Case studies B: Illustrative examples of the application of particular techniques

B.2. Illustrative examples of the application of individual ranking to enterprise microdata

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. In this document only the third stage will be discussed, assuming that the key variables were previously defined and that the disclosure risk of each unit was adequately estimated. Many statistical disclosure limitation methodologies may be used in order to ensure that the confidentiality of respondents cannot be breached. One of the simplest and well-known methods for continuous variables is the individual ranking. The aim of this work is to assess the impact of the individual ranking on both data utility and confidentiality. Three versions of the individual ranking are discussed here: a) independent on any categorical key (stratification) variable, b) depending on some of the categorical key (stratification) variables and c) depending on all categorical key (stratification) variables. It will be empirically shown that an approximate Risk-Utility map of the individual ranking should be similar to the one represented in Figure 1.

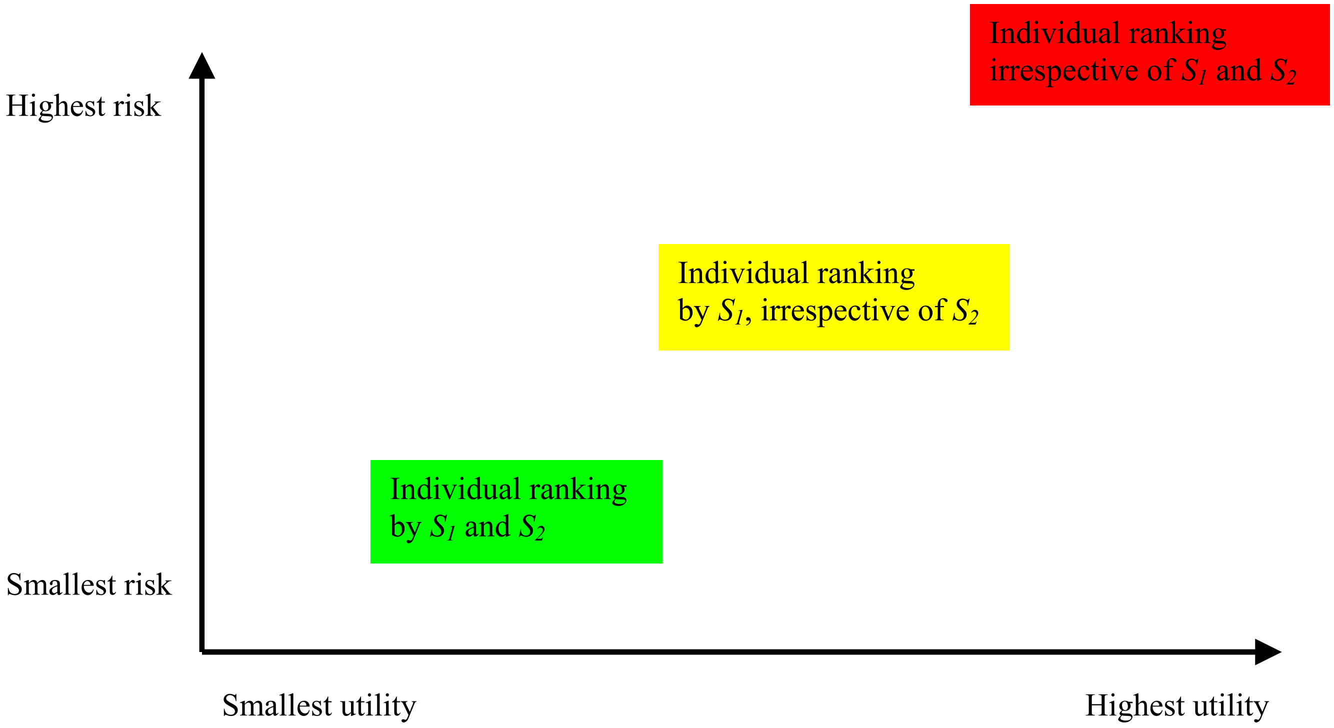
Highest risk
Smallest risk
Smallest utility
Highest utility
Individual ranking irrespective of S_1 and S_2
Individual ranking by S_1, irrespective of S_2
Individual ranking by S_1 and S_2

Figure 1 A Risk-Utility map for different versions of individual ranking.
S_1 and S_2 denote the two categorical key (stratification) variables.

In Section 2 the individual ranking method is briefly illustrated. The dataset used in this simulation is described in Section 3. Sections 4, 5 and 6 present the results of the application of the individual ranking using three versions: a) without stratification, b) partially stratified and c) completely stratified. Very simple statistics[1] were used in order quantify the risk of disclosure and the data utility.

2. Individual ranking

Micro-aggregation is a very well-known perturbation method introduced in Defays (1998). It aims at creating at least k equal units with respect to the values of the continuous key variables. The records subject to a micro-aggregation process are clustered in groups of at least k similar units. A methodology to achieve an optimal clustering is described in Domingo-Ferrer (2002). In a second stage, 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 the removal of the re-identification risk by means of a perturbation. The underline idea is that *all* values are changed so as to prevent from *exact disclosure*. However in microaggregation there is a clear lack of risk definition and assessment. Moreover, the quantity of perturbation induced by micro-

[1] Univariate distributions of *TURN02*, *TURN04* and *TURN02/TURN04* and variances and correlations.

aggregation is not at all related to the risk of disclosure. That's why only the *exact disclosure* might be avoided by micro-aggregation.

When micro-aggregation is applied in real case-studies there are several issues that ought to be discussed.

Firstly, the choice of the parameter k. Intuitively, it should be derived from the dissemination policy of the NSI collecting the data. In the simulations presented in this paper, k was always set equal to 3.

Secondly, the way in which the continuous key variables are micro-aggregated has to be tackled. A possible strategy is to apply a multivariate micro-aggregation, i.e. all continuous key variables are simultaneously micro-aggregated. In practical situations, this strategy is not very much used because it might produce a significant information loss, independently on the way the information is quantified. An alternative would be the individual application of micro-aggregation on each continuous key variable independently from the others. This approach is called individual ranking and it is the only methodology studied in this paper.

The final issue to be addressed when applying the micro-aggregation is the treatment of the categorical key variables. This paper is entirely dedicated only to the assessment of this third issue. A first possibility is to completely ignore the categorical key variables. In this case the continuous key variables could be readily perturbed by micro-aggregation, independently on categorical key variables. In practice, this means that all the units in the sample are subject to a unique micro-aggregation process, i.e. all the units in the microdata file are simultaneously perturbed. An alternative could be the micro-aggregation application with respect to (some of) the categorical key variables. In practice, the dataset is divided in a certain number of sub-datasets, according to the number of domains defined by the combinations of the categorical key variables. Then, in each domain, the micro-aggregation (or the individual ranking) is applied to the continuous key variables, independently on the information contained in the other domains.

By definition, the micro-aggregation, and the individual ranking too, aims at preventing the disclosure by creating (at least) k units with the same value on the continuous key variables, i.e. the micro-aggregation aims at satisfying the k-anonymity principle. It should be anyway noted that the k-anonymity principle is defined with respect to the entire set of key variables, both continuous and discrete. That is, in order for the k-anonymity principle to be satisfied, there should exist, in the disseminated dataset, at least k units having, simultaneously, the same values on all the key variables. It follows that, if a categorical key variable exists, the micro-aggregation should be applied with respect to the domains defined by this categorical key variable. Otherwise, the k-anonymity principle would not be satisfied. Moreover, if the exact disclosure is not the only issue, the problem is even more complicated. For example, when approximate values are sufficient for re-identification, the micro-aggregation could not be sufficient at all. Such situations occur in practice when we have to deal with economic continuous key variables with very skew distributions. In Winkler (2004), it was noted that the individual ranking does not offer any degree of protection, even for higher values of k. Similar results based on a ranking approach for risk assessment of micro-aggregated microdata was reported in Leppälahti (2007).

3. Simulated data

A dataset containing about 20000 records was simulated. Four key variables were generated: two continuous key variables and two categorical key variables. We tried to artificially create a situation similar to the Italian Community Innovation Survey (CIS) microdata file. The two categorical key variables, called *NACE* and *SIZE,* had 43 and 3 categories respectively. The two continuous key variables were called *TURN02* and *TURN04*. These types of variables may be observed in most of the enterprise surveys. Moreover, in real case studies, the "real" *NACE* and *SIZE* variables are structural variables. Consequently, it is hard to assume that the possible intruder would not use this readily available information for the identification of the enterprises (the "real" *NACE* and *SIZE* variables are very easily available in various kinds of public registers; they are also quite accurate). The two continuous key variables, *TURN02* and *TURN04*, intended to simulate the two turnover variables registered in the Italian CIS4 microdata file.

This document will show that, if *NACE* and *SIZE* are correlated with *TURN04* and/or *TURN02*, the micro-aggregation process should be applied with respect to the domains defined by cross-classifying *NACE* and *SIZE*. In practical situations, like in CIS, it is known that the turnover is correlated with the principal economic activity (the "real" *NACE*) and especially with the dimension of an enterprise (the "real" *SIZE*). Moreover, as it was already observed, the principal economic activity and the dimension of an enterprise are two structural variables. It follows that any application of individual ranking not stratified by *NACE* and *SIZE* does implicitly ignore any intruder a-priori knowledge.

In the simulations presented in this document the individual ranking was applied to *TURN02* and *TURN04* in three possible ways: a) irrespective of *NACE* and *SIZE* (Section4), b) with respect to *NACE* domains only (Section 5) and c) with respect to *NACE* and *SIZE* (Section 6). Simulations with other categorical and/or numerical variables are not presented here. Anyway, the used methodology may be easily applied to other categorical and/or numerical variables.

4. Individual ranking applied irrespective of the categorical variables (unconstrained individual ranking)

Individual ranking was firstly applied to each *TURN02* and *TURN04* variable, regardless of the combinations of *NACE* and *SIZE*.

4.1 DATA QUALITY

Figure 2 shows the distribution of original and perturbed values for *TURN04* and *TURN02*. The green solid line represents the original data; the red dashed line represents the perturbed data. A logarithmic transformation was used to improve the graphical presentation. It may be observed that the univariate distributions are almost exactly preserved. Even the tails of the distributions are very well preserved by the individual ranking. The same qualitative result was observed for each stratification domain derived from the categorical variables (*NACE*, *SIZE*).

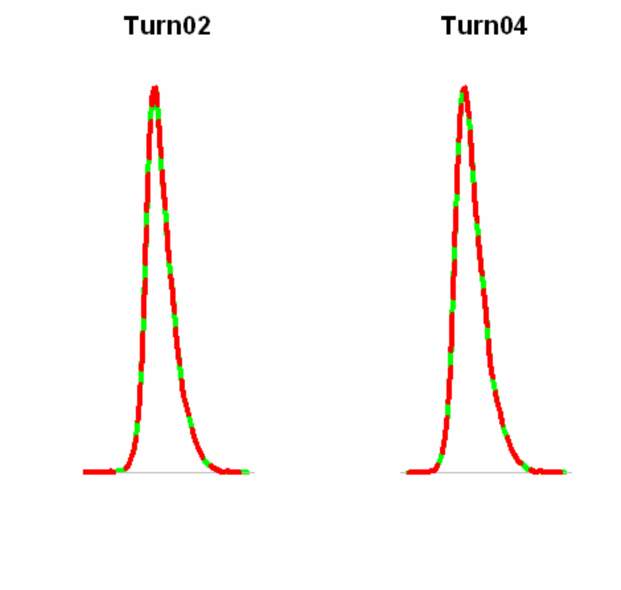

Figure 2 Density plots for the original and perturbed microdata.

To further assess the data quality, the variances and correlations were computed for both original and perturbed data. These statistical indicators were evaluated for each combination of *NACE* and *SIZE*. Figure 3 shows the variance comparison, while figure 4 shows the correlations comparison. On the horizontal axis the original values are presented, while the variances and the correlations computed on the perturbed microdata are presented on the vertical axis.

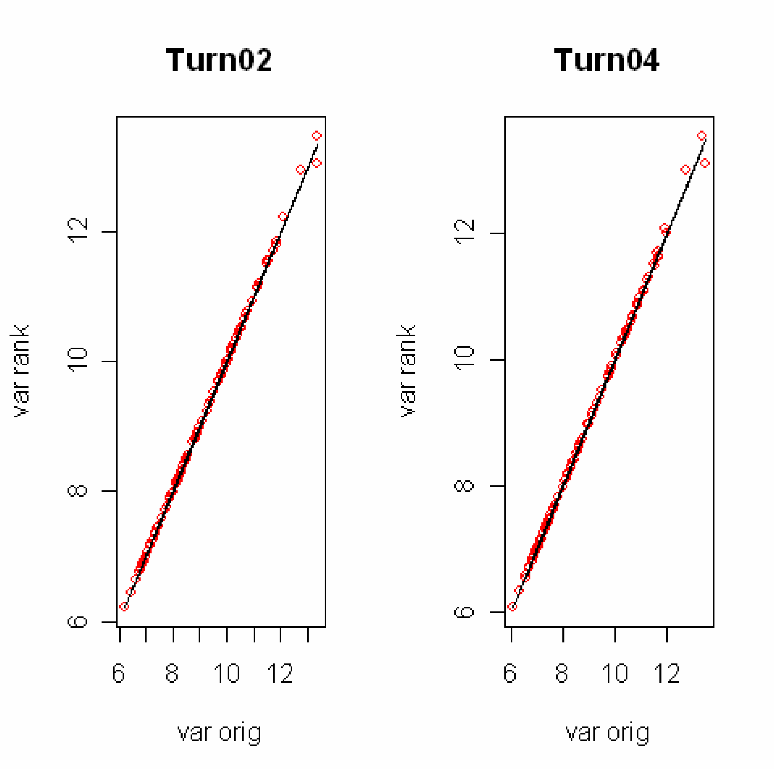

Figure 3 Variance comparison for the original and perturbed microdata, by combinations of *NACE* and *SIZE*.

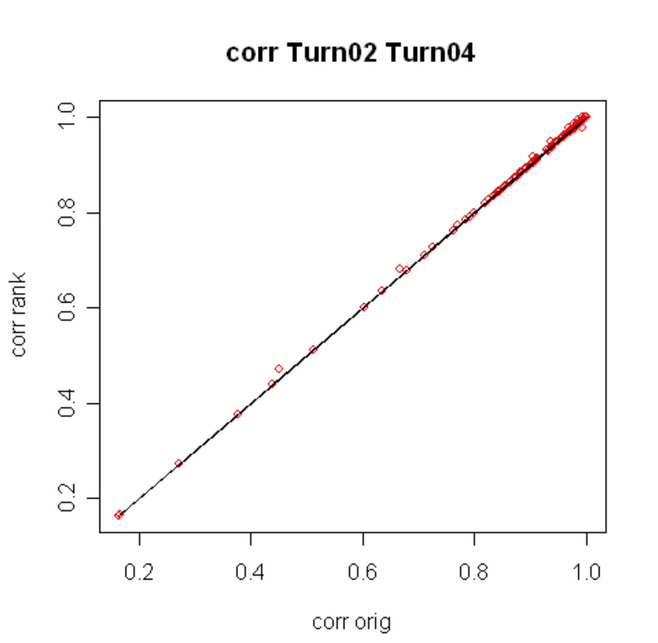

Figure 4 Correlations comparison for the original and perturbed microdata, by combinations of *NACE* and *SIZE*. The graphic shows the coefficient of correlations between *TURN02* and *TURN04*.

Figure 5 shows the distribution of the derived variable *TURN02/TURN04*. The green solid line represents the original values, while the red dashed line represents the perturbed values.

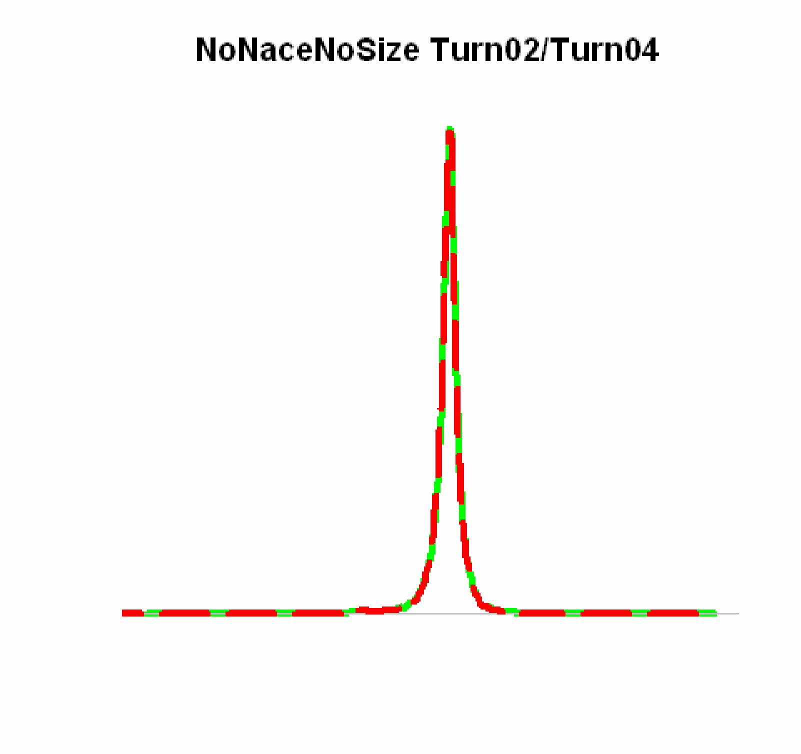

Figure 5 Density of *TURN02/TURN04* for the original and perturbed microdata.

Figure 6 shows a dot plot of the values of *TURN02* and *TURN04*, original and perturbed values.

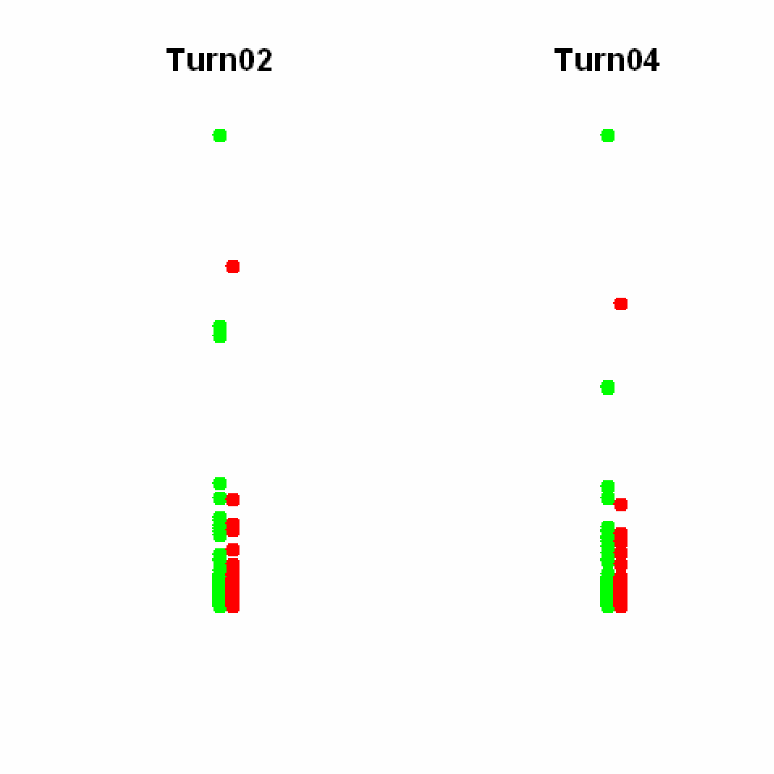

Figure 6 Dot plot of the original (green) and perturbed (red) microdata.

In Figures 3, 4, 5 and 6 only the results for the overall data are presented, but similar effects were observed for each combination of *NACE* and *SIZE*.

It should be clear from the results presented in this section that such an application of the individual ranking, irrespective of the stratification domains, is likely to produce excellent results from the information preservation point of view.

4.2 DATA SAFETY

The data safety was assessed by means of the absolute relative perturbations (percentages) of *TURN02* and *TURN04*. These values were computed as $abs\left(\frac{X_{orig} - X_{perturbed}}{X_{perturbed}}\right) * 100$, where X may be *TURN02* or *TURN04*.

In about 33% of cases the *TURN02* values were unchanged. What happens if a unit at risk of re-identification is among this 33% of units? This problem is not solved by the individual ranking because, as previously observed, the quantity of perturbation induced by the individual ranking is not at all related to the (degree of the) risk of re-identification of a unit.

In 75% of cases, the absolute relative perturbation of *TURN02* was smaller than 0.04%. Moreover, in 99% of cases, the same absolute relative perturbation was smaller than

0.58%. The perturbation of *TURN04* shown the same characteristics. This means that, for 75% of units, the probability of getting insufficient protection is quite high. One might argue that this extremely small variation of the *TURN04* values is typical for the units having very common values of *TURN04*. This is not always the case because it depends on the meaning of "common values". Should we consider them as "common values" with respect to the whole population, to the entire sample or to the sample stratified by some other variable(s)? **Only if *TURN04* (or *TURN02* as well) were the unique key variable, such small perturbations could be regarded as sufficient, being very common.** But in presence of other key variables, either continuous or categorical, the previous statement does not hold anymore. In statistical disclosure control terminology, this means that the underlying k-anonymity principle was not achieved (it should be reminded here that the k-anonymity principle should be satisfied with respect to the entire set of key variables).

In presence of some additional a-priori knowledge, some units (which are enterprises in real situations) could be very easily identified. For example, only a very general knowledge like *NACE* about the phenomena under study could be (intentionally or unintentionally) used for the identification of some enterprises. Figure 7 shows the comparison between the original and the perturbed values. The entire data set is shown on the left side. The *TURN04* values corresponding to a single *NACE* category were selected and compared in the right side of the figure. It is clear that, in the selected *NACE* category, the dominant unit (6 times greater than the second greatest unit) maintains this characteristic (the dominance). It follows that the unique dominant unit in this *NACE* category may be identified with certainty, even if its exact value of the *TURN04* is (slightly) perturbed. This happens because the avoidance of only the exact disclosure is not sufficient in presence of very skew distributions. And it is known that such distributions are more likely to characterise the real business surveys.

From the users point of view, it should be noted that economical analyses are generally performed taking into account the *NACE* classification by sector. This means that, in order to analyse the data, the microdata file is, by default, divided according to the *NACE* categories. It follows that whatever kind of re-identification (e.g. exact, spontaneous or approximate) is more likely to occur inside the domains defined by the *NACE* categories.

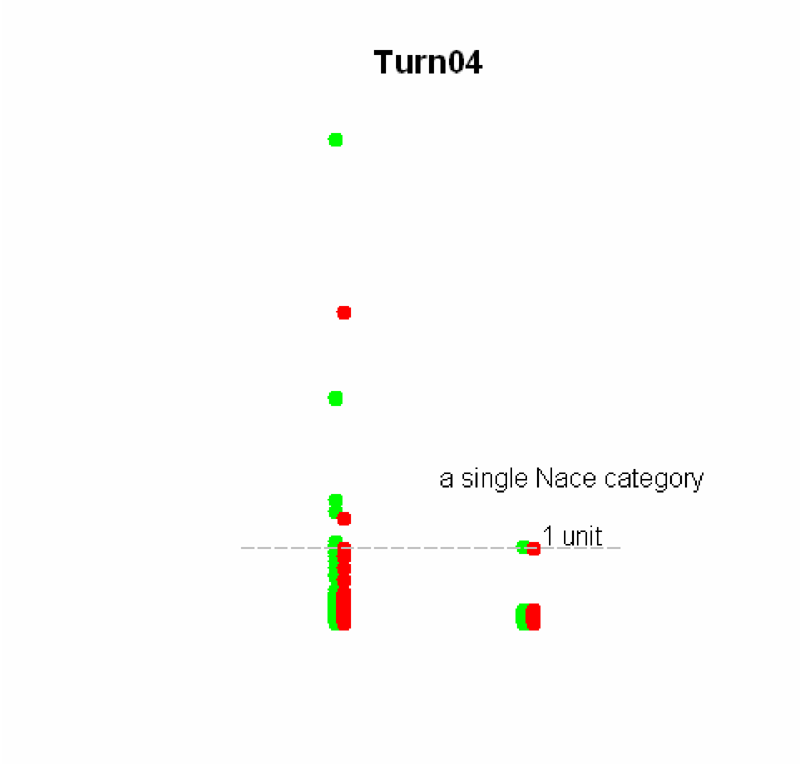

Figure 7 Dot plot of the original (green) and perturbed (red) microdata. One *NACE* category selected.

As it was empirically proved, the individual ranking applied irrespective of the categorical key variables preserved almost perfectly the information content of the microdata file. This is due to the fact that the relative perturbation was smaller than 0.6% in 99% of cases. At the same time, this "perfect information preservation" should warn us about the unchanged risk of re-identification. Indeed, in most of the cases an approximate disclosure is still possible and it has the same degree of difficulty as before the microdata perturbation. Moreover, even the exact disclosure is not avoided in all the cases.

In conclusion, if *TURN04* variable is an identifying (key) variable, it is clear that the applied protection method is not sufficient. This is mainly due to the fact that *NACE*, being a structural variable, should be considered a key variable, too. Consequently, the chosen protection method should be applied with respect to *NACE* too.

5. Individual ranking applied with respect to a single categorical variable (partially constrained individual ranking)

Individual ranking was applied to each *TURN02* and *TURN04* variable, taking into account only the stratification derived from the *NACE* categories. The domains that could be derived from the *SIZE* categories or by cross-classifying *NACE* and *SIZE* were simply ignored.

5.1 DATA QUALITY

Figure 8 shows the distribution of original and perturbed values for *TURN02* and *TURN04*. The green solid line represents the original data; the red dashed line represents the perturbed data. A logarithmic transformation was used to improve the graphical presentation. It may be observed that the univariate distributions are well preserved. As expected, only on the tails some differences between the corresponding distributions may be observed. The same qualitative result was observed for each stratification domain derived from the categorical variables *NACE* and *SIZE*.

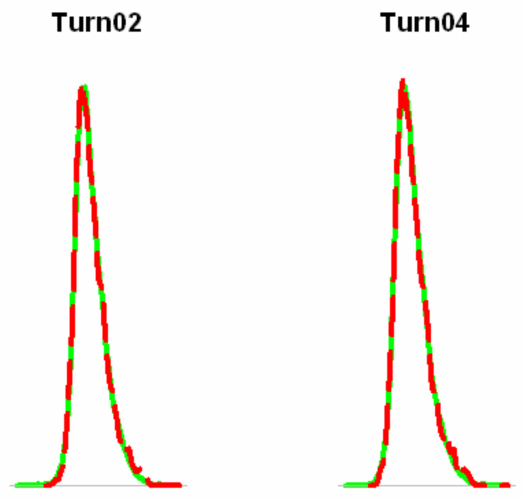

Figure 8 Density plots for the original (green) and perturbed (red) microdata.

To further assess the data quality, the variances and correlations were computed for both original and perturbed data. These statistical indicators were evaluated for each domain derived from the classifications of *NACE* and *SIZE*. Figure 9 shows the variance comparison, while figure 10 shows the correlations comparison. On the horizontal axis the original values are presented, while the variances and the correlations computed on the perturbed microdata are presented on the vertical axis. It should be observed that the variances computed for each domain defined by cross-classifying *NACE* and *SIZE* are increased or decreased. If the variances were computed for the domains defined by *NACE* only (i.e. only by the domains defined by the variable used when applying the individual ranking), the variances would be only decreased, see Defays (1998).

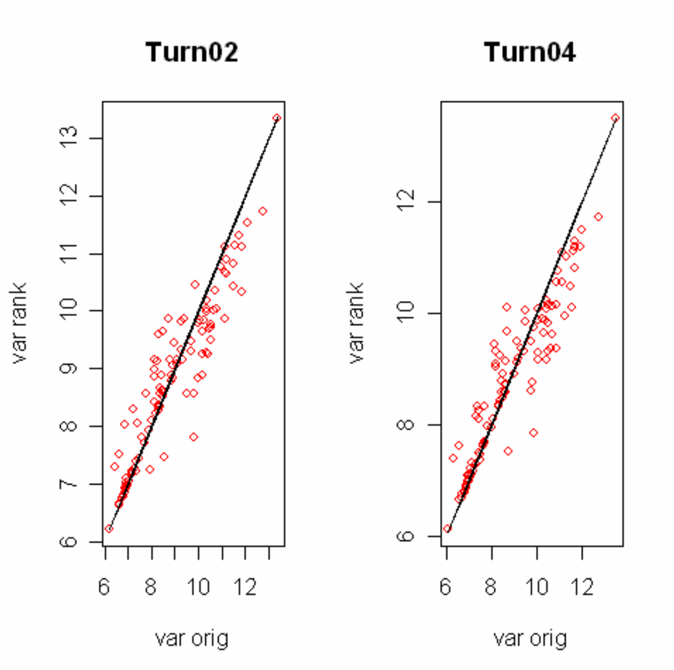

Figure 9 Variance comparison for the original and perturbed microdata, by combinations of *NACE* and *SIZE*.

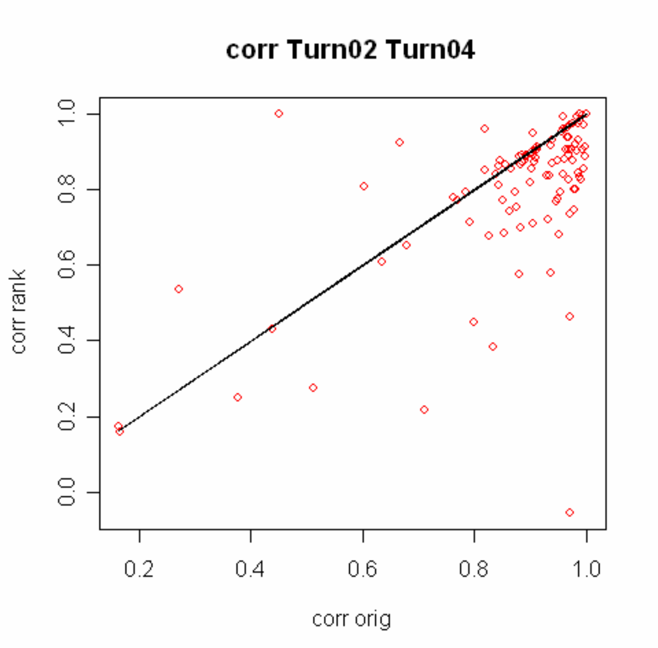

Figure 10 Correlations comparison for the original and perturbed microdata, by combinations of *NACE* and *SIZE*.

Figure 11 shows the distribution of the derived variable *TURN02/TURN04*. The green solid line represents the original values, while the red dashed line represents the perturbed values.

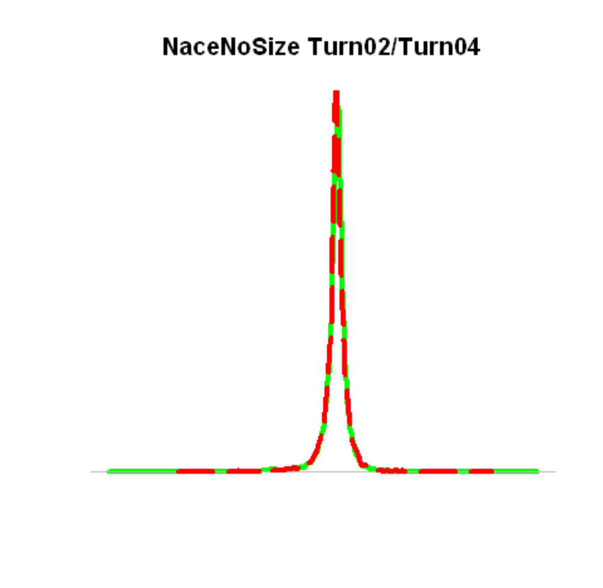

Figure 11 Density of *TURN02/TURN04* for the original and perturbed microdata.

Figure 12 shows a dot plot of the values of *TURN02* and *TURN04*, original and perturbed values. One of the main effects of the application of a stratified individual ranking may be immediately observed. In the central part of the overall distribution of either *TURN02* or *TURN04* there are more than *k* units belonging to different *NACE* strata. Each of these units is obviously averaged with units from the same strata. This is the reason why the units in the central part of the overall distribution are significantly lowered.

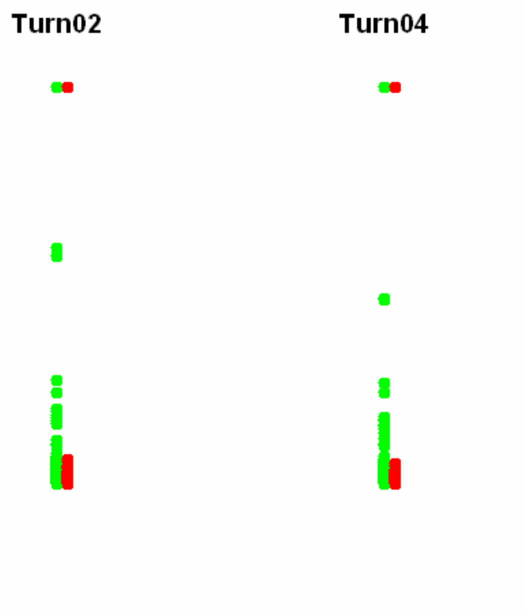

Figure 12 Dot plot of the original (green) and perturbed (red) microdata.

It should be clear from the results presented above in this section that such an application of the individual ranking, irrespective of the domains defined by *SIZE*, is likely to produce good results from the information content point of view.

5.2 DATA SAFETY

The absolute relative perturbations (percentages) of *TURN02* and *TURN04* were computed. These values were computed as $abs\left(\frac{X_{orig} - X_{perturbed}}{X_{perturbed}}\right)*100$, where X may be *TURN02* or *TURN04*.

In only 1.75% of cases the *TURN02* values were unchanged and this is an already a significant improvement with respect to the situation described in the previous section (individual ranking applied irrespective of the structural variables). In 75% of cases, the absolute relative perturbation of *TURN02* was smaller than 7.2%. The increment from 0.04% to 7.2% is another improvement with respect to the unconstrained individual ranking. Anyway, 25% of the units received an absolute relative perturbation smaller than 0.5%. Similar results were obtained for *TURN04*. Even if the confidentiality of the respondents is more protected, the general considerations given in Section 4.2 still hold.

In presence of some additional a-priori knowledge, some enterprises could be very easily identified. For example, only a very general knowledge like *NACE* and *SIZE* about the phenomena under study could be (intentionally or unintentionally) used for the identification of some enterprises. Figure 13 shows the comparison between the original and perturbed values. The overall *TURN04* values were compared in the left side of the figure. In the right side of the figure, the units belonging to a single *NACE* x *SIZE* category were selected and compared. It is clear that the dominant unit (1.5 times greater than the second greatest) maintains this characteristic (the dominance). It follows that the unique dominant unit in this *NACE* x *SIZE* category may be identified with certainty, even if its exact *TURN04* value is not known. This happens because the avoidance of only the exact disclosure is not sufficient in presence of very skew distributions. And it is known that such distributions are more likely to characterise the real business surveys, even if they are partially stratified.

It should be noted that economical analyses are often performed taking into account the *NACE* classification by sector and the enterprise dimension, expressed as the number of employees (variable *SIZE*). This means that, in order to analyse the data, the microdata file is, by default, divided according to the *NACE* x *SIZE* categories. It follows that whatever kind of re-identification (e.g. exact, spontaneous or approximate) is more likely to occur inside the domains defined by the combinations of *NACE* and *SIZE*.

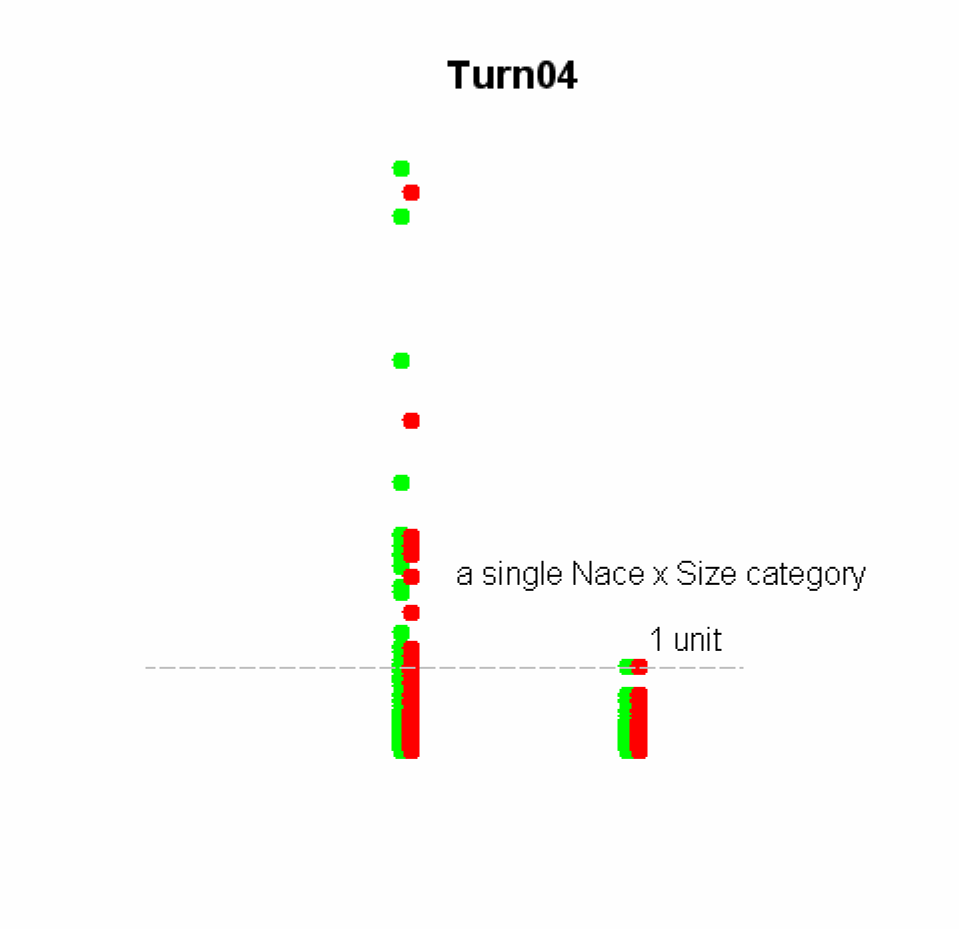

Figure 13 Dot plot of the original (green) and perturbed (red) microdata. One *NACE* x *SIZE* category selected.

Once the unit is identified by means of *TURN04*, *NACE* and *SIZE* values, the sensitive information about the enterprise becomes easily available, by looking at the values of the other variables.

As it was empirically proved, the individual ranking applied with respect to a single categorical key variable preserved very much the information content of the microdata file. This is due to the fact that the relative perturbation was inferior to 7.2% in 75% of cases. At the same time, this "almost perfect information preservation" should warn us about the unchanged risk of re-identification. Indeed, in most of the cases an approximate disclosure is still possible and it has the same degree of difficulty as before the microdata perturbation. Moreover, even the exact disclosure is not avoided in all the cases. It should be admitted that, with respect to the unconstrained individual ranking, the risk of re-identification is reduced and the information loss is not dramatically increased. Anyway, the reduction of the disclosure risk might not be sufficient when the partial constrained individual ranking is applied in real business surveys.

In conclusion, if *TURN04* variable is considered as one identifying (key) variable, it is clear that the applied protection method might not always be sufficient. This is mainly due to the fact that *NACE* and *SIZE*, being structural variables, should be both considered key variables. Consequently, the chosen protection method should be applied with respect to the combinations of *NACE* and *SIZE*.

6. Individual ranking applied with respect to both categorical variables (constrained individual ranking)

Individual ranking was applied to each *TURN02* and *TURN04* variable, taking into account the stratification derived from both *NACE* and *SIZE*.

6.1 DATA QUALITY

Figure 14 shows the distribution of original and perturbed values for *TURN04* and *TURN02*. The green solid line represents the original data; the red dashed line represents the perturbed data. A logarithmic transformation was used to improve the graphical visualisation. It may be observed that the univariate distributions are well preserved. As expected, only on the right tails some differences may be observed. This effect is more evident when the distributions are studied for each stratification domain.

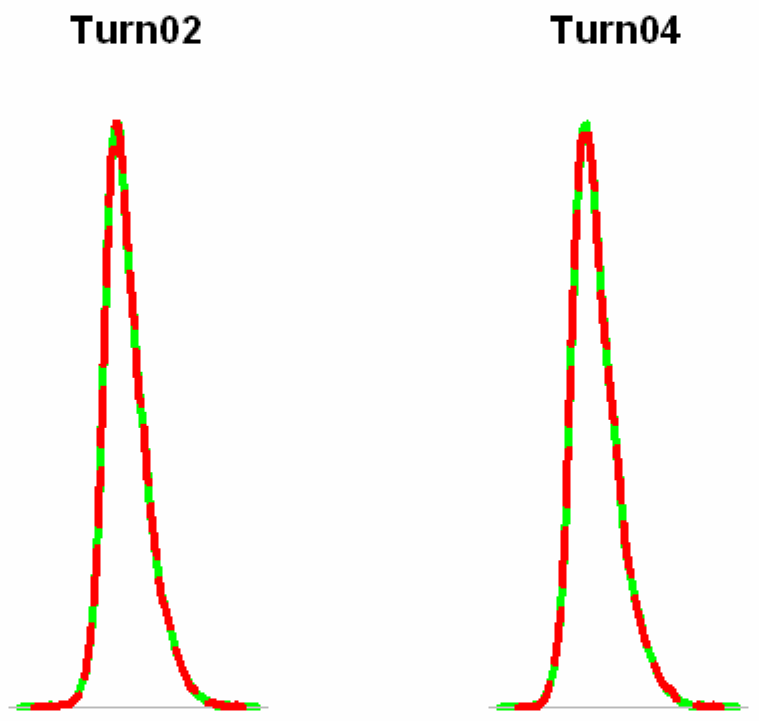

Figure 14 Density plots for the original (green) and perturbed (red) microdata.

To further assess the data quality, the variances and correlations were computed for both original and perturbed data. These statistical indicators were evaluated for each domain derived from the classifications of *NACE* and *SIZE*. Figure 15 shows the variance comparison, while figure 16 shows the correlations comparison. On the horizontal axis the original values are presented, while the variances and the correlations computed on the perturbed microdata are presented on the vertical axis. It should be observed that the variances computed for each domain defined by *NACE* and *SIZE* are always decreased because the individual ranking was applied taking into account the stratification derived from *NACE* and *SIZE* categories themselves.

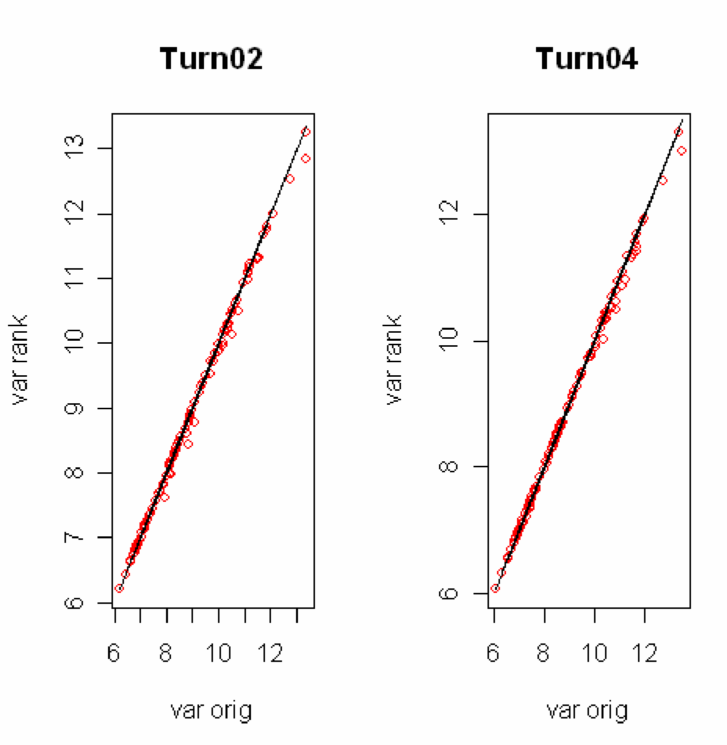

Figure 15 Variance comparison for the original and perturbed microdata, by combinations of *NACE* and *SIZE*.

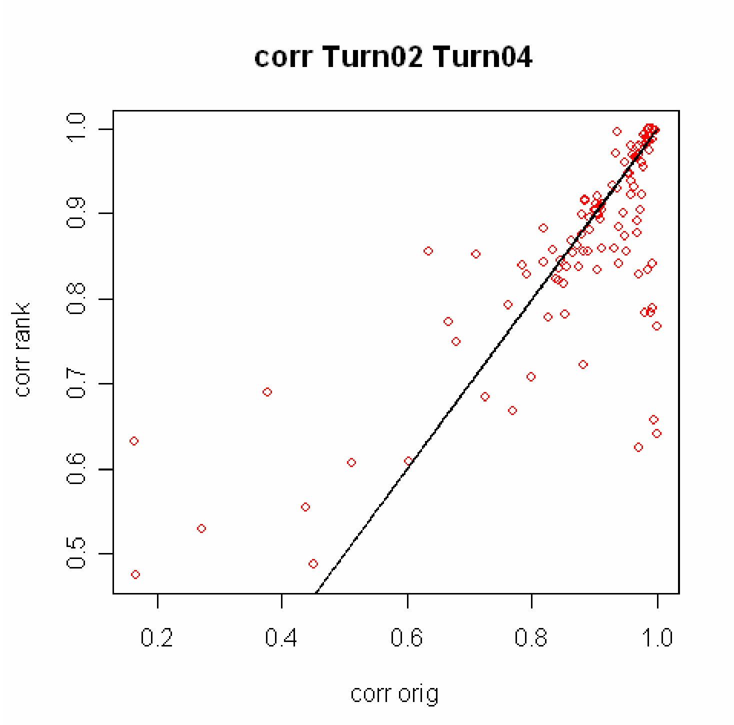

Figure 16 Correlations comparison for the original and perturbed microdata, by combinations of *NACE* and *SIZE*.

Figure 17 shows the distribution of the derived variable *TURN02/TURN04*. The green solid line represents the original values, while the red dashed line represents the perturbed values.

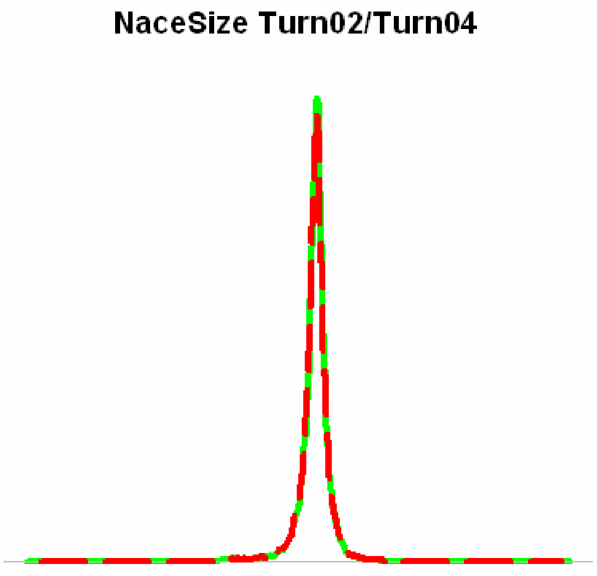

Figure 17 Density of *TURN02/TURN04* for the original and perturbed microdata.

Figure 18 shows a dot plot of the values of *TURN02* and *TURN04*, original and perturbed values. One of the main effects of the application of a completely stratified individual ranking may be immediately observed. On the right tail (upper part) of the overall distribution of either *TURN02* or *TURN04* there are more than *k* units belonging to different *NACE* strata. Each of these units is obviously averaged with units from the same strata. This is the reason why the values of the units on the right tail (upper part) of the overall distribution are significantly lowered.

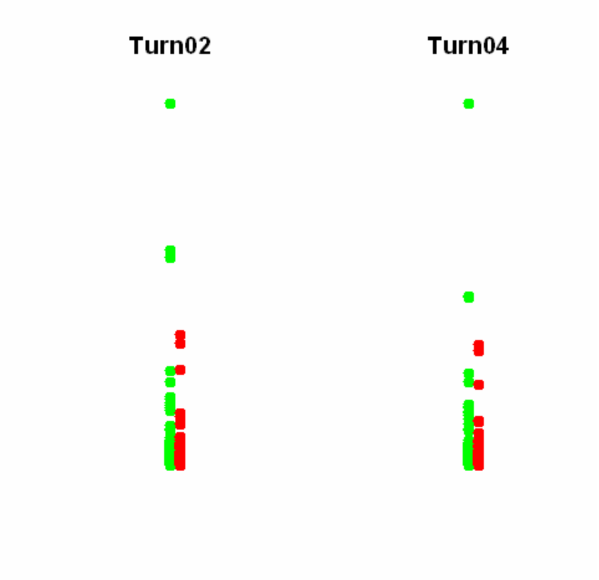

Figure 18 Dot plot of the original (green) and perturbed (red) microdata.

From the point of view of the information preservation, the difference between the three versions of the individual ranking is now obvious. It should be clear from the results presented above in this section that such an application of the individual ranking, with respect to the domains defined by *NACE* and *SIZE*, is likely to produce acceptable results.

6.2 DATA SAFETY

The absolute relative perturbations (percentages) of *TURN02* and *TURN04* were computed. These values were computed as $abs\left(\frac{X_{orig} - X_{perturbed}}{X_{perturbed}}\right) * 100$, where X may be *TURN02* or *TURN04*.

Anyway, in 75% of cases, the absolute relative perturbation of *TURN02* was smaller than 36%. This is a significant improvement with respect to the application of individual ranking irrespective of the stratification domains. Only in 0.09% the original and the perturbed values were equal. Only 1.8 % of the units received an absolute relative perturbation smaller than 0.5%. This percentage (1.8%) could really be associated to the very "common" values, also because the perturbation method (individual ranking) was applied with respect to the combinations of *NACE* and *SIZE*. In other words, more protection was introduced for the units where it was necessary, while the "common " units were not perturbed too much. Similar results were obtained for *TURN04*.

Obviously, this time, the k -anonymity principle is satisfied with respect to the entire set of key variables.

7. Conclusions

Three versions of application of individual ranking were considered, approximating the Risk-Utility map presented in Figure 19. Without being exhaustive, the main features of the individual ranking were underlined. The best results, from a data quality point of view, might be obtained if the individual ranking is applied irrespective of the categorical key variables. At the same time, the safest results might be obtained if the individual ranking is applied for each stratification domain defined by the categorical key variables. In the simulation presented, the categorical variables were also structural ones. If the data protector considers that the structural variables, e.g. *NACE* and *SIZE*, are not identifying, he might choose to apply the individual ranking irrespective of the structural variables. Instead, if the data protector deems that the structural variables are identifying, he should be aware that, in order to guarantee the confidentiality of respondents, a loss in data quality/utility has to be accepted and that a constrained individual ranking should be applied.

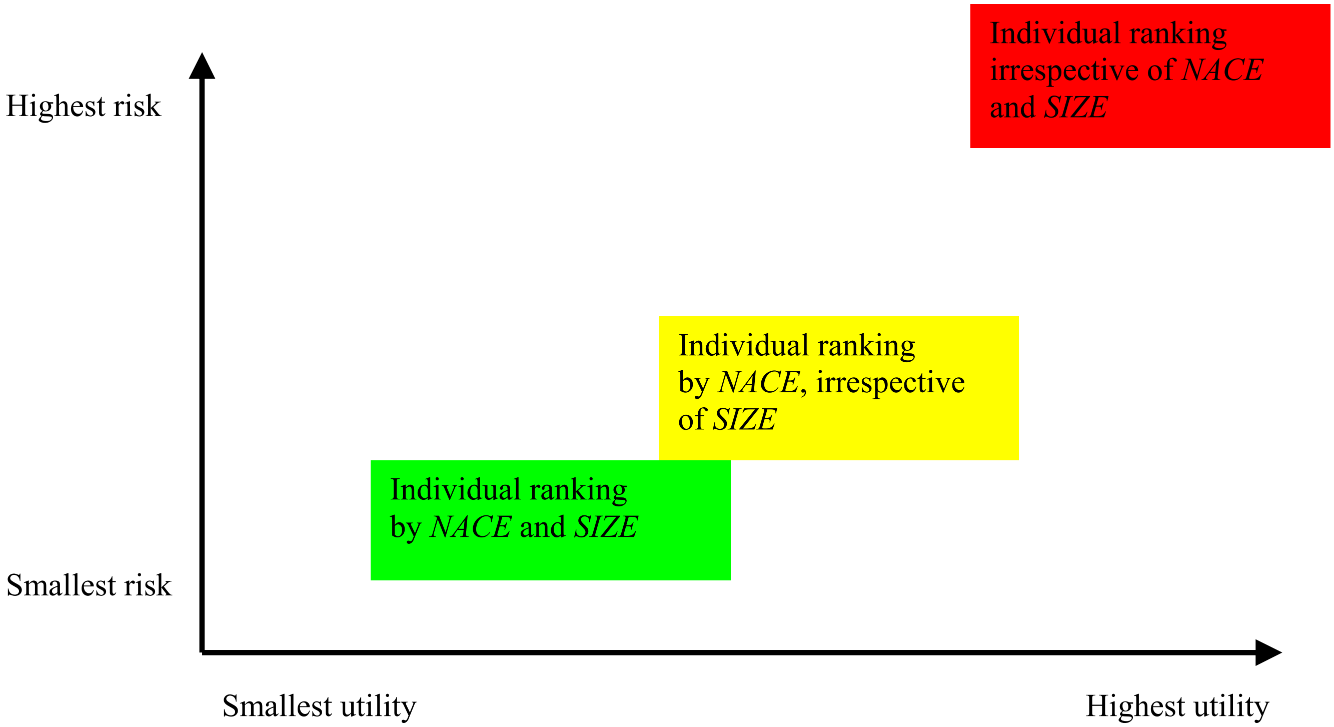
Individual ranking irrespective of NACE and SIZE
Highest risk
Individual ranking by NACE, irrespective of SIZE
Individual ranking by NACE and SIZE
Smallest risk
Smallest utility
Highest utility

Figure 19 A Risk-Utility map for different individual ranking versions.

8.References

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

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002), “Practical Data-Oriented Microaggregation for Statistical Disclosure Control”, IEEE Transactions on Knowledge and Data Engineering, 14 (1), 189-201.

Leppälahti, A. and Teikari, I. (2007) “Problems with micro-data from small countries”, 32nd CEIES Seminar Innovation Indicators - more than technology.

Winkler, W. (2004) “Re-identification methods for masked microdata.” In Privacy in Statistical Databases., Eds. J. Domingo-Ferrer and V. Torra, 216-230.