WP. 23
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (iv): Balancing data quality and data confidentiality

Comparison of perturbation methods based on pre-defined quality indicators

Prepared by Matthias Templ, Vienna University of Technology and Statistics Austria

WP.
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (iv): Balancing data quality and data confidentiality

COMPARISON OF PERTURBATION METHODS BASED ON PRE-DEFINED QUALITY INDICATORS [1]

Invited Paper

Submitted by the Department of Statistics and Probability Theory, Vienna University of Technology, and the Department of Methodology, Statistics Austria[2]

ABSTRACT

Ichim and Franconi (2010) outlined a concept of minimum quality requirements for public and scientific use files. For such data sets that will be sent to Eurostat, 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.
In this contribution a small set of perturbation is applied. They are evaluated if they fulfill pre-defined benchmark statistics that are briefly descibed in this contribution.
For evaluation, the Structural Earnings Survey is chosen. This data set includes confidential variables on enterprise and employment level.

I. STRUCTURAL EARNINGS SURVEY (SES)

1. One aim of the project "*ESSnet on common tools and harmonised methodology for SDC in the ESS*" is to propose quality guidelines for anonymised data, whereas estimates from the Structural Earnings Survey (SES) are in focus. The most important estimates from the anonymised survey should be close to the estimates from the confidential data.

2. The Structural Earnings Survey (SES) is conducted in almost all European Countries, and the most important figures are reported to Eurostat. The anonymised microdata should be send to Eurostat. However, many countries do not agree with the proposed rules for anonymisation communicated by Eurostat, nor they can allow remote access systems like the PiEP Lissy project (Marsden 2010) because of restrictions in national laws.

[1] This work was funded by Eurostat and Statistics Netherlands within the project *ESSnet on common tools and harmonised methodology for SDC in the ESS.*). Visit http://neon.vb.cbs.nl/casc/ESSNet2index.htm for more information on the project.

[2] Prepared by Matthias Templ (templ@tuwien.ac.at), visit http://www.data-analysis.at.

3. 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.
In the Austrian case, for example, the sample has only $2,4\%$ non-response. Regression imputation is applied by using tax data to replace these missing values.
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).

4. SES includes information from different perspectives and sources:

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.

5. SES Microdata from Czech Republic, Hungary, Ireland, Italy, Latvia, Lithuania, Netherlands, Norway, Portugal, Slovakia and Spain can be analysed via the Piep Lissy remote access system using `Stata` whereas some commands (12 commands in summary) are blocked by the system to prevent listing of individuals. However, since the user can produce almost any output, confidentiality cannot be provided by the system, i.e. it is an easy task to write some code around to extract confidential information on individual level.

II. INDICATORS

6. From SES data the most important analysis is related to

Gender wage gap: The gender wage gap is nowadays the most important indicator obtained from SES in many European countries (for Education and the Labour Market 2009). In Austria, for example, a lot of publications about the gender wage gap are published by Statistics Austria and the national authorities (Stockinger 2010). The topic *Women and Equality* is of central interest not only for the Federal Minister for Women and the Civil Service, and socio-economic studies are carried out with support from the state (Geissberger 2010).

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 (Caju et al. 2010; Caju et al. 2009; Caju et al. 2009; Messina et al. 2010; Dybczak and Galuscak 2010; Simón 2010; Pointner and Stiglbauer 2010).

Low-pay dynamics: In some countries, great changes in the distribution of earnings is observed (Dell'Aringa et al. 2000; Geissberger 2009) with a widening of inequality and an increase in dispersion. The Gini-index and the quintile share ratio is one of the main indicators to estimate the inequality (Graf et al. 2011; Kolb et al. 2011).

Enterprise characteristics that effects earnings or profit: The differentials that describes the profit of an enterprise is one interesting aspect. How flexibility, information sharing and the size of the enterprise influences the profitability of an enterprise? On the other hand, it is of interest to investigate in predicting pay flexibility with the size of the enterprise, level of competition, training, job rotation, time flexibility, etc. (Marsden 2010).

Collective bargaining: Due to the importance of unions on wage determination, to measure the extent of the union-non union wage gap is of interest (Edwards 2010; Fitzenberger et al. 2006).

Average Earnings: Average earnings in enterprises as indicator for productivity or performance (Winter-Ebmer and Zweimüller 1999; Marsden 2010). The idea is that in a competitive market environment in which employees' pay corresponds to the value of their output, i.e. deviations from this position would lead to difficulties in recruitment and retention. In branches with high output, the earnings would therefore be higher as within enterprises categorised in low-productive economic branches.

Occupation and tenure: Other interesting analysis includes the difference in income for different occupation levels or by the length of tenure.

In this contribution we investigate in the gender pay gap and in one model-based prediction on employment level.

7. The gender pay gap in unadjusted form is defined on population level as the difference between average gross earnings of male paid employees and of female paid employees divided by the earnings of mail paid employees (EU-SILC 2009). Since the gender wage gap is usually estimated with survey information, sampling weights have to be considered in order to ensure sample represantivity. The gender pay gap is usually estimated at domain level like economic branch, education and age groups (Geissberger 2009). In addition, the variance of that estimations are important to estimate.
The estimates given in Listing 1 and 2 are used for benchmarking perturbation methods.

Listing 1. Estimation of the gender pay gap including breakdown by education.

```
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
```

Listing 2. Estimation of the variance of the gpg including breakdown by 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
5       ISCED 5B 0.2370654

Variance by stratum:
        stratum          var
1 ISCED 0 and 1 1.205312e-03
2        ISCED 2 6.806878e-05
3 ISCED 3 and 4 1.472970e-05
4       ISCED 5A 2.984983e-04
5       ISCED 5B 1.325028e-04

Confidence interval by stratum:
        stratum      lower     upper
1 ISCED 0 and 1 0.10061805 0.2462696
2        ISCED 2 0.07719914 0.1126314
3 ISCED 3 and 4 0.20757235 0.2222908
4       ISCED 5A 0.20296804 0.2710628
5       ISCED 5B 0.19859306 0.2461893
```

Note that these estimates may differ from the original one because a sample of the SES data are used for estimation to avoid discussions about differences with official figures estimated by other software.

8. Respectively for all model-based estimations at employment level we choose a model described in Marsden (2010), Dybczak and Galuscak (2010) applied within the PiEP Lissy project. They fit OLS regression models where they modeled the gross hourly earnings of workers in enterprises using age, age^2, sex, education and occupation as predictors.
The log hourly earnings for each country are predicted with the following predictors:

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

The numbers in brackets correspond to the number of categories for binary or categorical variables. The summary statistics of the model are presented in Table 1. It is easy to see that all predictors are highly significant and explains the log hourly earnings well, i.e. sex, age, age^2 and education gives a good explanation of the log hourly earnings.

III. CONFIDENTIALITY ISSUES

9. The identification of an enterprise may leads to information about their employee's.

10. Key variables at enterprise level might be NUTS 1 (3), NACE 1-digit level, Size (5), public/privat (2) and white/blue colour worker/rest (3). In brackets the number of categories are given. Categorical key variables at employment level might be NUTS 1 (3), age class (6), education(7), occupation (8), part time/full time (2), gender (2). This leads to 4032 stratas which is too much and less key variables have to be chosen like NACE, NUTS 1, size and age, which is also proposed by (Ichim and Franconi 2007).
Continuous key variables at employment level might be the gross earnings and special payments.

11. Anonymised SES 2002 and 2006 data (Eurostat) from 23 countries can be can be accessed for research purposes through the safe centre at the premises of Eurostat. Anonymisation is done by

Table 1. Output from the underlying regression model with Multiple R-squared: 0.5604, Adjusted R-squared: 0.5603.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.1666	0.0302	38.65	0.0000
Sexmale	0.1958	0.0018	106.28	0.0000
age	0.0691	0.0004	178.91	0.0000
I(age^2)	-0.0007	0.0000	-145.56	0.0000
educationISCED 2	-0.0230	0.0117	-1.97	0.0483
educationISCED 3 and 4	0.1707	0.0116	14.66	0.0000
educationISCED 5A	0.3269	0.0123	26.64	0.0000
educationISCED 5B	0.2407	0.0121	19.83	0.0000
Occupation12	0.1094	0.0273	4.01	0.0001
Occupation21	-0.1513	0.0281	-5.39	0.0000
Occupation22	-0.0540	0.0290	-1.86	0.0624
Occupation23	-0.1023	0.0285	-3.59	0.0003
Occupation24	-0.2111	0.0278	-7.59	0.0000
Occupation31	-0.2254	0.0272	-8.29	0.0000
Occupation32	-0.3013	0.0277	-10.86	0.0000
Occupation33	-0.3722	0.0279	-13.35	0.0000
Occupation34	-0.2280	0.0275	-8.30	0.0000
Occupation41	-0.3591	0.0271	-13.23	0.0000
Occupation42	-0.3614	0.0272	-13.31	0.0000
Occupation51	-0.7088	0.0272	-26.10	0.0000
Occupation52	-0.5988	0.0272	-21.98	0.0000
Occupation71	-0.5363	0.0272	-19.72	0.0000
Occupation72	-0.5220	0.0272	-19.22	0.0000
Occupation73	-0.4522	0.0281	-16.10	0.0000
Occupation74	-0.7088	0.0272	-26.06	0.0000
Occupation81	-0.4173	0.0277	-15.04	0.0000
Occupation82	-0.5160	0.0273	-18.92	0.0000
Occupation83	-0.7264	0.0272	-26.66	0.0000
Occupation91	-0.7373	0.0272	-27.12	0.0000
Occupation93	-0.6184	0.0272	-22.76	0.0000

recoding of NACE, NUTS and size, removing citizenship and building six age classes, microaggregation (individual ranking) for absence days and earnings and removing the sampling weights (Eurostat).

12. Scenario 1 (employment) with categorical key variables NUTS1, age classes, education and size indicates that 39 observations do not fulfill 2-anonymity (see Listing 3).

Listing 3. 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
```

13. 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 (see Listing 4).

Listing 4. 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
```

IV. PERTURBATION METHODS

14. Two possibilities (amongs others) for anonymisation:

a) 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 variables (Alfons et al. 2011). Simulation of all variables by drawing from predictive distributions. Note, that it is a gain in knowledge when only simulating gross earnings and taking the categorical key variables unchanged. By identifying, for example, the age of the person, information about the person is identified even all continuous scaled key variables like earnings are perturbed.

V. DATA UTILITY

15. The utility measures chosen are based on the benchmarking indicators defined in Section II, namely

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

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

- 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.

VI. RESULTS

16. Table 2 shows the ARB and model errors whereas the overall estimate is shown and the mean of the domain (sex × age class) estimates. It is easy to see that microaggregation (here for computational reasons, method 'pca' from package `sdcMicro` (Templ 2010; Templ 2008) was chosen) provides much better results than the correlated noise method (Brand 2004) using the default parameters of `sdcMicro`. The overlap in the confidence intervals is zero for the correlated noise method. The best results obtained from deletion and imputation where the hot deck method from `R` package `VIM` (Templ, Alfons, and Kowarik 2011; Templ, Alfons, and Filzmoser 2009) was used. Here, 10% of the data were deleted and imputed.

Table 2. Errors in percentages with reduced risk by 89.75 percentage.

		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		

VII. CONCLUSION AND FUTURE WORK

17. This contribution serves as a starting point for the evaluation of disclosure methods on defined benchmarking indicators. We have briefly shown few of these benchmarking indicators and evaluated few methods on relatively simple data utility measures.

Therefore, future work is to evaluate all popular microdata protection and synthetic data generation methods. Here, some additional work to properly define data utility and disclosure risk measures is necessary.

References

Alfons, A., S. Kraft, M. Templ, and P. Filzmoser (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods & Applications*. DOI 10.1007/s10260-011-0163-2, to appear.

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

Caju, P., C. Fuss, and L. Wintr (2009). Understanding sectoral differences in downward real wage rigidity: workforce composition, institutions, technology and competition. Working paper series no 1006, European Central Bank. Wage dynamics network.

Caju, P., F. Rycx, and I. Tojerow (2009). Inter-industry wage differentials: how much does rent sharing matter? *Journal of the European Economic Association 79*(4), 691–717.

Caju, P., F. Rycx, and I. Tojerow (2010). Wage structure effects of international trade: evidence from a small open economy. Working paper series no 1325, European Central Bank. Wage dynamics network.

Dell'Aringa, C., P. Ghinetti, and C. Lucifora (2000). Pay inequality and economic performance in italy: a review of the applied literature. In *Proceedings of the LSE conference on 3-4 November 2000*, London.

Dybczak, K. and K. Galuscak (2010). Changes in the czech wage structure: Does immigration matter? Working paper series no 1242, European Central Bank. Wage dynamics network.

Edwards, C. (2010). Public sector unions and the rising costs of employee compensation. Technical Report 1, Cato Institute, Washington, D.C.

EU-SILC (2009). 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.

Eurostat. Anonymisation method for ses 2002 and 2006 microdata – synthesis.

Fitzenberger, B., K. Kohn, and A. Lembcke (2006). Union wage effects in germany: Union density or collective bargaining coverage? Research Report FSP 1169, DFG research programme.

for Education, R. C. and M. U. the Labour Market (2009). Development of econometric methods to evaluate the gender pay gap using structure of earnings survey data. Research paper no. ks-ra-09-011-en-n, European Commission.

Geissberger, T. (2009). *Verdienststrukturerhebung 2006, Struktur und Verteilung der Verdienste in Österreich*. Statistik Austria.

Geissberger, T. (2010). Frauenbericht. teil 4: Sozioökonomische studien. Technical Report 4, Federal Minister for Women and the Civil Service of Austria, Wien.

Graf, M., A. Alfons, C. Bruch, P. Filzmoser, B. Hulliger, R. Lehtonen, B. Meindl, R. Münnich, T. Schoch, M. Templ, M. Valaste, A. Wenger, and S. Zins (2011). State-of-the-art of laeken indicators. Research Project Report WP1 – D1.1, FP7-SSH-2007-217322 AMELI.

Ichim, D. and L. Franconi (2007). Disclosure scenario and risk assessment: structure of earnings survey. In *Joint UNECE/Eurostat work session on statistical data confidentiality*, Manchester. DOI: 10.2901/Eurostat.C2007.004.

Ichim, D. and L. Franconi (2010). Strategies to achieve sdc harmonisation at european level: Multiple countries, multiple files, multiple surveys. In *Privacy in Statistical Databases'10*, pp. 284–296.

Kolb, J.-P., R. Münnich, S. Beil, A. Chatziparadeisis, and J. Seger (2011). Policy use of indicators on poverty and social exclusion. Research Project Report WP9 – D9.1, FP7-SSH-2007-217322 AMELI.

Marsden, D. (2010). Pay inequalities and economic performance. Technical Report PiEP Final Report V4, Centre for Economic Performance London School of Economics, London.

Messina, J., M. Izquierdo, P. Caju, C. Duarte, and N. Hanson (2010). The incidence of nominal and real wage rigidity: an individual-based sectoral approach. *Journal of the European Economic Association 8*(2-3), 487–496.

Pointner, W. and A. Stiglbauer (2010). Changes in the austrian structure of wages. Working paper series no 1268, European Central Bank. Wage dynamics network.

Simón, H. (2010, 06). International differences in wage inequality: A new glance with european matched employer-employee data. *British Journal of Industrial Relations 48*(2), 310–346.

Stockinger, S. (2010). Frauenbericht 2010. Technical report, Federal Minister for Women and the Civil Service of Austria, Wien.

Sweeney, L. (2002). *k*-anonymity: a model for protecting privacy. *Int J Uncertain Fuzziness Knowl Syst 10*(5), 557–570.

Templ, M. (2008). Statistical disclosure control for microdata using the R-package sdcMicro. *Transactions on Data Privacy 1*(2), 67–85.

Templ, M. (2010). *sdcMicro: Statistical Disclosure Control methods for the generation of public- and scientific-use files. Manual and Package.* R package version 2.6.6.

Templ, M., A. Alfons, and P. Filzmoser (2009). Visualization of missing values before imputation using the R-package VIM. *submitted for publication*.

Templ, M., A. Alfons, and A. Kowarik (2011). *VIM: Visualization and Imputation of Missing Values*. R package version 2.1.1.

Winter-Ebmer, R. and J. Zweimüller (1999). Firm size wage differentials in switzerland: Evidence from job changers. *American Economic Review 89*(2), 89–93.