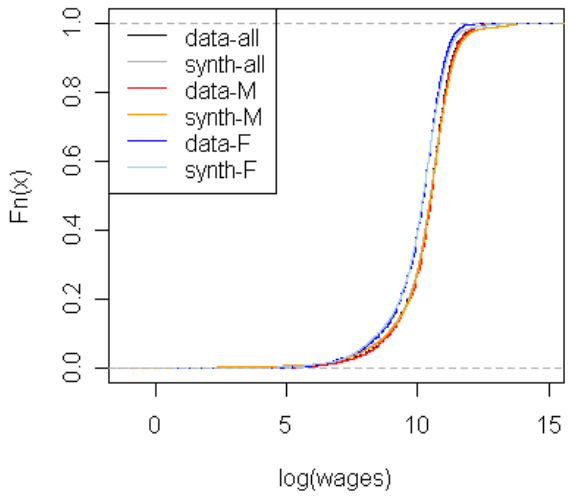

Figure 2: Empirical cumulative distribution functions of l.wages from original and synthetic data. Simulated using quantile regression equation $Q_{l.wages}(\tau|x)$.

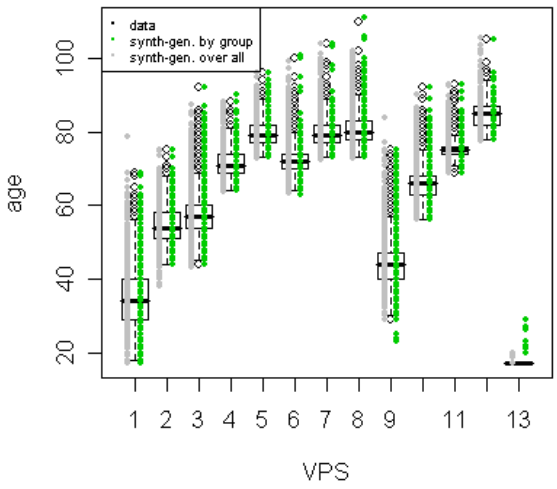

Figure 3: Box plots of age within VPS from original and synthetic data. Generated using quantile regression across all records and within VPS categories.

to illustrate this in Figure 3 in Section 3.4, where we discuss some concerns about consistency of the synthetic data.

3.3 Discussion

A practical concern the Census Bureau has about releasing synthetic data is ensuring consistency within individual records. Consider age and Veteran Period of Service (VPS), for instance. An example of one such inconsistency is a record containing a synthetic value of 17 for age, say, and a value corresponding to World War II for VPS. We examine the distributions of values for age within VPS categories. Synthetic age values rarely fall outside the range of data values, thus few inconsistencies exist. To ensure the resulting synthetic age values fall within the range of values in the data, we consider generating age values within VPS categories using the quantile regression method. The box plots in Figure 3 show the ranges of age values in the original and synthetic data simulated over all records and within VPS categories. We examine synthetic age values that fall outside of ranges in the data to determine if such inconsistencies are truly nonsensical or if they are plausible values. If they are plausible, then it might be acceptable to allow them. Further investigation to this topic should be considered.

Another concern is that the synthetic data be consistent with published tables. If synthetic data do not produce consistent records or counts, the validity of their data products may come under question. The Census Bureau receives frequent requests for tables of counts in categories defined by levels of VPS, race, and 10-year age intervals, so we consider simulating values for age within these categories. This ensures that the number of records in each category is the same in the original and synthetic data. In Figure 4, we see that the distribution of synthetic age is more like the distribution of values in the data when values are simulated under these restrictions. The concerns about consistency of the synthetic data with published tables could lead to more restrictions for simulations, potentially becoming quite cumbersome to implement. Further investigation to determine which consistencies are crucial should be considered.

From the results presented in 3.2.1, our SDL method using regression quantiles to simulate values shows promise to preserve important characteristics in the data and simultaneously protect confidentiality. Hot deck imputation with rank swapping will also be applied to variables in the ACS veterans data to study the methods proposed in Sections 2.2 and 2.3. In an application to individual income tax records at the Iowa Department of Revenue (IDR), we see similar quantile regression results as well as results from an application of hot deck and rank swapping; this work is summarized in Huckett and Larsen (2007) and Huckett (2006).

4. Summary and Future Work

In this paper we have presented an option for creating full synthetic data files for public release from a government agency. We study combining quantile regression, hot deck imputation, and additional confidentiality-preserving methods to produce releasable, usable data. The result is a releasable data set containing original values for a few key variables, synthetic values for several variables, and perturbed values for remaining variables. We present results from an application of generating synthetic values using quantile regression for veterans data in the American Community Survey at the U.S. Census Bureau to show that the procedure provides quality data to the user. Further assessment of data utility should be considered. By

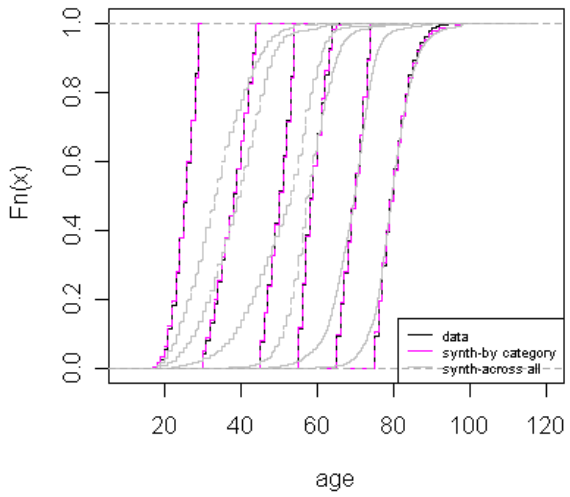

Figure 4: Empirical cumulative distribution functions of age within VPS category from original and synthetic data. Generated using quantile regression accross all records and within VPS, race, and 10-year age intervals.

creating synthetic data for release we believe that this method protects the confidentiality of the respondents, though it remains to be shown quantitatively through disclosure risk analysis.

Acknowledgments

This work was supported in part by a Census Dissertation Fellowship (Census Task Order YA132307SE0304). This article is the responsibility of the authors and does not necessarily reflect the opinion of the U.S. Census Bureau or any of its employees. Special thanks go to Sam Hawala, Laura Zayatz, and Tommy Wright for their time and insight into ACS veterans data, disclosure limitation practices, and the Statistcal Research Division at the U.S. Census Bureau.

REFERENCES

2006 Data Users Handbook: The American Community Survey. (2006), U.S. Census Bureau, Washington D.C., (http://www.census.gov/acs/www/).

Abowd, J. M. and Woodcock, S. D. (2004), "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data," *in (J. Domingo-Ferrer and V. Torra, eds.) Privacy in Statistical Databases*. New York: Springer.

ACS Design and Methodology Paper. (2006), U. S. Census Bureau, Washington D.C., (http://www.census.gov/acs/www/).

Bassett, G. and Koenker, R. (1978), "Asymptotic Theory of Least Absolute Error Regression", *Journal of the American Statistical Association, 73*, 363, 618–622.

Dalenius, T. and Reiss, S.P. (1982) "Data Swapping - A Technique for Disclosure Control", *Journal of Statistical Planning and Inference*, 6, 73-85.

Dandekar R.A., Domingo-Ferrer, J., and Sebé, F. (2002), "LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection", *Inference Control in Statistical Databases*, 153-162.

Domingo-Ferrer, J. (2002), "Inference Control in Statistical Databases: From Theory to Practice," *Lecture Notes in Computer Science*, Springer, New York.

Domingo-Ferrer, J., Torra, V. (2004), *Privacy in Statistical Databases: CASC Project International Workshop*, Lecture Notes in Computer Science, Springer, New York.

Domingo-Ferrer, J. and Franconi, L. (2006), *Privacy in Statistical Databases: CENEX-SDC Project International Conference*, LNCS, Springer, New York.

Doyle, P., Lane, J.I., Theeuwes, J.J.M., and Zayatz, L.V. (Editors), *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier Science B.V. in conjunction with the U.S. Bureau of the Census.

Duncan, G.T., Jabine, T.B., and de Wolf, V.A. (1993), *Private Lives and Public Policies: Confidentiality and Accessibility of Fovernment Statistics*, National Academy Press, Washington, D.C.

Federal Committee on Statistical Methodology. (1994), Report on Statistical Disclosure Limitation Methodology, Statistical Policy Working Paper 22, Office of Management and Budget, Washington, D.C.

Haberman, H. (2006), "Ethics, Confidentiality, and Data Dissemination," *Journal of Official Statistics*, 22, 599-614.

Hawala, S. (2003), "Microdata Disclosure Protection Research and Experiences at the U.§. Census Bureau", *Workshop on Microdata*, Stockholm, Sweden.

Huckett, J.C., "Technical Report, Study of a Proposal to Simulate Tax Records for the State of Iowa", Iowa State University and Iowa Department of Revenue, Iowa.

Huckett, J.C. and Larsen, M.D. (2007), "Microdata Simulation for Confidentiality of Tax Returns Using Quantile Regression and Hot Deck", *2007 Proceedings of the Third International Conference on Establishment Surveys*, American Statistical Association, Statistical Computing Section.

Kinney, S.K. and Reiter, J.P. (2007), "Making Public Use, Synthetic Files of the Longitudinal Business Database", *Joint Statistical Meetings proceedings [CD-ROM]*, American Statistical Association, Government Statistics Section.

Koenker, R. and Hallock, K., (2001), "Quantile Regression", *Journal of Economic Perspectives*, 15, 73, 143-156.

Koenker, R., *Quantile Regression*, Econometric Society Monograph Series, Cambridge University Press, New York, NY.

Little, R.J.A., and Liu, F. (2002), "Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata," *Proceedings of the Section on Survey Research Methods, CD-ROM*, American Statistical Association.

Little, R.J.A., and Liu, F. (2003), 'Comparison of SMIKe with Data-Swapping and PRAM for Statistical Disclosure Control of Simulated Microdata," *Proceedings of the Section on Survey Research Methods, CD-ROM*, American Statistical Association.

Little, R.J.A., Rubin, D. (2002), *Statistical Analysis with Missing Data, Second Edition*, Wiley Series in Probability and Statistics, John Wiliey & Sons, Inc., NY.

Moore, R.A. (2006), "Controlled Data-Swapping Techniques for Masking Public Use Microdata Sets", *Research Report Series*, 2005-04, U.S. Census, Washington D.C.

Raghunathan, T.E., Reiter, J. P., and Rubin, D.R. (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, 531-543.

Reiter, J.P. (2003), "Inference for Partially Synthetic, Public Use Data Sets," *Survey Methodology*, 29, 181-189.

Reiter, J.P. (2005), "Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study," *Journal of Royal the Statistical Society*, Series A, 168, 185-205.

Rubin, D.B. (1993), "Satisfying Confidentiality Constraints through the Use of Synthetic Multiply-Imputed Microdata, *Journal of Official Statistics*, 9, 461-468.

Willenborg, L.C.R.J. and de Waal, T. (1996), *Statistical Disclosure Control in Practice*, Lecture Notes in Computer Science, New York: Springer.

Willenborg, L.C.R.J. and de Waal, T. (2001), *Elements of Statistical Disclosure Control*, Lecture Notes in Computer Science, Springer, New York.

Winkler, W. (2006), "Modeling and Quality of Masked Microdata", *Research Report Series*, 2006-01, U.S. Census Bureau, Washington D.C.

Zayatz, L.V. (2005), "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update," *Research Report Series*, 2005-06, U.S. Census Bureau, Washington D.C.