

# **PARTIALLY SYNTHETIC DATA VIA EXPERT KNOWLEDGE, MODELING, AND MATCHING**

Sam Hawala<sup>1</sup>

## **ABSTRACT**

This paper discusses the combined use of model fitting to data and matching using the model predicted values to reproduce the aggregate behaviour and the main features of a data set. We approximate the data using semi parametric regression models combining a simple additive structure with the flexibility of the nonparametric approach. Then we use the model predicted values to obtain hot-deck imputations therefore simulating data to preserve confidentiality. In the case of highly sensitive data this method provides the necessary protection from disclosure.

KEY WORDS: Disclosure, Partially synthetic data

## **RÉSUMÉ**

Le point central de mon article est une méthode pour produire des données synthétiques grâce à l'utilisation combinée de l'expertise sur le genre des données en question, des modèles appropriés, et l'utilisation des valeurs prédites pour trouver des donateurs. Ces trois éléments jouent un rôle important dans le succès de reproduction du comportement général et les caractéristiques principales de l'ensemble des données. Les analystes de l'US Census Bureau fournissent des connaissances spécialisées qu'on introduit dans la synthèse afin que les données synthétiques réussissent la phase d'éditions des données. Le Bureau s'appuie sur les données synthétiques en combinaison avec des techniques traditionnelles pour la confidentialité des données.

MOTS CLÉS : Confidentialité; données synthétiques.

## **1. INTRODUCTION**

### **1.1 Description of the Problem**

The technique I discuss in this paper was born out of the Census Bureau's need to protect data confidentiality for institutionalized and non-institutionalized populations living in collective dwellings or Group Quarters (GQ), which are included in the American Community Survey (ACS). These populations play an important part in the accuracy of national and sub-national population survey estimates.

The Census Bureau would have used a swapping technique to protect ACS-GQ data, but the scarcity of record pairs that we could swap called into question the efficacy of the swapping technique. Swapped records have to agree on some key characteristics including geography and GQ type. There are not enough records in the same geography and GQ type that match on other additional characteristics to form swapping pairs.

GQ records appear on ACS public use microdata samples (PUMS) and other data products. The geographic information on the PUMS files identifies Public Use Microdata Areas (PUMAs) of 100,000 people or more. Some of the 3000 PUMAs nationwide coincide with government entities such as counties or cities. This fact alone could make records for individuals in those areas more likely at risk of disclosure if they are linked to external datasets published by those government entities.

---

<sup>1</sup> Sam Hawala, U.S. Census Bureau, 4600 Silver Hill Rd, DID, Washington, DC 20233

## **2. APPLICATION TO THE AMERICAN COMMUNITY SURVEY (ACS)**

The Census Bureau uses many modes of data publication for the ACS. We are interested mainly in protecting microdata. In the origin of the synthetic data idea for confidentiality protection, the focus is on protecting microdata files. We also note that data products, such as tables and profiles, have built-in confidentiality protection if they are extracted from partially synthetic microdata files.

For some background on the ACS, we mention that it is a relatively new survey conducted by the Census Bureau. The ACS uses a series of monthly samples to produce annually updated data for the same small areas (census tracts and block groups) as the decennial census long-form sample formerly surveyed. The ACS collects data monthly, and the Census Bureau publishes data annually on 3,000,000 people covering every county in the U.S.

From a user's perspective, it is important to keep in mind the context in which we apply the partially synthetic data method, in particular other ACS data releases. The Census Bureau publishes over 800 ACS Base Tables. These tables provide the most detailed data on all topics and geographic areas. An example might be a table that provides Poverty

Status in the past 12 Months of Unrelated Individuals, age 15 Years and over, by Sex, by Age. Tabular Profiles are extracted from the Base Tables. They summarize key demographic, social, economic, and housing characteristics for the nation, states, and other geographic areas that exceed certain population count thresholds (at least 1,000,000 people, or sub-groups of at least 65,000.) The published counts in the tables are weighted counts.

ACS public-use microdata files show a broad range of responses made on individual questionnaires. For example, how one household or one household member answered questions on occupation, place of work, and so forth. The files contain records for a sample of all housing units, with information on the characteristics of each unit and the people in it.

In addition to removing all identifying information from data releases, the Census Bureau continues to use several procedures to protect data such as truncating, top and bottom coding, and data swapping for records on individuals living in households with or without other people. For details on these and other techniques, the reader is referred to Willenborg and De Waal (2001.)

## **3. PARTIAL SYNTHESIS PROCEDURE**

We select vulnerable records and select key variables to synthesize. The selection of records proceeds by classifying variables appearing on the microdata file into two groups: Identifying Variables (IVs) that are commonly assessed in government and private surveys such as age, and sex, and non-Identifying Variables (non-IVs). We condense the values or categories of the IVs and build a multi-way contingency table with cells. We consider cells that contain less than a small threshold number of cases as sensitive to disclosure. The records belonging to these cells are the ones we partially synthesize. We do not synthesize the non-IVs. They are in general not publicly available and by themselves do not identify individuals. The non-IVs may contain sensitive information about respondents, such as some components of the respondent's income information.

We first delete the values of an IV for the vulnerable records. Then we replace the deleted values by a random sample from the remaining values of the IV. Next we specify and estimate a formula, or prediction equation that relates the IV to predictors: other IVs and non-IVs. We obtain model-based estimates for all the values of the IV. Finally, through predictive mean matching we find replacement (synthetic) values for the values of the IV that we deleted. We use the predicted values from the models to find donors from the remaining observed data after deletion of the IV values for the at-risk records. A synthetic value is the observed value of a respondent having closest predicted value, in absolute value distance, to the predicted value of the respondent with the deleted value.

We make use of the models only to define a criterion for matching complete donor records (donors with non-deleted values of the IV) and incomplete recipient records (recipients with deleted values of the IV). We transfer the IV values from the donor records to recipient records.

The models only approximate the data-generating processes. The goal is not to understand or study these processes but rather to mimic, or synthesize them into mathematical models, or prediction equations, which we use in a predictive mean matching process. There is no restriction on including predictors that are exogenous to the original data set to be synthesized, such as the survey sampling design variables. For our application the predictive power is the most important aspect of the models.

Another aspect of the procedure is that we impose logical constraints on the models to avoid impossible combinations of values in the partially synthetic data. For example, it is impossible to have a 10 year old in a nursing home. The logical constraints are determined by subject-matter experts, rather than by the statisticians making the synthesis models.

We avoid re-identification disclosures from the released partially synthetic data since the records that were at risk no longer refer to survey participants. Those individuals that were at risk of disclosure are represented in the released data only through their non-identifying data elements that we did not synthesize. Moreover, from the user's perspective, the synthetic records fit into the overall data in a way that preserves the main features of the originally collected data.

The performance of our procedure is a function of several factors. In particular the choice of IVs, and the percentage of records that are synthesized. These factors determine the degree of confidentiality protection and the extent of data distortion. Detailed information on the rate of synthesis will remain confidential until further research shows that disclosing this information to the public does not compromise data confidentiality.

### 3.1 Synthesizing Models Specifications

If the variable  $Y$  is continuous then we fit a generalized additive model to relate  $Y$  to a set of predictors  $\underline{X}$ . The models are semi-parametric and additive. Given a set of predictor variables  $\underline{X} = (X_1, X_2, \dots, X_p)$ , we assume the target (response) variable follows a distribution from the exponential family, with mean

$$E(Y | \underline{X}) = \alpha + \sum_{i=1}^p g_i(X_i)$$

Where  $g_i \quad i = 1, \dots, p$  are unknown, smooth functions and  $\alpha$  is an unknown parameter. This model combines the simple additive structure of the parametric regression model with the flexibility of the nonparametric approach. The additivity allows for the inclusion of more predictors than otherwise is possible (Venables (2002)). We do not impose any strong shape restrictions on the functions that determine how the variables  $X_i$ s influence the mean regression of  $Y$ . The additive components are approximated using an iterative *backfitting* procedure.

If the variable  $Y$  is an ordered categorical variable with  $J$  categories, then we fit a *proportional odds logistic regression* model; see details in Agresti (2007).

$$\text{logit}[P(Y \leq j | \underline{X})] = \alpha_j + \beta^T \underline{X} \quad j = 1, \dots, J-1$$

We obtain estimates of the cell probabilities  $P(Y = j)$  and draw prediction values (categories) according to these estimated probabilities. In particular if  $Y$  has two categories ( $J = 2$ ) this model is the usual linear logistic model.

If the variable  $Y$  is an unordered categorical variable, then we fit a logit model for multinomial responses.

$$\log \frac{P(Y = j | \underline{X})}{P(Y = J | \underline{X})} = \alpha_j + \beta_j^T \underline{X} \quad j = 1, \dots, J-1$$

We do not check for the adequacy of the models, since any attempt to do that will almost certainly lead to rejecting the models. However, this does not mean that the models are not useful descriptions of the data.

The modeling strategy in our application is not the usual approach that seeks the simplest possible model to explain the relationships between variables. Our objective is prediction not causal inference. We use the largest possible number of covariates  $\underline{X}$  that can produce estimates of the coefficients  $\beta$  and smoothers  $g$ . Information on the covariates improves the accuracy of the predictions. It is, however, computationally impossible to include all possible predictors in the models. The predicted values are sensitive to the choice of variables that we include in  $\underline{X}$ . We expect a weakening of the associations between the predicted response variable  $Y$  and those variables omitted from the model. We also expect some predicted values to be unsuitable (out of range) for imputation, especially in the low and high ranges of the variable

$Y$ . For this reason, we do not directly use the predicted values to impute for the deleted values. We match each record to a record with the closest predicted mean and then impute the  $Y$  value directly from the match.

### 3.2 Multiple Imputations

To obtain valid inferences Rubin (1986) proposed multiple imputations to estimate the added variance from estimating deleted (missing) values. Reiter (2003) extended the theory to measure the added variance due to partial data synthesis. One implicit assumption we made to conform to the theory is that the selection of records to synthesize is done at random (MAR) conditional on the predictors used in the synthesizing models. We could conceivably control the choice of records to synthesize to satisfy this assumption. However we did not do this. The author is not aware of research on the sensitivity of inferences due to departure from the MAR assumption.

There is a concern, at the Census Bureau, about releasing several partially synthetic data sets, which are called implicates. Together several published implicates may, in turn, create a new disclosure risk problem. Using the implicates, an intruder could easily determine which records had not been altered, discovering in this way the exact records and variables that were deemed sensitive to disclosure risk. Suppose, for example, the data consist of one thousand records, fifty of which were found to be at risk on five variables. We model those variables for those fifty at-risk-records. Suppose we release four partially synthetic implicates differing only in the fifty records. By making simple comparisons on the implicates an intruder can easily find the fifty records and the five variables. With some form of reverse engineering an intruder could discover the very sensitive information the agency was trying to protect. For this reason, more research is needed on this question. In view of this problem the Census Bureau releases only one implicate of partially synthetic data (a single imputation), with the appropriate adjustments to the survey weights for more accurate estimation of variances.

## 4. Assessment

Ideally after the implementation of a synthetic data procedure users should be able to reproduce all key outputs from many statistical analyses of the original data set. This is not possible to achieve, so we perform some diagnostics to check for data quality. We evaluate before and after means and variances. We look at bivariate distributions to evaluate the extent to which they differ.

We cannot measure the goodness of fit of the partially synthetic data; we can only measure the absence of badness of fit. Our aim is to not substantially change the original data, while at the same time protect the records that are at risk. Our example shows that this can be done.

The synthetic values fit with the rest of the data well. They are plausible values, i.e. values really observed in the surveyed population. This aspect of our procedure guarantees the credibility of univariate distributions. For instance, synthetic values are never out of range. The role of the models was only to provide a suitable distance function to be used to find the best donor for a particular recipient.

## 5. CONCLUSION

The Census Bureau continues working on meeting users' demands for microdata. Our research is geared towards improving the quality of the data while at the same time protecting data confidentiality. We have developed a procedure to create partially synthetic data. Our procedure fits in well with other processes that survey data have to undergo, such as editing and imputation for missing items. More research is needed to improve synthetic data procedures so that many more synthetic data inferences are the same as original data inferences. Also, work is needed to show that publicly disseminating details on the procedure does not compromise data confidentiality.

## 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.*
- Agresti, A. (2007), *An Introduction to Categorical Data Analysis* (2nd Edition), New York, N.Y.: J. Wiley.

- D’Orazio, M., Di Zio, M., and Scanu, M. (2006), “Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints”. *Journal of Official Statistics*, 22 (1), 137-157.
- 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*.
- Little, R., (1988) “Missing data in large surveys”, *Journal of Business and Economic Statistics* 6:287-301.
- 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.
- Rubin DB, (1986), “*Multiple Imputation in Sample Surveys and Censuses*”, New York: John Wiley and Sons.
- Rubin DB, (1997), “Estimating causal effects from large data sets using propensity scores.” *Annals of Internal Medicine* 127:757–763.
- Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S* (4<sup>th</sup> ed.) New York: Springer-Verlag.
- Willenborg, L. and de Waal, T. (2001). *Elements of statistical disclosure control*, New York: Springer-Verlag.
- Winkler, W. E. (2007), “Analytically Valid Discrete Microdata Files and Re-identification,” technical report, presented at the 2007 Annual Meeting of the American Statistical Association, available at <http://www.census.gov/srd/www/byyear.html>.
- Cochran, W.G. (1977). *Sampling Techniques* (3rd. ed.). New York: John Wiley.
- Rubin, D.B. (1976). “Inference and missing data”. *Biometrika*, **63**, 593-604.