WP. 24
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
(Geneva, Switzerland, 9-11 November 2005)

Topic (iii) Confidentiality aspects of statistical information taking into account register-based data

ESTIMATED RECORD LEVEL RISK FOR THE CVTS

Supporting Paper

Submitted by Statistics Norway and Eurostat[1]

[1] Prepared by Liv Belsby (lbe@ssb.no) and Alexander Stuart McAllister (Alexander.MC-ALLISTER@cec.eu.int).

Estimated record level risk for the CVTS

Liv Belsby [*] and Alexander Stuart McAllister [**]

[*] Statistics Norway, lbe@ssb.no
[**] Eurostat, Alexander.MC-ALLISTER@cec.eu.int

Abstract: We estimate the record level disclosure risk for the anonymised EU *Continuing Vocational Training Survey,* (CVTS). CVTS covers companies in all the MS and in the EFTA countries. NACE and size group of the company are regarded as identifying variables and these two variables are also in business registers. Consequently, the total number of companies for the combinations of these two variables will be known. Additionally we include the variable *has been involved in a take over or not during the reference year,* i.e., up to three identifying variables. The estimates are the conditional expectations of the inverse of the totals, given the totals in the strata and the sample. Our estimates indicate the data is not sufficiently anonymised.

1 Introduction

The CVTS is one of the four surveys covered by the Commission Regulation no. 831/2002 on ‘‘access to confidential data for scientific purposes’’. Moreover, Article 6 in the Regulation requires ‘‘ … that the methods of anonymisation applied to these microdata sets minimise in accordance with current best practice the risk of identification of the statistical units concerned, in accordance with Regulation (EC) No 322/97’’.

This analysis aims to assess the degree of anonymisation that was agreed on between Eurostat and the National Statistical Institutes. The goal of the anonymisation was to produce a Microdata File for Researchers (MFR) and not a public use file.

The approaches to assess the disclosure risk are generally based on estimating the number of “rare” observations with respect to characteristics given in the both the data file and are known for the population.

The disclosure risk is the probability of identifying a company correctly in the dataset. This is often denoted the *record level risk*. The person who attempts to do disclose data is called an *intruder*, see e.g. Benedetti *et al* (2004). By *intruder scenario* we mean the conditions and the type of information under which the identification occurs. We assume that the intruder has available an external database or public registers, e.g. via Internet, with identifiers such as name of the company and other identifying variables which are also in the CVTS dataset. The NACE codes, size of the company measured by number of employees or turnover are examples of identifying variables for companies. The European Business Registers

(Internet site www.ebr.org,) is an example of such a register. This register include business registers from Belgium, Austria, Denmark, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Latvia, Norway, Spain and Sweden. The type of information and detail level vary to some extent from country to country.

Furthermore, we assume that the identifying variables in the register or the database are identical to the identifying variables in the CVTS dataset, i.e., that they are reported without measurement error and refer to the same period. In this analysis we apply the common practice of combining up to three identifying variables at one time. This is motivated by the assumption that if the intruder knows more identifying variables, then he or she is one who knows the company and the information we seek to protect.

In the following we will denote the combination of the identifying variables as a *key*. The individual risk is then the probability of linking the company in the register or database correctly with the company in the CVTS file, given the key.

As pointed out by Polettini (2003), the record level risk has the advantage of allowing for selective protection. The estimated disclosure risk for all the companies in the CVTS file gives a detailed picture of the how safe the data is. Additionally the estimated risks indicate which variables should be further recoded or whether some of the observations should be suppressed to avoid disclosure. The record level risk approach is suggested by Benedetti and Franconi (1999) and is implemented in μ-Argus (2002).

On the other hand the *global risk* approach focuses on population uniques, see Bethlehem *et. al.* (1990). Moreover the global risk approach seeks to classify the whole data file as safe or not. For the CVTS data we would like to obtain a more detailed picture. For example we shall try to determine if there is a difference between countries. Both the record level risk approach by Benedetti and Franconi (1999) and the global risk approach by Bethlehem et. al. (1990) include estimation of the population totals, F_k, as well as similar model assumptions, see Rinott (2003) and Skinner *et. al.* (1994) for more details. Moreover Skinner and Elliot (2002) propose a new measure: the probability that a unique match between a microdata record and a population unit is correct.

The CVTS dataset also has a legal protection in that contracts are signed both by Eurostat and the Institution where the researcher is employed. Before the signing of a contract all data providing countries are consulted. Thus the anonymisation intends exclusively to protect against spontaneous recognition. Note that the legislation in Netherlands does not allow business data to be given out as MFR.

2 The data

All MS and EFTA countries provide data to Eurostat. The sampling unit is the company. The strata are defined by NACE and *size*, i.e., how many employees the company has. The size variable is coded into the groups of 10-49, 50-249 and 250 and more employees. The NACE variable was collected with four digits NACE-code, but anonymised by recoding it into 20 groups. See Appendix A for more details. The data, which we base the analysis on, was sampled in 2000/2001 with reference year 1999. The CVTS data from this year is called CVTS2.

The sampling fraction is the same within one stratum, but varies for the different strata. The number of companies in the strata is known from registers. This will be utilized in the estimation.

NACE	*Size group*	*Stratum*	*Take over or not*	*Number of observations*	*Number of observations sampled from the stratum*	*Number in the strata*
Mining	1	1	no	14	15	279
″	1	1	yes	1	15	279
″	2	2	no	7	11	134
″	2	2	yes	4	11	134
″	3	3	no	6	10	79
″	3	3	yes	4	10	79
.	.		.	.	.	.
.	.		.	.	.	.

Table 1. The structure of the CVTS data

3 Record level risk and the estimation

We illustrate the record level risk with a simple example: Assume that the intruder finds a company in the file, which he suspects is a company he knows. Furthermore, assume that there are two companies in his register (the population) with the same key as the company, which has caught his interest. We assume that both these two companies are in his register, and have the same probability of being the identical company as in his CVTS file. Consequently, we assume that the probability of linking the company in the CVTS file with the right one in his register is simply ½. More generally, we assume that the record risk is the inverse of the total number in the population with the same key as this record. Often the total number in the population with a certain key *k*, say F_k, is unknown. Denoting the number in the sample with key *k* $f_{k,}$ a common estimate for the record level risk is the conditional expectation $E\ \left\{\hat{F}_k^{-1} \middle| f_k\right\}$. Benedetti and Franconi (1999) base the estimation on the assumption that the population total with key k, F_k, given f_k is negative-binomial distributed. Moreover different models have been discussed, see Stander (2003).

Bethlehem *et.al.* also assume that the population totals F_k 's are stochastic variables, and that the parameters, say $\Pi_{k,}$ in the distribution for F_k are stochastic. Furthermore, they suggest a gamma distribution for the Π_k. The conditional distribution $F_k \mid \Pi_k$ is assumed to be a Poisson distribution. Consequently the marginal distribution for F_i is a negative-binomial distribution. Rinott (2003) shows that when the selection probabilities are equal, the model by Benedetti and Franconi (1999) can be regarded as embedded in the model by Bethlehem *et.al.*.

In our estimation approach we will utilize the fact that, as shown in the table above, the total number of companies within each stratum defined by NACE and size denoted by F_s, is known. Furthermore, the two stratification variables are assumed to be identifying variables. The third identifying variable is if variable *has been involved in a take over or not during the reference year*. This means that the key k, will consist of these two stratification variables and a third identifying variable, indicated by $k=\{i,s\}$, where $i=0$ or $i=1$ and where s is the index for the stratum. Consequently, the estimator we use is the conditional expectation, given both the number in the sample with key k, i.e., f_k and the number in the stratum F_s, expressed $E\ \left\{\hat{F}_k^{-1} \middle| f_k, F_s\right\}$

Given the totals in the strata it is not advantageous to model them as stochastic variables as they are then fixed numbers. This is different from the two approaches described above, and it simplifies the estimation of the record level risk.

We illustrate with an example for $s=1$ in table 1. For $k=\{0,1\}$ and $k=\{1,1\}$, respectively we have that f_{01} equals 14, f_{11} equals = 1, F_1 equals 179 and f_{01} + f_{11} is the number in the sample and equals 15. The possible values for F_{01} are {14, … , 278}. To simplify we rather consider the difference F_{01}-f_{01} and denote it by x. This variable can be considered as a sum of 264 independent, binomial experiments. The binomial variable is zero if the company has not been involved in ‘a take over’ and 1 otherwise. The probability of being involved in a ‘take over’ or ‘not’ is estimated from the sample by MLE from the sample.

Thus we assume that the companies in the strata, which are not in the sample, are binomially distributed. Furthermore we estimate the record level risk by the conditional expectation given f_k and F_s as follows,

$$
\begin{aligned}
&E_{p_k}\left\{ {}^{1}\!\!/\!_{F_k} \mid f_k, F_s \right\} \\
&= \sum_{x=0}^{F_s - f_s} {}^{1}\!\!/\!_{(f_k + x)} \cdot \Pr(F_k - f_k = x) \\
&= \sum_{x=0}^{F_s - f_s} {}^{1}\!\!/\!_{(f_k + x)} \cdot \Pr(X = x) \\
&= \sum_{x=0}^{F_s - f_s} {}^{1}\!\!/\!_{(f_k + x)} \cdot Bin(F_s - f_s, x, p_k).
\end{aligned}
$$

As mentioned above, the p_k is estimated by the ratio $f_k/\ \ {}_k f_k$, which is the MLE, i.e., by the relative frequency with key *k* in the sample selected from stratum *s*. Of course the variance of the estimator will be strongly influenced by the size of the sample. We have not performed any estimation of the variance in this study. The estimate will generally not be an unbiased estimate of $1/F_s$.

The estimation has conducted using SAS, utilizing among other things the cumulative binomial formula implemented in SAS BASE. Our estimator is simple to implement, as is illustrated by the program in the appendix.

4 Estimated record level risk

The estimated record level risks are high for many of the records. Fig 1 below shows the estimates for the key NACE**size*" has been involved in a take over or not during the reference year"*. Most of the countries have some records with estimated risk above 50%. However, to make the frequency bars more visible, we limit the scale up to 0.20**.**

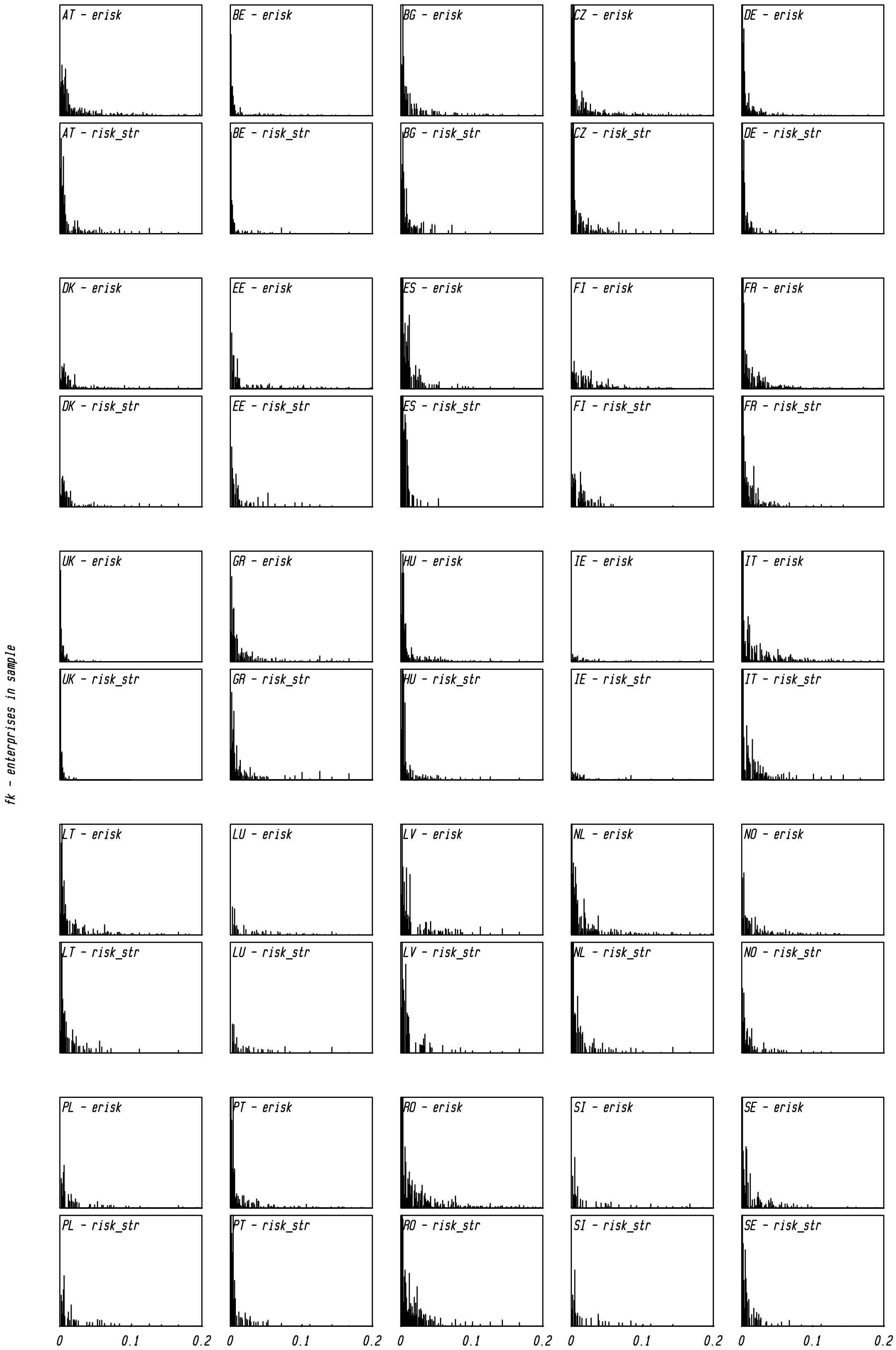
AT - erisk
BE - erisk
BG - erisk
CZ - erisk
DE - erisk
AT - risk_str
BE - risk_str
BG - risk_str
CZ - risk_str
DE - risk_str
DK - erisk
EE - erisk
ES - erisk
FI - erisk
FR - erisk
DK - risk_str
EE - risk_str
ES - risk_str
FI - risk_str
FR - risk_str
UK - erisk
GR - erisk
HU - erisk
IE - erisk
IT - erisk
UK - risk_str
GR - risk_str
HU - risk_str
IE - risk_str
IT - risk_str
LT - erisk
LU - erisk
LV - erisk
NL - erisk
NO - erisk
LT - risk_str
LU - risk_str
LV - risk_str
NL - risk_str
NO - risk_str
PL - erisk
PT - erisk
RO - erisk
SI - erisk
SE - erisk
PL - risk_str
PT - risk_str
RO - risk_str
SI - risk_str
SE - risk_str
fk - enterprises in sample
0
0.1
0.2
0
0.1
0.2
0
0.1
0.2
0
0.1
0.2
0
0.1
0.2

Fig. 1. Fig 1 The estimated disclosure risk for the companies in the CVTS2 using the key NACE_*size*`` has been involved in a take over or not during the reference ear´´_ - erisk. Additionally the disclosure risk using the key NACE_*size - risk_str_.

We see that record level risk are high also when the key consists of NACE*_size_. These probabilities are based on known figures from the registers. Thus they do not have the uncertainty of the estimates for the key NACE_*size *" has been involved in a take over or not during the reference year"_. There are singletons, which of course correspond to record level risk equal to one. In addition many of the records have risk probability of more than 20%.

5 Some concluding remarks

Our estimates clearly show that with a standard approach such as record level risk, the data cannot be considered safe for all countries. The estimated record level risk for NACE20_*size*" has been involved in a take over or not during the reference year"_ indicate that spontaneous recognition may occur. Also when the key consists of only NACE20 and _size_ there are already many records with high risk. As mentioned before, for this key the totals are known for the population.

It is recommended that the data could be more extensively anonymised by for example not releasing either _size_ or NACE. But of course both these variables are an important basis for many analyses of _CVTS_ data.

The disclosure scenario selection of key variables is of course a very important factor in the assessment of how safe the data is. These depend heavily on the availability of registers, which varies to a large extent from country to country, due in part to differences in national legislation. In Norway the NACE code (a modified version is used) is by law public and available in "The central coordinating register of legal entities" together with an identity number and the name of the company. The number of employees is also often available too, and there is a proposal from the authorities this official. Additionally many companies have their financial report available on their Internet site so that potential investors have easy access to this data for their financial analysis.

Another discussion centres around the question "if data is older than2000/2001 is it of any interest to an intruder". For an intruder, who is interested in gaining information useful for improving his financial analysis for investing in the stock market, the data is certainly too old. On the other hand the age of the data is not so important for a journalist acting as an intruder to spread negative publicity for some NSI.

Our results actualise the discussion to what degree the data protection should rely on the anonymisation and on the legal protection of the data, respectively. Business data tend to have higher risk of disclosure than personal and household data. This may be taken, as an argument for the case that access to business data should be treated differently. Currently Regulation (EC) No 831/2002 covers access to both these types of data. One possibility is to adjust the legal framework by increasing the screening of the researches and put more weight on the legal protection and so be less strict with the anonymisation of the business data.

Appendix A. NACE codes used

NACE-categories in CVTS2 based on NACE Rev. 1

NACE 20	Section/ Sub-section	Division	Description
01	C/CA, CB	10-14	Mining and quarrying
02	D/DA	15-16	Manufacture of food products, beverages and tobacco
03	D/DB, DC	17-19	Manufacture of textiles and textile products; Manufacture of leather and leather products
04	D/DE	21-22	Manufacture of pulp, paper and paper products; Publishing, printing and reproduction of recorded media
05	D/DF to DI	23-26	Manufacture of coke, refined petroleum products and nuclear fuel; Manufacture of chemicals, chemical products and man-made fibres; Manufacture of rubber and plastic products; Manufacture of other non-metallic mineral products
06	D/DJ	27-28	Manufacture of basic metals and fabricated metal products
07	D/DK, DL	29-33	Manufacture of machinery and equipment n.e.c.; Manufacture of electrical and optical equipment
08	D/DM	34-35	Manufacture of transport equipment
09	D/DD, DN	20, 36-37	Manufacture of wood and wood products; Manufacturing n.e.c.
10	E	40-41	Electricity, gas and water supply
11	F	45	Construction

12	G	50	Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel
13	G	51	Wholesale trade and commission trade, except of motor vehicles and motorcycles
14	G	52	Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods
15	H	55	Hotels and restaurants
16	I	60-63	Land transport; transport via pipelines; Water transport; Air transport; Supporting and auxiliary transport activities; activities of travel agencies
17	I	64	Post and telecommunications
18	J	65-66	Financial intermediation, except insurance and pension funding; Insurance and pension funding, except compulsory social security
19	J	67	Activities auxiliary to financial intermediation
20	K; O	70-74; 90-93	Real estate, renting and other business activities; Other community, social, personal service activities

Appendix B. SAS program to estimate record level risk[1]

```
data checksafe;
merge strata help;
by country NACE_SP SIZE_SP ;
drop x1;
status='dontknow';
if fk gt T then status='safe';
else if fk le T then status='unsafe';

pk=fk/nsample;
N_est=pk*NSTRA_SP;
risk=1/N_est;

* Estimating the E(1/N ¦ fk,NSTRA_SP), p 277 MOS ws on conf;

data checksafe;
set checksafe;

nn=NSTRA_SP-nsample;

prob=probbnml(pk,nn,0);
sumprob=prob;
```

[1] This is the version to check the NACE*SIZE*TAKEOVER

```
Erisk=prob/fk;

do x=1 to nn by 1;
   xmin=x-1;
   probx=probbnml(pk,nn,x);
   probxmin=probbnml(pk,nn,xmin);
   prob= probx-probxmin;
   Erisk=Erisk+prob/(fk+x);

   sumprob=sumprob+prob;
   end;

* The record level risk is the inverse of the number in the strata;
data checksafe;
set checksafe;
risk_str=1/NSTRA_SP;

proc sort;
by country;

proc print;
var NACE_SP A5C Nsample  fk NSTRA_SP risk_str Erisk risk;
by country;

proc sort data=checksafe;
by country;

proc plot;
plot risk_str*erisk;
by country;

file 'riskest.out';
Put country NACE_SP A5C Nsample  fk NSTRA_SP risk_str Erisk risk;

proc univariate data=checksafe;
var Erisk risk;
by country;
run;
```

References

Benedetti, R. and Franconi, L. (1998), ‘An estimation method for individual risk of disclosure based on sampling design’

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990), Disclosure control of microdata, *Journal of the American Statistical Association* Vol 85, 38-45.

Di Consiglio, L., Franconi, L. & Seri, G. (2003). Assessing individual risk of disclosure: an experiment, *Proceedings from the Joint ECE/Eurostat work session on Statistical Data Confidentiality (Luxembourg, 7-9 April 2003*). MOS Eurostat.

Hundepool, A., Wetering, A. van de, Franconi, L., Capobianchi, A. & Wolf, P.P. de (2002b). *μ-ARGUS, user's manual*, version 3.1.

Pannekoek, J., (1996). 'Statistical methods for some simple disclosure limitation rules', Statistica Neerlandica, Volume 53, Nr. 1 - March 1999, sider 55-67.

SAS Institute (1999): SAS Language References: Dictionary, Version 8.

Skinner, C. J. , and Elliot, M. J. (2002), *A measure of disclosure risk for microdata,* Journal of the Royal Statistical Society, Series B, Methodological, 64 (4) , 855-867

Skinner, C. J., Marsh C., Openshaw S., and Wymer, C. (1994), ``Disclosure control for Census microdata", *Journal of Official Statistics, Vol.10* , 31-51

Stander, Julian (2003). Discussion of Topic (v) : Risk Assessment, *Proceedings from the Joint ECE/Eurostat work session on Statistical Data Confidentiality (Luxembourg, 7-9 April 2003).* MOS Eurostat.

Willenborg, L.C.R.J. & Waal, T. de (1996), Statistical disclosure control in practice,*Lecture Notes in Statistics.* New York: Springer-Verlag.

Willenborg, L.C.R.J. & Waal, T. de (2001), Elements of S*tatistical disclosurecontrol, Lecture Notes in Statistics.* New York: Springer-Verlag.