Select Synthetic Data Generation Methods Toward a Public Use File for a Longitudinal Survey

Minsun Riddles[1], Thomas Krenzke[1], Natalie Shlomo[2], Wan-Ying Chang[3], Angela Chen[1], Robyn Ferg[1], Lin Li[1], Medha Uppala[1]

[1]Westat [2]University of Manchester
[3]National Center for Science and Engineering Statistics, U.S. National Science Foundation

Abstract

Uses of synthetic data have been consistently increasing as the demand for access to microdata and privacy concerns grow. For example, synthetic data are seen as a solution for sharing vast amounts of health data for the purpose of developing machine learning models and speeding up research on health data while protecting privacy. Challenges to generating synthetic data are balancing reduction of disclosure risk to a tolerable level and retention of the statistical integrity of the original data (e.g., maintaining the aggregates, distributions, and associations between variables). To address these challenges, one may synthesize variables of lower analytical interest in records with high disclosure risks, referred to as the "select" data synthesis approach, to reduce the disclosure risk while limiting the extent of data synthesis, reducing the chance that the integrity of the data is compromised.

We will describe challenges and demonstrate solutions to generating select synthetic data in the context of a national longitudinal survey. We will outline the details of the research conducted to develop public use longitudinal microdata and summarize the findings from the research. It will include the rationale behind choosing the selected aspects of data synthesis, the extent of data synthesis, and the approach for data synthesis. The comparison of each option in consideration will be described, along with details on the measures used in the comparison. Variance estimation methods will be discussed, which will allow data users to obtain valid inferences. The selected approach for the synthetic public use file will be discussed as well as how the decisions are made.

Details of the study are as follows: During a research phase, we evaluated the additional disclosure risk associated with releasing longitudinal data alongside publicly available cross-sectional data to identify cases with high disclosure risks, which were synthesized at higher rates. Subsequently, we synthesized a subset of variables in the longitudinal survey of interest systematically with a pair of parameters: one for the extent of data synthesis (varying overall treatment rates) and the other for specifying the approach for data synthesis. We assessed both the disclosure risk and utility of multiple synthetic data files to select an approach to produce the public use file for the longitudinal survey. Below are brief descriptions of the three approaches.

- Imputation approach extension to synthetic data generation leverages a sequential hot-deck imputation method the survey employs to handle item nonresponse in the cross-sectional data and unit nonresponse in the longitudinal data. Hotdeck imputation is a statistical technique that identifies for each respondent with a missing survey item another respondent (called a "donor") with complete data for that item. The complete data is used for imputation of the missing data. The respondent with the imputed survey data is referred to as a "recipient". Carefully chosen class and sort variables associated with the variable at hand are used to find a donor similar to a given recipient. As the goal of imputation differs from the goal of data synthesis, we modified the process to introduce noise into the resulting synthetic data, rather than focusing on accuracy.

- Model-Assisted Constrained Hotdeck (MACH) runs its process sequentially through each target variable. For each target variable, hotdeck cells are created from model predictions along with other variables. The general approach for creating prediction groups to be used in forming the hotdeck cells was influenced by a sequential imputation procedure that was initially designed for handling non-monotone (Swiss cheese) missing data patterns in complex questionnaires (Judkins, et al., 2007) while addressing skip patterns and different variable

types. The MACH methodology is described in detail in Krenzke, Li and McKenna (2017). It has been used successfully to generate synthetic data on the American Community Survey data, which is used for the special tabulation for the Census Transportation Planning Products.

- Synthpop R package (Nowok, Rabb, and Dibben, 2016) is an R package that enables a user to produce synthetic data from original data using various parametric and non-parametric models. The package was produced under the SYLLS (Synthetic Data Estimation for UK Longitudinal Studies) project funded by the UK Economic and Social Research Council. The motivation of the project was to produce synthetic data from UK Longitudinal Studies and share it with researchers after mitigating privacy concerns.

Across all combinations, after all target variables were synthesized, weight calibration was applied to ensure consistency with specified totals through calibration. The weights were calibrated to align the synthetic data with the raw data with respect to the key domains. Multiple sets of synthetic data were generated to facilitate an examination of inference and the additional variation arising from the synthesis process. Various options for variance estimation were explored to allow data users to obtain valid inferences. In order to measure the cost and impact associated with reducing the risk through data synthesis, such as how much the precision of estimates reduced and how data synthesis affected the key estimates, the utility of each synthetic data file was compared with regard to a comprehensive list of utility measures, grouped into four broad groups:

• QC checks:

(1) percent of records changed and

(2) check for atypical value combinations across variables

• Weighted Frequency Checks:

(1) distance from the original variable,

(2) high-utility crosstabs,

(3) confidence interval overlap, and

(4) check for confidence intervals and point estimates

• Measures of Associations and Longitudinal Trend Checks:

(1) pairwise associations,

(2) significance of regression coefficients, and

(3) U-statistic

• Measures of Variance Checks:

(1) between implicate variance

Along with the reduction in risk, the utility measures were used to determine the specifics for the official public-use file for the longitudinal survey.