Some Results Obtained from the Application of the ISTAT CIS Anonymisation to Three Member States: Italy, United Kingdom and Portugal

Daniela Ichim[1], Maria Teresa Crespo[2], Philip Lowthian[3]
Servizio Metodi, Strumenti e Supporto metodologico per i processi di produzione statistica,
1 Istituto Nazionale di Statistica, Via Cesare Balbo, 16, 00184 Roma, Italia, e-mail: ichim@istat.it
2 Instituto Nacional de Estatística, Av. António José de Almeida 1000-043 Lisboa, Portugal
3 Office for National Statistics, 1 Myddelton Street , London, EC1R 1UW, UK

Version: October 2009

Abstract
The Community Innovation Survey (CIS) is one of the surveys mentioned in the Commission Regulation 831/2002. For the dissemination of a microdata file for research purposes, ISTAT developed a statistical disclosure control methodology. Its flexibility was tested using microdata files stemming from the British, Portuguese and Italian CIS 4 surveys. This document presents only some comparisons between the application of the Italian dissemination methodology to the British, Portuguese and Italian microdata files. Considering both the risk of disclosure and data utility requirements, several dissemination/quality indicators were calculated.

Acknowledgement
The experimentation on the UK CIS4 microdata was performed at the Virtual Microdata Laboratory. At the processing time (February 2008), the UK CIS4 microdata were not the final validated microdata; consequently, the results presented here might differ from the indicators computed on the official UK CIS4 microdata file.

ISTAT, INE, ONS and VML are not responsible for any views or results presented in this paper. The author was supported by the ESSnet-SDC project.

1. Introduction

Microdata files are one of the products disseminated by the National Statistical Institutes (NSI) in order to satisfy the information need of the users. Anyway, the NSI should also guarantee that information about the respondents could not be too accurately inferred. The dissemination process of a microdata file may be summarized in three steps: 1) definition of a disclosure scenario, including the specification of the key variables, 2) risk assessment and 3) reduction of the disclosure risk.

The CIS disclosure limitation methodology developed by ISTAT (the Italian National Statistical Institute) takes into account both economic features of the data and the dissemination policy of the National Statistical Institute. An important key point is the fact that the final protected data set would be released for research purposes, and hence subject to a signed contract. Consequently, a rigorous study of possible disclosure scenarios was carried out in order to define the identifying variables. Two spontaneous identification scenarios based on structural and non structural information were modelled. Then a careful risk assessment analysis was performed to single out the records at risk. The risk assessment was performed considering both the economic classification and size classes, when these were considered identifying variables. In the disclosure scenario based on structural information, the basic idea is that a value of an identifying variable (or indeed the values from a set of identifying variables) is considered at risk if it is isolated i.e. the “density” of the points around this value is not deemed sufficient (below a certain threshold). The "density" concept is defined using both the distance between points and number of points in a neighbourhood. The "distance" used may be easily extended to a multivariate situation[1]. With respect to the used thresholds, the number of isolated (hence at risk of identification) units would tend to increase or decrease. These thresholds may be defined according to the observed phenomenon, assumed disclosure scenario and national dissemination policy. A disclosure scenario was also modelled in order to take into account some qualitative information a possible intruder might have. Justified by the research purposes of the microdata file release, **only** the key variables and **only** the records at risk of identification should be perturbed whereas the rest of the file should be released unchanged. Perturbation is mainly achieved by an imputation from the nearest clustered unit and a particular microaggregation. For extreme thresholds choice in the identification phase, the method proposed by ISTAT reduces to microaggregation. A deterministic adjustment procedure is proposed to maintain the published totals. Modifying only the key variables and only for the records at risk of identification, a lot of variables (including the sampling weights) would remain unchanged, hence coherence with many already published aggregate statistics would be naturally achieved.

The microdata disclosure limitation methodology is based on the following eight steps:

1. Definition of the disclosure scenarios.
    a. spontaneous identification scenario based on structural information
    b. spontaneous identification scenarios based on non-structural information
        i. dominance scenario
        ii. uniqueness scenario
2. Preliminary work on variables.
    a. variable suppression
    b. global recoding
    c. preliminary rounding
3. Risk assessment: identification of units at risk.

[1] considering both continuous and categorical variables.

   a. spontaneous identification scenario based on structural information (clustering)
   b. spontaneous identification scenarios based on non-structural information (uniqueness)
4. Microdata protection
   a. imputation from the nearest clustered unit
   b. micro-aggregation on tails
   c. multiplicative perturbation
   d. micro-aggregation of some records of variables related to the first year of the reference period
5. Adjustment to preserve published totals.
6. Audit strategies.
   a. negative values
   b. insufficient protection
   c. overprotection
7. Information loss assessment.
   a. variance comparison
   b. correlations comparison
   c. users perspective.
8. Description of the microdata file to be released.

More details on the ISTAT disclosure limitation methodology may be found in [1, 2, 3]. In this work only the flexibility of the methodology is discussed and assessed. This characteristic was evaluated through the application of the same methodology to the UK, Portuguese and Italian CIS4 microdata. In this document the comparisons between several dissemination/quality indicators are indicated.

In Section 2 the UK, Portuguese and Italian Community Innovation Surveys 4 are very briefly described. In Section 3, some details on the various applications of the ISTAT disclosure limitation methodology are given. In Section 4, the dissemination/quality indicators considered in this work are briefly introduced and the obtained results are presented in a comparative manner.

2. The Community Innovation Surveys

This section briefly illustrates the main characteristics of the microdata files that were analyzed.

2.1 European CIS4

The Community Innovation Survey (CIS) is a survey of innovation activity in enterprises covering EU Member States, candidate countries, Iceland and Norway. The data are collected on a two-yearly basis (from 2004 onwards). The latest survey (CIS 4) was carried out in 25 Member States, candidate countries, Iceland and Norway in 2005 based on the reference year 2004.

In order to ensure comparability across countries, Eurostat, in close cooperation with the EU Member States and other countries, developed standard core questionnaires for CIS 4, with an accompanying set of definitions and methodological recommendations. CIS 4 is based on the Oslo Manual (2nd edition, 1997), which gives methodological guidelines and defines the concept of innovation, and on Commission Regulation No 1450/2004.

More details on the European CIS4 may be found in [7].

STATISTICAL UNITS

The main statistical unit for CIS 4 was the enterprise, as defined in Council Regulation No 696/1993 on statistical units or as defined in the national statistical business register. EU Regulation No 2186/1993 requires Member States to set up and maintain a register of enterprises, as well as associated legal units and local units.

TARGET POPULATION

The population of CIS 4 is determined by the size of the enterprise and its principal activity. At least all enterprises with 10 or more employees in any of the specified sectors were included in the statistical population. The target population of CIS 4 was the total population of enterprises with mostly the following market activities: mining and quarrying (NACE 10-14), manufacturing (NACE 15-37), electricity, gas and water supply (NACE 40-41), wholesale trade (NACE 51), transport, storage and communication (NACE 60-64), financial intermediation (NACE 65-67), computer and related activities (NACE 72), architectural and engineering activities (NACE 74.2) and technical testing and analysis (NACE 74.3).

TYPE OF SURVEY

Most Member States and other countries carried out CIS 4 by means of a stratified sample survey, while a number of countries used a census or a combination of both.

The CIS 4 data are organised in the Eurostat reference database following broadly the same structure as the harmonised survey questionnaire.

The enterprise size classes referred to in this publication are:

- small: 10-49 employees;
- medium-sized: 50-249 employees;
- large: 250+ employees.

REFERENCE PERIOD

For CIS 4 the observation period covered was 2002-2004 inclusive, i.e. the three-year period from the beginning of 2002 to the end of 2004. The reference period for CIS 4 was the year 2004. All countries covered collected data for this observation period; only the Czech Republic took 2003-2005 as the observation period.

DEFINITIONS (Oslo Manual, 1997)

Innovation: a new or significantly improved product (good or service) introduced to the market or a new or significantly improved process introduced within an enterprise. Innovations are based on the

results of new technological developments, new combinations of existing technology or the utilisation of other knowledge acquired by the enterprise.
Enterprises engaged in innovation activity (propensity to innovate): enterprises that introduce new or significantly improved products (goods or services) to the market or enterprises that implement new or significantly improved processes. Innovations are based on the results of new technological developments, new combinations of existing technology or the utilisation of other knowledge acquired by the enterprise. The term covers all types of innovator, i.e. product innovators, process innovators and enterprises with only ongoing and/or abandoned innovation activities.
An *organisational innovation* is the implementation of new or significant changes in firm structure or management methods that are intended to improve the firm's use of knowledge, the quality of its goods and services, or the efficiency of work flows.

The main research and development expenditure variables are:

RTOT	= total expenditure on research and development;
RrdInX	= Expenditure in intramural R&D;
RrdExx	= Expenditure in extramural R&D;
RMacX	= Expenditure in acquisition of machinery;
ROekX	= Expenditure in other external knowledge;
RPreX	= Expenditure in other preparation;
RTrX	= Expenditure in training and
RMarX	= Expenditure in market introduction of innovation

More details on the Community Innovation Survey may be found at the Eurostat web-site.

2.2 The United Kingdom CIS4[2]

The UK Innovation Survey was funded by the Department of Trade and Industry (DTI), at the moment known as UK Department for Business Innovation and Skills (BIS). The survey was conducted on behalf of the DTI by the Office for National Statistics (ONS), with assistance from the Northern Ireland Department of Enterprise, Trade and Investment (DETI). The UK Innovation Survey is part of a wider Community Innovation Survey (CIS) covering European countries. The survey is based on a core questionnaire developed by the European Commission (Eurostat) and Member States. This is the fourth iteration of the survey (CIS 4) – CIS 3, covering the period 1998 to 2000, was carried out in 2001 and the results form part of various EU benchmarking exercises (see www.cordis.europa.eu.int/en/home.htm). The UK Innovation Survey 2005 sampled over 28,000 UK enterprises. The survey was voluntary and conducted by means of a postal questionnaire. A copy of the questionnaire used can be found at www.dti.gov.uk/iese/cis4quest.htm (or http://www.berr.gov.uk/).

The survey covered enterprises with 10 or more employees in sections C to K of the Standard Industrial Classification (SIC) 2003. The 2005 survey included additional sector: Sale, maintenance and repair of motor vehicles (SIC 50), Retail trade (SIC 52) and Hotels and restaurants (SIC 55) excluded from the 2001 survey. The sample was drawn from the ONS Inter-Departmental Business Register (IDBR) in December 2004. Details can be found at www.dti.gov.uk/iese/cis4sample.htm (or http://www.berr.gov.uk/).

[2] Source: ONS.

The questionnaires from the initial survey were distributed on 31 March 2005. Valid responses were received from 16,446 enterprises to give a response rate of 58 per cent.

More details and analyses may be found in [4].

2.3 The Portuguese CIS4

CIS4 was conducted by OCES, an entity from the Ministry of Science, Technology and Superior Education.

The definition and selection of the sample was done in INE-PT, using as sampling frame the FUE, the statistical register of enterprises (Ficheiro de Unidades Estatísticas – statistical units register). The sample size was determined according to the precision requests of the methodological recommendations for CIS4 given by *EUROSTAT*. The economic activity, geographical location and size of the enterprises were the stratification variables.

Stratification variables:

NACE - two digits for the most part of the economic activities, but at 3 digits level for a few activities, namely 742 and 743.
NUT2 - five regions in the mainland and Açores and Madeira islands
The stratification by NUT2 was done only to guarantee the geographical distribution of the sample and not to be representative at this level, because it would result in a very large sample.
SIZE classes of number of employees - [10; 49],[20; 249] and [250 + [workers. The last stratum was all observed. For some activities it was also considered the size class [5; 9], but not used in this study.

Universe - 26 723 units
Sample size - 7 370 units
Number of strata – 639
Response rate – 35%

Weights calculation – No calibration method was used.

Final weighting factors – ratio of the number of enterprises in the universe in each stratum to the number of enterprises in the same stratum, in the realised sample.

More details on the Portuguese CIS4 may be found at www.ine.pt.

2.4 The Italian CIS4

The industries mentioned in section 2.1 were included in the core target population of the CIS4.“Non-core” industries that were covered in addition are: construction (NACE 45); motor trade (NACE 50); retail trade (NACE 52); hotels and restaurants (NACE 55); real estate activities (NACE 70); renting of machinery and equipment without an operator (NACE 71); research and development (NACE 73); other business activities: legal, accounting, book-keeping and auditing activities; tax consultancy; market research and public opinion polling; business and management consultancy; holdings (NACE 74.1); advertising (NACE 74.4); labour recruitment and provision of personnel (NACE 74.5); investigation and security activities (NACE 74.6); industrial cleaning (NACE 74.7); miscellaneous business activities n.e.c. (NACE 74.8).

All enterprises included in the target population follow the minimum coverage which was all enterprises with 10 employees or more.

The survey was based on a one stage stratified simple random sample. At least 6 enterprises in each stratum were selected. In the case of less than 6 enterprises in a stratum, a full census was conducted. The target population of the CIS4 was broken down into similar structured subgroups or strata (which should be as homogeneous as possible and form mutually exclusive groups).
The stratification variables to be used for the CIS4, i.e. the characteristics used to break down the sample into similarly structured groups, were:

- The economic activities (in accordance with NACE).

  Stratification by NACE has been done at least at two-digit (division) level, except for NACE 24, 35, 74. NACE 24.4, 35.3, 74.2 and 74.3 groups were treated as separate NACE sectors while the remaining groups of NACE 24, 35 and 74 were treated as single NACE sectors.
- Enterprise size according to the number of employees.

  The size-classes used were the following:
  - *10-49 employees;*
  - *50-249 employees;*
  - *250+ employees.*
- Regional aspects at NUTS 2 level.

  The regional allocation of the sample was taken into consideration when sampling. In particular, the breakdown of national territory into regions was performed on the basis of the NUTS level 2. (Regulation EC No 1059/2003 of the European Parliament and of the Council of 26 May 2003 on the establishment of a common classification of territorial units for statistics - NUTS)

The official, up-to-date, statistical business register, called ASIA (Archivio Statistico delle Imprese Attive - statistical business register of active enterprises) was used. It provided both the key variables for the stratification (number of employees, NACE economic activity, NUTS geographical information) and the identification characters (enterprise name, address, etc.).
The Italian CIS4 sample included 44,571 enterprises out of a population of about 193,300 enterprises with 10 employees or more and potentially active in the year 2004 and the average response rate turned out to be 49%.

Data were collected through a combination of census and sample survey. The census concerned all the enterprises with more than 249 employees: they were around 3,300 enterprises out of a target population of more than 193,300 units of target population.

Calibration estimators methodology, currently applied at ISTAT, were used for the estimation process. For CIS, as well as for most of the business surveys, number of enterprises and number of employees were used as auxiliary variables, according to the information provided by the Italian Official Business Register ASIA.

More details on the Italian CIS4 may be found in [5].

3. Application of the ISTAT CIS Statistical Disclosure Limitation Methodology

In this section the settings of the statistical disclosure methodology as applied to the UK, PT and IT CIS4 microdata are described.

3.1 UK CIS4

a) The variable RTOT was computed as sum of the variables RrdInX, RrdExx, RMacX, ROekX, RPreX and RMarX.
b) For the 2004 Turnover (TURN 2004) and number of employees an ad-hoc imputation procedure was applied. Whenever the TURN 2004 and/or number of employees were missing or equal to zero, their values were imputed using two variables derived from a linking operation with an ONS enterprise archive.
c) The principal economic activity was recoded following the NACE rev1.1 – 2 digits classification. This recoding was performed even for NACE 74 category.
d) The large enterprises were considered to be those enterprises with more than 250 employees.
e) The 2004 number of employees was recoded in 3 classes: E1: 10-49 employees, E2: 50-249 employees, E3: +250 employees. Large enterprises were considered to be those having more than 250 employees.
f) Several NACE 2-digit categories were aggregated in order to take into account the survey features and the national dissemination policy. These aggregations are illustrated in Section 4. For some particular NACE categories, the size classes were aggregated in order to have a minimum number of units in each domain. These further recodings are illustrated in section 4, too.
g) The clustering algorithm was applied using NACE and Size classes as stratification variables. The minimum number of units required in the vicinity of a units in order to declare it "not-at-risk of disclosure" was set equal to 5 (the parameter MinPts). If in a domain (strata or combination) the number of units was smaller than 15, all the units were considered at risk of disclosure (the clustering algorithm was not applied). The threshold on distances was defined by the third quantile (it was not possible to use the change point criterion). Generally the quantile criteria worked quite well in the sense that the identified point was the true abrupt change point. Anyway, there were some cases where the quantile criterion was not a good substitute of the change point criterion, e.g. the domain defined by the NACE 23 and size class E1. In figure 1, examples of right and false abrupt change point identification are presented.
h) In the uniqueness scenario, the final weights were used instead of the direct weights that were not registered in the UK CIS4 microdata file.
i) In the dominance scenario, the variables RTOT, RRDINX and RMARX were used.
j) When microaggregation was applied on the tails of the distribution, the individual ranking parameter was set equal to 5.

The settings of the parameters and choices not mentioned in this subsection were indicated in [1].

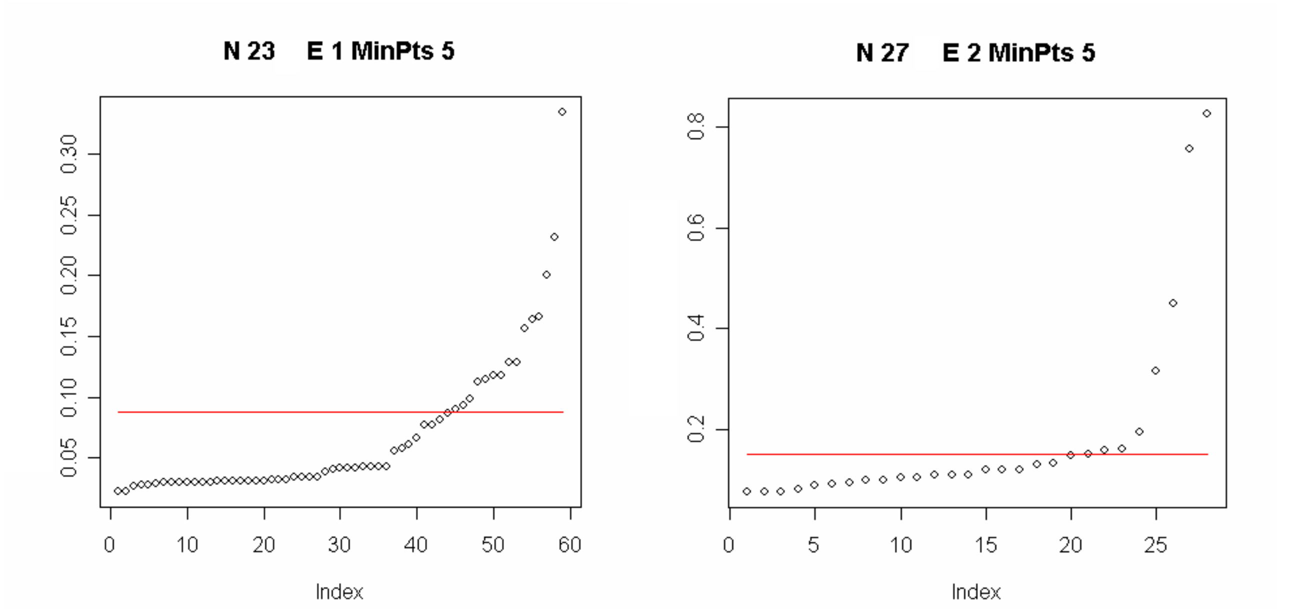


Figure 1. UK CIS4 microdata. Assessment of the third quantile criterion. On the vertical axes the distance from the fifth unit is represented. On the left, domain NACE 23, SIZE CLASS E1, a situation where the quantile criterion didn't work well. On the right, domain NACE 27, SIZE CLASS E2, a situation where the quantile criterion identified the true change point.

3.2 PT CIS4

a) The principal economic activity was recoded following the NACE rev1.1 – 2 digits classification.
b) The large enterprises were considered to be those enterprises with more than 250 employees.
c) The 2004 number of employees was recoded in 3 classes: E1: 10-49 employees, E2: 50-249 employees, E3: +250 employees. Large enterprises were considered to be those having more than 250 employees.
d) Several NACE 2-digit categories were aggregated in order to take into account the survey features and the national dissemination policy. These aggregations are illustrated in Section 4.
e) The clustering algorithm was applied using only NACE as stratification variable. The minimum number of units required in the vicinity of a units in order to declare it "not-at-risk of disclosure" was set equal to 5 (the parameter MinPts). Due to the NACE aggregations performed, the minimum number of units in any strata was set equal to 15. Consequently, the clustering algorithm was applied in all domains. The threshold on distances was generally defined by the change point criterion, except for the NACE categories 40, 55, 61, 62 and 65. Indeed, for these categories, the change point criterion didn't produce reliable results (assessed by graphical inspection). In these cases, the third quantile criterion was used.
f) In the dominance scenario the variables RTOT and RRDINX were used.
g) When microaggregation was applied on the tails of the distribution, the individual ranking parameter was set equal to 3.

The settings of the parameters and choices not mentioned in this subsection were indicated in [1].

3.3 IT CIS4

a) The principal economic activity was recoded following the NACE rev1.1 – 2 digits classification. This recoding was not performed for the NACE 742 and 743 categories.

b) The large enterprises were considered to be those enterprises with more than 250 employees.
c) The 2004 number of employees was recoded in 3 classes: E1: 10-49 employees, E2: 50-249 employees, E3: +250 employees. Large enterprises were considered to be those having more than 250 employees.
d) Several NACE 2-digit categories were aggregated in order to take into account the survey features and the national dissemination policy. These aggregations are illustrated in Section 4. For some particular NACE categories, the size classes were aggregated in order to have a minimum number of units in each domain. These further recodings are illustrated in section 4, too.
e) The clustering algorithm was applied using NACE and Size classes as stratification variables. The minimum number of units required in the vicinity of a units in order to declare it “not-at-risk of disclosure” was set equal to 3 (the parameter MinPts). If in a domain (strata or combination) the number of units was smaller than 10, all the units were considered at risk of disclosure (the clustering algorithm was not applied). The threshold on distances was defined by the change point criterion.
f) In the dominance scenario the variables RTOT, RRDINX and RMARX were used.
g) When microaggregation was applied on the tails of the distribution, the individual ranking parameter was set equal to 3.

The settings of the parameters and choices not mentioned in this subsection were indicated in [1].

3.3 SUMMARY OF SETTINGS

A summary of parameters settings is presented in table 1. It should be noted that some of these settings were derived from the discussions with the British and Portuguese colleagues.

	UK	**PT**	**IT**
Domain definition	NACE, SIZE CLASS	NACE	NACE, SIZE CLASS
Minimum number of units in a domain	15	15	10
Minimum number of units in the vicinity (MinPts)	5	3	3
Threshold definition	quantile	change point, quantile	change point
Microaggregation parameter	5	3	3

Table 1. Parameters settings for the application of the Italian disclosure limitation methodology on the UK, PT and IT CIS4 microdata files.

4. Comparison of the Results Obtained from the Analysis of the UK, PT and IT CIS4 Microdata Files

4.1 NACE Aggregations

The performed NACE aggregations are illustrated in table 2.

UK		PT		IT	
old	**new**	**old**	**new**	**old**	**new**
NACE 10-14	NACE 14	NACE 11-14	NACE 11	NACE 10-14	NACE 14
NACE 15-16	NACE 15	NACE 15-16	NACE 15	NACE 15-16	NACE 15
NACE 23-24	NACE 23	NACE 23-24	NACE 23		
		NACE 30-31	NACE 30		
NACE 40-41	NACE 40			NACE 40-41	NACE 40
		NACE 72-73	NACE 72		

Table 2. The NACE 2-digit aggregations performed to account for the survey feature and the characteristics of the national dissemination policy.

4.2 Size Class Aggregations

For each appropriate NACE 2-digit category, the performed size classes aggregations are illustrated in Table 3. Except for the Eurostat standards, other size classes aggregations were not necessary for the Portuguese CIS4 microdata.

UK		IT	
NACE	**Aggregated size classes**	**NACE**	**Aggregated size classes**
19	E1, E2 and E3		
		37	E1, E2 and E3
62	E1, E2 and E3	62	E1, E2 and E3
		64	E1, E2 and E3
18	E2 and E3		
20	E2 and E3	20	E2 and E3
		23	E2 and E3
30	E2 and E3	30	E2 and E3
37	E2 and E3	67	E2 and E3
40	E2 and E3	73	E2 and E3
61	E2 and E3		

Table 3. The size classes aggregations performed in order to have a minimum number of units (a-priori derived from the national dissemination policy) in each domain.

4.3 Percentages of Isolated Units on the Tails

In this section information of the number of isolated units on the tails are given (TURN 2004). More precisely, the outcomes of the spontaneous identification scenario based on structural information are presented.

In figure 2, the percentages of number of units on the left tail are shown. No particular order was used. In table 4 some descriptive statistics of the percentages of isolated units on the left tail are illustrated. For each domain, the percentages of isolated units on the left tail were computed over the total number of observations belonging to the corresponding domain. Q1 represents the first quantile, Q2 represents the second quantile (the median) and Q3 represents the third quantile.

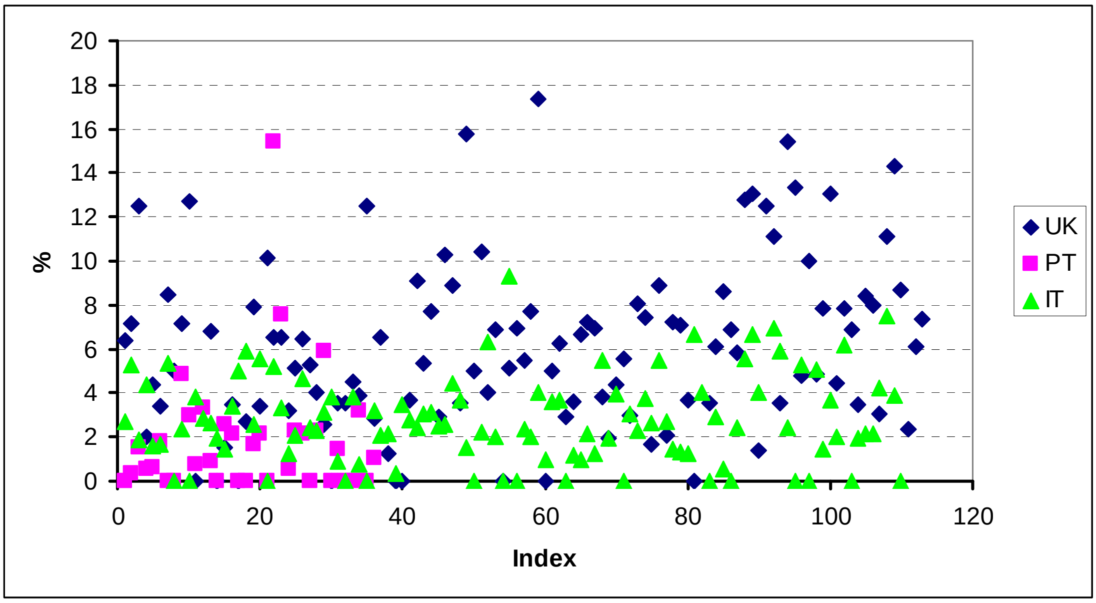


Figure 2. Percentages of isolated units on the left tail (TURN 2004). (a zoom on the most significant region)

	Min	Q1	Q2	Mean	Q3	Max
UK	0.00	3.45	5.83	7.70	8.00	100.00
PT	0.00	0.00	0.97	1.89	2.29	15.38
IT	0.00	1.44	2.48	2.95	3.96	21.43

Table 4. Descriptive statistics of the percentages of isolated units on the left tail. (TURN 2004)

In figure 3, the percentages of number of units on the right tail are shown. No particular order was used. In table 5 some descriptive statistics of the percentages of isolated units on the right tail are illustrated. For each domain, the percentages of isolated units on the right tail were computed over the total number of observations belonging to the corresponding domain. Q1 represents the first quantile, Q2 represents the second quantile (the median) and Q3 represents the third quantile.

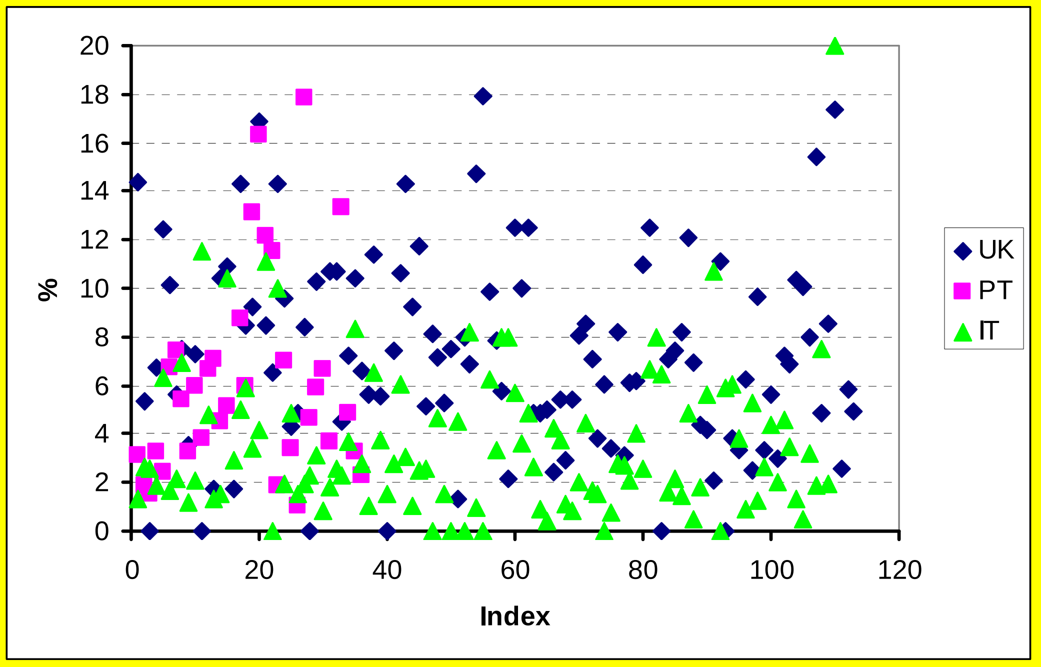


Figure 3. Percentages of isolated units on the right tail (TURN 2004). (a zoom on the most significant region)

	Min	Q1	Q2	Mean	Q3	Max
UK	0	4.861	7.143	9.036	10.26	100
PT	1.064	3.259	5.659	7.138	7.79	23.4
IT	0	1.549	2.635	3.582	4.865	20

Table 5. Descriptive statistics of the percentages of isolated units on the right tail (TURN 2004).

In figure 4, the percentages of the total number of units are shown. No particular order was used. In table 6 some descriptive statistics of the percentages of the total number of isolated units are illustrated. For each domain, the percentages of isolated units were computed over the total number of observations belonging to the corresponding domain. Q1 represents the first quantile, Q2 represents the second quantile (the median) and Q3 represents the third quantile.

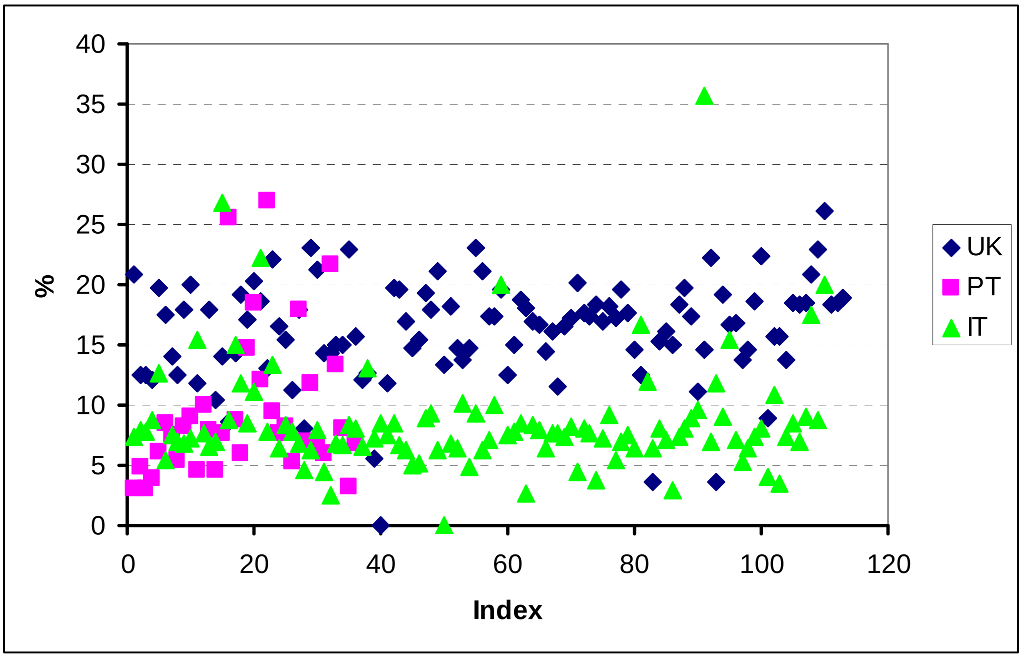


Figure 4. Percentages of the total number of isolated units (TURN 2004). (a zoom on the most significant region)

	Min	Q1	Q2	Mean	Q3	Max
UK	0.00	14.29	16.94	17.66	18.75	100.00
PT	3.06	5.80	7.83	9.44	10.44	26.92
IT	0.00	6.60	7.63	8.62	8.86	35.71

Table 6. Descriptive statistics of the percentages of the total number of isolated units (TURN 2004).

4.3 Number of Units at Risk of Disclosure in the Spontaneous Identification Scenario Based on Non-structural Information

In this section the results of the application of the dominance and uniqueness scenarios are illustrated.

In the uniqueness scenario, for the UK CIS4 microdata, no unit was found at risk of identification. The only enterprise having a weight smaller than 1.5 was not a large enterprise. Considering only enterprises with more than 250 employees, for the Italian CIS4 survey, no unit at risk of re-identification was found in the uniqueness scenario. Only 1 enterprise was found at risk of re-identification in the uniqueness scenario when the Italian statistical disclosure methodology was applied to the Portuguese CIS4 microdata.

In table 7 the number of units at risk in the dominance scenario is presented. This information is not available for the Portuguese CIS4 microdata.

	UK	IT
RTOT, RRDINDX, RMARX	21	25
Turn2002	3	12
Turn2002/Turn2004	8	18

Table 7. Number of units at risk in the dominance scenarios.

4.4 Percentages of TURN 2004 Values Modified by the Statistical Disclosure Limitation Methodology

In figure 5 the percentages of modified TURN 2004 values are shown. No particular order was used. For each domain, the percentages were computed as the number of modified values (independently on the disclosure control stage) over the total number of units belonging to the domain. In table 8 some descriptive statistics of the percentages of modified TURN 2004 values are shown.

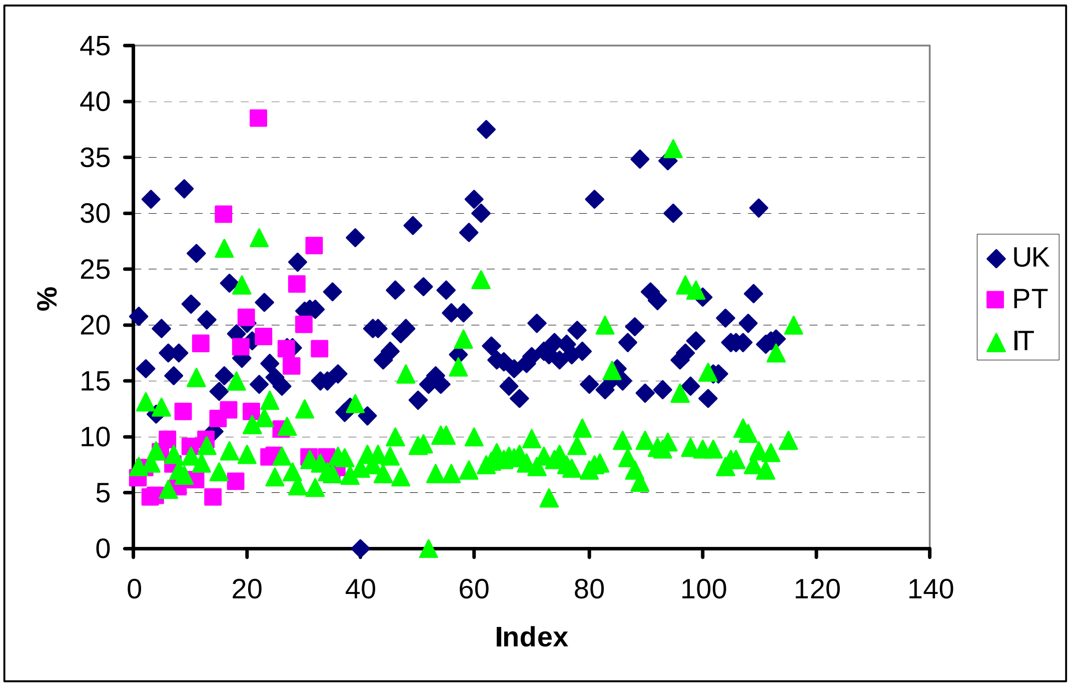


Figure 5. Percentages of the modified TURN 2004 values. (a zoom on the most significant region)

	Min	Q1	Q2	Mean	Q3	Max
UK	0.00	15.64	18.24	20.69	21.43	100.00
PT	4.55	7.16	9.75	12.80	17.90	38.46
IT	0.00	7.47	8.52	14.94	11.89	100.00

Table 8. Percentages of modified TURN 2004 values.

4.5 Variances

For each domain, the ratios of the variances of the perturbed to variances of the original TURN 2004 were computed. The trends of these ratios are illustrated in figure 6 for the UK, PT and IT CIS4 microdata. No particular order was used. In table 9 some descriptive statistics of the ratios of the variances of the perturbed to variances of the original TURN 2004 are shown.

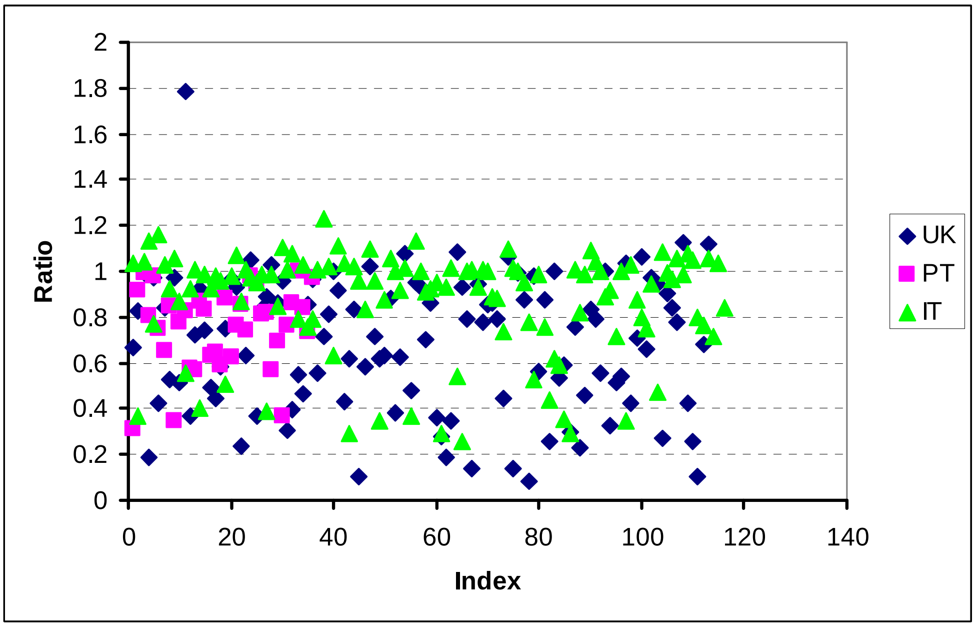


Figure 6. Ratios of the variances of TURN2004 before and after protection, by domain.

	Min	Q1	Q2	Mean	Q3	Max
UK	0.08	0.46	0.72	0.69	0.93	1.78
PT	0.32	0.64	0.79	0.75	0.86	1.00
IT	0.25	0.79	0.96	0.87	1.02	1.23

Table 9. Ratios of the TURN 2004 variance, before and after the perturbation.

4.6 Correlations between RTOT and TURN2004

For each domain, the ratios of the original correlation between TURN 2004 and RTOT to the correlation derived from the perturbed data were computed. The trends of these ratios are illustrated in figure 7 for the UK, PT and IT CIS4 microdata. No particular order war used. In table 10 some descriptive statistics of the ratios of the original correlation between TURN 2004 and RTOT to the correlation derived from the perturbed data are shown.

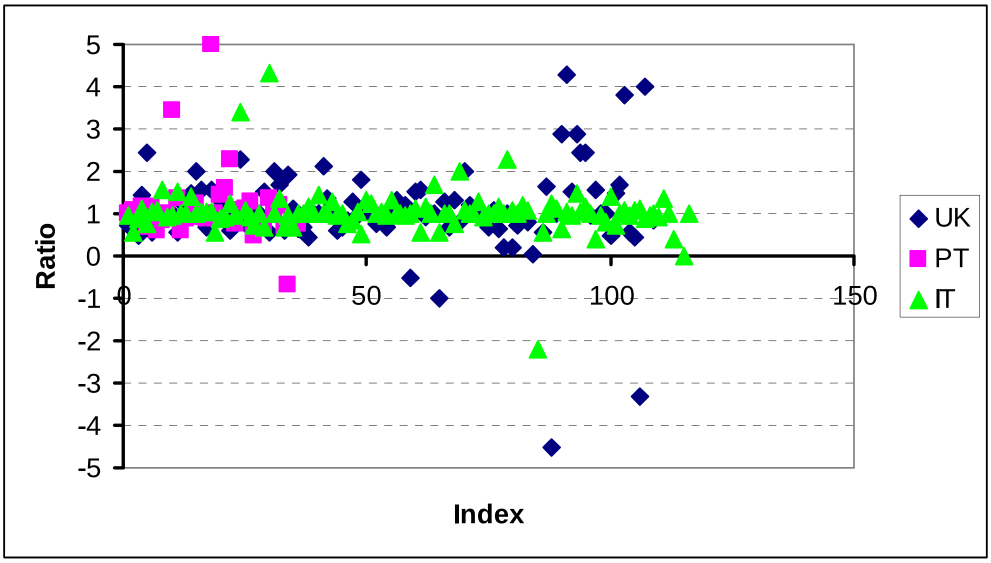


Figure 7. Ratios of the original correlations between TURN2004 and RTOT to the correlations derived from the perturbed data, by domain. (a zoom on the most significant region)

	Min	Q1	Q2	Mean	Q3	Max
UK	-4.50	0.73	1.00	1.29	1.53	9.00
PT	-0.67	0.78	1.02	1.55	1.26	15.00
IT	-6.33	0.93	1.00	1.04	1.10	9.00

Table 10. Ratios of the original correlations between TURN2004 and RTOT to the correlations derived from the perturbed data, by domain. (a zoom on the most significant region)

4.6 Comparisons with Individual Ranking

In figure 8, the correlation between RTOT and TURN 2004 was compared to the individual ranking (IR), see [6]. The correlation between RTOT and TURN 2004 was considered in all cases. The blue circles correspond to the Italian statistical disclosure methodology. The red squares correspond to the individual ranking. The black line is the baseline. The microaggregation parameters were those indicated in section 3. “AllReks” means that the individual ranking was applied irrespective of the innovation attitude of the enterprises. “NACE” means that the IR application domain was the principal economic activity; “NaceEmp” means that the IR application domain was defined by the combinations of NACE categories and size classes; If neither “NACE” nor “NaceEmp” are indicated, the IR was applied irrespective of any stratification.

In figure 9, the variances of TURN 2004 were compared to the individual ranking (IR), see [6]. The blue circles correspond to the Italian statistical disclosure methodology. The red squares correspond to the individual ranking. The black line is the baseline. The microaggregation parameters were those indicated in section 3. “AllReks” means that the individual ranking was applied irrespective of the innovation attitude of the enterprises. “NACE” means that the IR application domain was the principal economic activity; “NaceEmp” means that the IR application domain was defined by the combinations of NACE categories and size classes; If neither “NACE” nor “NaceEmp” are indicated, the IR was applied irrespective of any stratification.

In figure 10, the distributions of RTOT were compared to the individual ranking (IR), see [6]. The blue lines correspond to the Italian statistical disclosure methodology. The red lines correspond to the individual ranking. The green lines correspond to the original distribution. The microaggregation parameters were those indicated in section 3. “AllReks” means that the individual ranking was applied irrespective of the innovation attitude of the enterprises. “NACE” means that the IR application domain was the principal economic activity; “NaceEmp” means that the IR application domain was defined by the combinations of NACE categories and size classes; If neither “NACE” nor “NaceEmp” are indicated, the IR was applied irrespective of any stratification. The graphics for NACE 37, IT CIS4 was not significant, so another NACE category was choose.

In figure 11, the GINI coefficients were compared to the individual ranking (IR), see [6]. The blue circles correspond to the Italian statistical disclosure methodology. The red squares correspond to the individual ranking. The black line is the baseline.

Generally the same effects may be observed on the distributions of the other perturbed variables or their ratios.

UK

PT

IT

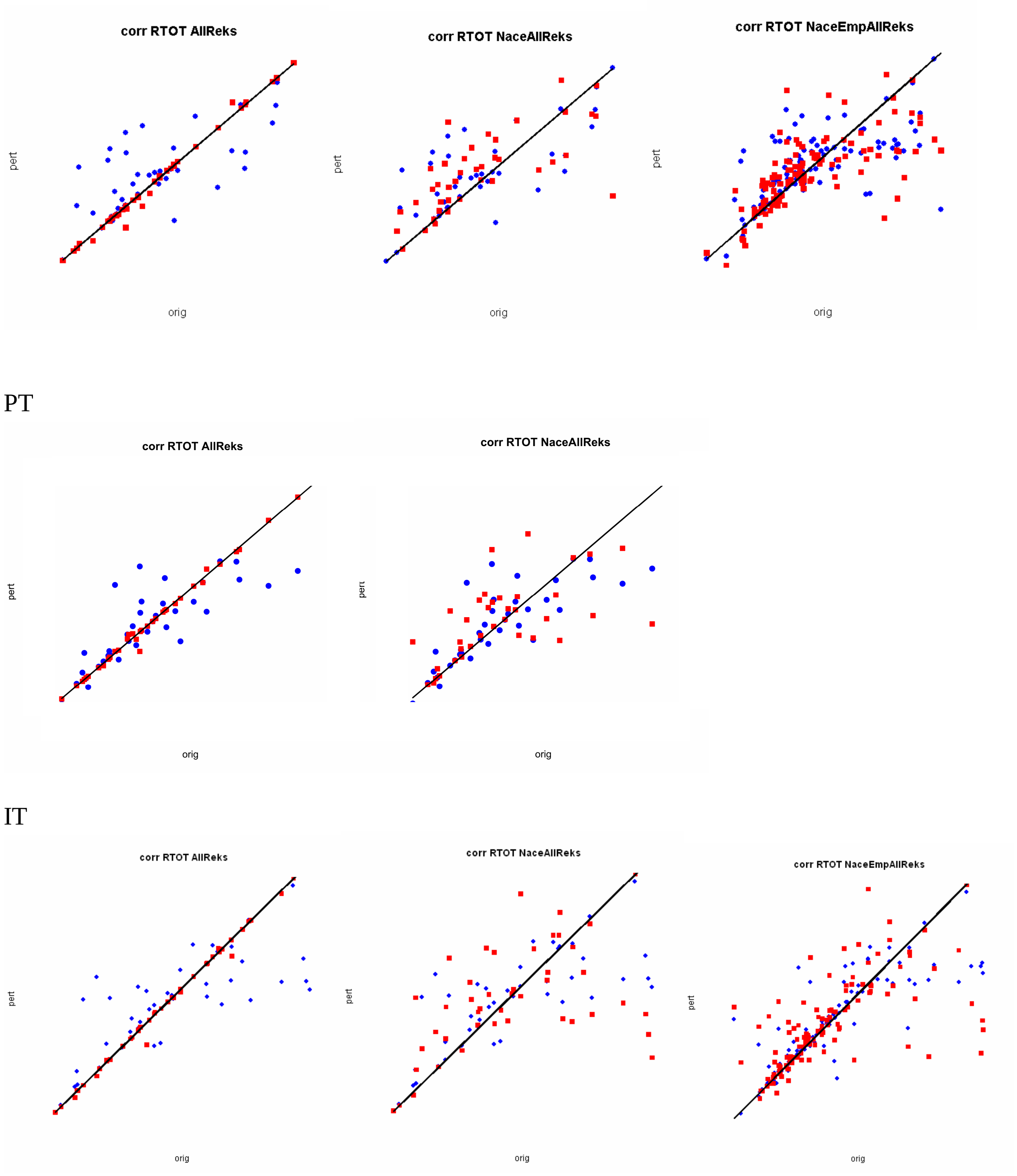
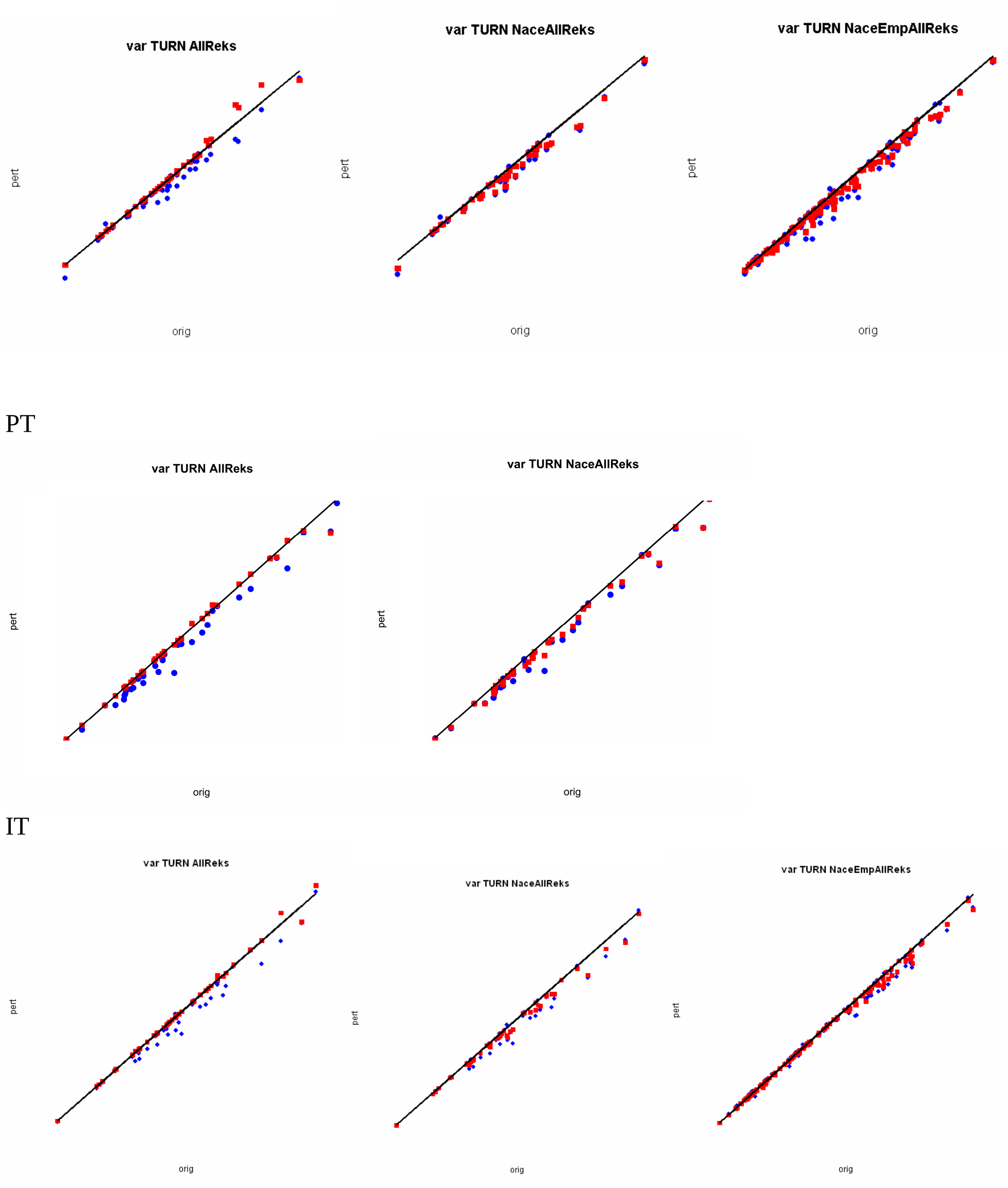


Figure 8. Comparisons with the individual ranking. Correlations between RTOT and TURN 2004.

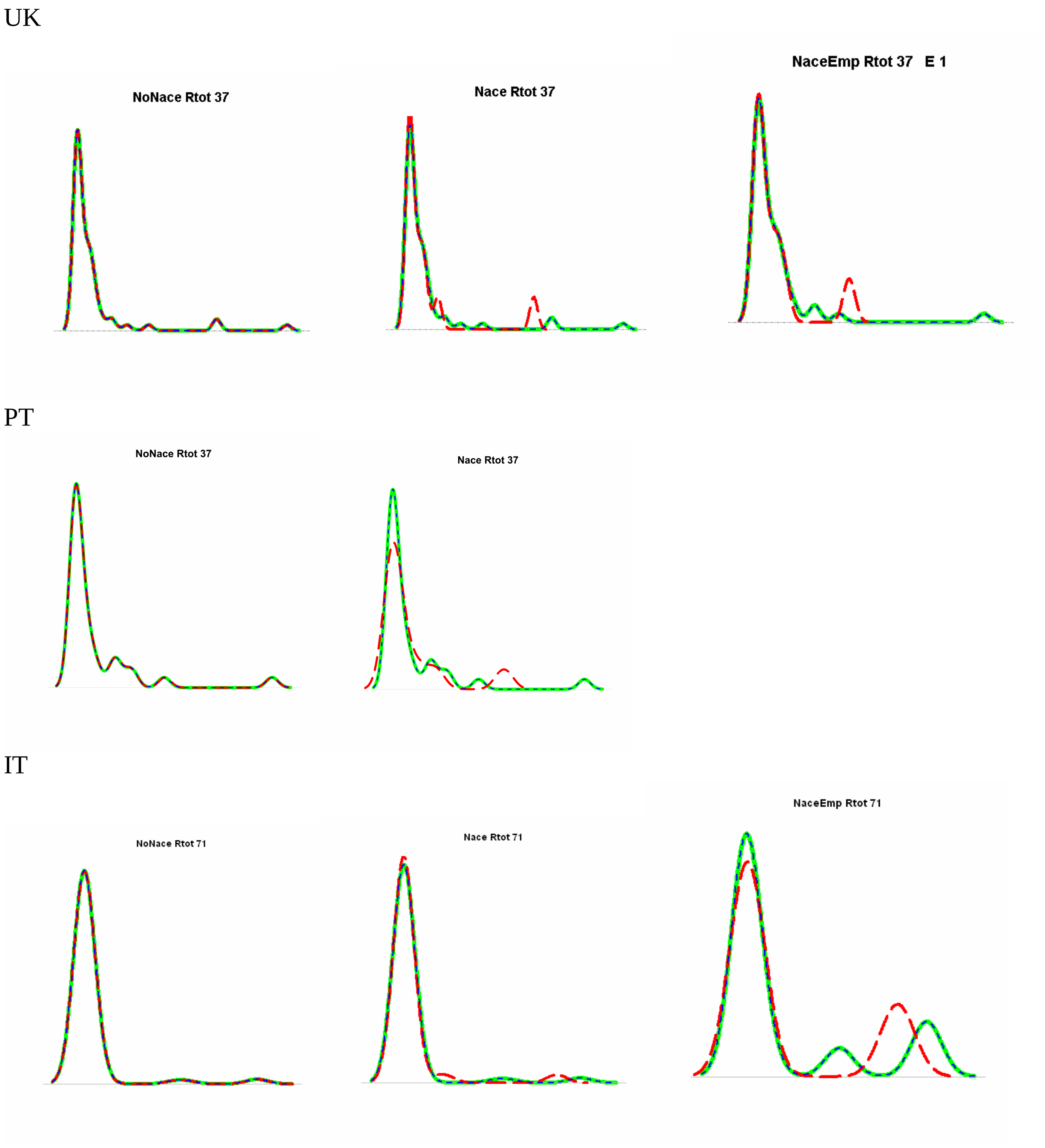
UK



PT

IT

Figure 9. Comparisons with the individual ranking. Variances of TURN 2004.

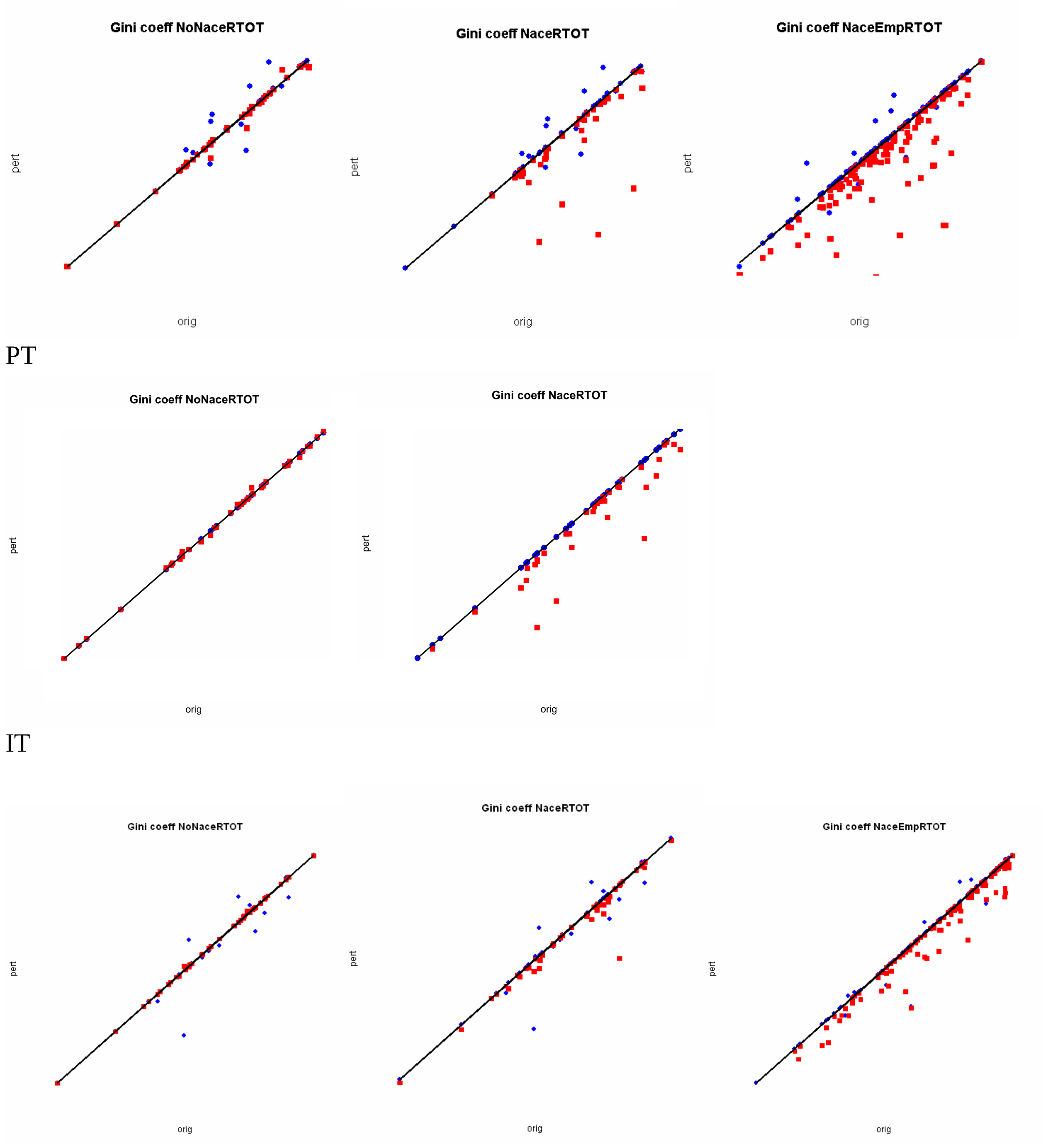


UK

PT

IT

Figure 10. Comparisons with the individual ranking. Distributions of RTOT.


UK

PT

IT

Figure 11. Comparisons with the individual ranking. Gini coefficients of RTOT.

5. Conclusions

The flexibility of the CIS ISTAT statistical disclosure control methodology was tested. Microdata stemming from the British, Portuguese and Italian surveys were used. Different settings were used in order to test the adaptability of the methodology: different stratification domains, different thresholds criteria, different microaggregation parameters, etc. Finally, a comparison with the individual ranking was performed.

In conclusion, the ISTAT disclosure control methodology proved to be easily adapted to different dissemination policies.

6. References

1. Hundepool, A. *et al.* (2008), "Handbook on Statistical Disclosure Control. Case Study A1: The release of microdata for research purposes. Community Innovation Survey 4", available at http://neon.vb.cbs.nl/casc/handbook.htm (deliverable of the ESSnet on SDC 2008-2009)

2. Ichim, D. (2008), "Community Innovation Survey: a Flexible Approach to the Dissemination of Microdata Files for Research", deliverable of the ESSnet on SDC 2008-2009, available at http://neon.vb.cbs.nl/casc/ESSNetDeliv.htm.

3. Ichim, D. (2009), "Disclosure Control of Business Microdata: A Density-Based Approach", International Statistical Review, 77 (2), 196-211.

4. http://www.dius.gov.uk/science/science_and_innovation_analysis/cis.

5. ISTAT, (2008) "Statistiche sull'innovazione delle imprese", Collana *Informazioni*, N. 1.

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

7. Eurostat (2008), "Science, technology and innovation in Europe", available at epp.eurostat.ec.europa.eu/cache/ITY.../KS-EM-08-001-EN.PDF.