Microdata protection: a method that combines subsampling and calibration

Maxime Bergeat*

* Insee, French National Institute for Statistics and Economic Studies, Statistical Methods Division, maxime.bergeat@insee.fr

Abstract. This paper draws a comparison between several anonymization methods. A method based on subsampling to reduce disclosure risk and calibration of sampling weights to increase data utility is introduced. This method enables to produce a k-anonymized file. An experiment is made on the French household survey about victimisation "Vols, violence et sécurité". Two k-anonymized files (the first one is obtained with local suppression, the second one after subsampling and calibration) are compared in terms of data utility. Some multivariate statistics are computed and regression analysis is performed on the two k-anonymized datasets. Results are compared with results obtained from the original file.

1 Introduction

In the Open Data context, National Statistical Institutes (NSIs) are encouraged to disseminate more and more data to their users, including microdata sets. Eurostat has launched in beginning 2015 a grant about statistical disclosure control. One goal of this 4-year project is to define harmonized methodologies in order to produce anonymized microdata sets.

Files disseminated by NSIs need to be anonymized in order to keep trust of survey respondents. The anonymization process can be summed up in a few steps. First of all the risk model must be defined: some objectives for disclosure risk reduction are presented in Section 2. Then some methods are applied in order to fulfill anonymization goals. A traditional approach that mixes global recoding and local suppression is recalled in Section 3. In Section 4 we present a method based on subsampling and calibration. After generation of an anonymized dataset we need to estimate loss of data utility implied by anonymization procedures. An application based on a French household survey is made on Section 5 where two k-anonymized datasets are compared for some utility metrics. In Section 6 some concluding comments are given.

2 Objectives for disclosure risk

In this Section, we give some examples of disclosure risk measures and objectives for disclosure risk reduction. Risk models described in this Section are based on the distinction in the dataset between direct identifiers (best option is to remove them or replace them by a non-significant variable), quasi-identifiers (used to estimate disclosure risk) and other non-identifying variables, that could be sensitive. For instance an intruder able to identify someone in the data (identity disclosure) will be able to deduce (possibly sensitive) information for this individual about non-identifying variables if they are not perturbed (attribute disclosure). In the following quasi-identifiers are categorical variables. See Hundepool (2012) for more details.

Full name	Sex	Age category	Favourite meal	Sampling weight
Ambre Marval	Female	-24	raclette	1 000
Cathy Pradel	Female	-24	fish soup	1 500
France Jabot	Female	25-49	sauerkraut	2 000
Ghislaine Metayer	Female	+50	calf's head	1 100
Mireille Henri	Female	+50	calf's head	1 400
Robert Briton	Male	-24	fish soup	800
Louis Brandt	Male	25-49	raclette	1 100
Jean Achard	Male	25-49	beef stew	1 900
Jacques Crillou	Male	+50	sauerkraut	1 200

Figure 1: Example of non-anonymized dataset

In the example given in Figure 1, the full name is the direct identifier. This variable should be removed or replaced in an anonymized dataset. There are two quasi-identifiers: sex and age category. When combined, these two variables may enable an intruder to identify one record. In this case if an intruder knows that my neighbour France Jabot was surveyed and knows her age, this intruder can deduce the favourite meal of France Jabot. There is disclosure of the "Favourite meal" attribute for this record.

In the remaining part of this paper, let c be an identification key (combination of values taken by the quasi-identifiers). Let f_c denote the frequency of occurrence of c in the observed sample, and F_c the frequency of occurrence of c in the reference population. F_c is unknown if we consider a sample.

First option is to estimate disclosure risk using sample observation.

Definition 1 k**-anonymity**
A file is said to be k-anonymous if $f_c \geq k \, \forall \, c$.

Definition 2 l**-diversity**
A file is said to be l-diverse if, for each identification key c, there are at least l "well represented" different values for each sensitive non-identifying attribute.

l-diversity is stronger than k-anonymity. A l-diverse dataset is necessarily l-anonymous. l-diversity prevents against group disclosure: if a group of individuals

who share the same identification key have a characteristic in common there is a risk of group disclosure. In the dataset described Figure 1, all surveyed women who are more than 50 years old love calf's head. An intruder can say that the favourite meal of Mireille Henri is calf's head if (s)he knows that Mireille has been surveyed.

Another possibility is to estimate disclosure risk using an estimate of F_c, the frequency of occurrence of the key c in the reference population. Sampling weights are used in order to make the estimation. According to these models, the individual disclosure risk (for a record whose identification key is c) can be estimated by:

$$r_c = \mathbb{E}\left(\frac{1}{F_c}|f_c\right)$$

The goal for disclosure risk reduction could be in this case:

$$r_c \geq \text{ threshold } \forall\, c$$

Different hypotheses can be made in order to estimate F_c. A Bayesian approach is presented in Benedetti and Franconi (1998) and an estimation using Poisson modelling is introduced in Eleamir and Skinner (2006). This kind of risk model is not discussed further in this paper but it can be used to identify records at-risk. In the application presented in Section 5, the only objective for disclosure risk reduction is 3-anonymity.

3 Traditional approach

After definition of an objective for disclosure risk reduction (that could be k-anonymity, l-diversity or a maximum threshold for the estimated individual risk r_c), we should apply some anonymization methods in order to reach the anonymization goal. A traditional approach consists in mixing global recoding (aggregation of categories of modalities for quasi-identifiers) and local suppression. Local suppression consists in suppression of modalities for some quasi-identifiers for records that do not fulfill initially the objective.

Some algorithms have been developed in order to minimize loss of information implied by local suppression. A cost is assigned to each quasi-identifier. The procedure minimizes the cost of suppression under the constraint of the fixed anonymization goal. We can use the software μ-Argus or sdcMicro where some optimization algorithms are implemented. See Hundepool (2008) or Templ (2015) for more details.

Dataset presented Figure 2 is obtained from the example of Figure 1 (where global recoding has already been applied) after local suppression. In this example the cost assigned to "Sex" variable is higher than the cost of suppression for "Age category" variable.

This example shows a possible weakness of local suppression. The suppression process may be reversed when quasi-identifiers have no missing value in the non-anonymized dataset. For this case if an intruder knows the optimization algorithm used for local suppression and the anonymization goal (2-anonymity here), (s)he can deduce proceeding by elimination that the third record of the dataset is aged between 25 and 49 years.

Sex	Age category	Favourite meal	Sampling weight
Female	-24	raclette	1 000
Female	-24	fish soup	1 500
Female	-	sauerkraut	2 000
Female	+50	calf's head	1 100
Female	+50	calf's head	1 400
Male	-	fish soup	800
Male	25-49	raclette	1 100
Male	25-49	beef stew	1 900
Male	-	sauerkraut	1 200

Figure 2: Example of 2-anonymized dataset after applying local suppression

4 A method based on subsampling

In this Section, a method based on subsampling is developed. The method consists of two steps:

- Suppression of all "at-risk" records. Instead of suppressing some variables for records that do not fulfill anonymization goals, these records are simply deleted in the resulting file.
- In order to restore data utility for the anonymized file, weight calibration is made. Marginals used in the calibration process can be related to quasi-identifying variables or non-identifying variables. Marginals are computed with the original non-anonymized dataset. This step can be seen as a correction for non-response where non-response has been artificially generated after data collection in order to have a low level of disclosure risk in the "anonymized" dataset.

Let S denote the original sample with all respondents and S' the subsample after deletion of at-risk records. In the remaining part of this Section, we use the following notations for a record k:

- x_k is the value taken for k for a calibration variable.
- d_k^{init} is the weight used to initialize the calibration process.
- d_k^{fin} is the "final" weight (in the original dataset) after correction for non-response and potential consideration of auxiliary information with previous calibration. This weight is used to compute marginals in the calibration procedure after suppression of at-risk records.

The goal of calibration in this method is to compute the weights $w_k \, \forall \, k \in S'$ verifying, for all calibration variables:

$$\sum_{k \in S'} w_k x_k = \sum_{k \in S} d_k^{fin} x_k$$

The margin $\sum_{k \in S} d_k^{fin} x_k$ is computed with the original non-anonymized dataset. If we want to use a categorical variable X for calibration, the quantity $\sum_{k \in S} d_k^{fin} x_k$ is defined for each category.

It is possible to control distortion of calibrated weights by using bounded calibration techniques. We can define lower and upper bounds (denoted by L and U) and compute w_k verifying:

$$L \leq \frac{w_k}{d_k^{init}} \times \frac{\sum_{k \in S'} d_k^{init}}{\sum_{k \in S} d_k^{fin}} \leq U$$

For more information about calibration techniques we can refer to Deville and Särndal (1992).

Sex	Age category	Favourite meal	Calibrated weight
Female	-24	raclette	**1 400**
Female	-24	fish soup	**1 900**
Female	+50	calf's head	**1 700**
Female	+50	calf's head	**2 000**
Male	25-49	raclette	**2 100**
Male	25-49	beef stew	**2 900**

Figure 3: Example of 2-anonymized dataset with deletion of at-risk records and calibration

Figure 3 presents an example of a 2-anonymous dataset obtained from the dataset introduced Figure 1 with application of the method described in this Section. Variables used for calibration are "Sex" and "Age category". It is not possible in this example to use "Favourite meal" as a calibration variable given all sauerkraut lovers are suppressed in the "anonymized" dataset. It is not possible to use for calibration variables for which some empty domains are created following deletion of at-risk records.

Other anonymization techniques considering sampling weights have been developed. In Casciano, Ichim and Corallo (2011), a method based on balanced subsampling is presented. With this method probability of selecting at-risk records is not null. The method presented in this Section is more radical: there is a total control about the selected subsample, but loss of data utility may be very high. In next Section we will present an application for a French household survey and give a few utility results to compare two k-anonymized datasets.

5 Application

In this Section an application of the methods presented previously is presented. After presentation of the original dataset, two 3-anonymous datasets are produced

and compared in terms of data utility.

5.1 Original microdata

The experiment is based on the French household survey "Vols, violence et sécurité (Thefts, violence and safety)" (VVS). This experimental survey (web and paper) took place in 2013, in parallel of the French yearly victimisation survey (face-to-face) "Cadre de vie et sécurité (Living environment and safety survey)" (CVS). The CVS survey is realised since 2007 and it is close to the Eurostat SASU project. The main goal of CVS survey is to estimate rates of victimisation in the French population.

The main goal of VVS experiment is to estimate the potential bias of an auto-administered survey on victimisation, given an important concern of this kind of survey is the self-selection issue. In the VVS experiment, two different protocols are tested to handle problems of selection within the household. The questionnaire covers different forms of victimisation (violence, thefts, insults...). The original sample consists of 12 901 individuals.

In the application presented in this Section we do not take into account mode of data collection. The objective is to compare "anonymized" datasets with original VVS microdata.

5.2 Production of 3-anonymous datasets

The goal of this experiment is to produce 3-anonymous datasets. We do not check for l-diversity. The quasi-identifiers are for this test (after global recoding):

- Gender
- Income (5 categories)
- Age (6 categories)
- Size of the urban unit where the respondent lives (4 categories)
- Highest qualification achieved (5 categories)
- Living in a couple or not
- Household size (4 categories)

Original microdata includes certain imputations to correct for item non-response, especially for the qualification variable. For all quasi-identifiers, we consider the "Not applicable / Missing" category as non-identifying. We build two 3-anonymous datasets for this application.

The first file is produced with μ-Argus software. 3-anonymity is reached with local suppression. Costs of suppression assigned to each quasi-identifier and number of deletions are summarized in Figure 4. We have made 3 178 deletions involving 3 033 records.

Quasi-identifier	Cost of suppression	Number of deletions
Gender	70	0
Income	60	4
Age	50	9
Size of urban unit	40	25
Qualification	30	493
Lives in a couple	20	351
Household size	10	2 296

Figure 4: Cost of suppression and number of deletions for each quasi-identifier

The second file is obtained with the method described in Section 4. First all records that do not fulfill 3-anonymity (records with an identification key that belongs to less than 3 individuals) in original microdata are deleted using a SAS procedure. In the subsample (9 868 records) weights are calibrated. The weight used to initialize the calibration process is the weight after correction for total non-response. We choose this variable because its distribution is not too sparse. We use the CALMAR2 SAS macro in order to make the calibration: see Sautory and Le Guennec (2005).

Calibration variables are close to the ones used for calibration in original microdata for computation of first results: see Razafindranovona (2014). Moreover we also use one victimisation variable (denoted "synthetic indicator of victimisation" in the following). The synthetic indicator of victimisation has 3 categories: people who were not victims of any delinquency act (among these acts: theft in the housing, vehicle theft, other theft with violence, other theft without violence, physical violence, threat) in 2011 and 2012, people victims of exactly one delinquency act among the previous list, and people victims of two or more delinquency acts (denoted "multi-victims" in the following). In the end calibration variables are:

- Cross-variable gender $\times$ synthetic indicator of victimisation
- Cross-variable Age category $\times$ synthetic indicator of victimisation
- Cross-variable Size of urban unit $\times$ synthetic indicator of victimisation
- Cross-variable Highest qualification achieved $\times$ synthetic indicator of victimisation
- Cross-variable Household size $\times$ synthetic indicator of victimisation

There is no item non-response for all calibration variables. The calibration enables to preserve victimisation rates (considering the synthetic indicator of victimisation) by gender, age category, size of urban unit, qualification and household size in the final 3-anonymous dataset.

We have decided to use the raking ratio method to compute calibrated weights. This unbounded calibration technique gives positive calibrated weights. A bounded (enabling to control dispersion of the ratio $\frac{\text{weight after calibration}}{\text{weight before calibration}}$) calibration

technique (logit approach with bounds $L = 0.25$ and $U = 3$) has been tested but this approach was not retained : see also Figure 7 in the Appendix.

5.3 Comparison between original microdata and the two 3-anonymous datasets

In this Subsection two 3-anonymous datasets obtained with methods described in Sections 3 and 4 are compared with original microdata for a few results.

Descriptive statistics

Figure 5 presents victimisation rates according to household size.

Original microdata			
Household size	Non-victim	Victim of one delinquency act	"Multi-victim"
1	87.3%	9.7%	3.0%
2	84.4%	12.1%	3.5%
3-4	82.5%	12.3%	5.2%
5+	77.9%	16.0%	6.1%

3-anonymous dataset after local suppression			
Household size	Non-victim	Victim of one delinquency act	"Multi-victim"
1	88.3%	9.0%	2.7%
2	85.7%	11.1%	3.2%
3-4	82.4%	12.2%	5.4%
5+	77.6%	14.6%	7.8%

3-anonymous dataset after subsampling and calibration			
Household size	Non-victim	Victim of one delinquency act	"Multi-victim"
1	87.3%	9.7%	3.0%
2	84.4%	12.1%	3.5%
3-4	82.5%	12.3%	5.2%
5+	77.9%	16.0%	6.1%

Figure 5: Victimisation rates according to the household size (number of persons in the household)

We consider the synthetic indicator of victimisation presented in previous Subsection. Given the variable "household size × synthetic indicator of victimisation" has been used for calibration, rates are obviously the same in the calibrated subsample compared to original microdata. There are slight differences between the file with local suppression and original dataset, especially for rare categories (large households and people who are victims of two or more delinquency acts).

Figure 6 shows results for a variable that is not used in the calibration procedure. Results for the file produced with local suppression are very close to results from original microdata because the "living in a couple" variable has a higher cost of suppression than "household size" and consequently there are only a few deletions

Original microdata			
Lives in a couple	Non-victim	Victim of one delinquency act	"Multi-victim"
Yes	85.6%	11.1%	3.3%
No	79.5%	14.2%	6.3%

3-anonymous dataset after local suppression			
Lives in a couple	Non-victim	Victim of one delinquency act	"Multi-victim"
Yes	85.6%	11.2%	3.2%
No	79.3%	14.3%	6.4%

3-anonymous dataset after subsampling and calibration			
Lives in a couple	Non-victim	Victim of one delinquency act	"Multi-victim"
Yes	84.9%	11.6%	3.5%
No	79.2%	14.5%	6.3%

Figure 6: Victimisation rates according to "living in a couple" variable

of this variable in the 3-anonymized dataset. Results in the dataset obtained after subsampling and calibration are slightly different compared to original victimisation rates.

We also have noticed that the calibration after subsampling improves results in this case. For instance overall victimisation rate for single people is 20.8% in the dataset after subsampling and calibration, compared to 20.5% in non-anonymized microdata. In the file where risky records are deleted and without calibration, the overall victimisation rate is 19.3%. We obtain similar results when studying victimisation rates according to income.

Finally we have done a multiple correspondence analysis considering as active variables 6 victimisation variables (theft in the housing, vehicle theft, other theft with violence, other theft without violence, physical violence, threat). Sampling weights are taken into account in the multivariate analysis. At first glance, when looking at the two principal axes of inertia interpretation is very similar between original microdata and both 3-anonymous datasets. Then we have looked at the projections of modalities of quasi-identifiers onto the two first axes resulting from the analysis. Results are slightly closer to original results for the file produced after local suppression compared to the second 3-anonymouse file.

Modelling

A logistic regression analysis is performed on original dataset and the two 3-anonymized files in order to compare results. Interest variable is:

$$Y = \begin{cases} 1 & \text{, if the respondant declares having been} \\ & \text{victim of at least one delinquency act} \\ 0 & \text{, if not} \end{cases}$$

Considered delinquency acts are the same than previously: theft in the housing, vehicle theft, other theft with violence, other theft without violence, physical

violence, threat. Selection of explanatory variables is made with the original non-anonymized dataset. We use normalised weights in all models. Records who have a local suppression for an explanatory variable are not suppressed from the analysis: a "missing" category is created. We do not take into account variables with item non-response in original microdata. Explanatory variables retained finally are:

- Age category (6 categories)
- Highest qualification achieved (5 categories)
- Household size (4 categories)
- Size of the urban unit where the respondent lives (4 categories)

We consider for this model as a data utility metric the confidence interval overlap defined in Drechsler (2009).

Definition 3 Confidence Interval Overlap
Let $[L_{\text{ORIGINAL}}, U_{\text{ORIGINAL}}]$ the 95%-confidence interval for a given parameter in original dataset. Let $[L_{\text{ANONYMIZED}}, U_{\text{ANONYMIZED}}]$ the corresponding interval in the "anonymized" dataset. We denote the intersection of the two intervals by :

$$[L_{\text{INTER}}, U_{\text{INTER}}] = [L_{\text{ORIGINAL}}, U_{\text{ORIGINAL}}] \cap [L_{\text{ANONYMIZED}}, U_{\text{ANONYMIZED}}]$$

The Confidence Interval (CI) overlap is given by:

$$\text{CI overlap} = \frac{1}{2} \times \left(\frac{U_{\text{INTER}} - L_{\text{INTER}}}{U_{\text{ORIGINAL}} - L_{\text{ORIGINAL}}} + \frac{U_{\text{INTER}} - L_{\text{INTER}}}{U_{\text{ANONYMIZED}} - L_{\text{ANONYMIZED}}} \right)$$

When the intervals are identical in both compared datasets, CI overlap = 1. When the intervals do not overlap at all, CI overlap = 0. The second term in the sum is included to avoid to give the *maximum* utility score if:

$$[L_{\text{ANONYMIZED}}, U_{\text{ANONYMIZED}}] \supset [L_{\text{ORIGINAL}}, U_{\text{ORIGINAL}}] .$$

This correction is particularly important if subsampling is applied (mechanically size of confidence intervals will increase in the "anonymized" dataset).

Confidence Interval overlaps for odds ratios in this model are given in Figure 8 in the Appendix. Results are better for the file anonymized with local suppression: this result seems logical given width of confidence intervals will generally increase when we look at a smaller sample. Overlaps obtained for the dataset produced with subsampling and calibration seem however acceptable: the mean confidence interval overlap is here 76.3%.

6 Concluding comments

In this paper we present a method to reduce disclosure risk based on deletion of risky records and calibration of sampling weights in the "anonymized" subsample. The main advantage of this technique is that the disclosure risk is entirely controlled. In the application presented in Section 5 we see that for some statistics computed on the two 3-anonymized datasets we have quite similar results between original microdata and both 3-anonymous datasets.

In the application presented in Section 5 a simple criteria (k-anonymity) is used as a goal for disclosure risk reduction. We can however use more sophisticated methods for risk estimation, like the measures presented in Benedetti and Franconi (1998) or Eleamir and Skinner (2006).

Microdata protection always consists in a trade-off between disclosure risk reduction and loss of data utility. Dissemination of anonymized datasets is also an important question and choice of anonymization techniques generally depends on who will get access to the anonymized file and how. The method presented in Section 4 controls in the first place the disclosure risk and seems quite adapted for public use files that are available for everyone.

Finally documentation of disseminated datasets is essential. It is important to wonder which amount of information about anonymization process is given to the final user. Giving too much detail can enable a user to reverse the anonymization process. It is also important to give guidelines about how "anonymized" datasets should be used.

References

Benedetti, R. and Franconi, L. (1998) *Statistical and technological solutions for controlled data dissemination*, Pre-proceedings of New Techniques and Technologies for Statistics, **1**, 225–232.

Casicano, C. and Ichim, D. and Corallo, L. (2011) *Sampling as a way to reduce risk and create a Public Use File maintaining weighted totals*, UNECE/Eurostat work session on statistical data confidentiality.

Deville, J.-C. and Särndal, C.-E. (1992) *Calibration estimators in survey sampling*, Journal of the American Statistical Association, **87**, 376–382.

Drechsler, J. and Reiter, J.-P. (2009) *Disclosure Risk and Data Utility for Partially Synthetic Data: An Empirical Study Using the German IAB Establishment Survey*, Journal of Official Statistics, **25**, 589–603.

Hundepool, A. *et al.* (2008) *μ-Argus User's manual*, available online.

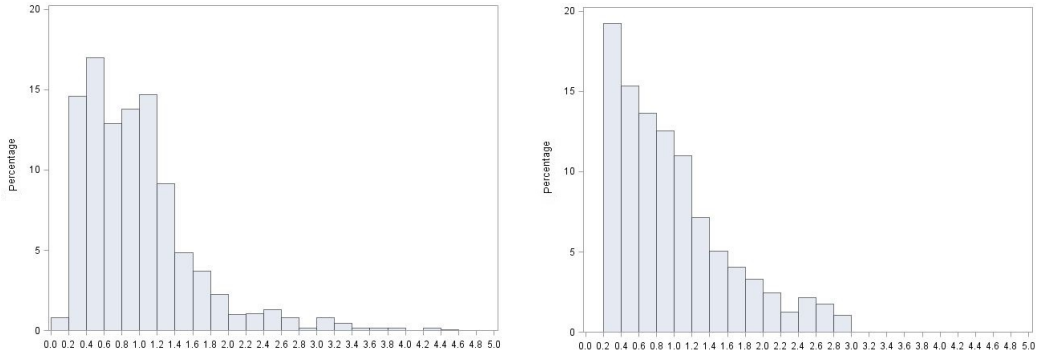
Hundepool, A. *et al.* (2012) *Statistical disclosure control*, Wiley Series in Survey Methodology.

Razafindranovona, T. and Dietsch, B. and Burricand, C. and de Peretti, G. (2014) *Le multimode pour mesurer la victimation : est-on dans la zone de sécurité ?*, 8th French Colloquium in Survey Sampling, in French.

Sautory, O. and Le Guennec, J. (2005) *La macro CALMAR2 - Redressement d'un échantillon sur marges*, in French.

Templ, M. and Meindl, B. and Kowarik, A. (2015) *Guidelines for statistical disclosure control using sdcMicro*, sdcMicro vignette, available online.

Appendix



Raking ratio (unbounded) Logit method (bounded)

Figure 7: Distribution of ratios $\frac{w_k}{d_k^{init}} \times \frac{\sum_{k \in S'} d_k^{init}}{\sum_{k \in S} d_k^{fin}}$ according to the calibration technique

Variable	Odds ratio	Local suppression	Subsampling and calibration
Age category	Age1	94.8%	87.7%
	Age2	90.9%	76.5%
	Age3	90.8%	78.7%
	Age4	98.0%	92.5%
	Age5	88.3%	86.8%
Qualification	Qua1	86.6%	78.2%
	Qua2	92.3%	75.0%
	Qua3	93.3%	75.8%
	Qua4	87.6%	63.0%
Household size	Hsize1	91.3%	54.3%
	Hsize2	89.2%	52.5%
	Hsize3	88.9%	76.8%
Size of urban unit	UU1	94.6%	80.1%
	UU2	95.3%	92.5%
	UU3	96.2%	74.0%

Figure 8: Confidence Interval Overlaps for odds ratios of the model described in Section 5