

## Chapter 2

# Background on Multiply Imputed Synthetic Datasets

### 2.1 The history of multiply imputed synthetic datasets

In 1993, the *Journal of Official Statistics* published a special issue on data confidentiality. Two articles in this volume laid the foundation for the development of multiply imputed synthetic datasets (MISDs). In his discussion “Statistical Disclosure Limitation,” Rubin (1993) for the first time suggested generating synthetic datasets based on his ideas of multiple imputation for missing values (Rubin, 1987). He proposed to treat all the observations from the sampling frame that are not part of the sample as missing data and to impute them according to the multiple imputation framework. Afterwards, simple random samples from these fully imputed datasets should be released to the public. Because the released dataset does not contain any real data, disclosure of sensitive information is very difficult. On the other hand, if the imputation models are selected carefully and the predictive power of the models is high, most of the information contained in the original data will be preserved. This approach is now called generating fully synthetic datasets in the literature.

In the same issue, Little (1993) suggested a closely related approach that is also based on the idea of replacing sensitive information by multiple imputation. The major difference is that only part of the data are replaced. The replaced data could either be some sensitive variables, such as income or turnover, or key variables such as age, place of birth, and sex that could be jointly used to identify a single unit in the dataset. With this approach, now called generating partially synthetic datasets, it is not mandatory to replace all units for one variable. The replacement can be tailored only to the records at risk. It might be sufficient for example to replace the income only for units with a yearly income above 100,000 euros to protect the data. This method guarantees that only those records that need to be protected are altered. Leaving unchanged values in the dataset will generally lead to higher data quality, but releasing unchanged values obviously poses a higher risk of disclosure.

Fienberg (1994) proposed a related approach for data confidentiality. He suggested generating synthetic datasets by bootstrapping from a “smoothed” estimate

of the empirical cumulative density function of the survey data. This approach was further developed for categorical data in Fienberg et al. (1998).

Ten years after the initial proposal, the complete theory for deriving valid inferences from multiply imputed synthetic datasets was presented for the first time. Raghunathan et al. (2003) illustrated why the standard combining procedures for multiple imputation (Rubin, 1987) are not valid in this context and developed the correct procedures for fully synthetic datasets. The procedures for partially synthetic datasets were presented by Reiter (2003). One year earlier, Liu and Little (2002) had suggested the selective multiple imputation of key variables (SMIKe), replacing a set of sensitive and nonsensitive cases by multiple draws from their posterior predictive distribution under a general location model.

Reiter also demonstrated the validity of the fully synthetic combining procedures under different sampling scenarios (Reiter, 2002), derived the combining procedures when using multiple imputation for missing data and for disclosure avoidance simultaneously (Reiter, 2004), developed significance tests for multi-component estimands in the synthetic data context (Reiter, 2005c; Kinney and Reiter, 2010), provided an empirical example for fully synthetic datasets (Reiter, 2005b), and presented a nonparametric imputation method based on CART models to generate synthetic data (Reiter, 2005d). Recently he compared CART models with imputation models based on random forests (Caiola and Reiter, 2010). Further work includes suggestions for the adjustment of survey weights (Mitra and Reiter, 2006), selecting the number of imputations when using multiple imputation for missing data and disclosure control (Reiter, 2008b), measuring the risk of identity disclosure for partially synthetic datasets (Reiter and Mitra, 2009; Drechsler and Reiter, 2008), a two-stage imputation strategy to better address the trade-off between data utility and disclosure risk (Reiter and Drechsler, 2010), and an alternative approach for generating public use microdata samples (PUMS) from Census data called sampling with synthesis (Drechsler and Reiter, 2010).

A new imputation strategy based on kernel density estimation for variables with very skewed or even multimodal distributions has been suggested by Woodcock and Benedetto (2009), while Winkler (2007a) proposed the use of different EM algorithms to generate synthetic data subject to convex constraints. The attractive features of synthetic datasets are further discussed by Fienberg and Makov (1998); Abowd and Lane (2004); Little et al. (2004); An and Little (2007), and Domingo-Ferrer et al. (2009).

It took several years before the groundbreaking ideas proposed in 1993 were ever applied to any real dataset. The U.S. Federal Reserve Board was the first agency to protect data in its Survey of Consumer Finances by replacing monetary values at high risk of disclosure with multiple imputations, releasing a mixture of these imputed values and the unreplaced, collected values (Kennickell, 1997). Abowd and Woodcock (2001) illustrated the possibilities of protecting longitudinal, linked datasets with data from the French National Institute of Statistics and Economic Studies (INSEE). A very successful implementation of a partially synthetic dataset is the data behind *On the Map*, illustrating commuting patterns (i.e., where people live and work) for the entire United States via maps available to the public on the

Web (<http://lehdmapp.did.census.gov/>). Since the point of origin (where people live) is already in the public domain, only the destination points are synthesized. Machanavajjhala et al. (2008) developed a sophisticated synthesizer that maximizes the level of data protection based on the ideas of differential privacy (Dwork, 2006) while still guaranteeing a very high level of data utility. The most ambitious synthetic data project to date is the generation of a public use file for the Survey of Income and Program Participation (SIPP) funded by the U.S. Census Bureau and the Social Security Administration (SSA). The variables from the SIPP are combined with selected variables from the Internal Revenue Service's (IRS) lifetime earnings data and the SSA's individual benefit data. Almost all of the approximately 625 variables contained in this longitudinal, linked dataset were synthesized. In 2007, four years after the start of the project, a beta version of the file was released to the public ([www.sipp.census.gov/sipp/synthdata.html](http://www.sipp.census.gov/sipp/synthdata.html)). Abowd et al. (2006) summarize the steps involved in creating this public use file and provide a detailed disclosure risk and data utility evaluation that indicates that confidentiality is guaranteed while data utility is high for many estimates of interest.

The Census Bureau also protects the identities of people in group quarters (e.g., prisons, shelters) in the public use files of the American Community Survey by replacing demographic data for people at high disclosure risk with imputations. The latest release of a synthetic data product by the Census Bureau is a synthetic version of the Longitudinal Business Database (Kinney et al., 2011) that is available as a public use dataset through the VirtualRDC's Synthetic Data Server located at Cornell University (<http://www.vrdc.cornell.edu/news/data/lbd-synthetic-data/>). Partially synthetic, public use datasets are in the development stage in the U.S. for the Longitudinal Employer–Household Dynamics survey and the American Community Survey veterans and full sample data. Recently, a statement by the American Statistical Association on data access and personal privacy explicitly mentioned distributing synthetic datasets as an appropriate method of disclosure control (<http://www.amstat.org/news/statementondataaccess.cfm>).

Outside the U.S., the ideas for generating multiply imputed synthetic datasets were ignored for many years, except for some small simulation studies at ISTAT in Italy (Polettoni, 2003; Franconi and Stander, 2002, 2003; Polettoni et al., 2002). They suggest generating model-based synthetic datasets. The main difference from the methods described in this book is that they do not propose multiple imputation and therefore do not correct for the additional variance from imputation. In 2006, the German Institute for Employment Research launched a research project to generate synthetic datasets of its longitudinal establishment survey for release as a scientific use file. In the first phase of the project, the fully and partially synthetic approaches were tested on a subset of the data (Drechsler et al., 2008b,a; Drechsler and Reiter, 2009). Drechsler et al. (2008a) also discuss the advantages and disadvantages of the two approaches in terms of data utility and disclosure risk. Since the evaluations during the first stage of the project indicated that the dataset could be sufficiently protected by the partial synthetic approach, the second stage of the project focused on the generation of a partially synthetic dataset for the complete 2007 wave of the

survey. This dataset, the first outside the U.S., was released in 2011. The growing interest in synthetic datasets in Europe is also documented by the report on synthetic data files requested by Eurostat 2008 and published by Domingo-Ferrer et al. (2009). Outside Europe, statistical agencies in Australia, Canada, and New Zealand (Graham and Penny, 2005; Graham et al., 2009) also are investigating the approach.

## 2.2 Advantages of multiply imputed synthetic datasets compared with other SDC methods

MISDs provide a number of advantages over other methods that are discussed in the statistical disclosure control (SDC) literature

**First**, the aim of any SDC method should be to preserve the joint distribution of the data. But most data perturbation methods either preserve only univariate statistics or some predefined multivariate statistics such as the mean and the variance-covariance matrix in previously defined subgroups. However, SDC methods are used to generate datasets for public release on the microdata level, and it is impossible to anticipate all analyses potential users will perform with the data. For example, one analyst might remove some outliers before running her regressions, and it is completely unclear what the effects of SDC methods that only preserve statistics in predefined subsets of the data will be for this reduced dataset. Besides, for some analyses it might be desirable to preserve more than just the first two moments of the distribution (e.g., maintain interaction and nonlinear effects).

**Second**, many SDC methods are only applicable either to categorical variables or continuous variables. This means that often a combination of different techniques is required to fully protect a dataset before release. Methods based on multiple imputation, on the other hand, can be applied to categorical and continuous variables likewise, rendering the use of different methods that might require different adjustments by the data analyst unnecessary.

**Third**, most of the data collected by agencies are subject to nonresponse, and besides the fact that missing data can lead to biased estimates if not treated correctly by the analyst, many SDC methods cannot be applied to datasets containing missing values. Since generating multiply imputed synthetic datasets is based on the ideas of multiple imputation for handling item nonresponse in surveys, it is straightforward to impute missing values before generating synthetic datasets. Reiter (2004) developed methods for simultaneous use of multiple imputation for missing data and disclosure limitation.

**Fourth**, model-based imputation procedures offer more flexibility if certain constraints need to be preserved in the data. For example, non-negativity constraints and linear constraints such as *total number of employees*  $\geq$  *number of part-time employees* can be directly incorporated at the model-building stage. Almost all SDC methods fail to preserve linear constraints unless the exact same perturbation is applied to all variables for one unit, which in turn significantly increases the risk of disclosure.

**Fifth**, skip patterns (e.g. a battery of questions are only asked if they are applicable) are very common in surveys. Especially if the skip patterns are hierarchical, it is very difficult to guarantee that perturbed values are consistent with these patterns. With the fully conditional specification approach (see also Section 3.1.2) that sequentially imputes one variable at a time by defining conditional distributions to draw from, it is possible to generate synthetic datasets that are consistent with all these rules.

**Lastly**, as Reiter (2008a) points out, the MI approach can be relatively transparent to the public analyst. Metadata about the imputation models can be released, and the analyst can judge based on this information whether the analysis he or she seeks to perform will give valid results with the synthetic data. For other SDC approaches, it is very difficult to decide how much a particular analysis has been distorted.

On the other hand, as with any perturbation method, limited data utility is a problem of synthetic data. Only the statistical properties explicitly captured by the model used by the data protector are preserved. A logical question at this point is, why not directly publish the statistics one wants to preserve rather than release a synthetic micro-dataset? Possible defenses against this argument are:

- Synthetic data are normally generated by using more information on the original data than is specified in the model whose preservation is guaranteed by the data protector releasing the synthetic data.
- As a consequence of the above, synthetic data may offer utility beyond the models they explicitly preserve.
- Not all users of a public use file will have a sound background in statistics. Some of the users might only be interested in some descriptive statistics and won't be able to generate the results if only the parameters are provided.
- The imputation models in most applications can be very complex because different models are fitted for every variable and often for different subsets of the dataset. This might lead to hundreds of parameters just for one variable. Thus, it is much more convenient even for the skilled user of the data to have the synthesized dataset available.
- The most important reason for not releasing the parameters is that the parameters themselves could be disclosive on some occasions. For that reason, only some general statements about the generation of the public use file should be released. For example, these general statements could provide information about which variables were included in the imputation model but not the exact parameters. So the user can judge whether his analysis would be covered by the imputation model, but he will not be able to use the parameters to disclose any confidential information.

Synthetic Datasets for Statistical Disclosure Control  
Theory and Implementation

Drechsler, J.

2011, XX, 138 p. 19 illus., Softcover

ISBN: 978-1-4614-0325-8