Model Based Disclosure Avoidance for Data on Veterans

Sam Hawala & Jeremy Funk
U.S. Census Bureau, Sam.Hawala@census.gov, Jeremy.M.Funk@census.gov

Abstract:
The U.S. Department of Veterans Affairs (VA) routinely requests from the U.S. Census Bureau special tabulations on veterans to study socio-economic characteristics, demographic characteristics and the geographic distribution of the veteran population. Due to confidentiality concerns, the Census Bureau may apply a number of different disclosure avoidance techniques to protect the data collected through the variety of its surveys. Techniques include rounding, top-coding, data swapping, imposing geographic thresholds, introducing random noise, cell suppression and complementary cell suppression. In the case of VA data, the techniques we use induce enough loss of detail in the data necessitating some counter-measures, such as "de-rounding". VA employs de-rounding to adapt the tabulations they obtain from the Bureau to their needs. The confidentiality edits also result in discrepancies between the estimates obtained from the special tabulations and what the Census Bureau publicizes on the web. We examine the use of partially synthetic data and, to a limited extent, other disclosure avoidance methods to produce a veterans' microdata file that is "disclosure proof", from which VA will be able to obtain all cross tabulations and run analyses, without the need for "de-rounding".

1. Introduction

1.1. Purpose

The aim of our model based disclosure avoidance technique is to produce multiple synthetic versions (implicates) of an original microdata file. The release of the synthetic data in lieu of the original data is to protect respondents' confidentiality. From the synthetic data, users could generate their own tabulations or select their own microdata samples to carry out their analyses. To estimate population characteristics, users could obtain standard errors and should use the collection of implicates (Reiter, 2003; Abowd & al., 2001; Ragunathan & al., 2003) to incorporate the effects of the disclosure avoidance procedure on the data.

We applied the technique to the American Community Survey data on veterans in the United States. The Department of Veterans Affairs (VA) requests the Census Bureau to produce hundreds of tabulations after each decennial census so that they can study veterans' socio-economic and demographic characteristics and their national geographic distribution.

All data collected or maintained by the Census Bureau under Title 13 of the U.S.Code need disclosure protection. The Census Bureau collects data under the promise that the data will be used only for statistical purposes, and that any specific respondent's identity will not be made public.

The Disclosure Review Board (DRB) reviews all data products for disclosure, including all VA custom tabulation requests. Up until the most recent census in 2000, VA custom tabulations were generated from the same decennial census base files, but whenever VA needed to perform new analyses VA had to come back to the DRB and ask that more data be released.

The creation of a data product releasable to VA, with adequate protection from disclosure and from which VA can generate tabulations on its own whenever it needs them, will make more efficient use of DRB time and other resources at both the Census Bureau and VA. This paper describes our work on this project, which is still in progress.

1.2. Overview of the model based disclosure avoidance technique

We replace the original data on some attributes for the *uncommon* records with values we obtain through statistical models. The uncommon records mostly belong to individuals who have a high probability of being re-identified from the original data. Those records tend to be unique in the data based on a set of characteristics (this set is called the key). They may represent individuals who are unusual in the general population; however they may also be simply unique in the sample. In the released synthetic data, the uncommon records no longer belong to real individuals, and in this way we protect the confidentiality of the original respondents. To maintain data quality, we work to reproduce the means and variances of variables, and the correlations between variables.

Our procedure is similar to the "selective multiple imputation of keys (SMIKE)" method of Little and Liu (2003). We choose the attributes in the key to be variables readily available to data intruders or commonly assessed in surveys, such as age, sex, etc.
We define data intruders as in Fuller (1993), i.e. persons using the data to discover the identity or to estimate the characteristics of a particular individual. In contrast, data users who estimate population parameters are called analysts. The Census Bureau's goal is to produce and release data useful to analysts but not to intruders.

Our synthesis procedure targets variables that could be in the key or could be other sensitive variables. We first delete the values of the target variable for the vulnerable records and then we build flexible prediction models relating the target variable to as many predictors as can be tolerated by the structure of the data. Our models are semi-parametric, and additive. The models use cubic spline regression – or all over continuous, piecewise regression models. Then through predictive mean matching, we use the predicted values from the model to find donors from the observed data: we replace a deleted value with the observed

value of a respondent having the closest predicted value to the predicted value of the respondent with the deleted value.

Every time we construct a model, we do it on a new bootstrap sample of the data, but for the same set of at-risk records. For each original value that we delete, we produce several synthetic values. One synthetic value cannot itself represent any uncertainty about which value to replace: If one value were really adequate, then that value is the original deleted value. So, analyses that treat synthetic values just like observed values tend to systematically underestimate uncertainty.

1.3. Releasing information about our technique

A data intruder using the several implicates of the partially synthetic data could compromise the confidentiality of the data. Since the "implicates" differ, at most only in the values that we synthesized, the intruder could reverse the synthesis process to rediscover the original values. This issue is discussed later in this paper and some preliminary results are presented. Detailed information on the parameters of our procedure –such as how we designated records for synthesis will remain confidential throughout this paper until further research shows that data confidentiality is not compromised by disclosing this information to the public.

When data users carry out estimation of population quantities using data that were subject to distortion for disclosure avoidance, details on the disclosure avoidance technique employed to protect the data can be very useful. The description of our synthesis technique in this paper is a way to release information about it to aid data users in their analyses. We believe transparency helps the Census Bureau gain more support among data users, which may ultimately increase respondents' cooperation and participation in surveys and censuses.

2. Further details on our technique

2.1. Designating records to be synthesized

Our technique is a partially synthetic data technique. This means that we designate for synthesis only a subset of the records, and a subset of variables for those records. Our principle is to select records that are easily distinguishable (uncommon) from other records in the data. Once we have decided upon the set of key variables, we select records as follows: We concatenate a record's values on the key variables. This subset is called the record's "fingerprint" with respect to the chosen key variables. We obtain a count of fingerprint frequencies across all records. If a given fingerprint has only *one* associated record (uncommon in the implementation we describe in this paper means a unique record), we flag the record as being at risk of disclosure. We select for synthesis those variables that could be advantageous for re-identifying respondents.

In this first implementation, we flagged too many records. It is likely that there were too many variables in the key, or the variables we chose had too many categories. To illustrate with an example: if age is in the key and it is represented in single years instead of age intervals, then it singles out many records, especially in combination with other characteristics. In future implementations, we have plans to address this issue (see concluding remarks.)

2.2. Sub-setting the data

The imputation procedure used to produce the synthetic data values uses a Predictive Mean Matching procedure that draws from other values in the data set. This ensures that the synthetic values are realistic from a univariate perspective. For example, if you were to impute for a variable such as income, you would never wind up with a synthetic value of $50M if there wasn't originally at least one real person in the data set with an income value of $50M.

The modeling procedure used to predict the missing values (set to missing in order to impute) is designed to preserve relationships between the variable being imputed and the rest of the variables in the data set. How well this procedure works is dependent upon how well the models fit the data in question. If the models are very good and able to predict the missing values with strong accuracy, then the relationships should be preserved quite well. Since the modeling works to ensure that the cross distributions of variables are kept generally intact we will refer to this as 'weak relationship preservation.'

There may be some cases of variable interaction in which weak preservation is not sufficient. For example, assume we would like to impute age values for the designated subset of at risk records. Assuming the model that is used predicts age well, we expect that it will preserve correlations between age and the other variables used in the model. However, there may be many cases in which certain age values don't make sense. In the VA data there are variables that indicate in which time periods the respondent served in the military. It could be the case that in order to have served during a certain time period a person must be, for example, at least 37 years old. To take a person whose actual value is very close to 37, say 37 to 39, to a value that is less than 37, say 35 or 36, would be structurally incorrect. Even though that amount of change is not likely to effect the univariate distribution of age or the correlation between age and other predictor variables it represents a technically incorrect microdata record. As this would never occur in the real data, it is unacceptable in the synthetic data. To account for this type of situation, we currently process the data set in subgroups. In contrast to the modeling procedure mentioned above, this is 'strong relationship preservation.' The challenge in this step was to determine which subgroups to use. We based our choices on previous VA data requests, which show how they group the data and what types of analyses they typically perform. Thus we broke the data into 12 subgroups based on age brackets and sex.

2.3. Synthesis procedure

For each variable to be synthesized:

1) Delete values of target variable for the at-risk records (treat them as missing).
2) Impute the "missing" values with a random sample from the available values.

3) Step 3 is the modeling procedure and has two parts:
   a. Draw a sample with replacement from the data of step 2, fit a semi-parametric additive model to predict the target variable, and use this model to predict all of the original observations.
   b. Find transformations of the target variable and of all the predictors while fitting the model.

4) Impute each deleted value of step 1 using predictive-mean matching.

5) Repeat steps 3 & 4 to generate several imputations.

3. Evaluation of the model based disclosure avoidance technique

We evaluate our technique based on the data confidentiality protection it offers and the quality of the data it produces. Any change or distortion, to which we subject the original data, will diminish their quality. However, the original data are most vulnerable to disclosure. So we have two competing objectives: data confidentiality and data quality. Synthetic data strike a balance between the two objectives. We will be satisfied if the essential statistical information (means, variances, correlations) contained in the released synthetic data are the same as that contained in the original data, and the release of these synthetic data do not present a disclosure risk.

3.1. Disclosure Risk

In recent years there have been many proposed methods for measuring the amount of protection that a given disclosure avoidance technique provides for a data set. One commonality is that in order to measure protection one must define risk. In our work, we have applied a somewhat simple framework that we feel allows for a meaningful and understandable quantification of risk. As we mentioned in the introduction, the purpose of our work is to release a veterans microdata set that can be used for investigating many characteristics of the United States veteran population. It will contain geographic information such as State and PUMA (Public Use Micro Data Area), qualitative values such as race and gender, and quantitative information such as income and age. We would like to provide as much of this information to analysts as accurately as possible without directly divulging any information about a particular identifiable individual. For example, let's say that an intruder had some information about a particular veteran. It is likely that knowing a veteran casually would allow the intruder to identify the veteran's gender, race, and possibly what military endeavor the veteran served in. The confidentiality requirement is that the intruder should not be able to take this set of information and identify which record in the released microdata set

belongs to the veteran, and find out information the intruder did not already know (such as income or age or marital status).

Sampling can be thought of as a primary level of protection. If the intruder does not know whether a veteran was in the survey sample and if the sample rate is 1 in 40, the intruder has, at most, a 1 in 40 chance of correct re-identification. Each record on the microdata file theoretically represents 40 people in the population, so the intruder cannot be sure whether the record represents the veteran he or she is looking for. The proposed VA microdata file is derived from a sub-sample of the ACS data. In its full implementation, ACS surveys 2.5% of the population. The original data therefore incorporate a first level of confidentiality protection.

Even though sampling provides a powerful source of protection, it is necessary to further protect those individuals who stand out in the data. These are people with unusual characteristics or combinations of characteristics who may be unique both in the sample data and in the population. To identify these individuals, we first decide what characteristics could possibly be available to the public that allow someone to identify an individual. We then find the subset of records that are unique in the data according to these values, and categorize them as the set of at risk records. There are varying levels of risk using this technique,

1. The more characteristics it takes a record to be unique, the less likely it is that someone would be able to identify this person, so it is considered at lower risk than those that are uniquely identifiable across only a few characteristics.
2. Certain characteristics are considered more readily available to the public, and therefore being unique on these values is considered higher risk than less available attributes.
3. Certain characteristics are considered more identifiable than others, and thus increase the level of risk more than less identifiable ones.

In our analysis of the protection that partially synthetic data provided to the VA microdata set, we categorized the subset of at risk records into four levels of risk, based on the criteria above. We calculated the minimal number of variables that makes a record unique, along with the specific combination of variables which caused them to be unique. We defined the following levels:

0 – No Risk
1 – Low Risk
2 – Medium Risk
3 – High Risk

After determining the level of risk for each of the records, we imputed several variables that we thought would provide protection for each of the at risk records. Ideally we would have imputed for the exact variables that put each of these records at risk, individually by record, and only for the values that made them stand out. For practical reasons this is not yet a feasible option for us. We did, however, impute the same set of variables for each of the records. To provide a simple and clear measure of how much protection this provided we simply looked at whether or not the values changed. Keep in mind there is no requirement that a synthetic value be different from its original value. We then looked at the number of variables for which each record actually changed. Table 1 below represents a cross tabulation of 'Risk Level' by 'Values Changed On' for the subset of at risk records. For presentation purposes we chose four of the imputed variables that we felt were the most important for protecting these records and only included these in our analysis. The values in the table are represented as percentages instead of numbers so that we do not divulge too much information about the implementation of this disclosure avoidance procedure.

Table 1: Number of variables changed by risk levels (Percentages represent percent of row total)

	Number of Variables Changed				
	0	1	2	3	4
Low Risk	1.97%	20.90%	43.67%	27.72%	5.74%
Medium Risk	1.66%	20.56%	46.66%	26.71%	4.41%
High Risk	1.29%	20.26%	56.44%	18.98%	3.03%

Risk Level	Percent of Risk Set
Low	16.32%
Medium	60.40%
High	23.28%

This table tells us that the amount of change is fairly consistent between risk levels. We would hope that in general, the higher the risk the more change that occurs, but this does not appear to be the case. Since high risk means highly unusual characteristics, we would also expect that when these characteristics are imputed the values are likely to change; following from the Predictive Mean Matching procedure that draws from actual within sample values. While currently there is no direct requirement that certain variables change, and no weighting of the amount of change by level of risk, we may in the future include additional constraints in the sub-setting and modeling procedures to obtain more favorable results. In particular, we want to see the high risk records change more and the low risk records change less.

3.2. Data Quality

We measure data quality by investigating the extent to which our synthesis procedure preserves marginal and multivariate distributions. We calculate the values of a Wald type statistic that compares the synthetic and original distributions.

$$X^2 = (S-O)^t \left[diag(S+O) - \mathbf{T} - \mathbf{T}^t \right]^{-1} (S-O)$$

where for a categorical variable with $c+1$ categories,

$$S \text{ is the } c-\text{vector of synthetic counts}$$
$$O \text{ is the } c-\text{vector of original counts}$$
$$T \text{ is the } c \times c-\text{matrix of crosstabulation between the synthetic and original variables}$$

Although this statistic is best suited for categorical variables, it can be used for continuous variables such as age by grouping the continuous variable into various percentiles.

Under the assumptions (Chambers 2001):

a. the synthesis process is stochastic and
b. the synthesized and original values are independently distributed conditional on the observed data,

the large sample distribution of X^2 is chi-square with c degrees of freedom χ_c^2, and so a statistical test of whether the synthesis method preserves the distribution of the categorical variable of interest can be carried out.

Current results indicate that the distributions of synthetic and original data are not homogenous. We have ideas on how to fix the problems to obtain better synthetic veterans data.

3.3 Effect of Releasing Multiple Implicates on Data Protection

As mentioned earlier in this paper, one of the criticisms of releasing multiple implicates of a partially synthetic data set is that a data user may be able to identify which records, and values, have been synthesized. The possible threat of this is twofold; would an intruder be able to use the knowledge of which records and values were changed to compromise the confidentiality of any respondent, and would they be able to use the set of multiple values for each record to better estimate the respondent's true value? The first part is a bit complicated. The reason is that if we properly protect these at risk records, then they will no longer represent real people in the population and it shouldn't matter if they can be identified.

The second issue is more applied and slightly easier to test. To do so, we performed a simple study using 10 implicates of our synthetic data set. We assumed that a user would be able to identify exactly which records and variables were synthesized. For each synthesized value we then took the mean of the 10 implicate values for numerical data and the mode of the 10 implicates for categorical variables. If there was more than one mode then we simply picked one at random. We used this single set of derived values and performed the same risk assessment as we presented earlier for one single set of synthetic data. The table below represents the results of this analysis for the same data and risk set.

Table 2: Number of variables changed by risk levels for 10 implicates Mean/Mode data
Note: Percentages represent percent of row total.

	Number of Variables Changed				
	0	1	2	3	4
Low Risk	1.97%	20.24%	49.41%	24.71%	3.67%
Medium Risk	2.48%	22.18%	50.19%	22.08%	3.08%
High Risk	3.42%	26.43%	53.97%	14.46%	1.72%

Comparing this table to the table derived from a single set of synthetic data indicates that there is a very minimal decrease in the level of protection resulting from the availability of multiple implicates. There may be several reasons for this; however it is likely that the models are consistent in producing predicted values.

4. Concluding Remarks

In the future, we intend to begin the technique with a better estimate of the at-risk set of records. We will estimate the conditional probability of being unique in the population given that it is a sample unique. We will adjust our set of at-risk records using this probability. We will also mix in some *common* records with the uniques so that from the final set of synthesized records it is uncertain which records are unique and which are not. We also plan to incorporate the quantile regression method, which is another procedure to synthesize variables that have highly skewed distributions. All of this will help us achieve a better balance between data confidentiality and data quality.

5. References

Abowd, J. M. And Woodcock, S. D. (2001). 'Disclosure limitation in longitudinal linked data.' In Doyle, P. Lane, J. I., Theeuwes, J. J., and Zayatz, L. V. (Eds), *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Chapter 10, pp. 215-278. North-Holland.*

Chambers, R., (2001) "Evaluation Criteria for Statistical Editing and Imputation", *Office for National Statistics, Report 28.*

Fuller, W. A., (1993) "Masking Procedures for Microdata Disclosure Limitation", *Journal of Official Statistics, 9(2), 383-406.*

Little, R., & Liu., F., (2003) "Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata", *The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 6.*

Ragunathan, T., Reiter, J., and Rubin, D. (2003). 'Multiple imputation for statistical disclosure limitation.' *Journal of Official Statistics, 19(1), 1-16.*

Reiter, J. P. (2003). Inferences for Partially Synthetic, Public Use Microdata Sets. *Survey Methodology 29, 181–188.*