



Special Issue on the Relationship between Imputation and Confidentiality

Message from the Editor ↓

Contents

Whenever there is imputation, data are changed in some ways. Either from missing to an estimated value, or from a (rejected) value to an estimated (or altered) one. If we consider confidentiality procedures, we also see data that is changed. In this case, it is either from a value to a missing one or from a value to another (estimated/altered) one. A parallel can be made between the two approaches. In fact, they can sometimes become one and the same! That is why we are devoting this entire issue to the relationship between imputation and confidentiality.

In the first paper, Jerome P. Reiter from Duke University explains how to create synthetic data using multiple imputation. The approach can be used to produce a file with only synthetic data or one containing both respondent and synthetic data. In the conclusion, he stresses the challenge and importance of correctly specifying the imputation model.

Then, Laura Zayatz, from the US Census Bureau, presents an overview of the disclosure avoidance techniques that are in use at the US Census Bureau, noting the parallel with imputation techniques.

Following is a paper by Jean-Louis Tambay from Statistics Canada exploring the spectrum of imputation approaches that can be used to protect the confidentiality of data. He explains how some approaches can be passive while some others may be more aggressive in the way they perturb the data.

Finally, Eric Schulte Nordholt and Peter-Paul de Wolf from Statistics Netherlands discuss specifically the links that exist between imputation and confidentiality methods.

Of course, imputation and confidentiality procedures are not fully equivalent and as the authors say, there is still a lot to be researched in this respect.

> MESSAGE FROM THE EDITOR	1
> CONTACT INFORMATION	2
> EDITORIAL TEAM	2
> STATISTICS CANADA'S COMMITTEE ON PRACTICES IN IMPUTATION (COPI)	2
> PROTECTING DATA CONFIDENTIALITY IN PUBLIC RELEASE DATASETS: APPROACHES BASED ON MULTIPLE IMPUTATION	3
> IMPUTATION FOR DISCLOSURE AVOIDANCE	7
> USING IMPUTATION TO PROTECT DATA CONFIDENTIALITY	12
> THE LINKS BETWEEN CONFIDENTIALITY AND IMPUTATION	17
> PAST TOPICS	21
> RECENT PAPERS	24
> STATISTICS CANADA INTRANET SITES	25
> INTERNET SITES	25
> NEWS GROUP	25

Enjoy your reading

Eric Rancourt

 **Contact Information** ↓

The *Imputation Bulletin* is published twice a year
For more information, please contact:

The Imputation Bulletin
Methodology Branch
Statistics Canada
R.-H. Coats Building
Ottawa, Ontario K1A 0T6
Tel : (613) 951-5046

Please note that authors in the *Imputation Bulletin* are from
Statistics Canada unless stated otherwise.

Intranet Site: http://method/BiblioStat/Research/TechCom/CoPI/imputationBulletin_e.htm
e-mail: imputationbulletin@statcan.gc.ca

 **Editorial Team** ↓

Sophie Arsenaault sophie.arsenaault@statcan.gc.ca

Jean-François Beaumont jean-francois.beaumont@statcan.gc.ca

Céline Ethier celine.ethier@statcan.gc.ca

David Haziza david.haziza@statcan.gc.ca

Eric Rancourt eric.rancourt@statcan.gc.ca

 **Statistics Canada's Committee on Practices** ↓
in Imputation (COPI)

Membership List

Eric Rancourt (President)
Sophie Arsenaault, Mike Bankier, Jean-François Beaumont,
Jean-Luc Bernier, Joël Bissonnette, Yves Deguire, Jean-Marc Fillion, Lyne Guertin,
David Haziza, Rob Kozak, Chantal Marquis, Christian Nadeau,
Claude Poirier, Joe Sun, Wesley Yung

Volume 8, Number 2, 2008
The Imputation Bulletin



Protecting data confidentiality in public release datasets: ↓ Approaches based on multiple imputation

1. Introduction

Statistical agencies that disseminate data to the public are ethically and often legally required to protect the confidentiality of respondents' identities and sensitive attributes. To satisfy these requirements, Rubin (1993), Little (1993), and Fienberg (1994) proposed that agencies utilize multiple imputation. For example, agencies can release the units originally surveyed with some values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. Multiple imputation for protecting confidentiality is often called the synthetic data approach.

In recent years, agencies have begun to use synthetic data approaches to create public use data for major surveys. In 2007, the U.S. Census Bureau released a synthetic, public use file for the Survey of Income and Program Participation that includes imputed values of social security benefits information and dozens of other highly sensitive variables. The Census Bureau also plans to protect the identities of people in group quarters (*e.g.*, prisons, shelters) in the next release of public use files of the American Communities Survey by replacing demographic data for people at high disclosure risk with imputations. Synthetic, public use datasets are in the development stage in the U.S. for the Longitudinal Business Database, the Longitudinal Employer-Household Dynamics survey, and the American Communities Survey veterans and full sample data. Statistical agencies in Australia, Canada, Germany, and New Zealand also are investigating the approach.

This article reviews the ideas underpinning synthetic data methods. It outlines the potential benefits of synthetic data approaches and the practical challenges to realizing these benefits. It does not review the combining rules for obtaining inferences from the multiple synthetic datasets. These rules differ from the usual combining rules for multiple imputation for missing data. For a summary of inferential methods for various adaptations of multiple imputation, see Reiter and Raghunathan (2007).

2. Description of synthetic data methods

Synthetic data approaches come in two main flavors: fully and partially synthetic.

2.1 Fully synthetic data

To illustrate how fully synthetic data might work in practice, we modify the setting described by Reiter (2004a). Suppose the agency has collected data on a random sample of 10,000 people. The data

comprise each person's race, sex, income, and indicator for the presence of a disease. The agency has a list containing all people in the population, including their race and sex. This list could be the one used when selecting the random sample of 10,000, or it could be manufactured from census tabulations of the race-sex joint distribution. The agency knows the income and disease status only for the people who respond to the survey.

To generate synthetic data, first the agency randomly samples some number of people, say 20,000, from the population list. The agency then generates values of income and disease status for these 20,000 people by randomly simulating values from the joint distributions of income and disease status, conditional on their race and sex values. These distributions are estimated using the collected data and possibly other relevant information. The result is one synthetic dataset. The agency repeats this process say ten times, each time using different random samples of 20,000 people, to generate ten synthetic datasets. These ten datasets are then released to the public.

To illustrate how a secondary data analyst utilizes these ten datasets, suppose that the analyst seeks to fit a logistic regression of disease status on income, race, and sex. The analyst estimates the regression coefficients and their variances in each simulated dataset using standard likelihood-based estimates and software. The analyst averages the estimated coefficients and variances across the simulated datasets. These averages are used to form 95% confidence intervals based on the formulas developed by Raghunathan *et al.* (2003).

Releasing fully synthetic data makes it difficult for data snoopers to identify originally sampled units and learn their sensitive values. Almost all of the released units are not in the original sample, having been randomly selected from the sampling frame, and their values of survey data are simulated. The synthetic records cannot be matched meaningfully to records in other datasets, such as administrative records, because the values of released survey variables are simulated rather than actual. Releasing fully synthetic data is subject to attribute disclosure risk – the risk that the released data can be used to estimate unknown sensitive values very closely – when the models used to simulate data are “too accurate.” For example, when data are simulated from a regression model with a very small mean square error, analysts can estimate outcomes precisely using the model, if they know predictors in that model. Or, if all people in a certain demographic group have the same, or even nearly the same, value of an outcome variable, the imputation models likely will generate that value for imputations. Agencies can reduce these types of risks by using less precise models when necessary.

Fully synthetic datasets can have positive analytic features. When data are simulated from distributions that reflect the distributions of the collected data, frequency-valid inferences can be obtained from the multiple synthetic datasets for a wide range of estimands. These inferences can be determined by combining standard likelihood-based or survey-weighted estimates; the analyst need not learn new statistical methods or software programs to adjust for the effects of the disclosure limitation. Synthetic datasets can be sampled by schemes other than the typically complex design used to collect the original data, so that analysts can ignore the design for inferences and instead perform analyses

based on simple random samples. Additionally, the data generation models can incorporate adjustments for nonsampling errors and can borrow strength from other data sources, thereby resulting in inferences that can be even more accurate than those based on the original data. Finally, because all units' data are simulated, geographic identifiers can be included in the synthetic datasets, facilitating estimation for small areas.

There is a cost to these benefits: the validity of fully synthetic data inferences depends critically on the validity of the models used to generate the synthetic data. This is because the synthetic data reflect only those relationships included in the data generation models. When the models fail to reflect accurately certain relationships, analysts' inferences also will not reflect those relationships. Similarly, incorrect distributional assumptions built into the models will be passed on to the users' analyses. This dependence is a potentially serious limitation to releasing fully synthetic data. Practically, it means that some analyses cannot be performed accurately, and that agencies need to release information that helps analysts decide whether or not the synthetic data are reliable for their analyses. For example, agencies can include the models as attachments to public releases of data. Or, they can include generic statements that describe the imputation models, such as "Main effects for age, sex, and race are included in the imputation models for education." Analysts who desire finer detail than afforded by the imputations may have to apply for special access to the original data.

2.2 Partially synthetic data

Partially synthetic data comprise the units originally surveyed with some collected values replaced with multiple imputations. To illustrate a partially synthetic strategy, we can adapt the setting used in Section 2.1. Suppose the agency wants to replace income when it exceeds \$100,000 and is willing to release all other values. The agency generates replacement values for the incomes over \$100,000 by randomly simulating from the distribution of income conditional on race, sex, and disease status. To avoid bias, this distribution also must be conditional on income exceeding \$100,000. The distribution is estimated using the collected data and possibly other relevant information. The result is one synthetic data set. The agency repeats this process multiple times and releases the multiple datasets to the public.

As with fully synthetic data, when the replacement imputations are generated effectively, analysts can obtain valid inferences for a wide class of estimands with simple combining rules (Reiter, 2003). An advantage of partially synthetic data relative to fully synthetic data is that only a fraction of the data are imputed, so that analysts' inferences are generally less sensitive to the agency's model specification. Unlike fully synthetic data, partially synthetic data must be analyzed in accordance with the original sampling design.

The protection afforded by partially synthetic data depends on the nature of the synthesis. Replacing key identifiers with imputations makes it difficult for users to know the original values of those identifiers, which reduces the chance of identifications. Replacing values of sensitive variables makes it difficult for users to learn the exact values of those variables, which can prevent attribute disclosures.

Nonetheless, there remain disclosure risks in partially synthetic data no matter which values are replaced. Analysts can utilize the released, unaltered values to facilitate disclosure attacks, for example via matching to external databases, or they may be able to estimate genuine values from the synthetic data with reasonable accuracy.

When some data are missing, multiple imputation can be used to fill in missing data and replace confidential values simultaneously with a two stage imputation approach. See Reiter (2004b) for details.

3. Concluding remarks

The main challenge to implementing fully and partially synthetic data approaches is specifying imputation models. To release high quality synthetic data, agencies require flexible and, ideally, automated synthesis methods. There has been some progress in this research direction; for example, Reiter (2005) synthesizes data with regression trees, which can preserve relationships in high dimensional data with categorical and continuous variables. The success of machine learning techniques in bioinformatics and data mining suggests that such techniques, suitably adapted, have great potential for generating model-free synthetic data with high analytic validity.

Given their reduced reliance on imputation models, partially synthetic data may be more appealing to agencies than fully synthetic data. With partial synthesis, agencies must decide which values to replace with imputations. General candidates for replacement include the values of identifying characteristics for units that are at high risk of identification, such as sample uniques and duplicates, and the values of sensitive variables in the tails of distributions. Confidentiality can be protected further by, in addition, replacing values at low disclosure risk. Guidance on selecting values for replacement is a high priority for research in this area.

As resources available to malicious data users continue to expand, the alterations needed to protect public use data with traditional disclosure limitation techniques – such as swapping, adding noise, or microaggregation – may become so extreme that, for many analyses, the released data are no longer useful. Synthetic data, on the other hand, has the potential to enable public use data dissemination while preserving data utility. Ultimately, a statistical disclosure limitation strategy that combines restricted data access for sophisticated analyses and synthetic data for a wide range of simple analyses, such as regressions and comparisons of means, should meet the needs of most secondary data users.

Jerome P. Reiter
Duke University
E-mail: jerry@stat.duke.edu

References

- Fienberg, S.E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Tech. rep., Department of Statistics, Carnegie-Mellon University.
- Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.
- Raghunathan, T.E., Reiter, J.P. and Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.
- Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 181-189.
- Reiter, J.P. (2004a). New approaches to data dissemination: A glimpse into the future (?). *Chance* 17, 3, 12-16.
- Reiter, J.P. (2004b). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30, 235-242.
- Reiter, J.P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, 21, 441-462.
- Reiter, J.P., and Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462-1471.
- Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462-468.



1. Introduction to confidentiality, census bureau data products, and a broad definition of imputation

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code. This prevents the Census Bureau from releasing any data "...whereby the data furnished by any particular establishment or individual under this title can be identified." In addition to Title 13, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) requires the protection of information collected or acquired for exclusively statistical purposes under a pledge of confidentiality. In addition, the agency has the responsibility of releasing data for the purpose of statistical analysis. Thus, the goal is to release as much high quality data as possible without violating the pledge of confidentiality. We apply disclosure avoidance techniques prior to publicly releasing our data products to protect the confidentiality of our respondents and their data. The most common forms of data release are microdata, frequency count data, and magnitude data.

The Census Bureau releases microdata files from our demographic surveys. A microdata file consists of data at the respondent level. Each record represents one respondent and consists of values of characteristic variables for that respondent (Federal Committee on Statistical Methodology, 1994). Typical variables for a demographic microdata file are age, race, sex, income, and occupation of a respondent.

The Census Bureau publishes frequency count data from the decennial census and the American Community Survey (ACS). Tables of frequency count data present the number of units in each table cell. For example, a table may have columns representing the marital status of respondents and rows representing their age groups. The cell values reflect the number of people in a given geographic area having the various combinations of marital status and age group.

The Census Bureau publishes magnitude data from its economic censuses and surveys. Tables of magnitude data often contain the frequency counts of establishments in each cell, but they also contain the aggregate of some quantity of interest over all units of analysis (establishments) in each cell. For example, a table may present the total value of shipments within the manufacturing sector by North American Industry Classification System code by county within state. The frequency counts in the tables are not considered sensitive because so much information about establishments, particularly classifications that would be used in frequency count tables, is publicly available. The magnitude values, however, are considered sensitive and must be protected. Disclosure avoidance techniques are used to ensure published data cannot be used to estimate an individual company's data too closely.

For this paper, we will view imputation as the substitution of one value for another (missing or present).

2. Using noise to protect establishment magnitude data

A noise addition technique is currently being used for our Quarterly Workforce Indicator data and will soon be adopted for several economic surveys. This technique results in all tables cells (and underlying microdata values) being replaced by another value (a form of imputation). Noise is added to the underlying microdata prior to tabulation (Evans, Zayatz, and Slanta, 1998). Each responding company's data are perturbed by a small amount, say 10% (the actual percent is confidential), in either direction. Noise is added in such a way that cell values that would normally be primary suppressions, thus needing protection, are changed by a large amount, while cell values that are not sensitive are changed by a small amount. Noise has several advantages over cell suppression (a method commonly used for this type of data). It enables data to be shown in all cells in all tables. It eliminates the need to coordinate cell suppression patterns between tables. It is a much less complicated and less time-consuming procedure than cell suppression. Because noise is added at the microdata level, additivity of the table is guaranteed.

To perturb an establishment's data by about 10%, we multiply its data by a random number that is close to either 1.1 or 0.9. We could use any of several types of distributions from which to choose our multipliers, and the distributions remain confidential within the agency. The overall distribution of the multipliers is symmetric about 1. The noise procedure does not introduce any bias into the cell values for census or survey data. Because we protect the data at the company level, all establishments within a given company are perturbed in the same direction. The introduction of noise causes the variance of an estimate to increase by an amount equal to the square of the difference between the original cell value and the noise added value. One could incorporate this information into published coefficients of variation. We are currently using the noise method to protect Quarterly Workforce Indicators, Non-Employer data, and Survey of Business Owners data, and intend to use the method more extensively in the future (Massell and Funk, 2007a and 2007b).

3. Disclosure techniques for microdata

Noise Addition

Noise is added to the age variable for persons in households with 10 or more people. Ages are required to stay within certain groupings so program statistics are not affected. Original ages are blanked, and new ages are chosen from a given distribution of ages within their particular grouping. Noise is also added to a few other variables to protect small but well defined populations, but we do not disclose those procedures.

Topcoding

Topcoding is used to reduce the risk of identification by means of outliers in continuous variables (for example someone with an income of five million dollars). All continuous variables (age, income amounts, travel time to work, etc.) are topcoded using the half-percent/three-percent rule. Topcodes for variables that apply to the total universe (for example age) should include at least 1/2 of 1 percent of all cases. For variables that apply to subpopulations (for example farm income), topcodes should include either 3 percent of the non-zero cases or 1/2 of 1 percent of all cases, whichever is the higher topcode. Some variables, such as year born, are likewise bottomcoded. Topcoded values are typically replaced with the mean or median of all topcoded values.

Data Swapping

We examine the records, looking for what are often called "special uniques" (Elliott, Skinner, and Dale, 1998). These are household records which remain unique based on certain demographic variables at very high levels of geography and, therefore, have a disclosure risk. Any such household we find is

swapped (or replaced) with some other household in a different geographic area. This typically does not effect many records, but those that it does need this added protection. See more on data swapping in the next section.

4. Using data swapping to protect frequency count data

The main procedure used for protecting Census 2000 tabulations was data swapping. It was applied to both the short form (100%) data and the long form (sample) data independently. It is also currently being used to protect American Community Survey tabulations. In each case, a small percent of household records is swapped. Pairs of households that are in different geographic regions are swapped across those geographic regions. The selection process for deciding which households should be swapped is highly targeted to affect the records with the most disclosure risk. Pairs of households that are swapped match on a minimal set of demographic variables. All data products (tables and microdata) are created from the swapped data files.

5. Synthetic data

Given a data set, one can develop posterior predictive models to generate synthetic data that have many of the same statistical properties as the original data (Abowd and Woodcock, 2001). Generating the synthetic data is often done by sequential regression imputation, one variable in one record at a time. Using all of the original data, we develop a regression model for a given variable. Then, for each record, we blank the value of that variable and use the model to impute for it. Then, we go to the next variable and repeat the process (Reiter, 2004).

Synthesizing data can be done in different ways and for different types of data products. One can synthesize all variables for all records (full synthesis) or a subset of variables for a subset of records (partial synthesis). If doing partial synthesization, we target records that have a potential disclosure risk and those variables that are causing this risk. We can synthesize demographic data and establishment data, though demographic data are easier to model and synthesize. We can synthesize data with a goal of releasing the synthetic microdata or some tabulation or other type of product (such as a map) generated from the synthetic microdata. And finally, we can generate one implicate which looks exactly like the original file, but with synthetic data; or we can generate several implicates that could be released together. Multiple synthetic replicates can be analyzed using multiple imputation analysis techniques. We are currently using synthetic data to protect our “On The Map” data product, Group Quarters data from the American Community Survey, and a file which links some of our data with data

from the Social Security Administration, and we anticipate more extensive use of it in the future (Hawala and Funk, 2007).

6. Conclusion

Using our broad definition of imputation (the substitution of one for another), we see that disclosure avoidance procedures very often involve imputation. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

The views expressed are those of the author and not necessarily those of the Census Bureau.

Laura Zayatz
U.S. Department of Commerce, U.S. Census Bureau, Statistical Research Division
E-mail: laura.zayatz@census.gov

References

- Abowd, J.M., and Woodcock, S.D. (2001), Disclosure Limitation in Longitudinal Linked Data. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, (Eds., Doyle, P., Lane, J., Zayatz, L. and Theeuwes, J.), Elsevier Science, The Netherlands, 215-277.
- Elliott, M.J., Skinner, C.J. and Dale, A. (1998). Special uniques, random uniques and sticky populations: Some counterintuitive effects of geographical detail on disclosure risk. *Proceedings of the 1st International Conference on Statistical Data Protection*. Lisbon, March 1998.
- Evans, B.T., Zayatz, L. and Slanta, J. (1998). Using noise for disclosure limitation for establishment tabular data. *Journal of Official Statistics*, Vol. 14, No. 4, 537-552.
- Federal Committee on Statistical Methodology (1994). *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*, Washington, DC, U.S. Office of Management and Budget.
- Hawala, S., and Funk, J. (2007). Model based disclosure avoidance for data on veterans. *Proceedings of the 2007 Federal Committee on Statistical Methodology Conference*.
- Massell, P., and Funk, J. (2007a). Protecting the confidentiality of tables by adding noise to the underlying microdata. *Proceedings of the 2007 Third International Conference on Establishment Surveys (ICES-III)*.
- Massell, P., and Funk, J. (2007b). Recent developments in the use of noise for protecting magnitude data tables. *Proceedings of the 2007 Federal Committee on Statistical Methodology Conference*.
- Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30, 235-242.

Using imputation to protect data confidentiality ↓

A statistical agency's ability to collect information depends on the goodwill of its respondents, be they individuals, businesses or organisations, and their trust that the agency will not reveal identifiable information about them that they have provided in confidence. The consequences of a breach in confidentiality can be devastating to the agency. At minimum, people or businesses will stop responding to surveys or will answer less truthfully – implying a severe loss in the quality and usefulness of the agency's data. Ironically, to maintain the quality of respondent data statistical agencies may resort to using perturbative methods to protect the confidentiality of their data – methods that have a negative impact on data quality. Many of the perturbative methods can be considered as types of imputations. This note deals with the use of imputation in protecting statistical data confidentiality. It starts by contrasting the roles of imputation and disclosure control methods.

Imputation's primary role is to improve data quality by correcting erroneous or missing data. It is also used to produce full (rectangular) datasets in order to facilitate the analysis of data. Some surveys, such as the Survey on Employment, Payrolls and Hours, may use mass imputation to populate a population dataset using auxiliary data from administrative sources. Strategies for imputation may vary depending on the nature of the data and on whether the objective is to produce aggregate estimates or to produce datasets for analysis. Two statistical issues involving imputation are how to use it to minimize nonresponse bias and how to account for imputation in variance estimation.

Disclosure control methods are used to protect the confidentiality of respondents' identities and data. Common data protection strategies involve data reduction, data perturbation, or both, and they typically entail some loss of data content, quality or both. Data analysis is not facilitated by disclosure control methods, as users may have to deal with incomplete datasets or account for the perturbative techniques used when interpreting results – and these are often not described in much detail. However, this is often considered as “the price to pay” for allowing greater access to data without jeopardizing confidentiality. Strategies for disclosure control vary depending on the nature of the data and on whether the information released is in the form of aggregates or microdata. The two main statistical questions when applying a disclosure control strategy are what is the resulting level of “safety” of the released data and what is the impact of disclosure control methods on data quality. The rest of this article deals primarily with the use of imputation in protecting microdata releases, which is where it is more likely to be used.

Imputation can be given a very “passive” role in disclosure control. For example one can simply take nonresponse imputation into account when determining the risk of disclosure. In microdata, the disclosure risk is often related to the likelihood of identifying a record on the datafile based on its values for a set of indirect identifiers, called a “key”. For a file of individuals, a key could include sex, age group, marital status, occupation and region. One risk measure, the probability that an individual is

identified in a microdata file given an attempt by an *intruder*, is taken as the product of the probabilities that (a) an individual is in the microdata, (b) the matching (key) variables for the individual are recorded identically on the microdata file and by the intruder, (c) the combination of key values for the individual is unique in the population and (d) the intruder is able to verify with high probability that the individual is correctly linked (Skinner, *et al.*, 1994). The first probability is affected by sampling and nonresponse. Imputation, but also response error and conceptual differences, will have a direct effect on the second probability and indirect ones on the next two. For example, a high imputation rate for the occupation code will diminish the probability of identification with a key that includes this variable. Note that, from a disclosure control perspective, it is preferable not to reveal which values on the file were imputed. Note also that the presence of imputed values does not necessarily make identification impossible, since record linkage methods that allow for some degree of “error” in the matching of datasets can be used (Winkler, 2004).

For survey data the estimation of the probability that a sample record with a unique key is also unique in the population ((c) above) is difficult. Measures have been proposed that are based on some underlying distribution of key frequencies (*e.g.*, Skinner and Holmes, 1998). Accounting for imputation or any data discrepancy between the survey and intruder data sources in these models can be difficult. Willenborg and de Waal (2001) provide a measure of the re-identification risk of an individual microdata record that takes into account misclassification in the matching key. One component of that risk is directly related to the non-misclassification rate.

At the opposite extreme, imputation can be used very aggressively to generate an entire synthetic dataset from a microdata file (*e.g.*, Abowd and Woodcock, 2001). The aim is to produce a “safe” file whose analytic properties are similar to those for the original dataset. For surveys that do release anonymized public-use microdata files mass imputation can also be used to replace those variables that were excluded from the file because their inclusion was considered to represent a significant disclosure risk. In both situations, and under certain conditions, multiple imputation can be used to allow the incorporation of the variance due to imputation during data analysis. The generation of synthetic data is done by a random process and is considered to represent almost no risk of disclosure. However, generating analytically useful synthetic data can be very laborious, particularly if one is to take into account the edit relationships between variables on the data.

Between these two extremes, the use of imputation to protect data confidentiality varies between methods applied globally (to certain variables in the entire dataset) and those that are applied locally (targeting at-risk records). Data perturbation can be thought of as a type of imputation – although the link is more tenuous when it is applied at the global level.

Among the data perturbation methods that are applied globally, one that is used for continuous data is the addition of noise. Noise addition can be used to prevent the identification of individual records (*e.g.*, if another organization that has the same variables uses them to link their data records with those

of the statistical agency) or of their values. For example, Fuller (1993) and Kim and Winkler (1995) propose additive noise that has zero mean and the same correlation structure as the original data. This method preserves the first order moments and the correlation structure for the domains whose correlation structure is used for the additive noise.

For categorical variables the Post-Randomisation Method (PRAM) is used to replace the reported response categories by others according to a predetermined set of probabilities. The method is described in Willenborg and de Waal (2001). The matrix of transition probabilities is passed to the user so that he can make an adjustment when analysing the data. Alternatively, probabilities of movement can be devised in such a way that the perturbed data have the same univariate distributional properties. Naturally, when perturbation is applied to variables independently the correlation between variables can be severely affected.

With global methods the statistical agency can control the level of noise or perturbation and provide information about it to the data user. The impact of global perturbation on analysis is an issue. Global methods can cause excessive perturbation, for example, if dealing with a population requiring little protection because it is fairly homogeneous.

Unlike with regular imputation, data users can be given the opportunity to validate their results by resubmitting the models that they developed with perturbed data against the actual data file. However, it should be noted that providing users with multiply-imputed copies of the perturbed file may pose a disclosure risk unless the entire datafile was generated artificially (as with synthetic data). This is because the outcomes can be pooled to get closer to the original values.

Global imputation methods can also be used when the objective is not the release of microdata. For example, additive noise can be added to establishment microdata to allow the release of tabular data without having to suppress values of confidential cells (Evans, Zayatz and Slanta, 1998). This spares the need to suppress nonconfidential cells to prevent the derivation of values for the confidential cells – a difficult problem that can become unwieldy when related tabular outputs are released. Microdata perturbation is also proposed for remote access systems. For example, Keller-McNulty and Unger (1998) propose imputing the value $y_i^* = y_i + X_i H_i \bar{y}_G$ for each record y_i in a query set G , where \bar{y}_G is the mean of the query set, X_i takes values among $\{-1, 0, 1\}$ with predetermined probabilities, and $H_i \sim \text{Uniform}(U, L)$ for $0 \leq L \leq U \leq 1$.

Finally, data perturbative methods can be applied locally to the set of records that present a higher risk of identification. The selection of records for data perturbation can be deterministic or it can be stochastic (*e.g.*, use perturbation rates that vary according to the identification risk). Local perturbation has the advantage that its “damage” is more limited than with global perturbation. With quantitative microdata outlier values may be considered to represent sensitive information and be top-coded or imputed. To preserve totals the largest n values in each domain can be replaced by their average value. For categorical data high risk records would be those that possess a combination of values for a set of identifying variables (key) that is unique in the population (called population uniques). On microdata

files from survey samples only a subset of the records with a unique combination of identifying variables in the sample (sample uniques) are also population uniques – but we do not know which subset.

The heuristic concept of multiplicity has been proposed to identify records presenting a greater risk of identification (*e.g.*, Boudreau, 1995). Multiplicity works on the assumption that units that are unique in the population as well as in the sample will usually be unique based on smaller subsets of variables than those that are unique in the sample only. If one has province, age group, sex and 15 other indirect identifiers, one generates all 455 six-dimensional tables involving the first three variables plus any three of the remaining 15 variables. A record's multiplicity value is the number of tables where the record appears as unique in its cell. The higher the multiplicity, the greater the likelihood that the record is also unique in the population.

One issue with multiplicity is knowing what threshold to set for the identification of high risk records. Another involves knowing which variables to perturb (impute) for those records. One option consists of changing the variable that is most responsible for the record's high multiplicity, *i.e.*, the one present the most often in tables where the record shows up as unique. Another would be to change the variables that would make identification least likely, without making the record inconsistent.

That data perturbation can introduce inconsistencies in the microdata is a problem in itself. But inconsistencies can also be exploited by hackers to determine if some high-risk records have been perturbed. Willenborg and Van den Houte (2006) propose a method for protecting microdata that eliminates unsafe (low frequency or unique) combinations by replacing them with combinations from a donor record that are consistent with other variables. Donors are chosen using direct matching or using some distance metric, as in nearest neighbour imputation.

The above has shown several strategies for using imputation to protect data confidentiality, each with a different impact on the resulting data. However, imputation may provide its greatest benefit for disclosure control when it is used alongside other strategies (including other imputation strategies). For example, after applying additive noise, such as in Kim and Winkler (1995), imputation may be used to further protect those few units which are not deemed "sufficiently protected" by the additive noise, perhaps because they are still far away from other units. This use of imputation allows the additive noise to be calibrated at a level that is sufficient to protect most, but not all, units. The negative impact on data quality is lessened.

A very good example of the integration of imputation with other strategies is the Micro Agglomeration, Substitution, Subsampling, and Calibration (MASSC) method (Singh, Yu and Wilson, 2004). MASSC partitions the data into risk strata (the highest risk stratum corresponding to units that are unique with respect to sets of core identifying variables), substitutes (imputes) values of key variables for a randomly selected subset of records with those of similar records (nearest neighbours), applies subsampling to further protect the data and calibration to reduce the variability introduced by this sampling. The substitution rate is highest in the high risk strata, which is what efficiency would

dictate, but all records are subject to substitution (even those in low risk strata). Due to the stochastic nature of substitution and subsampling, disclosure risk measures can be calculated.

An important benefit of applying sets of strategies for disclosure control is that it makes attempts to circumvent the disclosure control methods very difficult. And imputation can always be used to handle the residual disclosure risk – with minimal impact on the data. Hackers cannot model how the different methods interact, but the statistical agency, having access to the original data, can always assess the impact on data quality empirically.

The use of imputation to protect data confidentiality is still open to innovations and no doubt new types of methods remain to be developed. Imputation will certainly continue to play a key role in providing greater access to data while preserving confidentiality.

Jean-Louis Tambay
Statistics Canada
E-mail: jean-louis.tambay@statcan.gc.ca

References

- Abowd, J., and Woodcock, S. (2001). Disclosure limitation in longitudinally linked data. *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, 215-277. Amsterdam: North Holland.
- Boudreau, J.R. (1995). Assessment and reduction of disclosure risk in microdata files containing discrete data. *Proceedings of Statistics Canada Symposium 95: From Data to Information – Methods and Systems*, 143-153.
- Evans, T., Zayatz, L. and Slanta, J. (1998). Using noise for disclosure limitation establishment tabular data. *Journal of Official Statistics*, 14, 537-551.
- Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383-406.
- Keller-McNulty, S., and Unger, E. (1998). A database system prototype for remote access to information based on confidential data. *Journal of Official Statistics*, 14, 347-360.
- Kim, J., and Winkler, W. (1995). Masking microdata files. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 114-119.
- Singh, A.C., Yu, F. and Wilson, D.H. (2004). Measures of information loss and disclosure risk under MASSC treatment of micro-data for statistical disclosure limitation. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 4374-4381.
- Skinner, C., and Holmes, D. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14, 361-372.
- Skinner, C., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics*, 10, 31-51.

- Willenborg, L., and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. Lecture notes in Statistics, Volume 155, New York: Springer-Verlag.
- Willenborg, L., and Van den Hout, A. (2006). Peruco: A method for producing safe and consistent microdata. *International Statistical Review*, 74, 271-284.
- Winkler, W.E. (2004). Re-identification Methods for Masked Microdata. U.S. Census Bureau Statistical Research Division report #2004-03.

The links between confidentiality and imputation ↓

1. Introduction

In this paper we will provide an overview of topics that are related to both editing and imputation as well as statistical disclosure control (SDC). These two fields of research have a lot in common but differ on many points as well. We will describe some of the similarities as well as some of the differences. We will do this from an SDC viewpoint; *i.e.*, we will approach this paper using the two traditional SDC fields: microdata and aggregated data. We are aware of the fact that many of the topics that we will mention have been discussed before. However, we think that it is a good idea to give an overview of the current issues.

The information from statistics becomes available for the public in tabular and microdata form. Historically, only tabular data were available and National Statistical Institutes (NSIs) had a monopoly on the microdata. Since the eighties the PC revolution led to the end of this monopoly. Now also other users of statistics have the possibility of using microdata. These microdata can be conveyed with floppies, CD-ROMs, USB sticks and other means. Recently also other possibilities of getting statistical information have become more popular as remote access and remote execution. With these techniques researchers can get access to data that remain in a statistical office or can execute set-ups without having the data on their own PC. For very sensitive information some NSIs have the possibility to let bona fide researchers work on-site within the premises of the NSI.

The task of statistical offices is to produce and publish statistical information about society. The data collected are ultimately released in a suitable form to policy makers, researchers and the general public for statistical purposes. The release of such information may have the undesirable effect that information on individual entities instead of on sufficiently large groups of individuals is disclosed. The question then arises how the information available can be modified in such a way that the data released can be considered statistically useful and do not jeopardize the privacy of the entities concerned. The Statistical Disclosure Control (SDC) theory is used to solve the problem of how to publish and release as much

detail in these data as possible without disclosing individual information (Willenborg and De Waal, 1996 and 2001).

In this short paper imputation as a SDC method is discussed. In section 2 different kinds of microdata are discussed. Section 3 deals with tabular data. A short discussion and some conclusions can be found in Section 4.

2. Microdata

2.1 Introduction

Many National Statistical Institutes (NSIs) produce so-called microdata for research (MUCs). These MUCs aim at facilitating a select group of bona fide researchers. Some NSIs also produce public use files (PUFs), available to the general public. Both kinds of microdata have to be protected, by legal measures or SDC methods. SDC methods that can be used to protect microdata are described in the European Handbook on SDC (<http://neon.vb.cbs.nl/cenex/>) and implemented in statistical software packages like μ – ARGUS, see *e.g.* Hundepool *et al.* (2007).

All SDC techniques necessarily involve data manipulation or suppression and are likely to reduce the quality of estimates to be produced from the data. As a result, NSIs have begun to investigate other methods that allow use of data while protecting confidentiality of sensitive information given by respondents. Probably the most important access modality developed in the past decade is that of restricted access sites. This facilitates a possibility to individual researchers to perform their research on richer microdata on the premises of the NSIs. Bona fide researchers have the opportunity to work on-site in a secure area within the NSI. At Statistics Netherlands (SN) such an on-site facility is present. Researchers can do their research on only mildly protected microdata but their results may not leave the premises of SN, unless an officer of SN has checked and cleared those results on confidentiality. See Kooiman *et al.* (1999) for more information.

Very recent developments are remote access and remote execution. In principle, remote access is the same as the on-site situation, except that the researcher is connected with the NSI using a secure connection where the data (and the results) remain on the servers of the NSI. Again, approval from an officer is needed to get the results from those servers to the researchers' computer. Remote execution is more limited to the researcher: a script is sent to the NSI, the NSI runs the script, the result is checked on confidentiality and if approved the output is finally sent to the researcher.

2.2 Editing, imputation and SDC on microdata

In the statistical processes at NSIs, often imputation is applied before SDC is concerned. This raises some very interesting questions:

- Should SDC still be used when imputed (*i.e.* non-real) data is concerned? Is it legally allowed to publish imputed data on individual respondents?
- How does the imputation model influence the rules that determine whether or not the confidentiality is breached and SDC is needed?
- Should the imputation model take the SDC rules already into account at the beginning of the statistical process?
- Should SDC methods take the edit-rules into account? See *e.g.* Shlomo and De Waal (2008).

Whether or not imputed data are to be protected by SDC, also depends on the aim of the imputation process: best values on micro level or on macro level? The former would result in an imputation model that estimates the missing value as accurate as possible and hence would yield a value that would be very close to the true value. The latter however might result in data on micro level that are not even close to the true value, as long as the estimates on macro level are accurate enough. Apart from the legal aspects, this has implications on the need for SDC.

Using imputation itself as an SDC method is also an option. In the literature, many instances can be found where so called synthetic data are produced. To produce synthetic data, an imputation model is derived and used to either impute some of the sensitive data (partially synthetic data) or produce a file completely determined by the imputation model (fully synthetic data). Obviously, this method can have some drawbacks: a researcher using the synthetic data might just find the imputation model the NSI had used to produce the data. Usually, analyses only work within the framework of the assumptions of the applied imputation model. Any analysis outside that framework might lead to undesired results.

Both imputation and SDC are concerned with missing data: imputation techniques try to fill the gaps whereas SDC deliberately puts holes in the data. Imputation techniques often assume that the missing data are Missing (Completely) At Random (MAR or MCAR). SDC on the other hand makes a very selective subset of the data missing: *i.e.* this results in a dataset where the missing data are explicitly *not* MAR.

One could argue that after local suppression of data has been applied as an SDC technique, imputation could be used to fill the gaps again. However, this might just as well undo the confidentiality, especially when a ‘good’ imputation model is used.

SDC methods might learn from imputation methods (see *e.g.* Schulte Nordholt, 1998) as well; *e.g.*, let us consider a stochastic SDC method as PRAM (see *e.g.* Gouweleeuw *et al.*, 1998). This means that the resulting protected micro datafile is a realisation of a stochastic experiment. To assess the variance this method adds to estimators, a similar approach as the one used by multiple imputation might be used. In multiple imputation, several values are imputed for every missing value. The variation among all these values reflects the uncertainty under one model for nonresponse and across several models (Rubin, 1996). Each set of imputations is used to create a complete data set, each of which is to be analysed as if

there were no missing value. It can be shown that multiple imputation yields better inferences (Little and Rubin, 1989).

3. Tabular data

3.1 Introduction

Imputation is often only applied on microdata. However, it may have severe impact on aggregated data as well. Obviously, tables produced using imputed data might differ from tables produced on the basis of the original data.

3.2 Editing, imputation and SDC on tabular data

Since imputation is often applied before tabular data are produced, again the question is how the imputation model influences the rules that determine whether or not SDC is needed. When applying SDC to tabular data, the data of individual respondents is used to find the primary unsafe cells. Another question would be: is SDC needed when the underlying data is ‘not-real’ (*i.e.* imputed)?

An obvious link between editing and SDC on tabular data is given by linear restrictions. When applying SDC on tables, the structure of the table is defined by a set of linear relations between the cells of the table. Edit rules are often linear restriction as well.

An often used SDC method in tabular data is cell suppression. Confidentiality rules determine the primary unsafe cells. When they are suppressed, additional suppressions are needed to guarantee that these primary unsafe cells cannot be estimated too accurately using the table at hand. The way to look for the best possible way to suppress secondary cells is very similar to the way one could look for the variable that is most likely the one that violates an edit rule. Indeed, in De Waal (2003) this relation was already mentioned.

4. Discussion and conclusions

In this paper we have given a limited overview of some of the similarities as well as differences between the fields of imputation and SDC. Some of the things we mentioned are very well known among the two research communities. However, we feel that both research communities can still benefit a lot from each other. Moreover, the NSIs should be more aware of and pay more attention to some of the links we described in this paper.

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

Eric Schulte Nordholt and Peter-Paul de Wolf
Statistics Netherlands
E-mail: ESLE@CBS.NL and PWOFF@CBS.NL

References

- De Waal, T. (2003). *Processing of erroneous and unsafe data*, ERIM Ph.D. Series Research in Management 24, Rotterdam, The Netherlands, Erasmus University Rotterdam.
- Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J. and de Wolf, P.P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14, 463-478.
- Hundepool, A., van de Wetering, A., Ramaswamy, R., Franconi, L., Poletini, S., Capobianchi, A., de Wolf, P.P., Domingo, J., Torra, V., Brand, R. and Giessing, S. (2007). *μ - ARGUS, user's manual, version 4.1*. Voorburg, The Netherlands: Statistics Netherlands.
- Kooiman, P., Nobel, J.R. and Willenborg, L.C.R.J. (1999). Statistical data protection at Statistics Netherlands. *Netherlands Official Statistics*, 14, 21-25.
- Little, R.J.A., and Rubin, D.B. (1989). The analysis of social science data with missing values. *Sociological methods and research*, 18, number 2&3, 292-326.
- Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.
- Schulte Nordholt, E. (1998). Imputation: Methods, simulation experiments and practical examples. *International Statistical Review*, 66, 157-180.
- Shlomo, N., and de Waal, T. (2008). Protection of micro-data subject to edit constraints against statistical disclosure. *Journal of Official Statistics*, 24, to be published.
- Willenborg, L.C.R.J., and de Waal, T. (1996). *Statistical Disclosure Control in practice*, Lecture Notes in Statistics 111, New York: Springer-Verlag.
- Willenborg, L.C.R.J., and de Waal, T. (2001). *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics 155, New York: Springer-Verlag.



Volume 1, Number 1

- History of the word imputation
- Why do we impute?
- Overview of impudon

Eric Rancourt
David Haziza
Joël Bissonnette

Volume 8, Number 2, 2008
The Imputation Bulletin

Volume 1, Number 2

- The risks of imputation
 - Canadian census edit and imputation system (CANCEIS)
 - The connection between models and commonly used imputation methods
 - Good practices in imputation - Generic processing systems
- David Haziza
Mike Bankier
Jean-François Beaumont
- Matt Briggs and Colleen Martin

Volume 2, Number 1

- When are we in the presence of nonignorable nonresponse?
 - BANFF
 - Imputation classes
 - Imputation in the Survey of Employment, Payrolls and hours (SEPH)
- Jean-François Beaumont
Chantal Marquis and Robert Kozak
David Haziza
Sophie Arsenault

Volume 2, Number 2

- Distorsion of distributions
 - GENESIS : Generalized system for imputation simulation
 - What is nonresponse variance?
 - Imputation in the Longitudinal Survey of Immigrants to Canada (LSIC)
- David Haziza
David Haziza
Jean-François Beaumont
Sophie Arsenault

Volume 3, Number 1

- Imputation methods
 - The system for estimation of variance due to nonresponse and imputation (SEVANI)
 - Introduction to variance estimation in presence of imputation
 - Imputation in the Survey of Household Spending (SHS)
- Graham Kalton
Jean-François Beaumont
- Eric Rancourt
Sophie Arsenault

Volume 3, Number 2

- Variance estimation in the presence of imputation for item nonresponse
 - PROC MI and PROC MIANALYZE in SAS
 - A comparison of regression models to impute continuous variables of partially complete observations
- J.N.K. Rao
- David Haziza
Peter Wright

Volume 4, Number 1

- Variance estimation under the two-phase imputation model approach
 - The reverse approach to variance estimation from survey data with imputed values
 - Imputation in action
- David Haziza et Eric Rancourt
- Steve Matthews
- Sophie Arsenault

Volume 4, Number 2

- Variance due to ratio model imputation
 - The E-M algorithm as an imputation benefit
 - Imputation in action
- Daniel Hurtubise
Catalin Dochitoiu
Jean-François Beaumont and
Cynthia Bocci

Volume 5

- Substitution Imputation vs. Nearest Neighbour Imputation: An Application to Estimation of a Distribution
Ray Chambers
- Some Thoughts on Nearest-Neighbour Imputation
Jean-François Beaumont and
Cynthia Bocci
- Tax Replacement in Business Surveys: Variations on an Imputation Theme
Eric Rancourt

Volume 6, Number 1

- Adjusting for Measurement Error Using Predictive Mean Matching Imputation Methods
Gabriele B. Durrant
- Simulation Studies in the Presence of Nonresponse and Imputation
David Haziza
- Imputation in action
Karla Helgason and
Jean-Sébastien Provençal

Volume 6, Number 2

- How Can Data Editing Improve Quality?
John Kovar
- Edit Restrictions in Error Localization and Imputation
Ton de Waal
- Verification and Imputation at Statistics Canada
Claude Poirier
- Statistics Canada's Quality Guidelines: Editing

Volume 7

- Multiple Imputation of Missing Data in Surveys
Roderick Little and
Trivellore Raghunathan
- Frameworks for Variance Estimation in the Presence of Imputed Data
David Haziza
- Application of SEVANI to the 2005 Survey of Innovation
James Ahkong

Volume 8, Number 1

- Synthetic donor imputation
Joël Bissonnette
- Imputation in action
Sean Crowe
- Statistics Canada quality guidelines: Imputation

- Abayomi, K., Gelman, A. and Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society - Series C Applied Statistics*, vol. 57, 3, 273-291.
- Foulkes, A.S., Yucel, R. and Reilly, M.P. (2008). Mixed modeling and multiple imputation for unobservable genotype clusters. *Statistics in Medicine*, vol. 27, 15, 2784-2801.
- Heaton, T.J., and Silverman, B.W. (2008). A wavelet- or lifting-scheme-based imputation method. *Journal of the Royal Statistical Society. B*, 70, Part 3, 567-587.
- Huang, J., Lee, C. and Yu, Q. (2008). A generalized log-rank test for interval-censored failure time data via multiple imputation. New York: John Wiley & Sons, Inc., *Statistics in Medicine*, vol. 27, 17, 3217-3226.
- Kadilar, C., and Cingi, H. (2008). Estimators for the population mean in the case of missing data. *Communications in Statistics - Theory and Methods*, vol. 37, 14, 2226-2236.
- Reiter, J.P., and Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462-1471.
- Shao, J., and Wang, H. (2008). Confidence intervals based on survey data with nearest neighbor imputation. *Statistica Sinica*, vol. 18, 1, 281-298.
- Wood, A.M., White, I.R. and Royston, P. (2008). How should variable selection be performed with multiply imputed data? New York: John Wiley & Sons, Inc., *Statistics in Medicine*, vol. 27, 17, 3227-3246.
- Yang, X., Li, J. and Shoptaw, S. (2008). Imputation-based strategies for clinical trial longitudinal data with nonignorable missing values. *Statistics in Medicine*, vol. 27, 15, 2826-2849.
- Yucel, R.M., He, Y. and Zaslaysky, A.M. (2008). Using calibration to improve rounding in imputation. *American Statistician*, 62, 2, 125-129.

 **Statistics Canada Intranet Sites** ↓

[Generalized Edit and Imputation System \(BANFF\)](#)
[Nearest Neighbour Imputation Methodology \(NIM\)](#)
[UN/ECE Workshop on Editing](#)

 **Internet Sites** ↓

<http://edimbus.istat.it>
<http://www.cs.york.ac.uk/euredit/>
<http://www.multiple-imputation.com>
<http://www.unece.org/stats/archive/02.02.e.htm>

 **News Group**

[Impute listserv](#)

This group is active with questions/answers on imputation.

To subscribe, send an e-mail at listar@utdallas.edu with “subscribe impute” on the subject line. Once you receive a message, reply or forward it to the sender to complete your registration.