WP. 7
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (i): Harmonization of statistical data confidentiality – legal and methodological aspects

HARMONISATION OF ANONYMISATION PRACTICES THROUGH PARTIALLY SYNTHETIC FILES

Invited Paper

Prepared by Sébastien Pérez-Duarte, European Central Bank

Harmonisation of anonymisation practices through partially synthetic files[†]

Sébastien Pérez-Duarte[*]

[*] European Central Bank, Kaiserstrasse 29, 60311 Frankfurt, Germany. sebastien.perez-duarte@ecb.int

Abstract: The problem of ensuring the application of harmonised anonymisation measures to an international survey can often be solved by selecting the strictest anonymisation procedures in each country. The data will be comparable across countries at the expense of the level of detail or the amount of information available. Even if a compromise in the anonymisation practices acceptable to all countries were possible, some countries might object that the information content of their data is being reduced more than necessary. We describe a procedure using partially synthetic data where differing country practices in terms of data reduction techniques (recoding of continuous and categorical variables) are respected while the information content is preserved as much as possible.

1 Introduction

The work on harmonising a survey across countries does not stop with the data collection. Besides the need to ensure compatible ex-post data editing, imputation, and weighting procedures, the anonymisation of the data prior to dissemination must also be done in as much a comparable way as possible. However, in some cases harmonised anonymisation will not be possible due to differing country constraints. If a compromise on the anonymisation techniques is not possible, alternative solutions must be explored.

If harmonisation at any cost were the only objective and anonymisation techniques differed only in their degree of coarseness, then the combination of all the country-specific anonymisation procedures could be applied to all the country datasets. This might lead to an unbearable loss of information, in particular if the anonymisation procedures differ in specifically incompatible ways. If the techniques differed in structure as well, it might not even be sensible to apply all rules simultaneously. Another alternative would be to provide the data "as-is", i.e. including the data anonymised by the different country procedures unchanged in the cross-country dataset. This maximises the information content in the data while respecting country procedures, but at a cost to the final users in terms of a loss of comparability: indeed

[†] Comments on this paper by Claudia Biancotti, Arthur Kennickell, Carlos Sánchez Muñoz, Patrick Sandars, and Caroline Willeke are gratefully acknowledged. The views expressed in this paper are those of the author and do not necessarily reflect those of the European Central Bank.

the users would have then to deal with the differing anonymisation procedures and the resulting differences in the data, and may have less information than the producers of the data to do so (similarly to what Rubin [1996] has argued for imputation of missing data, compatible anonymisation could be argued to be the responsibility of the data constructors).

We introduce the concept of the *infimum* of a set of data reduction methods; the infimum is the coarsest data reduction compatible (in a specified sense) with the country-specific methods, and is a mathematically valid solution to the selection of a fully harmonised procedure. However, such a solution would not be acceptable by all countries. This paper provides an alternative to the anonymisation of cross-national surveys other than the straight suppression of information/variables, namely through the use of partially synthetic data generated parsimoniously to preserve country constraints in data disclosure. Section 2 sketches the problem of incompatible anonymisation techniques. Section 3 presents the models of synthetic data and provides a simple example. Section 4 concludes.

2 Statistical producers in different countries have differing anonymisation strategies and differing goals

Statistical disclosure control is driven by two opposite forces (Duncan et al., 2003): preserving the utility to the users and reducing the disclosure risk of the data. Both are in part driven by local factors: on the one hand, interesting topics are related to the situation of the country and its economic, social and institutional setup; on the other hand, disclosure risk hinges on the availability of external data that can be used for matching. Moreover, rare combinations of variables do not have to be the same across countries, resulting in different disclosure control techniques. These elements are set out in this section.

2.1 Local circumstances, local interests, and different attacker scenarios…

The regulation (European Commission) No 831/2002, implementing Council Regulation (EC) No 322/1997 as amended by (EC) No 223/2009 on Community Statistics and concerning access to confidential data for scientific purposes, states that " '*anonymised microdata*' [means] individual statistical records which have been modified in order to minimise, in accordance with current best practice, the risk of identification of the statistical units to which they relate". However, anonymisation practices implemented to reach this goal differ widely across countries. In part, this is the result of differences in sensibilities regarding the privacy of personal data, which will not be elaborated upon here. A second factor is the perceived or actual risk of disclosure of data posed by the national institutions, external data sources available, and national idiosyncrasies in behaviour or choices available to individuals. The "attacker scenarios" used to identify threats can differ across countries, resulting in

different determinations of which combinations of unaltered variables would pose the greatest risk if they were released. Some differences can also be explained by specific country characteristics, inasmuch as they affect the risk of indirect identification. As an immediate example, countries with public register data (e.g. as in Nordic countries for income and other characteristics) must take this fact into account. Even where the nature of risks is very similar in two countries, existing institutional practices may incline the data producers apply different procedures for disclosure control.

2.2 …result in different disclosure protection practices

In general, statistical disclosure control principles recommend that characteristics shared by a small fraction of the population, or of the sample, should not be available in the data released, as the risk of identification may be unacceptably high. This applies in particular to highly visible characteristics, such as age or occupation. The recommended procedure usually is to collapse the rare codes with other codes, through top or bottom coding, recoding, or other measures that coarsen the information. As the infrequent characteristics may not be the same ones across countries, this may lead to incompatible anonymisation procedures.
For example, the criteria listed in the *Checklist on Disclosure Potential of Proposed Data Releases* (Confidentiality and Data Access Committee, 1999) indicate a few rough guidelines for the anonymisation of data. All identified geographic areas should have more than 100,000 persons in the sampled area. Although "there are no hard and fast rules for determining which cutoffs to use in topcoding", the Census Bureau recommends topcoding at least the top 0.5% of the non-zero values. By contrast, and although the size of the country may play a role, Statistics Netherlands recommends that in the data made available to researchers regions that can be identified need to have at least 10,000 inhabitants, the maximum level of detail for occupation, firm and level of education needs to be limited depending on the geographic information available, and the combinations of "very identifying" variables (i.e., those easily externally observable) should occur at least 100 times in the population (Schulte Nordholt, 2003).

2.3 These practices are difficult to homogenise while preserving information content.

Data users would like to have both homogeneous data across countries and maximum information content.

As a simple hypothetical example, consider two countries disagreeing on the topcoding of age. Country A requires the data to be topcoded at 80, country B at 90. The non-harmonisation solution would leave the data from each country unchanged. Harmonising the anonymisation could be done by:

1. Topcoding at age 80

2. Topcoding at age 90
3. Topcoding at age 90 and recoding ages 80 to 89 as a single group
4. Providing a harmonised variable topcoded at age 80, and another one with the age not harmonised (i.e. with age details between 80 and 90 only for country B)

None of the first three solutions satisfies both countries in terms of the anonymisation applied and usability of the data: country A may object to solutions 2 and 3 because the information is not anonymised as would be required, while country B would object that solutions 1 and 3 limit the information content of the data. While solution 3 could be a compromise in between both country practices, eventually neither country would be fully happy with this solution for different reasons. Alternatively, both countries would agree on solution 4, but the problem would get transferred to users, who would then be confronted with the need to further edit and harmonise the data such that they can eventually feed in their econometric models; users may also specifically account for the differing truncations in the structure of their models.

The problem is even more acute if, in addition to the pooled harmonised dataset, country B releases its own version of its data, with aggregate statistics by age bracket. If the aggregate statistics followed the truncation in the national version of the file, users of the pooled and harmonised dataset may wish to be able to replicate this aggregation, but this might impose additional constraints on the common anonymisation procedure.

3 Partially synthetic files for the harmonisation of anonymised data

Synthetic files have been proposed as a solution for disclosure limitation since the seminal papers of Little (1993) and Rubin (1993), which have spawned a growing literature (*inter alia* Reiter [2002]) that it is not our intention to review here. Although fully synthetic files, where the full dataset has been replaced by multiply-imputed values, are rare in practice, partially synthetic datasets have already been used for disclosure control purposes. For example, as described in Kennickell (1997) for the U.S. Survey of Consumer Finances, the U.S. Federal Reserve Board replaces the monetary values of those observations carrying a high risk of disclosure with multiply-imputed values, and releases this mixture of observed and synthetic data. Reiter (2003) specifically addresses the problem of inference with partially synthetic data sets.

Another approach is taken in Little and Liu (2003), who describe a procedure of selective multiple imputation of key variables. Key variables (for example age and gender) are the characteristics of individuals that data intruders might be able to obtain from publicly available sources. Those key variables of observations with high

disclosure risk and of a selected sample of other observations are replaced with multiply-imputed values.

3.1 Setup and notation: data reduction as a partition of a set

We only consider in this paper the disclosure control methods known as "data reduction methods" (Eurostat 1996), which include suppression of variables, reduction in detail (merging codes or changing a variable from continuous to categorical), top and bottom coding, and suppression of extreme values. This excludes "data modification methods" (noise addition, data swapping, microaggregation) whose information loss cannot be codified as described below.

All these methods can be mathematically represented by the partitions of the set K of different values of a variable. For example, we would have $K = \{\text{male}, \text{female}\}$ in the case of the gender variable, or $K = \{0,1,2...,120\}$ for age in years. A partition of the set K is a division of the set in non-overlapping pieces (pairwise disjoint) that cover all of K. Any data reduction method considered here can be represented as a partition of this set, $P_i(K)$. In the case of age, top coding at age 80 is equivalent to the partition with the pieces {0}, {1},…, {79}, and {80, 81, 82,…,120}. Recoding continuous variables can also be described in this way with partitions, by using a set K with continuous values.

Therefore, the partition of a set, and the recoding of a variable taking values in that set, are two alternative but interchangeable facets of the same phenomenon, and we will use this correspondence in what follows. We introduce some further notations and definitions, in order to describe how partitions can be analysed, compared, and combined.

3.1.1 Finer and coarser partitions

The first question is, can two partitions be compared? If each piece in partition P is included in a piece in partition Q, the answer is yes, and we will write that $P \leq Q$ and will say that P is *finer* than Q, or that P is a refinement of Q (equivalently, Q is *coarser* than P). In terms of data reduction, this implies that P is less anonymised than Q and that a variable recoded as P could be recoded as Q.

The order $\leq$ is only a partial order (not every two partitions can be compared), as two partitions can be "incompatible" (neither variable typology can be recoded as the other).

3.1.2 Infimum of two partitions

Nevertheless, two partitions can always be "intersected". We will write $P \wedge Q$ the *infimum* partition, defined as the coarsest partition finer than P and Q. Since the set

of partitions of a set is a complete lattice, such an infimum always exists (at worst, the infimum is the set of singletons of K). In data reduction terms, the infimum $P \wedge Q$ specifies the coarsest recoding of the variable of interest that can still be recoded as P and Q.

As an example, consider the continuation of the example in section 2.3, of two countries disagreeing on the topcoding of the age variable. In the notation of this section, the partition for country *A*, topcoding at age 80, would be written $P_A = \{\{0\},\{1\},...,\{79\},\{80,81,...,120\}\}$. For country B, which topcodes at age 90, the partition would be $P_B = \{\{0\},\{1\},...,\{89\},\{90,91,...,120\}\}$. The infimum of these two partitions is $P_A \wedge P_B = \{\{0\},\{1\},...,\{89\},\{90,91,...,120\}\}$, which is identical to P_B. Since the topcoding of *B* is at a higher level than that of *A*, P_B is finer than P_A (P_B has more information content than P_A.

3.1.3 Disagreement of two partitions

We introduce a final definition and notation. Suppose that $P \leq Q$. Since P is finer than Q, Q may contain subsets that are not included in any subset of P. We will write $Q \setminus P$ the list of those subsets in Q that "disagree" with the sets in P, and will call $Q \setminus P$ *disagreement* of Q over P. This is defined as:

$$Q \setminus P = \{ X \in Q \mid \forall Y \in P,\ X \not\subset Y \}.$$

The disagreement of one partition over another can be empty if the two partitions are identical.

In the example with the top coding of age, since $P_B \leq P_A$, $P_A \setminus P_B$ is well-defined and is equal to $\{\{80,81,...,120\}\}$ (the double brackets specify that this is a list of sets, not a list of values).

Any value that is in one of the elements of the disagreement set of Q over P is called a *disagreeing value*. In other terms, a disagreeing value of Q over P cannot be straightforwardly recoded as a value in P.

3.1.4 Extension of notation to multiple countries

The definitions above can be generalised to a setting of *I* countries, indexed by *i*. The data reduction procedure of each country can be summarized by a partition of the set of values of that variable. Let P_i be the partition for country *i*. Different countries may have different data reduction methods. The infimum of all these partitions is defined as:

$$\overline{P} = \bigwedge_{i=1}^{I} P_i .$$

This infimum is the coarsest partition that is also a refinement of all the P_i partitions. In other words, it is the most disruptive data reduction method that can still be recoded according to any of the country specific methods.

3.2 Application to multi-country harmonisation of data reduction methods

Our approach is to treat the infimum of the different data reduction methods $\overline{P}$ as the desired output format of the variable of interest. Since it is finer than any of the country data reduction methods, it can be recoded into any of them, and it is thus compatible with all country practices. Furthermore, since it is the coarsest refinement, it is the procedure most compatible with all country practices. However, since it is finer than some country methods, it cannot be accepted as the harmonised anonymisation.

For each country *i*, since $\overline{P} \leq P_i$, we can construct the disagreement set of P_i over $\overline{P}$, $D_i = P_i \setminus \overline{P}$. The disagreement set D_i will include the values whose recoding differs between P_i and $\overline{P}$. The values not in the disagreement set are recoded in the same way in P_i and in $\overline{P}$, and country i will not object to them being released in this way.

Any disagreeing value in country i must be transformed in the harmonised released data. It is set as “sensitive” and is set to be multiply imputed. The inference formulas of Reiter (2003) are appropriate in this case, and are not recalled here.

Additionally, it is possible and desirable to restrict the generated values to lie in the appropriate disagreement subset. This ensures that, when recoded according to the country specific method, the values remain in the same group or category, and it avoids selection bias.

This method does not disclose in which of the subsets of $\overline{P}$ the original value lies, since all values in the disagreement set D_i for country i are replaced by multiply imputed values.

In the example of the top coding of age, all ages of 80 and above in country A are replaced by multiple imputed values drawn from the distribution of ages, conditional as well on the inclusion in the disagreement set. In this case, the imputed ages in country A can remain in the class of "80 years or more". Imputed values are then top coded at 90, as is required by the infimum of the two country procedures.[1]

3.3 Example – Harmonisation of anonymisation of the number of employees

As a more fully-fledged example, we consider the harmonisation of the anonymisation of the number of employees in a business owned by a household. Since most of the firms have few employees, this variable is seen as identifying in several countries. Country A anonymises this variable by recoding in intervals 0, 1 to 2, 3 to 9 and 10 and more. In country B, it is very important for analysts to be able to distinguish firms with 5 employees or less, so the variable is recoded as "5 or less, more than 5".

In the notation of this section, the anonymisation of each country can be represented by the following partitions:

$$P_A = \{\{0\},\{1,2\},\{3,...,9\},\{10,...\}\},$$

$$P_B = \{\{0,...,5\},\{6,...\}\}.$$

The infimum of these two partitions is:

$$\bar{P} = P_A \wedge P_B = \{\{0\},\{1,2\},\{3,4,5\},\{6,..,9\},\{10,...\}\}.$$

The disagreement set between the practices in each country and the infimum are:

$$D_A = P_A \setminus \bar{P} = \{\{3,...,9\}\},$$

$$D_B = P_B \setminus \bar{P} = \{\{0,...,5\},\{6,...\}\}.$$

Therefore the values of 3 to 9 of employees in country A are disagreeing values, and are set as sensitive, while this is the case for all numbers of employees in country B. Any observation with a sensitive value is set to be replaced: in this case, any value in country *A* between 3 and 9 (included) is multiply imputed from the Bayesian posterior predictive distribution of the replaced variable, conditional on all observed characteristics. In country *B*, all values are replaced.

[1] An alternative would be to simulate all the undisclosed values, eliminating the topcoding at 90 altogether.

4 Conclusion

We develop a technique to harmonise data reduction methods which also obeys national disclosure avoidance constraints. The standard data reduction methods are covered: top and bottom coding, recoding, and both categorical and continuous variables. Any data reduction method can be specified as a partition of the set of values, and the different national partitions can then be combined in a data reduction method finer than all the national partitions. The incompatible codes, specific to each country, are then replaced by multiply imputed values generated from the appropriate posterior predictive distribution. The inference can then be done according to the standard formulas for partially synthetic data.

References

Confidentiality and Data Access Committee (1999). *Checklist on Disclosure Potential of Proposed Data Releases*. http://www.fcsm.gov/committees/cdac/

Duncan, G., S. Keller-McNulty and S. L. Stokes (2003). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. *Working Paper, Carnegie Mellon's Heinz College.*

Eurostat (1996). *Manual on disclosure control methods*, Statistical Document, Luxembourg: Office for Official Publications of the European Communities

Kennickell, A. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. *Working Paper, Survey of Consumer Finances* http://www.federalreserve.gov/pubs/oss/oss2/method.html

Little R. and F. Liu (2003). Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata. *The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 6.*

Little, R. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, vol. 9, pp. 407–426.

Reiter, J. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, vol. 18, no. 4, pp. 531–543.

Reiter, J. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, vol. 29, no. 2, pp. 181–188.

Rubin, D. (1993). Satisfying confidentiality constraints through use of synthetic multiply-imputed microdata. *Journal of Official Statistics*, vol. 9, pp. 461–468.

Rubin, D. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, Vol. 91, No. 434, pp. 473–489.

Schulte Nordholt, E. (2003) "Applications of statistical disclosure control methods", *Proceedings of Statistics Canada Symposium.*