Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data

John M. Abowd[1] and Simon D. Woodcock[2]

[1] Cornell University, Ithaca, NY 14850, USA; CREST, NBER, and IZA
[2] Simon Fraser University, Burnaby, BC V5A 1S6, Canada

Abstract. This paper describes ongoing research to protect confidentiality in longitudinal linked data through creation of multiply-imputed, partially synthetic data. We present two enhancements to the methods of [2]. The first is designed to preserve marginal distributions in the partially synthetic data. The second is designed to protect confidential links between sampling frames.

1 Introduction

Statistical agencies are frequently confronted with the competing objectives of providing high-quality data to researchers and protecting the confidentiality of survey respondents. Numerous methods have been developed to protect confidential data without undue distortion to underlying relationships among variables. Commonly used methods include cell suppression, data masking, and data swapping (see e.g., [16] or the appendix to [2]). In general, the extent to which these methods succeed in protecting confidentiality and preserving the analyst's ability to obtain valid statistical inferences depends on the nature of the underlying data. Furthermore, downstream statistical analyses may require detailed knowledge of the disclosure control techniques or specialized software.

An alternate approach is to develop multiple synthetic data sets for public release. This approach stems from the related proposals [15] and [3]. [15] suggests generating synthetic data through multiple imputation;[3] [3] suggests generating synthetic data by bootstrap methods.[4] A decided advantage of the synthetic data approach is that valid inferences can be obtained using standard software and methods.[5] Furthermore, since the released data are synthetic, i.e., contain no data on actual units, they pose no disclosure risk.

In practice, generating plausible synthetic values for all variables in a database may be difficult. This has led several authors to consider the creation of multiply-imputed, partially-synthetic data sets that contain a mix of actual and imputed values. In partially synthetic data, confidential data are multiply-imputed, and

[3] This proposal is developed more fully in [8]. [9] provides a simulation study, [12] discusses inference, and [11] provides an application.

[4] [5] apply this method to categorical data; [4] use related concepts to develop a measure of disclosure risk

[5] In the case of multiply-imputed synthetic data, these methods are related to those applied to the analysis of multiply-imputed missing data, e.g., [14]. See [8] for details.

disclosable data are released without perturbation. [6] pioneered this approach in the Survey of Consumer Finances. [2] adopt this approach to protect confidentiality in longitudinal linked data. [10] develops methods for valid inference, and [13] presents a nonparametric method to generate multiply-imputed, partially-synthetic data.

We consider the case of longitudinal linked data. These are defined as microdata that contain observations from two or more related sampling frames, with measurements for multiple time periods for all units of observation. They can be survey or administrative data, or some combination thereof. Our prototypical example is longitudinal data on employers and employees. Employment relationships define the links between them. We are primarily interested in the problem of protecting confidentiality when data from all three sampling frames (employers, employees, and employment histories) are combined for statistical analysis, and when the links between sampling frames (a history of employment relationships) are deemed confidential. In [2] we considered the case where the links between sampling frames were disclosable. In this paper we discuss multiply-imputing confidential links. We also present an improved method for multiply-imputing confidential characteristics of the sampled units. We apply a nonparametric transformation to each continuous confidential variable to improve the fit of the imputation model, and to better preserve marginal distributions in the partially synthetic data.

Longitudinal linked data present particular challenges for statistical disclosure limitation. Like all longitudinal data, they are characterized by complicated dynamic relationships between variables. However when data from multiple related sampling frames are combined, these dynamic relationships span multiple frames. Furthermore, these data are generally composed of a mix of discrete and continuous variables, some with censored or truncated distributions. Finally, the links between sampling frames may themselves be deemed confidential. Protecting their confidentiality requires new methods.

The remainder of the paper is organized as follows. Section 2 introduces notation and discusses the [2] method for multiply-imputing confidential characteristics of units of observation when links between sampling frames are disclosable. Section 3 presents an improvement over these methods that better preserves marginal distributions. Section 4 considers the case where links between frames are confidential, and Sect. 5 concludes. Simulations and empirical results are forthcoming.

2 Concepts

2.1 Multiply-Imputed Partially Synthetic Data

Consider a database with confidential elements Y and disclosable elements X.[6] Both Y and X may contain missing data. Using standard notation from the

[6] The database in question is defined quite generally, and the discussion in this section is not necessarily limited to longitudinal linked data.

missing data literature, let the subscript mis denote missing data and the subscript *obs* denote observed data, so that $Y = (Y_{mis}, Y_{obs})$ and $X = (X_{mis}, X_{obs})$. We assume throughout that the missing data mechanism is ignorable.

The database is represented by the joint density $p(Y, X, \theta)$, where θ are unknown parameters. [2] suggest imputing confidential data items with draws $\tilde{Y}$ from the posterior predictive density

$$p(\tilde{Y}|Y_{obs}, X_{obs}) = \int p(\tilde{Y}|X_{obs}, \theta) p(\theta|Y_{obs}, X_{obs})\, d\theta \ . \tag{1}$$

The process is repeated M times, resulting in M multiply-imputed partially synthetic data files $\left(\tilde{Y}^m, X^m\right), m = 1, ..., M$. In practice, it may be easier to first complete the missing data using standard multiple-imputation methods and then generate the masked data as draws from the posterior predictive distribution of the confidential data given the completed data. For example, first generate M imputations of the missing data (Y^m_{mis}, X^m_{mis}), where each implicate m is a draw from the posterior predictive density

$$p(Y_{mis}, X_{mis}|Y_{obs}, X_{obs}) = \int p(Y_{mis}, X_{mis}|Y_{obs}, X_{obs}, \theta)\, p(\theta|Y_{obs}, X_{obs})\, d\theta \ . \tag{2}$$

With completed data $Y^m = (Y^m_{mis}, Y_{obs})$ and $X^m = (X^m_{mis}, X_{obs})$ in hand, draw the partially synthetic implicate $\tilde{Y}^m$ from the posterior predictive density

$$p(\tilde{Y}|Y^m, X^m) = \int p(\tilde{Y}|X^m, \theta) p(\theta|Y^m, X^m)\, d\theta \tag{3}$$

for each imputation m.

In practice, it can be very difficult to specify the joint probability distribution of all data, as in (1), (2), and (3). Instead, [2] approximate the joint densities using a sequence of conditional densities defined by generalized linear models. Doing so provides a way to model complex interdependencies between variables that is both computationally and analytically tractable. One can accommodate both continuous and categorical data by choice of an appropriate generalized linear model. The multiply-imputed partially synthetic data are drawn variable-by-variable from the posterior predictive distribution defined by an appropriate generalized linear model under an uninformative prior. If we let y_k denote a single variable among the confidential elements of our database, imputed values $\tilde{y}_k$ are drawn from

$$p(\tilde{y}_k|Y^m, X^m) = \int p(\tilde{y}_k|Y^m_{\sim k}, X^m, \theta_k) p(\theta_k|Y^m, X^m)\, d\theta_k \tag{4}$$

where $Y^m_{\sim k}$ are completed data on confidential variables other than y_k.

2.2 Longitudinal Linked Data

It is convenient to represent a longitudinal linked database as a collection $\mathcal{F}$ of files. Each file $F \in \mathcal{F}$ contains longitudinal data from a single sampling frame. Each file may contain both confidential and disclosable data elements. Observations in different files are linked by a series of identifiers. An example serves to illustrate the structure of a longitudinal linked database.

Our prototypical longitudinal linked database contains observations about individuals and their employers, linked by means of a work history. The work history contains data on each job held by an individual, including the identity of the employer. Suppose we have linked data on I employees and J employers spanning T periods. There are three data files. The first file $U \in \mathcal{F}$ contains longitudinal data on employees, with elements denoted u_{it} for $i = 1, ..., I$ and $t = 1, ..., T_i$. The second data file $Z \in \mathcal{F}$ contains longitudinal data on employers, with elements z_{jt}, for $j = 1, ..., J$ and $t = 1, ..., T_j$. The third data file $W \in \mathcal{F}$ contains work histories, with elements w_{ijt}. The data files U and W are linked by a person identifier. The data files Z and W are linked by a firm identifier, conceptualized by the link function $j = J(i, t)$ that indicates the firm j at which worker i was employed at date t. For simplicity, assume that all work histories in W can be linked to individuals in U and firms in Z and that the employer link $J(i, t)$ is unique for each (i, t).[7]

As discussed at length in [2], it is desirable to condition the imputation equations on all available data. In the context of longitudinal linked data, this includes data from all sampling frames. Thus when imputing variable y_k in file $F \in \mathcal{F}$, conditioning information should include not only data elements in F, but also data from other files $F' \in \mathcal{F}$. This helps to preserve relationships among variables in the various files. Inevitably, some data reduction is required. We conceptualize these data reductions by functions g of data in files $F' \in \mathcal{F}$.

It is frequently desirable to estimate separate imputation equations on subsets of the data, *e.g.*, separate models for men and women, full-time and part-time workers, *et cetera*. We conceptualize these subsets as data configurations, indexed by c. A given configuration may also reflect the structure of available data. For example, to impute earnings in some period t, we may wish to condition on past and future values of earnings at that employer. Such data may not be available for every observation because of "structural" aspects of the employment history, *e.g.*, the worker was not employed in the previous period.

Let $p\left(y_k^c|\cdot, \theta_k^c\right)$ represent the likelihood of an appropriate generalized linear model for configuration c of variable $y_k \in F$. Under an uninformative prior, imputations are drawn from the posterior predictive density

$$p\left(\tilde{y}_k^c|Y^m, X^m\right) = \int \begin{array}{c} p\left(\tilde{y}_k^c|Y_{\sim k}^m \in F, X^m \in F, g_k^c\left(Y^m \in F', X^m \in F'\right), \theta_k^c\right) \\ \times p\left(\theta_k^c|Y^m, X^m\right) d\theta_k^c \end{array} \quad (5)$$

[7] The notation to indicate a one-to-one relation between work histories and individuals when there are multiple employers is cumbersome. Our application properly handles the case of multiple employers for a given individual during a particular sample period.

where $Y_{\sim k}^m$ represents other confidential data in F, and F' denotes the complement of F in $\mathcal{F}$. Note $Y_{\sim k}^m$ may include measurements on y_k taken in other time periods.

3 Preserving Marginal Distributions in the Partially Synthetic Data

[2] discuss several enhancements to the above methods that improve the confidentiality protection or the analytic usefulness of the partially synthetic data. We present an additional one here, that helps preserve the marginal distributions of confidential variables.

Under the variable-by-variable imputation method described above, an appropriate generalized linear model defines a parametric distribution for the variable y_k under imputation, conditional on confidential and non-confidential data in all files. In many cases, the marginal distribution of y_k is unknown or differs from the parametric family of the posterior predictive distribution of the imputation model. This is problematic for generating multiply imputed, partially synthetic data using generalized linear models, since it can lead to discrepancies between the moments of the confidential data and the partially synthetic data. [2] found that their method preserved first and second moments of the confidential data. However, higher moments may be distorted if the posterior predictive distribution of the generalized linear model differs from the marginal distribution of y_k.

The usual solution, of course, is to take some analytic transformation of y_k. For example, it is frequently argued that the earnings of white males have an approximately lognormal distribution. Thus a suitable imputation model might be a normal linear regression of the logarithm of earnings on other data items. The imputed values are normally distributed. Exponentiation returns them (approximately) to the original location and scale.

There are two important limitations to such a strategy. First, any error in the analytic transformation biases the distribution of the imputed values.[8] Second, no convenient analytic transformation may be available. We suggest a nonparametric transformation that addresses these limitations.

Our transformation is conceptually simple, and is applicable to continuous variables in a variety of contexts. We consider the case where the imputation model is a normal linear regression, though other applications are possible. Under an uninformative or conjugate prior, the posterior predictive distribution is normal. If the marginal distribution of the confidential variable y_k differs greatly from normality, the distribution of the imputed values will differ from that of the confidential data. The idea is to transform the confidential data so that they have an approximately normal distribution, estimate the imputation model on

[8] Error in the transformation is any difference between the distribution of the transformed variable and the posterior predictive distribution of the generalized linear model.

the transformed data, and perform the inverse transformation on the imputed values. The first step is to obtain an estimate of the marginal distribution of y_k. Since we are in the case where the exact parametric distribution of y_k is unknown, we suggest a nonparametric estimate, e.g. a kernel density estimate $\hat{K}$. Provided sufficient data, this can be done for each data configuration c. For each observation y_k, compute the transformed value $y'_k = \Phi^{-1}\left(\hat{K}(y_k)\right)$, where Φ denotes the standard normal CDF. By construction, the y'_k have a standard normal distribution. Then estimate the imputation regression on y'_k, and draw imputed values $\tilde{y}'_k$ from the posterior predictive distribution. The imputed values are normally distributed with conditional mean and variance defined by the regression model. To return the imputed values to the original location and scale, compute the inverse transformation $\tilde{y}_k = \hat{K}^{-1}(\Phi(\tilde{y}'_k))$. The imputed values $\tilde{y}_k$ are distributed according to $\hat{K}$, preserving the marginal distribution of the confidential data.

There is one caveat to the above discussion. The transformation y'_k and its inverse depend on the data. That is, the transformation function depends on an estimate $\hat{K}(y_k)$, and hence contains model uncertainty (sampling error). To obtain valid downstream inference, we need to account for the additional uncertainty introduced by the transformation. A simple way to do this is to bootstrap the transformation. We therefore suggest an additional step. In each implicate m, draw a Bayesian bootstrap sample of values of y_k, denoted y_k^m, and compute the transformation $y_k^{m\prime} = \hat{K}^m(y_k^m)$. After drawing imputed values $\tilde{y}_k^{m\prime}$ from the appropriate posterior predictive distribution, perform the inverse transformation $\tilde{y}_k^m = \left(\hat{K}^m\right)^{-1}(\Phi(\tilde{y}_k^{m\prime}))$.

Simulation results and an empirical application of this method are forthcoming.

4 Protecting Confidential Links Between Sampling Frames

[2] considered the case where links between data files $F \in \mathcal{F}$ were among disclosable data elements X. In many situations this is unlikely to be the case. Returning to our prototypical longitudinal linked database, links between files constitute a history of employment relationships. From these one can compute a variety of statistics (*e.g.*, the number of jobs held by an individual in each period, firm employment in each period, *etc.*) that can be used to identify employers and employees in the partially synthetic data. Thus we now consider the case where links between data files are deemed confidential. Our suggestion is to treat these like other confidential data items, and multiply-impute them under appropriate generalized linear models. In the context of our prototypical longitudinal linked database on employers and employees, this amounts to imputing the link function $j = J(i,t)$. For a given worker i, this can be accomplished either by imputing the firm's identity j in some period t; imputing the dates t

associated with an employment spell at firm j; or both. We illustrate the method with an example taken from current research at the U.S. Census Bureau.

An application of the procedure described in this paper is currently underway at the U.S. Census Bureau, using data from the Longitudinal Employer-Household Dynamics (LEHD) program. These are administrative data built from quarterly unemployment insurance (UI) system wage reports. They cover the universe of employment at businesses required to file quarterly UI reports – estimated to comprise more than 96 percent of total wage and salary civilian jobs in participating states. See [1] or [7] for a detailed description of the data. The file structure corresponds to that of the prototypical database described in Sect. 2.2. To protect confidentiality of the employment history, our objective is to ensure that any person- or firm-level summary of the history is perturbed. To do so, it is sufficient to multiply-impute the identity of at least one employer in each individual's employment history, and the start and end dates of all employment spells.

To multiply-impute (at least) one employer's identity in the individual's employment history, we use a logistic regression model that conditions on establishment employment and detailed employer and employee geography. The set of candidate "donor" establishments is restricted to businesses operating in the same county and detailed industry, in some cases stratified by employment. Denote the set of such firms by $\Upsilon(i,t)$. Conditioning variables for the regression include establishment employment, characteristics of the within-establishment wage distribution, and the worker's physical proximity to the business. Denote the vector of these characteristics by x_{ijt}. The imputation model is based on

$$\Pr\left(J(i,t) = j | j \in \Upsilon(i,t)\right) = \frac{\exp\left\{\alpha_{jt} + x'_{ijt}\beta\right\}}{\sum_{k\in\Upsilon(i,t)} \exp\left\{\alpha_{kt} + x'_{ikt}\beta\right\}} \tag{6}$$

where α_{jt} is a firm and time specific effect.

We also multiply-impute the start and end date of each employment spell. We can represent the employment history of an individual at a particular employer by a binary string. Each digit of the string corresponds to one quarter in the sample period. It takes value 1 if the worker was employed at that business in that quarter, and 0 otherwise. The imputation model is a binary logit for employment in a given quarter, conditional on characteristics from all sampling frames and whether the individual was employed at that business in the four previous and subsequent quarters. We multiply-impute an individual's employment status at the business for each quarter in a window around the employment spell's start and end date. This perturbs the start and end dates of the spell, but constrains them to lie within a fixed interval of the true values. It can also fill or create short gaps in the employment spell, in a manner consistent with observed spells.

5 Summary

Multiply-imputed partially synthetic data hold great promise for statistical agencies and analysts alike. They satisfy the statistical agency's need to protect the

confidentiality of respondents' data, while preserving the analyst's ability to perform valid inference. In the context of longitudinal linked data, the synthetic data approach is particularly appealing. It is sufficiently flexible to maintain complex relationships between variables in various sample frames. As demonstrated in this paper, it is also possible to preserve the marginal distribution of confidential variables in the partially synthetic data. Furthermore, the synthetic data approach is adaptable to protecting confidential links between frames. The application of these methods to the LEHD database promises further refinement of the techniques discussed in this paper. This application will further demonstrate their ability to protect confidentiality and preserve valid inferences, and will demonstrate the practicality of the synthetic data approach.

References

1. Abowd, J. M., J. I. Lane, and R. Prevost: Design and conceptual issues in realizing analytical enhancements through linkages of employer and employee data. In *Proceedings of the Federal Committee on Statistical Methodology* (2000).
2. Abowd, J.M., and S. D. Woodcock: Disclosure limitation in longitudinal linked data In P. Doyle, J. I. Lane, J. J. Theeuwes, and L.V. Zayatz (eds), *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies* (2000) 215–278. North-Holland.
3. Fienberg, S. E.: A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Carnegie Mellon University Department of Statistics Technical Report No. 611 (1994).
4. Fienberg, S. E. and U. E. Makov: Confidentiality, uniqueness, and disclosure limitation for categorical data. J. Official Statistics **14**(4) (1998) 385–397.
5. Fienberg, S. E., U. E. Makov, and R. J. Steele: Disclosure limitation using perturbation and related methods for categorical data. J. Official Statistics **14**(4) (1998) 485–502.
6. Kennickell, A. B.: Multiple imputation and disclosure protection: the case of the 1995 Survey of Consumer Finances. SCF Working Paper (1997).
7. LEHD Program: The Longitudinal Employer-Household Dynamics program: Employment Dynamics Estimates project versions 2.2 and 2.3. LEHD Technical Paper 2002-05 (2002).
8. Raghunathan, T. J., J. P. Reiter, and D. Rubin: Multiple imputation for statistical disclosure limitation. J. Official Statistics **19**(1) (2003) 1–16.
9. Reiter, J. P.: Satisfying disclosure restrictions with synthetic data sets. J. Official Statistics **18**(4) (2002) 531–544.
10. Reiter, J. P.: Inference for partially synthetic, public use microdata sets. Survey Methodology **29** (2003) 181–188.
11. Reiter, J. P.: Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. J. Royal Statistical Society, Series A (forthcoming).
12. Reiter, J. P.: Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. J. Statistical Planning and Inference (forthcoming).
13. Reiter, J. P.: Using CART to generate partially synthetic public use microdata. Duke University working paper (2003).
14. Rubin, D. B.: *Multiple Imputation for Nonresponse in Surveys.* (1987) Wiley.

15. Rubin, D. B.: Discussion of statistical disclosure limitation. J. Official Statistics **9**(2) (1993) 461–468.
16. *Statistical Disclosure Control in Practice*. (1996) Springer-Verlag.

This article was processed using the LaTeX macro package with LLNCS style