Comparison of Perturbation Methods Based on Pre-defined Quality Indicators [1]

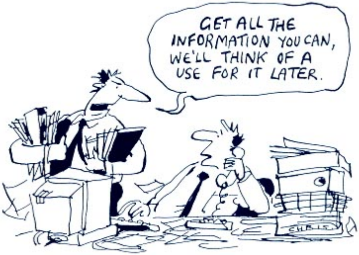
Matthias Templ[1,2,3], Bernhard Meindl[1,3]

[1] Department of Methodology, Statistics Austria
[2] Department of Statistics and Probability Theory, TU WIEN, Austria
[3] http://www.data-analysis.at/

[1] This work was partly funded by the ESSNet on common tools and harmonised methodology for SDC in the ESS

Preserving the User-Needs Estimates I

- If data are protected with no specific data use in mind,

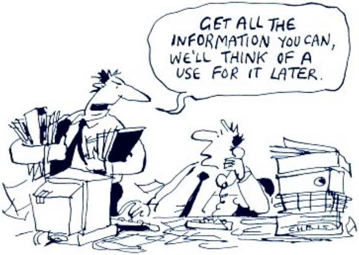

GET ALL THE INFORMATION YOU CAN, WE'LL THINK OF A USE FOR IT LATER.

protection methods are applied in general manner and the structure of the anonymised data should look as similar to the original data.

- If the estimators applied on the data is mostly (assumed to be) known, the protected data should give (at least) **precise estimates for pre-defined estimators**.

Preserving the User-Needs Estimates I

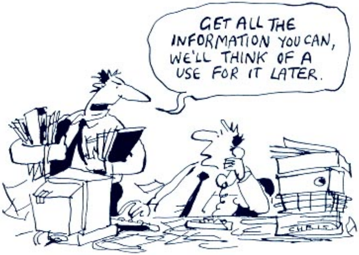
- If data are protected with no specific data use in mind, protection methods are applied in general manner and the structure of the anonymised data should look as similar to the original data.


GET ALL THE INFORMATION YOU CAN, WE'LL THINK OF A USE FOR IT LATER.

- If the estimators applied on the data is mostly (assumed to be) known, the protected data should give (at least) **precise estimates for pre-defined estimators**.

Preserving the User-Needs Estimates I

- If data are protected with no specific data use in mind,

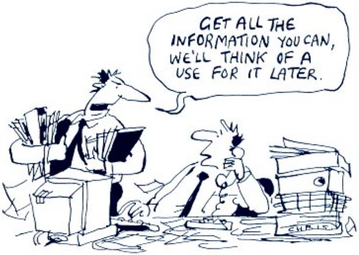

GET ALL THE INFORMATION YOU CAN, WE'LL THINK OF A USE FOR IT LATER.

  protection methods are applied in general manner and the structure of the anonymised data should look as similar to the original data.

- If the estimators applied on the data is mostly (assumed to be) known, the protected data should give (at least) **precise estimates for pre-defined estimators**.

Preserving the User-Needs Estimates I

- If data are protected with no specific data use in mind, protection methods are applied in general manner and the structure of the anonymised data should look as similar to the original data.


GET ALL THE INFORMATION YOU CAN, WE'LL THINK OF A USE FOR IT LATER.

- If the estimators applied on the data is mostly (assumed to be) known, the protected data should give (at least) **precise estimates for pre-defined estimators**.

Information Loss

If data are protected with **no** specific data use in mind, **general information loss measures** are applied.

Popular apprach, for continuous scaled variables:

```
require(sdcMicro)           ## load package
data(Tarragona)             ## load data
m1 <- microaggregation(Tarragona, method="mdav")$blowxm
m2 <- microaggregation(Tarragona, method="rmd")$blowxm
dUtility(Tarragona, m1)  ## IL1 method from
[1] 0.236
dUtility(Tarragona, m2)  ## Yancey, Winkler and Creedy
[1] 0.205
```

If data are protect with specific data use in mind, **analysis-depended** and **data-dependend measures** are more suitable.

Information Loss

If data are protected with **no** specific data use in mind, **general information loss measures** are applied.
Popular apprach, for continuous scaled variables:

```
require(sdcMicro)           ## load package
data(Tarragona)             ## load data
m1 <- microaggregation(Tarragona, method="mdav")$blowxm
m2 <- microaggregation(Tarragona, method="rmd")$blowxm
dUtility(Tarragona, m1)  ## IL1 method from
[1] 0.236
dUtility(Tarragona, m2)  ## Yancey, Winkler and Creedy
[1] 0.205
```

If data are protect with specific data use in mind, **analysis-depended** and **data-dependend measures** are more suitable.

Information Loss

If data are protected with **no** specific data use in mind, **general information loss measures** are applied.
Popular apprach, for continuous scaled variables:

```
require(sdcMicro)          ## load package
data(Tarragona)            ## load data
m1 <- microaggregation(Tarragona, method="mdav")$blowxm
m2 <- microaggregation(Tarragona, method="rmd")$blowxm
dUtility(Tarragona, m1)  ## IL1 method from
[1] 0.236
dUtility(Tarragona, m2)  ## Yancey, Winkler and Creedy
[1] 0.205
```

If data are protect with specific data use in mind, **analysis-depended** and **data-dependend measures** are more suitable.

Preserving the User-Needs Estimates II

- Ichim and Franconi (2010) outlined a concept of minimum quality requirements for public and scientific use files. (see also the *ESSnet on common tools and harmonised methodology for SDC in the ESS*)
- Data providers have to **guarantee sufficient precision** for a set of pre-defined **quality indicators** while the data providers have the freedom to select the SDC methods that are applied to their microdata.
- Benchmarking indicators have to be defined (and implemented) first (compromise between user needs and sensitivity of indicators/models)
- Some protetion methods will be evaluated if they fulfill these benchmark statistics.
- The evaluations based on indicators from **S**tructural **E**arnings **S**urvey data (**SES**)

Preserving the User-Needs Estimates II

- Ichim and Franconi (2010) outlined a concept of minimum quality requirements for public and scientific use files. (see also the *ESSnet on common tools and harmonised methodology for SDC in the ESS*)
- Data providers have to **guarantee sufficient precision** for a set of pre-defined **quality indicators** while the data providers have the freedom to select the SDC methods that are applied to their microdata.
- Benchmarking indicators have to be defined (and implemented) first (compromise between user needs and sensitivity of indicators/models)
- Some protetion methods will be evaluated if they fulfill these benchmark statistics.
- The evaluations based on indicators from Structural Earnings Survey data (**SES**)

Preserving the User-Needs Estimates II

- Ichim and Franconi (2010) outlined a concept of minimum quality requirements for public and scientific use files. (see also the *ESSnet on common tools and harmonised methodology for SDC in the ESS*)
- Data providers have to **guarantee sufficient precision** for a set of pre-defined **quality indicators** while the data providers have the freedom to select the SDC methods that are applied to their microdata.
- Benchmarking indicators have to be defined (and implemented) first (compromise between user needs and sensitivity of indicators/models)
- Some protetion methods will be evaluated if they fulfill these benchmark statistics.
- The evaluations based on indicators from Structural Earnings Survey data (**SES**)

Preserving the User-Needs Estimates II

- Ichim and Franconi (2010) outlined a concept of minimum quality requirements for public and scientific use files. (see also the *ESSnet on common tools and harmonised methodology for SDC in the ESS*)
- Data providers have to **guarantee sufficient precision** for a set of pre-defined **quality indicators** while the data providers have the freedom to select the SDC methods that are applied to their microdata.
- Benchmarking indicators have to be defined (and implemented) first (compromise between user needs and sensitivity of indicators/models)
- Some protetion methods will be evaluated if they fulfill these benchmark statistics.
- The evaluations based on indicators from **S**tructural **E**arnings **S**urvey data (**SES**)

Preserving the User-Needs Estimates II

- Ichim and Franconi (2010) outlined a concept of minimum quality requirements for public and scientific use files. (see also the *ESSnet on common tools and harmonised methodology for SDC in the ESS*)
- Data providers have to **guarantee sufficient precision** for a set of pre-defined **quality indicators** while the data providers have the freedom to select the SDC methods that are applied to their microdata.
- Benchmarking indicators have to be defined (and implemented) first (compromise between user needs and sensitivity of indicators/models)
- Some protetion methods will be evaluated if they fulfill these benchmark statistics.
- The evaluations based on indicators from **S**tructural **E**arnings **S**urvey data (**SES**)

SES Data

- SES is a complex survey of Enterprises and Establishments with more than 10 employees (e.g. 11600 enterprises in Austria), NACE C-O, including a large sample of employees (e.g., in Austria: 207.000).
- In many countries, a **two-stage design** is used whereas in the first stage a stratified sample of enterprises and establishments on NACE 1-digit level, NUTS 1 and employment size range is used, whereas large enterprises has higher inclusion probabilities. In stage 2, systematic sampling is applied in each enterprise using unequal inclusion probabilities regarding employment size range categories.
- Calibration is applied to represent some population characteristics corresponding to NUTS 2 and NACE 1-digit level, but also calibration is carried out for gender (amount of men and womens in the population).

SES Data

Information on enterprise level: Question batteries are asked to enterprises like if an enterprise is **private or public** or if an enterprise has a **collective bargaining agreement** (both binary variables). As a multinomial variable, the **kind of collective agreement** is included in the questionnaire.

Information on individual employment level: The following questions to employees comes with the standard questionnaire: **social identity number**, **date of being employed**, **weekly working time**, **kind of work agreement**, **occupation**, **time for holidays**, **place of work**, **gross earning**, **earning for overtime** and **amount of overtime**.

Information from registers: All other information may come from registers like information about **age**, **size of enterprise**, **occupation**, **education**, **amount of employees**, **NACE** and **NUTS** classifications.

Key Indicators

From SES data the most important analysis is related to

- Gender wage gap
- Inter-industry wage differentials. Differences in earnings for workers employed in different industries and occupations has long been recognised as an important issue for the labour market.
- Low-pay dynamics.
- Enterprise characteristics that effects earnings or profit.
- Collective bargaining.
- Average Earnings
- Occupation and tenure

I. Average Hourly Earnings

The average wage/earnings in enterprise j may be estimated by the weighted arithmetic mean,

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} w_{ij} x_{ij}}{\sum_{i=1}^{n_j} w_{ij}} \quad , \tag{1}$$

with w_{ij} and x_{ij} the sample weight and wage or earnings of employee i in enterprise j, respectively. n_j is the number of employees in enterprise j.

II. Gender Wage Gap

For the following definitions, let $\boldsymbol{x} := (x_1, \ldots, x_n)'$ be the hourly earnings with $x_1 \leq \ldots \leq x_n$ and let $\boldsymbol{w} := (w_i, \ldots, w_n)'$ be the corresponding sample weights on employment level, where n denotes the number of observations. Let $J^{(M)} := \{j \in \{1, \ldots, n\} \mid \text{worked as least 1 hour per week } \wedge (16 \leq \text{age} \leq 65) \ \wedge \text{person is male}\}$, and $J^{(F)}$ those index set which differs from $J^{(M)}$ in the fact that it includes all females instead of males.

With these index sets the gender pay gap in unadjusted form is estimated by

$$GPG_{(mean)} = \frac{\frac{\sum_{i \in J^{(M)}} w_i x_i}{\sum_{i \in J^{(M)}} w_i} - \frac{\sum_{i \in J^{(F)}} w_i x_i}{\sum_{i \in J^{(F)} w_i}}}{\frac{\sum_{i \in J^{(M)}} w_i x_i}{\sum_{i \in J^{(M)}} w_i}} \quad . \qquad (2)$$

III. GINI coefficient

The Gini coefficient according to EU-SILC (2004, 2009) is estimated by

$$\widehat{Gini} := 100 \left[\frac{2 \sum_{i=1}^{n} \left(w_i x_i \sum_{j=1}^{i} w_j \right) - \sum_{i=1}^{n} {w_i}^2 x_i}{\left(\sum_{i=1}^{n} w_i \right) \sum_{i=1}^{n} \left(w_i x_i \right)} - 1 \right] . \qquad (3)$$

IV. Model on Employment Level

The log hourly earnings for each country are predicted with the following predictors:

$$\mathit{log}(\texttt{earnings}) \sim \texttt{sex}\ (2) + \texttt{age} + \texttt{age}^2 + \texttt{education}\ (6) + \texttt{occupation}\ (23) + \text{error term} \quad .$$

The numbers in brackets correspond to the number of categories for binary or categorical variables.

Implementation in Software R

All mentioned estimations are implemented in R-package *laeken* (Templ and Alfons 2011) or can be easily carried out with R.

```
> g1 <- gpg(inc = "earningsHour", method = "median",
    gender = "Sex", weigths = "GrossingUpFactor.x",
    breakdown = "education", data = x)
> g1
g1
Value:
[1] 0.2092618

Value by stratum:
        stratum     value
1 ISCED 0 and 1 0.2116091
2       ISCED 2 0.1354932
3 ISCED 3 and 4 0.1898604
4      ISCED 5A 0.2769508
5      ISCED 5B 0.2370654
```

Implementation in Software R / package laeken

Variance of point estimates (via calibrated bootstrap):

```
variance("earningsHour", weights = "GrossingUpFactor.x",
    gender="Sex", data = x, indicator = g1,
    X = calibVars(x$Location), breakdown="education")
Value:
[1] 0.2092618

Variance:
[1] 1.853727e-05

Confidence interval:
    lower     upper
0.2069582 0.2247713

Value by stratum:
        stratum     value
1 ISCED 0 and 1 0.2116091
2       ISCED 2 0.1354932
3 ISCED 3 and 4 0.1898604
4      ISCED 5A 0.2769508
```

Risk 1

Scenario 1 (employment) with categorical key variables NUTS1, age classes, education and size indicates that 39 observations do not fulfill 2-anonymity (see Listing).

Listing 1: Frequency counts and individual risk. Scenario 1.

```
--------------------------
21 observation with fk=1
18 observation with fk=2
 --------------------------
(0,1]      (1,2]      (2,3]      (3,5]     (5,10] (10,1e+04]
 21         18         36         60        226     199548
----------------------------
indivRisk:
      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0000178 0.0000251 0.0000536 0.0002256 0.0001251 0.3783000
```

Risk 2

Scenario 2 (employment) with categorical key variables NACE, NUTS1, age classes, size, education, occupation, full/part time, sex indicates that 7993 observations do not fulfill 2-anonymity:

Listing 2: Frequency counts and individual risk. Scenario 2.

```
 --------------------------
4075 observation with fk=1
3918 observation with fk=2
 --------------------------
(0,1]      (1,2]      (2,3]      (3,5]     (5,10] (10,1e+04]
4075       3918       3591       6320      11926     170079
--------------------------
indivRisk:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0001896 0.0007078 0.0015460 0.0131700 0.0051480 1.1090000
```

Perturbation Methods

Two possibilities (amongs others) for anonymisation:

a) Recoding, local suppression to provide *k*-anonymity Sweeney (2002) for the categorical key variables (for enterprises, for employees). Microaggregation, adding (correlated) noise Brand (2004) or deletion and imputation for continuous variables.

b) Synthetic data generation of all important variables (Alfons et al. 2010). Simulation of variables by drawing from predictive distributions.

Via R-Package sdcMicro . . .

```
## Recoding of "Size" and "age"
levels(x$SizeRed) <- list(E10_49=c("E10_49"),
E50_249="E50_49", E250plus=c("E250_499","E500_999","E1000"))

x$age <- cut(2006 - x$birth, breaks=c(0,19,29,39,49,59,120))

## frequency and risk estimation:
keyEC <- c("Size", "age", "education", "Location", "
    economicActivity")
keyEC <- which(colnames(x) %in% keyEC)
frAfter <- freqCalc(x, keyVars=keyEC, w=41)

## perturbation: (or on subgroups via apply())
xm <- microaggregation(x[,c("earningsHour","earnings")])

## Pram, individual risk, adding noise, ..
## can also be done with the GUI of sdcMicro!
```

Data Utility

The utility measures chosen are based on the benchmarking indicators, namely

- about the difference in the estimation of the GPG and the GINI from the original and perturbed data defined for h domains:

$$ARB = \frac{|\frac{1}{h}\sum_{i=1}^{h}(\hat{\theta}_i - \theta_i)}{\theta_i} \quad . \tag{4}$$

- Additionally, one model is predicted and from the predicted values the average hourly earnings are estimated.
- Moreover, the variances are estimated and the **overlap of the confidence interval** of the perturbed and original data is evaluated and reported in percentages.

overlap of confidence intervals using package laeken. . .

```
v1 <- variance("earningsHour",
weights = "GrossingUpFactor.x", gender="Sex", data = x,
    indicator = g1, X =
calibVars(x$Location), breakdown="education", seed = 123)

v1a <- variance("earningsHourM",
weights = "GrossingUpFactor.x", gender="Sex", data = x,
    indicator = g1, X =
calibVars(x$Location), breakdown="education", seed = 123)

confcover(v1$ci, v1a$ci)
```

Preliminary Results

		GPG		GINI		MOD	
method	measure	overall	domain	overall	domain	overall	domain
microaggr.	ARB	4.73	8.66	2.72	4.17	17.45	13.68
microaggr.	overlap	94.86	65.63	0	30.48		
corr. noise	ARB	48.03	49.45	1.24	20.04	96.07	2465
corr. noise	overlap	0	5.38	0	2.1		
imputation	ARB	0.32	1.44	0.12	0.68	7.84	10.85
imputation	overlap	78.20	92.32	67.58	94.34		

Outlook

We identified the benchmarking indicators and implemented them in free and open-source software.

Future work includes:

- More sophisticated information loss measures for the point and variance estimates of regression coefficients.
- More perturbation methods will be evaluated.
- These methods will be compared with synthetic data generation methods.
- Recommondation on the use of perturbation methods for microdata.
- Providing all necessary code to the statistical agencies so that they can try out the methods and check the quality based on the pre-defined benchmarking estimators.

Outlook

We identified the benchmarking indicators and implemented them in free and open-source software.
Future work includes:

- More sophisticated information loss measures for the point and variance estimates of regression coefficients.
- More perturbation methods will be evaluated.
- These methods will be compared with synthetic data generation methods.
- Recommendation on the use of perturbation methods for microdata.
- Providing all necessary code to the statistical agencies so that they can try out the methods and check the quality based on the pre-defined benchmarking estimators.

References

A. Alfons, S. Kraft, M. Templ, and P. Filzmoser. Simulation of synthetic population data for household surveys with application to EU-SILC. Research Report CS-2010-1, Department of Statistics and Probability Theory, Vienna University of Technology, 2010. URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2010-1complete.pdf.

R. Brand. Microdata protection through noise addition. In Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer, pages 347–359, 2004.

EU-SILC. Common cross-sectional EU indicators based on EU-SILC; the gender pay gap. EU-SILC 131-rev/04, Working group on Statistics on Income and Living Conditions (EU-SILC), Eurostat, Luxembourg, 2004.

EU-SILC. Algorithms to compute social inclusion indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC). EU-SILC LC-ILC/39/09/EN-rev.1, Directorate F: Social and information society statistics Unit F-3: Living conditions and social protection, EUROPEAN COMMISSION, EUROSTAT,, Luxembourg, 2009.