WP.2
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
(Manchester, United Kingdom, 17-19 December 2007)

Topic (i): Microdata

COMMUNITY INNOVATION SURVEY: COMPARABLE DISSEMINATION

Invited Paper

Prepared by Luisa Franconi and Daniela Ichim, Istat, Italy

Community Innovation Survey: comparable dissemination

Luisa Franconi and Daniela Ichim

Istituto Nazionale di Statistica, via C. Balbo 16, Rome, Italy.
(franconi@istat.it, ichim@istat.it)

Abstract. The European Union is facing the problem of releasing microdata in a multinational setting i.e. microdata stemming from twenty seven member states. Different laws, methodologies, practices and cultural approaches to confidentiality may severely limit the possibility of obtaining comprehensive anonymised data sets. We recall the approach adopted in Europe to design and manage harmonised statistical surveys and claim that such an approach could be successfully applied in the release phase of a survey as well. An application to the Community Innovation Survey is proposed.

1 Introduction

Survey information is disseminated by National Statistical Institutes by means of different products, e.g. indices, tabular data or microdata files. With respect to timeliness, accuracy, level of detail and other quality indicators, each of the previous products has its own features. Each product should correspond to some well-defined users needs. Moreover, for either dissemination strategy, the national statistical institute has to guarantee that respondents confidentiality cannot be breached. The risk of re-identification of each unit should be evaluated. If needed, protection methods are commonly applied to reduce the risk of re-identification of respondents. Information loss criteria are then used to select among several protection achieving a pre-defined acceptable level of re-identification risk.

Dissemination of business microdata files for research (MFR) is one of the most delicate challenges the NSIs are facing nowadays. This is mainly due to two special characteristics of the business surveys. The first one is related to the data sampling strategies. Even if the business surveys are conducted as sample surveys, for some strata, all the units are included in the sample. Consequently, a census is actually conducted in several sampling strata. Moreover, the sampling frame content is often very similar to some external, publicly available, business register. The second special feature of business surveys is represented by the characteristics of the variables usually registered. Some economic variables like turnover or exports generally have very skew distributions. This means that only a small number of units significantly contribute to the overall phenomena. These units would then be more identifiable then others because such concentration is generally publicly known.

There aren't so many examples around the world of release of enterprise microdata. The Regulation CE 831/2002 establishes a list of european business surveys for whom access is granted for research purposes: the Community Innovation Survey, the Structure of Earnings Survey and the Continuing Vocational Training Survey. These surveys have undergone a complex process of harmonisation that inherently includes comparability as an important dimension of the quality framework. Comparability aims at measuring the impact of differences in applied statistical concepts and definitions on the comparison of

statistics between geographical areas, non-geographical domains, or over time. The factors that may cause several statistical figures to lose comparability are attributes of the survey that produces them. Such features may be grouped into two broad categories: the first one relates to survey concepts and the second one relates to measurement and estimation methodologies. To address the problems deriving from the first type of attributes, the approach usually taken at European level is via a regulatory framework where all the concepts of the survey are clearly defined and harmonised. This common framework clearly defines the phenomenon under study, target population, statistical units to be surveyed and all possible metadata descriptions for all the variables involved so as to avoid "structural" non-comparability. As far as the second group of issues is concerned, guidelines on the suggested methodologies for every survey phase are given: sampling design, data collection, weight calculation, imputation and so on. To improve standardisation on all phases, routines are provided by Eurostat for the use of member states. However, "member states are in general free to use whatever methods they prefer as long as some quality thresholds are met", see [6]. These guidelines coupled with quality thresholds represent the core of methodological comparability among member states. In fact, a decision process that gives some guarantee of reaching comparability judgements that are neither ad hoc nor arbitrary is necessary. This involves an assessment of the effects of different practices on predefined statistics, a threshold for determining when action is necessary and a process for choosing acceptable practices. We claim that this general framework adopted at European level to design and carry out harmonised surveys should be implemented also for the microdata release phase. We support this point by analysing the Community Innovation Survey.

In section 2 a description of the survey is given together with examples of how comparability has already been used in other phases of the survey and how it could be implemented in the release of microdata. Section 3 presents current situation as far as anonymisation strategies are concerned and suggests future evolution based on the comparability framework. Finally, some conclusions are given in section 4.

2 Survey comparability

The Community Innovation Survey (CIS) collects information on the innovation tendency at firm level. On each statistical unit, the enterprise, CIS registers information on the economic activity, geographical location, number of employees, expenditure on innovation and research, etc.. The latter is decomposed with respect to factors like intramural/extramural research, acquisition of machinery, acquisition of external knowledge, personnel training, etc. Various facets of innovation are also investigated, e.g. factors that determine or hamper innovation, number of employees with higher education, number of registered patents, etc.. A full survey description of the third wave of CIS (CIS3) is given in [4].

For the CIS in order to ensure what we have called "structural" comparability across countries, Eurostat, in close cooperation with the EU Member States, developed a standard core questionnaire, with an accompanying set of definitions and detailed metadata based on [16]. To address the type of incompatibilities due to issues of measurement and estimation, clear methodological recommendations were given at European level. If we take as an example the sampling design, the general methodological indications were to break down the target population into similar structured subgroups or strata which should be as homogeneous as possible and form mutually exclusive groups. The stratification variables to be used, i.e. the characteristics used to break down the sample into similarly structured groups, were: the economic activities (in accordance with NACE), enterprise size (given

Country	Strata criteria	N° N	N° S	N° R	Sample rate	Response rate
BE	N, S, R	4	5	3	32%	30%
DK	N, S	18	4		39%	30%
DE	N, S, R	21	3	2	12%	21%
EL	N, S, R	20	3	3	30%	62%
FR	N, S	Mixed	3 - 5		12%	82%
IT	N, S, R	46	5	20	20%	62%
LU	N, S	9	3		45%	72% - 73%
NL	N, S	41	3		43%	55%
AT	N, S, R	16	5	3	22%	43%
PT	N, S	42	3		19%	46%
FI	N, S	23	4		35%	50%
SE	N, S	37	6		27%	48%
IS	Census	Census	Census	Census	Census	93%*
NO	N, S	41	5		40%	94%

Table 1: CIS3: stratified sampling for each Member State. N = NACE, S = firm size, R = region. * for a pre-survey by phone.

as the number of employees) and regional aspects. The sample selection should be based on random sampling techniques, with known selection probabilities, applied to strata. In Table 1 we replicate the table presented in [4] on the practices followed by several countries for CIS3.

It is clear that, although the sampling design adopted by member states was broadly a stratified random sampling one, the number of strata as well as the hierarchical level of the stratifying variables were different. This may be due to peculiar characteristics of the phenomenon and distribution of enterprises among NACE in each country as well as well current implemented practices in member states. In any case, the European regulation on innovation sets clear quality thresholds on predefined statistics in order to comply with the comparability dimension of the European data set. Whatever practice was adopted, to be considered comparable, the sample should be carried out in order to achieve a predefined level of precision with regards to the following indicators: the percentage of innovators, the share of new or improved products in total turnover and the total turnover per employee. In particular, it is recommended that the 95% confidence interval for the first two indicators should be within ±5%. For the last indicator the confidence interval should be within ±10% of the estimated indicator. So the process can be summarised in three steps: development of general methodological guidelines, definition of benchmarking statistics and assessment of the effects of different practices on such statistics and, finally, the definition of a threshold for determining when an action is necessary.

The application of such framework in the context of anonymisation procedures for the release of microdata files for research would imply, to start with, the indication of the methodological paradigm of statistical disclosure as described for example in [9]. Such paradigm states the definition of a disclosure scenario, subsequent definition of risk, a measure to assess it, procedures to reduce the risk and finally, but absolutely crucial for the whole process, measures of data utility allowing the final users to judge how poor/good the results of his analysis on the anonymised microdata would be. Such utility measures represent the benchmarking statistics for comparability. In fact, as in the sampling example

the aim was releasing estimates that exhibit certain characteristics, again in the anonymisation phase one main goal should be the production of anonymised data sets sharing certain statistics with the original microdata. The key of the whole process should then be the definition of protection methods that maintain such statistics or the customisation of existing procedures to guarantee pre-selected characteristics.

We now analyse in detail each step of the statistical disclosure paradigm.

2.1 Disclosure scenario

A disclosure scenario defines the users of the released microdata and describes possible ways a malicious user could try to re-identify a unit in the released file. It also examines which variables can be used for re-identification purposes leading to the definition of the so called identifying variables. The scenario highlights the possible disclosure content, too. For any MFR the possible user is a very well-defined one: a scientific researcher who signs a contract impeding him to disclose any individual information from the released file. Consequently, the bona fide of such user may be readily assumed. A researcher has an optimal knowledge about the studied phenomenon, not necessarily about each unit. Anyway, in a business framework, information on the greatest enterprises is well known. We believe that, in this context, the user (although not malicious) has some a priori knowledge of a few large and publicly known enterprises. The disclosure scenario should then prevent spontaneous identification.

With respect to the CIS data, spontaneous identification might be based on combinations of variables included in external business registers such as the main economic activity (NACE), the geographical location (NUTS), the number of employees (EMP) and the total turnover (TURN). For example, the enterprise with the maximum value of turnover in a given NACE code could be publicly known, independently on the exact value of TURN. All these variables are to be considered as identifying variables.

Moreover, due to their skew distributions, other economic numerical variables could be subject to spontaneous identification. For example, among a small group of innovating firms, the one having a dominant investment in research and development, could be known either. RTOT (total expenditure on research and development) and RRDINX (expenditure on intramural R&D) were indicated as variables possibly subject to spontaneous identification and therefore identifying. We notice that some key variables are categorical (NACE and EMP or indeed an indication of the enterprise size) whereas the others are continuous (turnover and expenditures on innovation).

2.2 A general definition of disclosure risk

If we assume a scenario of spontaneous identification then the units at risk will be those that cannot be mistaken for others taking into account some reasonable knowledge (for example economic activity and size). If a unit u belongs to a very dense cloud of units similar to it, it may be supposed that its re-identification wouldn't be of interest: the potential gain would be too little with respect to resources needed. Moreover, since there are other units similar to u, due to the existence of measurement errors, the re-identification would still be uncertain. Instead, if a unit is very isolated with respect to its closest neighbours, it would be more easy to recognise it with some confidence. The latter units should be then considered at risk of re-identification.

2.3 Risk assessment

Once a definition of risk has been given, a measure or estimate of it is necessary. As first step, a specification of the level of detail of the categorical identifying variable is needed as these act as stratifying variables, i.e. defining sub-domains for the analysis of the users. The minimal user requirements are: a 2-digit NACE code for the economic activity variable and indication of the firm size. So, in accordance to the sampling scheme and researcher needs, a variable was produced that groups the number of employees in three classes -49, 50-249, 250+. The geographical variable NUTS was recoded at national level since at macroregional level would have led to an unacceptable number of extremely identifiable enterprises. The definition of these possible sub-domains (main economic activity at 2-digit NACE code and size of enterprise) allows now to concentrate on the continuous identifying variables.

As NACE and size are deemed very reliable variables, they may represent the a priori knowledge of users on the structure of the economy. The risk of re-identification with respect to the numerical identifying variable turnover and innovation expenditures could then be estimated in each sub-domain i.e. for each combination of these two categorical key variables. Based on its own dissemination policy, each National Statistical Institute should define the minimum number of units to which a unit u should be close to in order to be considered safe.

2.4 Microdata protection

Dissemination of business microdata files for research should involve a dedicated anonymisation method in order to avoid the most obvious identifications. By modifying only the records at risk of re-identification and only for the key variables, an optimal trade-off between protection and information loss could be achieved. For microdata files for research purposes, the information quality constraints are much more tight with respect to other dissemination products. Researchers naturally require a high quality data in order to perform reliable analysis and derive correct conclusions. Even if some statistical disclosure limitation method is applied, preservation of the most important statistics is highly desirable. In sections 3.1 and 3.2 two current protections methods will be analysed in more details.

Once the statistical disclosure paradigm has been set up, the definition of benchmarking statistics to check whether different disclosure procedures may be considered comparable from the final user point of view is necessary. Such benchmarking statistics should be related to their needs, i.e. data utility. In the next section we try to address this issue taking as a starting point several analyses carried out on CIS data in recent years.

2.5 Choosing benchmarking statistics: remarks on analyses performed on CIS data

Whatever protection method is applied, some information loss is *unavoidable*. Preservation of the most used analytical properties of the microdata file is possible only if the data protector is aware of the possible data usages. Even if some important statistics cannot be exactly maintained, the information loss with respect to these indicators should be at least quantified. For coherence reasons, from the point of view of the data producer, preservation of already published statistics (mainly tabular data) would be desirable. From the user point of view, data utility measures are essential in knowing the difference with

original data.

In order to gain some insight on possible statistical usages of CIS data a brief review on the scientific literature based on such data was carried out. An example of such review is provided in [1]. Below few common characteristics of several analyses based on CIS data are given.

Analyses are commonly performed at NACE 2-digit level, using the data at national level. This proves the strategic importance of the economic variable. Consequently, dissemination of CIS data at a more aggregated level of the economic activity would be almost useless.

A relationship between companies economic performance and their innovation attitude is commonly investigated. The economic performance may be modeled, for example, through turnover, employment and their variations. Examples of studied statistics are the innovation intensity (expenditure per employee on innovation linked to employment growth, by internal or external innovation) or the share of turnover that is due to new or improved products (quantifying the economic relevance of innovations). Each registered component of the expenditure on innovation is equally used to analyse the innovation phenomenon. Correlations and ratios involving these components and the ones expressing the economic performance seem to be particularly important. Such analyses may be found, for example, in [8, 12, 15, 14, 2].

As usual in survey statistics, weighted means are widely used. Besides being part of the already published tabular data, weighted means were found to be involved in the majority of analysis. For example, any share is expressed through the weighted totals. Consequently, preservation of such statistics is crucial.

As a result of this short overview we can cast a possible list of statistics to be used for benchmarking purposes in data utility. Ratios of innovation variables as a mean to analyse scaled quantities seems predominant. Also the change in turnover with respect to the first year of the reference period (a recorded variable) seems relevant. The next step in this approach is the definition of thresholds on such statistics in order to define comparability of anonymisation methods.

3 Towards comparable dissemination

The current situation regarding CIS microdata access at European level is quite varied. Eurostat has suggested, for CIS3, a micro-aggregation based anonymisation strategy for numerical variables, see [5]; such procedure has recently been reviewed for the CIS4, see [7], but maintaining the same approach. The idea of micro-aggregation and some features of the current implementation are commented in section 3.1. Some member states have accepted this proposal while other for different reasons haven't done so. Following a series of experiences on business microdata anonymisation, Istat is going to release microdata files for research stemming from both CIS3 and CIS4. The essential features of the procedure are described in [10, 11]; in section 3.2 the basic ideas and some results of its application are reported. In both approaches the categorical key variables are recoded following data utility reasonings. The applied recoding is generally given by the minimal user requirements, see section 2.3. However,the two approaches differ a lot on recognising these variables as structural information: the consequences of the different reasonings will be discussed in both section 3.1 and 3.2. In section 3.3 a view on how to gather the current approaches into a unique framework is presented and a way to proceed in order to reach comparable dissemination for CIS microdata is proposed.

3.1 Micro-aggregation based strategy

Micro-aggregation is a very well-known perturbation method introduced in [3]. It aims at creating at least k equal units with respect to the numerical key variables of a statistical unit. The records subject to a micro-aggregation process are clustered in groups of at least k similar units. Then, the value taken by a variable on a unit is replaced by the mean of the group to which the unit belongs to. Instead of the mean, other statistical indicators, like median or weighted mean, may be used.

The basic idea of micro-aggregation is removal of the re-identification risk by means of perturbation. The simplest way to use micro-aggregation is to indistinctly apply it to all numerical variable and to all units. The underline idea is that all values are changed (and there exists always at least k equal values in the released file) so as to prevent of exact disclosure. However in microaggregation there is a clear lack of risk definition and assessment.

When micro-aggregation is applied in real case-studies there are several issues that ought to be discussed. Firstly, the choice of k should be derived from each NSI dissemination policy. In the following examples reporting results on the Italian CIS4 data, k was set equal to three. Secondly, the way in which the numerical variables are micro-aggregated has to be tackled. A possible strategy is to apply a multivariate micro-aggregation, i.e. all numerical identifying variables are simultaneously micro-aggregated. In practical situations, this strategy is not generally used because it was proved that it produces a significant information loss. An alternative would be the individual application of micro-aggregation on each numerical variable independently from the others. This approach is called individual ranking and was actually adopted by Eurostat. The final issue to be addressed in the application of micro-aggregation is the treatment of categorical identifying variable. These could be totally ignored in which case the numerical identifying variables could be readily perturbed by micro-aggregation, independently on such categorical variables, i.e. independently on the specification of important sub-domains. In practice, this means that all sample units are subject to the same micro-aggregation process which produces almost no information loss. Application of micro-aggregation irrespective of the structural economic variables (which are generally represented by the categorical identifying variables) was widely reported in statistical disclosure control. In [17], it was noted that micro-aggregation does not offer any degree of protection, even for higher values of k. A ranking approach for risk assessment of micro-aggregated CIS microdata was reported in [13]. For what attains the disclosure scenario described in section 2.1 we simply notice that individual ranking not stratified by NACE and EMP does implicitly ignores any intruder *a-priori* knowledge.

These issues may be observed in figure 1. The upper panel shows the correlations between RTOT and TURN, for each combination of NACE and EMP. The preservation of this statistic is quite remarkable and the same effect was observed for other statistics taken into consideration; little information loss should be reported. The lower panel proves the inefficiency of such micro-aggregation in protecting visible dominant values: an extreme isolated outlier maintains this property when individual ranking is applied without taking into consideration the categorical identifying variables NACE and EMP. Setting k equal to 5, didn't seem to improve too much the protection level achieved by this type of micro-aggregation. The weighted mean usage instead of the average, didn't significantly change the outcome of this simulation, too.

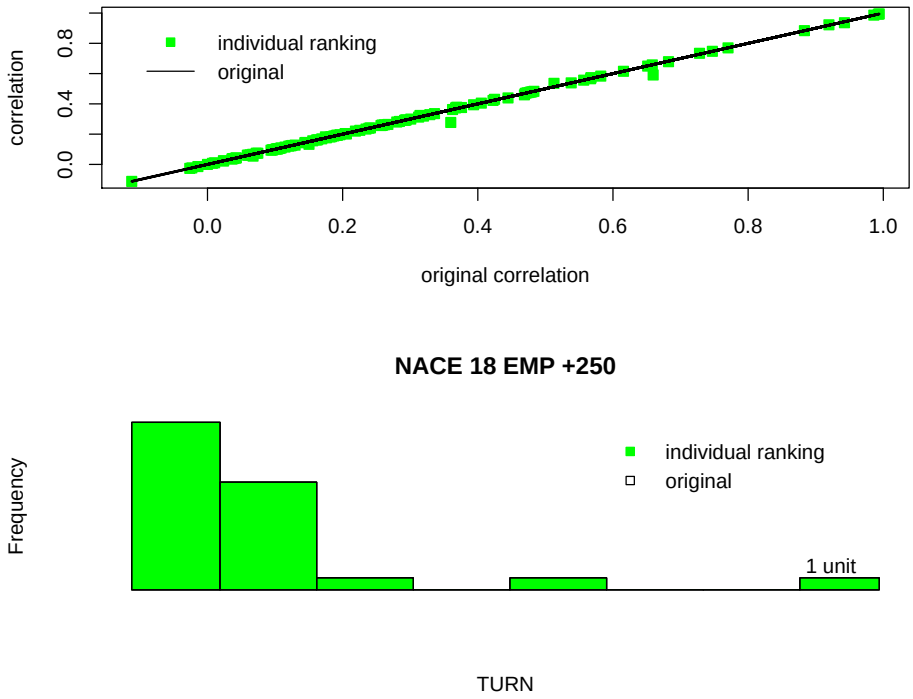

Figure 1: Results obtained when individual ranking is applied to TURN. $k = 3$.

3.2 Stratified selective masking

An alternative approach has been followed in [10, 11] where the paradigm of statistical disclosure has been followed. In strata defined according to the disclosure scenario chosen the risk of re-identification has been estimated following a density approach. Indeed, the density of units around a unit u may be quantified by the local outlier factor. The latter is a relative measure of the degree of isolation of u. Then, by setting a threshold, the units at risk of re-identification in each stratum may be singled out. In [11] an automatic method for this latter threshold setting is also discussed.

As far as protection is concerned different methods are proposed according to the nature of the identifying variables: categorical or continuous. For categorical identifying variables representing structural information of the phenomenon (NACE, size and Nuts) no perturbation is performed. If necessary, i.e. when only a very small number of enterprises are present in each combination of such variables, a recoding is suggested. The combinations of the levels of the resulting variables are the strata where the successive protection of the continuous identifying variables will be performed. For such variables a selective masking method based on the same uncertainty principle is used; moreover, by modifying only the records at risk of re-identification the information loss could be reduced. The selective masking proposed in [11] is based on the nearest neighbour imputation and micro-aggregation. The first perturbation method is applied only to those units at risk whose nearest neighbour is not at risk of re-identification. Then, micro-aggregation is applied to the remaining units at risk.

For the Italian CIS4, this protection method was applied to TURN for each combina-

tion of the categorical identifying variables. RTOT and RRDINX values for the units at risk of re-identification in the spontaneous identification were also perturbed. A change proportional to the change introduced in the corresponding TURN value was aimed. In order to preserve the relationship between variables, the other components of the expenditure in innovation and research were modified, too.

Several statistics were taken into account to perform a comparison between individual ranking and the uncertainty based selective protection method. For each combination of NACE and EMP, the preservation of TURN distribution was firstly addressed. The figure 2 shows a typical result. As expected, the individual ranking reduces the skewness. The comparison of variances and correlations is reported in figure 3. Generally, the selective protection method preserves better these statistical indicators. As mentioned in section 2, the expenditure in innovation and research variables are very concentrated in some units. Figure 3 also shows the performance of the two protection methods with respect to the Gini concentration coefficient. Since the ratios like RTOT/TURN are very much used in data analysis, the perturbation effect on such distributions was assessed; a typical outcome is presented in figure 3.

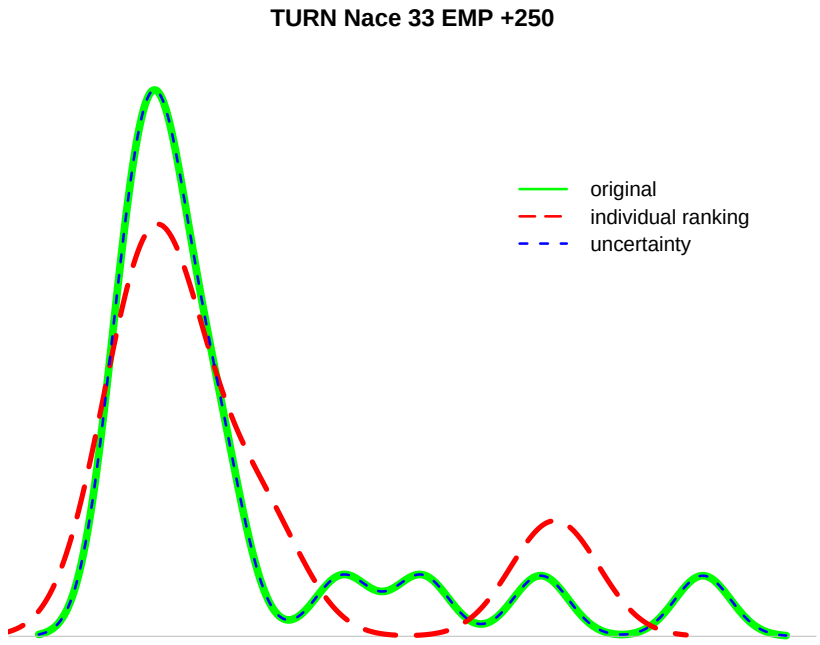

Figure 2: Comparison of individual ranking and uncertainty-based selective masking; both were applied for each combination of NACE and EMP.

3.3 Reaching comparability

An interesting feature of the anonymistion procedure outlined in 3.2 is that for extreme choice of the parameters in the risk assessment phase the protection process is equivalent to individual ranking. In fact, if a degenerate distance is considered for the categorical

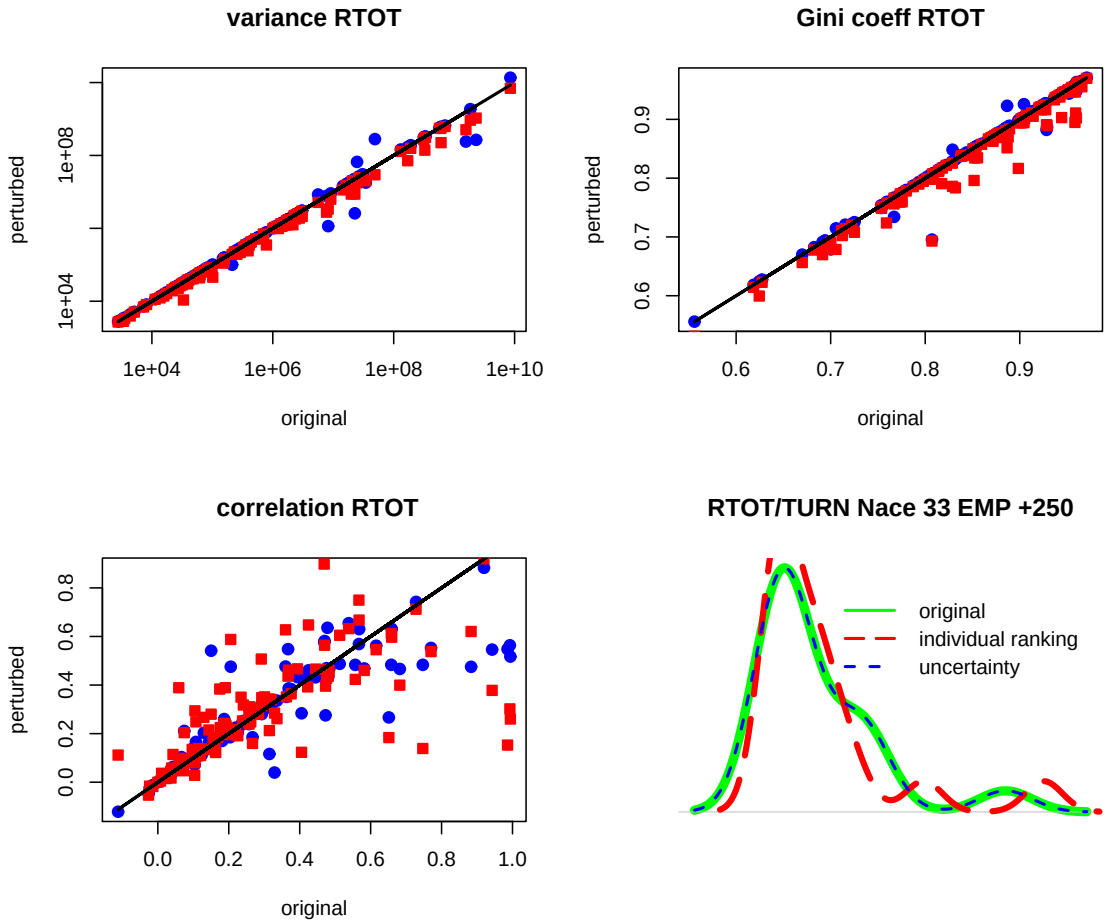

Figure 3: Comparison of some analytical properties. The red squares represent the individual ranking; the blue circles represent the selective uncertainty-based method; the black lines indicate the original values.

identifying variables, the method would be applied irrespective of NACE and EMP. Additionally, if an extreme threshold (zero) were used, all units would be considered at risk of re-identification. According to the procedure, all these units will be protected using micro-aggregation. An evolution of the current situation could see selective masking as possible framework for choosing different grades of anonymisation. Eurostat could provide member states guidelines on the release of CIS microdata along the lines with what is currently done for the other phase of the survey (eg. sampling). CIS experts and users could define further benchmarking statistics useful to measure relevant data utility and set thresholds to guarantee a common *baseline* quality for anonymous microdata. Careful definition and tuning of benchmarking statistics coupled with clear threshold setting would allow comparability of analysis among different methods and different parameter choices in different member states. After a period of training of member states and the preparation of suitable routines implementing different methods and evaluating the benchmarking statistics it would be possible to move towards a complete harmonisation on this subject.

4 Conclusions and further research

To obtain comparable and high quality survey data, a great effort was already made at EU level to harmonize CIS data collection and processing. Ideally, dissemination of microdata would have to be harmonized, too. Unfortunately, depending on the national situation, e.g. law restrictions or availability and quality of external business registers, each National Statistical Institute must face its own problems. Nonetheless, harmonisation of the anonymisation methodology may be achieved twofold. Firstly, different intensity parameters may be used to reach some required, pre-defined, protection levels. Secondly, the analytical issues should be put in evidence.

In this paper we claim that the implementation of the classical statistical disclosure paradigm for enterprise microdata is indeed possible. Moreover, we propose a flexible framework for developing different anonymisation procedures suitable for different member states, guaranteeing the final users on data quality. This is achieved through the use of the comparability concept. Careful definition of relevant statistics for the type of data under analysis is a key issue for defining data utility measures. Given the assurance of a pre-defined acceptable re-identification risk level, preservation of benchmarking statistics should then be the primary objective, independently on the anonymisation methodology. Further research will be devoted into developing guidelines, setting the list of key statistics and implementing the whole framework investigating the degree of flexibility of the protection methodology to reach the required level for the data utility indicators. Cooperation between survey experts and methodologists is strategic for improving the current situation by both increasing the number of microdata offered and improving the quality of the released data.

Acknowledgment

The views expressed in the paper are those of the authors and do not necessarily reflect Istat policies. The authors thank Valeria Mastrostefano for helpful suggestions.

References

[1] Arundel, A. and Bordoy, C. (2005) "The 4^{th} Community Innovation Survey: Final Questionnaire, Supporting Documentation, and the State-of-Art for the Design of the CIS", working paper, available on request.

[2] Belderbos, R., Carree, M. and Lokshin, B. (2004) "Cooperative R&D and Firm Performance", *Conference on Industrial dynamics, innovation and development.*

[3] Defays, D. and Anwar M.N. (1998), "Masking Microdata Using Micro-Aggregation", *Journal of Official Statistics*, **14 (4)**, 449–461.

[4] Eurostat (2004) "Innovation in Europe: Results for the EU, Iceland and Norway", *Panorama of the European Union, Theme 9 Sciences and technologies, European Communities.*

[5] Eurostat (2005) "The Third Community Innovation Survey. Methodology of Anonymisation."

[6] Eurostat (2006) " The Fourth Community Innovation Survey (CIS4). Methodological recommendations", Doc. Eurostat/F4/STI/CIS/2b.

[7] Eurostat (2007) "The CIS4 micro-data anonymisation method".

[8] Evangelista, R. and Mastrostefano, V. (2006) "Firm size, sectors and countries as sources of variety in innovation", *Economics of Innovation and New Technology*, **15 (3)**, 247–270.

[9] Hundepool, A. *et. al.* (2006) "Handbook on Statistical disclosure control", available at http://neon.vb.cbs.nl/casc/.

[10] Ichim, D. (2007)"Microdata anonymisation of the Community Innovation Survey data: a density based clustering approach for risk assessment", *Documenti Istat*, 2, available at www.istat.it.

[11] Ichim, D. (2007) "Disclosure control for business microdata: a density-based approach", submitted.

[12] Klomp, L. and van Leeuwen, G. (2001) "Linking innovation and firm performance: a new approach", *International Journal of the Economics of Business*, **8(3)**, 343–364.

[13] Leppälahti, A. and Teikari, I. (2007) "Problems with micro-data from small countries", *32nd CEIES Seminar Innovation Indicators - more than technology.*

[14] Loof, H. and Heshmati, A. (2002) "Knowledge capital and performance heterogeneity: a firm level innovation study", *International Journal of Production Economics*, **76(1)**, 61–85.

[15] Mastrostefano, V. and Pianta, M. (2007) "Innovation Dynamics and Employment Effects", submitted.

[16] OECD and Eurostat (1997) "Oslo-Manual, Proposed Guidelines for Collecting and Interpreting Technological Innovation Data", *Organisation for Economic Co-Operation and Development*, Paris.

[17] Winkler, W. (2004) "Re-identification methods for masked microdata." In *Privacy in Statistical Databases.*, Eds. J. Domingo-Ferrer and V. Torra, 216–230.