Chapter 2

Micro-aggregation Techniques

2.1 Introduction

As the role of the European Community grows more complex, the demand for detailed information grows also. The Commission increasingly relies on statistics to allocate funds, promote competitiveness in industry and to evaluate the impact of the current Community programmes. A colossal amount of data is needed some of which are highly sensitive or capable of causing harm to the respondent if disseminated in its original form. Hence, the Member States are not always willing to provide micro-data for confidentiality reasons. On the other hand the European authorities and the analysts must obtain sufficient information in order to carry out all relevant statistical analysis efficiently.

The purpose of this chapter is to give a solution to the conflict between confidentiality and data dissemination. In other words we will try to examine methods, for protecting micro-data, that respect as much as possible the confidentiality aspect and at the same time distort the structure (quality) of the original data set in a minimal way.

The confidentiality of the individual data is not a new problem. Various methods have been proposed and applied. The existing techniques can be classified into two categories. The perturbation ones where the whole data set is provided with a modification and the reduction ones where the data set is exact but certain sensitive values or variables are excluded. In this chapter we focus on the micro-aggregation techniques which belong to the perturbation category and which are mainly used by Eurostat for producing confidential data.

The idea behind the micro-aggregation methods, is the separation of the original data set into homogeneous groups (clusters) of k (defined in the sequel as the threshold) units. The final step of creating a 'confidential' data set is achieved by replacing the individual values with a measure of central tendency of the cluster that they belong. The differences between the micro-aggregation methods basically concern the number of variables that they treat in order to do the clustering, the measure of proximity (sorting criterion) that is used, the decision for the impose of a fixed size or of a minimal size constraint and the aggregation statistic that is used for replacing the individual values at the final step of the micro-aggregation process. Different methods are applied depending on the nature of the variables (metric, ordinal or nominal).

Starting this chapter a brief description of the confidentiality techniques is given. In the sequel we start studying the micro-aggregation techniques according to the nature of the variables that they are applied. In the case of quantitative variables the following methods are examined: the single axis method, the first principal component method, the sum of Z-scores method, the adaptation of Hanani's algorithm and of the Ward's algorithm, the individual ranking method and the weighted moving averages method. In the Hanani's algorithm the n x k-grouping problem in R^m and the 2 x k-grouping problem in R^m are studied and an improved algorithm is proposed. In the individual ranking method, the estimation of loss of information is studied under specific distribution assumptions. In the case of qualitative variables three micro-aggregated methods are described. The snake method applied in the case of multivariate ordinal variables, the similarity of distributions method applied in the case of binary nominal variables and the entropy applied both for ordinal and nominal variables. Some artificial examples are also given for a better illustration of the methods described. In the last part of this chapter criteria for evaluating the performance of the methods are proposed. These are classified into criteria for evaluating the confidentiality of the data and criteria for accessing the quality of the micro-aggregated data. In the first category, criteria like the value of the threshold, the predominance rule and the indicator of data perturbation are described. In the second category, the use of summary statistics the loss of information criteria, and the further processing ability indicator are examined. Last but not least the micro-aggregation procedures in Eurostat, in Spain, in Italy and in Russia are described and some proposals for the adoption of better practices are given.

2.2 Confidentiality Techniques

There are two major types of approaches to solve the problem of the dissemination of confidential micro-data.

- Reduction of the quantity of the provided data: the data are exact but certain sensitive cases or variables are missing or re-coded

- Perturbation of data: the whole data set is provided with a preliminary modification.

The methods belonging to each category can be viewed in the following schematic representation.

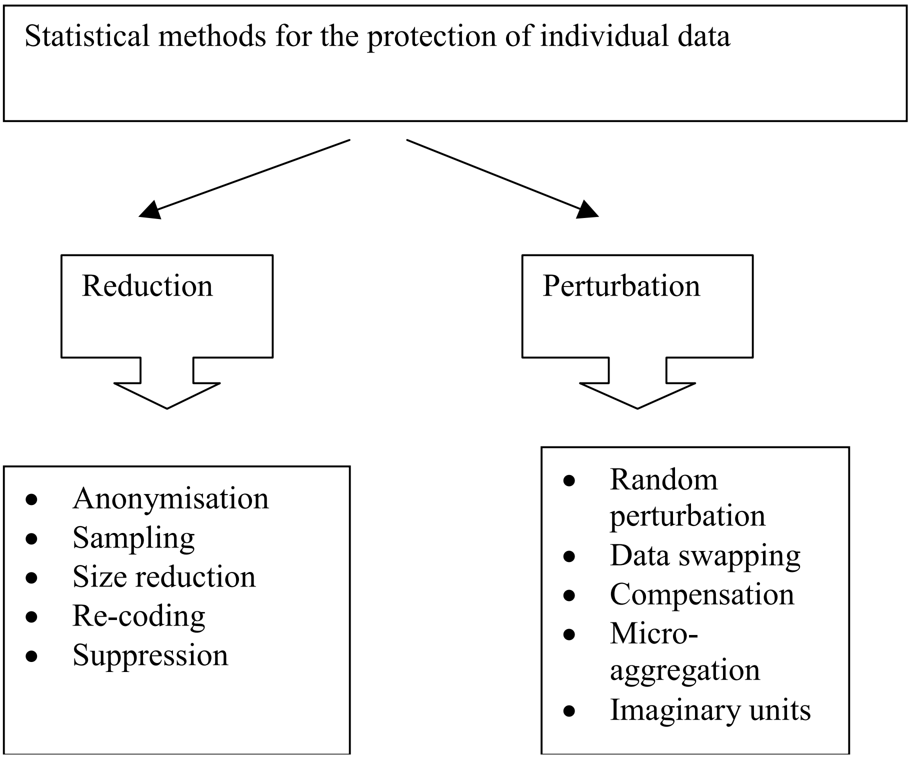

Fig 2.1: Graphical representation of the statistical methods for the protection of individual data.

Concerning the reduction techniques we can say the following. The anonymisation is the fact of extracting from the file the variables that serve as identifiers. Sampling consists of taking a population sub-sample, in order to reduce the risk of disclosure. The size reduction can take either the form of re-coding a variable (i.e. replace variable category by an interval) or the form of the reduction of the number of categories. The re-coding method resembles the previous one but it is used only for extreme categories and only if the frequencies for the first or the last categories are low. The data suppression method is the fact of suppressing a record that contains extreme values. This means that either one replaces these values by missing data or suppresses entirely the vector that shows a single categories combination.

With respect to the methods of modification of the data, the following can be said. The random perturbation method relies on adding a random distributed noise only to the numeric variables of the original data set. It appears that the larger the variance of the noise the higher the disturbance of the data and the lower the quality. The data swapping is based on a matrix transformation, which maintains a number of characteristics. The perturbation by compensation is based on the sorting of observations according the various categories of the nominal variables. Those that have the same combination of categories are aggregated and one preserves the frequency of the macro-cells, the sum and the mean. The imaginary units technique is based on the creation of fictitious observations that follow the same distribution.

In this chapter, we are going to deal with the micro-aggregation techniques. These techniques aim at partitioning the population into groups of k units. Afterwards, the individual values belonging to a group are replaced by the mean of the group or another measure of central tendency (i.e. median, mode, weighted average).

2.3 Micro-aggregation Methods

The micro-aggregation techniques are a group of methods aiming at obtaining data that respect a sufficient confidentiality level and at the same time making possible to carry out relevant analyses. The main idea of the micro-aggregation techniques is influenced by the work of Strudler, Lock, Oh and Scheuren on the Tax model at the US Internal Revenue Service.

In the majority of the micro-aggregation methods, one sets up fixed or variable size groups that are summarised by means of summary statistics. The various methods differ from each other by the method used for making the aggregation of the units (mean, median, mode) and especially the choice of the method for sorting the data. The sorting method has to set up groups as homogeneous as possible, in order not to lose too much information.

The micro-aggregation methods belong to different categories depending on the types of the variables that they are applied (numerical, ordinal or nominal). In this chapter the methods are studied according to this classification.

2.4 Methods Applicable To Quantitative Variables

Three types of techniques are developed. The first takes as a basis a single axis to sort the data. The second type uses classification methods and the third type is based on multivariate methods in which each variable is used successively to sort the data or to apply a moving average.

2.4.1.1 Sorting By A Single Axis

The simplest manner of sorting the observations is to choose a variable and to sort the individuals according to the values of this variable in an ascending or descending order. One groups the sorted units into groups of k units, where k is the threshold[11]. Afterwards, the value of each individual is replaced by a summary statistic such as the mean, the mode, the median, or a weighted average of the group that it belongs. The following artificial example illustrates the single axis technique.

Table 2.1: Artificial example for the application of the single axis method

Company	Number of employees	Turnover	Number of sites
1	12	1000	2
2	21	1500	6
3	39	2000	5
4	40	3000	3
5	42	1000	4
6	47	2000	10
7	53	1500	11
8	58	1500	10
9	60	3000	14

[11] The statistical law which defines the requirements in order a data set to be confidential , usually sets the threshold equal to 3.

Assume that the observations are sorted in ascending order according to the variable "number of employees". The clusters of companies that we obtain under this classification are the following: $\{1,2,3\},\{4,5,6\},\{7,8,9\}$. The mean value of each group for the variable "number of employees" is $\{24,43,57\}$. Similarly, the mean value of each group for the variable "turnover" is $\{1500,2000,2000\}$and for the variable "number of sites" is $\{4.33,5.66,11.66\}$. After the previous computations, the table containing the micro-aggregated data takes the following form.

Table 2.1a: Results from the application of single axis method

Company	Number of employees	Turnover	Number of sites	Groups
1	24	1500	4.33	1
2	24	1500	4.33	1
3	24	1500	4.33	1
4	43	2000	5.66	2
5	43	2000	5.66	2
6	43	2000	5.66	2
7	57	2000	11.66	3
8	57	2000	11.66	3
9	57	2000	11.66	3

The advantages of this approach are obvious for units ranked as homogeneous as possible within each group and heterogeneous without. However, by replacing the original values with micro-aggregates we risk losing the underlying structure of the population and therefore any statistical inferences drawn from the aggregated data could be misleading. In addition to the choice of the ranking basis we are also interested to investigate the degree of information that is retained or equivalently that is lost by using the micro-aggregation techniques. In order to study this we will use an analysis of variance notation. The original information that one variable contains can be presented by the original variance of that variable i.e.

Total Variance of one variable $= \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$. If there is a total of p variables the total variance becomes $\text{Total Variance of p variables } = \dfrac{\sum_{q=1}^{p}\sum_{i=1}^{n}(X_{iq} - \bar{X}_{\bullet q})^2}{n}$. Assume now that we apply the micro-aggregation process and we obtain a modified data set. The inter variance or in other words the variance between the clusters obtained from the micro-aggregation procedure is given by the

Inter Variance of the aggregated data $= \dfrac{\sum_{q=1}^{p}\sum_{j=1}^{C}(\bar{X}_{j,q} - \bar{X}_{\bullet q})^2}{n}$ where $\bar{X}_{j,q}$ is the mean of the jth grouping for variable q and C is the total number of groupings. The variance within the clusters obtained from the micro-aggregation procedure or in other words the intra variance is given by

Intra Variance of the aggregated data $= \sum_{q=1}^{p} \dfrac{\dfrac{\sum_{j=1}^{C}\sum_{i=1}^{k}(X_{i,j,q} - \bar{X}_{\bullet j,q})^2}{k}}{c}$ where $X_{i,j,q}$ is the ith of the jth grouping of the qth variable, $\bar{X}_{\bullet j,q}$ is the mean of the jth grouping of the qth variable, k is the number of observations per grouping and C is the total number of groupings. After the previous definitions, the information contained in the micro-aggregated data is given by the ratio $\dfrac{\text{InterVariance}}{\text{TotalVariance}}$. Equivalently, a measure of the loss of information is given by the Loss of Information Ratio $= \dfrac{\text{IntraVariance}}{\text{TotalVariance}}$. The degree to which the information is retained will also depend greatly on our choice of the variable. If we had chosen the variable “turnover” instead of the variable “number of employees” then we would have obtained different groupings and consequently a different information retention level.

Special attention should be paid in the way that we choose to make the ranking of the observations i.e. in descending or ascending order. The choice of this has a great influence on the results since in ascending order the outlying observations are sometimes in groupings of k units and at other times, if there are less than k observations in the last grouping, they are put in the penultimate grouping to form a larger than k observations grouping. If the ranking is reversed then the smaller observations are grouped with the penultimate grouping. It is clear that the two ranking orders will produce different losses of information. A general rule can be formulated for the sorting strategy in the single axis method depending on the position of the majority of the outliers. More specifically, if the majority of variables have outliers to the far left of the distribution curve then we should rank them in ascending order, and if the majority of variables have outliers to the far right of the distribution curve then we should rank them in descending order.

2.4.1.2 Sorting By The First Principal Component

Any ranking by one single variable will nearly never be optimal. Indeed there is, as already mentioned, no guarantee that proximity on one variable means also proximity on the others. A better solution is to rank the units according to the first principal component.

The principal component technique was first introduced by Karl Pearson, who apparently believed that this method could give solutions to some problems that were of interest to biometricians at that time. The object of the analysis is to take p correlated variables $X_1, X_2, ...X_p$ and find combinations of these that produce indices $Z_1, Z_2, ..., Z_p$ that are uncorrelated and called principal components. The lack of correlation for the principal components is a useful property because this means that the principal components are expressing different dimensions. When applying the principal component analysis there is always a hope that the variances of the most of the principal components, Z_i, will be so low as to be negligible. In that case the variation in the data set can be adequately described by the few linear combinations, Z_i, with non-negligible variances. In other words by using the principal component analysis we achieve a reduction of the dimensionality of the problem from p to usually 1, 2, or 3 dimensions by taking at the same time into account all the variables of the initial data set.

A principal component analysis starts with the data in p dimensions for n individuals. The first principal component is then a linear combination, of the p variables, of the following form

$$Z_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + ... + \alpha_{1p}X_p \text{ subject to}$$
$$\alpha_{11}^2 + \alpha_{12}^2 + ... + \alpha_{1p}^2 = 1$$

In practice the principal component analysis involves finding the eigenvalues and eigenvectors of the covariance or the correlation matrix of the initial p variables. Assuming that the eigenvalues of the covariance or the correlation matrix are denoted by $\lambda_1, \lambda_2, ..., \lambda_p$ $(\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p) \geq 0$ then $\text{var}(Z_i) = \lambda_i$ and $\alpha_{i1}, \alpha_{i2}, ..., \alpha_{ip}$ are the elements of the corresponding ith eigenvector.

Compared to the previous methods this one sorts the observations according to an axis which represents better all the variables and not only one. The use of first principal component gives a reasonable objective solution to the problem. This expectation is consoled by an interesting property on the projection on the first principal component. Suppose that n individual in our population are characterised by p quantitative variables, $X_1, X_2, ... X_p$. The distance between two individuals i and i' of our population at the beginning can be described as $d(i, i') = \sqrt{\sum_{q=1}^{p} (X_{iq} - X_{i'q})^2}$. Once it is projected on another axis in R^p the distance on this new axis that is denoted by δ, is, in general, different from the initial distance d. Consequently it is logical to search for the minimisation of the differences between distances $D^2(d, \delta) = \sum_{i,i'} (d(i,i') - \delta(i,i'))^2$. The axis that minimises this distance is given by the first principal component. As a result, if we aggregate using as a sorting variable the scores on the first principal component, we will obtain a solution which, even if it is not an optimal one , will not be far from the optimal.

After applying the principal component analysis to the standardised data set, the observations are ordered in descending order of their principal component scores, and then aggregated into groups of k units. Then each individual value is replaced by a summary statistic of the group that it belongs. The first principal component micro-aggregation technique is illustrated by using the following artificial example.

Table2.2: Artificial example for the application of the first principal component method

Company	Number of employees	Turnover	Number of sites	1st principal component score
1	12	1000	2	-2.4516
2	21	1500	6	-1.1941
3	39	2000	5	-0.3220
4	40	3000	3	0.0285
5	42	1000	4	-0.9596
6	47	2000	10	0.7402
7	53	1500	11	0.8237
8	58	1500	10	0.8740
9	60	3000	14	2.4611

Sorting the companies in descending order according to their 1st principal component score we obtain the following groups of companies: $\{9,8,7\},\{6,4,3\},\{1,2,5\}$. After the previous classification the means for all variables per group are calculated and the individual values are replaced. The following table includes the micro-aggregated data under the first principal component approach.

Table2.2a: Results from the application of the first principal component method

Company	Number of employees	Turnover	Number of sites	Groups
1	25	1167	4	1
2	25	1167	4	1
3	42	2333	6	2
4	42	2333	6	2
5	25	1167	4	1
6	42	2333	6	2
7	57	2000	11.67	3
8	57	2000	11.67	3
9	57	2000	11.67	3

By definition principal component analysis studies the underlying multidimensional correlation structure of the data, to produce principal components. Methods of ranking based on such an approach, therefore, utilise the correlation matrix. These types of methods are very useful with data that are very highly correlated, and the higher the correlation, the lower will be the loss of information. However, not all data are highly correlated. This does not mean that these data sets do not contain useful information. Under such conditions, it is likely, that these types of procedures fail to retain the required level of information.

Since the principal component analysis depends on a linear combination of all available variables, it should produce in theory better results than the single variable technique.

2.4.1.3 Sum of Z-scores

This approach resembles the first principal component method as far as its purpose is to consider as many quantitative variables of the data vector as possible. The difference between the two techniques is that the sum of Z-scores gives equal importance to the constituent variables whereas for the ranking according to the first principal component the weighting depends on the correlation structure of the underlying data. In other words, the sum of Z-scores method is similar to if we had calculated the first principal component using all the variables and all the weights had been the same.

Since the values of the variables in a data vector can significantly vary in magnitude from other variables in the same data vector, the chosen variables are standardised. A sum of Z-scores is calculated by adding across all variables the standardised values for each observation. This score is then used as a measure of "size" and the observations are ranked according to this score. The last step in the micro-aggregation process is to replace each individual value by a summary statistic of the group that it belongs.

The Z-score of an observation represented by one variable is $Z = \left(\frac{X_i - \bar{X}}{\sigma} \right)$ where X_i is the ith value of variable X, $\bar{X}$ is the mean of variable X and σ is the standard deviation of variable X . This can be modified as follows in the case that we have a total of p variables.

$$Z - \text{Score of an observation when we have a total of p variables} = \sum_{q=1}^{p} \frac{\left(X_{i,q} - \bar{X}_{.q} \right)}{\sigma_q}$$

where $X_{i,q}$ is the ith value of qth variable, $\bar{X}_{.q}$ is the mean of the qth variable, and σ_q is the standard deviation of the qth variable.

The sum of Z-scores method is illustrated using the same artificial example that we have used also for the previous methods.

Table2.3: Artificial example for the application of the sum of Z-scores method

Company	Number of employees	Turnover	Number of sites
1	12	1000	2
2	21	1500	6
3	39	2000	5
4	40	3000	3
5	42	1000	4
6	47	2000	10
7	53	1500	11
8	58	1500	10
9	60	3000	14

Table2.3a: Standardised values

Company	Z-Number of employees	Z-Turnover	Z-Number of sites	Sum of Z-scores	Groups
1	-1.82093	-1.11111	-1.25939	-4.19143	1
2	-1.26233	-0.44444	-0.29475	-2.00143	1
3	-0.14485	0.22222	-0.58359	-0.45854	2
4	-0.08277	1.55556	-1.10182	0.45455	2
5	0.04138	-1.11111	-0.77707	-1.84680	1
6	0.35177	0.22222	0.66989	1.24388	3

7	0.72423	-0.44444	0.91105	1.19084	2
8	1.03462	-0.44444	0.66989	1.26006	3
9	1.15877	1.55556	1.63453	4.34886	3

Sorting the companies in descending order according to their sum of Z-scores we obtain the following groups of companies: $\{9,8,6\},\{7,4,3\},\{1,2,5\}$. After the previous classification the means for all variables per group are calculated and the individual values are replaced. The following table includes the micro-aggregated data under the sum of Z-scores approach.

Table2.3b: Results from the application of the sum of Z-scores method

Company	Number of employees	Turnover	Number of sites
1	25	1167	4
2	25	1167	4
3	44	2167	6.33
4	44	2167	6.33
5	25	1167	4
6	55	2167	11.33
7	44	2167	6.33
8	55	2167	11.33
9	55	2167	11.33

There are also inherent deficiencies in this approach since the decision on selecting a set of variables for composing the Z-scores is made on an arbitrary basis. Consequently, this approach suffers from the same impotencies as the rest of the single axis techniques. Generally there is no guarantee that the axis chosen will provide an optimal clustering, and the methods are very sensitive to the set of variables chosen. This means that the removal of just one variable or of a relatively few key observations can change the results.

2.4.2 Classification Methods

The micro-aggregation methods described in this part are derived from the automatic classification methods. The aim is to partition the data set in the most homogeneous groups. This kind of techniques has to be adapted to the problem of micro-aggregation where groups must be of fixed size or of minimum size. The methods presented here are Hanani's algorithm and the modified Ward's algorithm.

2.4.2.1 Adaptation of Hanani's Algorithm

The aim of the micro-aggregation techniques, which is to build groups of the k most homogeneous elements, approximates that of the techniques of automatic classification, which aim at partitioning a data-set in classes by using a homogeneity criterion. The criterion most usually used is the minimisation of the intra group variance.

The problem is to partition the whole population in clusters in an optimal way. The averages of the "optimal" clusters would define the fictitious data whose elements will be transmitted.

Suppose that the total population Ω is formed by N units, and to each unit ω it corresponds a vector X of m variables, i.e. $X = X(\omega) = (X_1(\omega),...,X_m(\omega)),\ \omega \in \Omega$. The objective is to divide the set of Ω into $n = N/k$ groups of k units $G_1,...,G_n$ in such a way that the n groups are as homogeneous as possible.

To define homogeneity between groups we need a notion of proximity or distance $d(\omega, \omega')$ of the points in Ω, which has to depend the observed variables, and a derived distance $D(G, G')$ between groups of points in Ω. Formally, the quantity that we have to minimise is a kind of within-group variance like $\Psi(G_1,...,G_n) = \sum_{i=1}^{n} \psi(G_i)$ where $\Psi(G_i) = \sum_{\omega \in G_i} P(\omega) D(\omega, G_i)^2$ and P is a probability measure. The quantity $\Psi(G_1,...,G_n)$ has to be minimised under the constraints $P(G_i) = p_i\ (i = 1,...,n)$. Assume P to be a uniform probability measure. In this case, choosing $p_i = k/N\ (i = 1,...,n)$ we have the $n \times k$ grouping problem.

The n x k-Grouping Problem in R^m

The most important case is when $X \in R^m$, the latter endorsed with the Euclidean distance. This assumes the following distances between points in Ω and between groups of points in Ω: $d(\omega,\omega^{`}) = \|X(\omega) - X(\omega')\|$ and $D(G_i, G_j) = \|m_i - m_j\|$ where m_i is the average of X with respect to P over G_i. In that case the quantity that we have to minimise is the following: $\Psi(G_i) = \sum_{\omega \in G_i} P(\omega) D(\omega, G_i)^2 = \sum_{x \in G_i} P(x)\|x - m_i\|^2$.

Thus, using the conditional expectation of X over the field generated by the partition $G = \{G_1,...,G_n\}$, we have that $\Psi(G) = \Psi(G_1,...,G_n) = E(\|X - E(X/G)\|^2)$ where $E(X/G) = \sum_i m_i I_{G_i}$ with I_{G_i} being the indicator of the set G_i. If we assume further that the vector X is centred, i.e. $E(X) = 0$ we take that

$$\Psi(G) = \Psi(G_1,...,G_n) = E(\|X - E(X/G)\|^2) = E(\|X\|^2) - E(\|E(X/G)^2\|).$$

Consequently, the problem of minimisation of $\Psi(G)$ takes the form of maximising the $E(\|E(X/G)\|)$ subject to constraints $P(G_i) = p_i \ (i = 1,...,n)$.

The 2 x k-Grouping Problem in R^m

In the case that $n = 2$ the partition G is generated by $G = \{A, A^c\}$ and the quantity to maximise is given by $\xi(A) = \|E(I_A X)\|^2$ with the constraint $P(A) = p$. In the case that the dimensionality $m = 1$ the quantity that we have to maximise becomes $\xi(A) = (\int I_A X dP)^2$. From the lemma of Neyman-Pearson on the construction of the most powerful test, it follows that the maximising set is of the form $A^* = \{\omega \in \Omega / X(\omega) \geq \lambda\}$ for some constant λ. This in practice means that A^* is composed of the N_p elements of Ω with the largest values of X. In the general case where the dimensionality is greater than one, then exists vector $c \in R^m$ and a constant λ such that the solution which maximises the desired quantity is of the form $A^* = \left\{\omega \in \Omega / \sum_{i=1}^{m} c_i X_i(\omega) \geq \lambda\right\}$ with constraint $P(A^*) = p$.

The Hyperplane in the 2 x k-Grouping Problem in R^m

The most difficult aspect in this maximisation problem is to determine the vector c. It can be proved that the vector $c \in \mathrm{R}^m$ which defines the optimal set $A^* = \left\{ \omega \in \Omega / \sum_{i=1}^{m} c_i X_i(\omega) \geq \lambda \right\}$ satisfies $c = E(I_A X)$. Consequently, the vector c that defines the optimal set A coincides with the mean of X over A with respect to P. Since the overall mean $E(X) = 0$, the optimal hyperplane $< c, x >= \lambda$ separating the two groups $G = \{A, A^c\}$ is perpendicular to the line joining their centres of gravity.

In the general case with a larger number of groups, i.e. $n > 2$, it is obvious that every pair of groups (G_i, G_j) has to be separated by a hyperplane perpendicular to the line joining the centres of gravity.

The Algorithm

The separation of every pair of groups by a hyperplane perpendicular to the line joining the centres of gravity provides an improvement to Hanani's algorithm.

The approach followed by Hanani uses the following idea. Let $G^{(0)}$ be an arbitrary partition of Ω obeying the equal size constraint, i.e. $G^{(0)} = \{G_1^{(0)}, \ldots, G_n^{(0)}\}$ with $|G^{(0)}| = k \; (i = 1, \ldots, n)$. Let $L_1^{(0)}$ be set of n spheres, centred on the centres of gravity of the classes $G^{(0)}$ and having radius calculated to contain exactly k points of Ω i.e. $L_1^{(0)} = \{S_1^{(0)}, \ldots, S_n^{(0)}\}$. Let $L_2^{(0)}$ be the set of centres of gravity $m_i^{(0)}$ of the members of the partition $G^{(0)}$ i.e. $L_2^{(0)} = \{m_1^{(0)}, \ldots, m_n^{(0)}\}$. Define a new partition $G^{(1)}$ associated to $L_1^{(0)}$ and $L_2^{(0)}$ in the following way:

$$x \in G_i^{(1)} \text{ if and only if}$$
$$\begin{cases} \text{either } x \in S_i^{(0)} \text{and} \left\| x - m_i^{(0)} \right\| \leq \left\| x - m_j^{(0)} \right\| \text{for all } j \\ \text{or there is no k such that } x \in S_k^{(0)} \text{ and } \left\| x - m_i^{(0)} \right\| \leq \left\| x - m_j^{(0)} \right\| \text{ for all } j \leq n \end{cases}$$

The new partition will not necessarily obey the equal size constraint. From $G^{(1)}$, two new sets, $L_1^{(1)}$ and $L_2^{(1)}$ are constructed. A new partition $G^{(2)}$ associated to $L_1^{(1)}$ and $L_2^{(1)}$ can be obtained. The algorithm stops at step r if $G^{(r)} = G^{(r+1)}$

This method can be improved when it is realised that the regions that contain the clusters are not spheres but areas limited by hyperplanes. For instance, when $n = 2$ $L_1^{(0)}$ is a partition defined by hyperplanes orthogonal to the line that joins $m_1^{(0)}$ with $m_2^{(0)}$ and which separates Ω in two groups of equal size. For $n > 2$, this method can be generalised by considering each pair (G_i, G_j) defining a hyperplane orthogonal to the line joining the centres of gravity. The algorithm can be described by the following steps:

(1) set r=0

(2) Draw a random partition $G^{(0)}$ in n groups of equal size.

(3) Set i=1, j=2

(4) Find the hyperplane orthogonal to the line joining $m_1^{(r)}$ with $m_2^{(r)}$ and separates $G_i^{(r)}$ and $G_j^{(r)}$ into two groups of equal size.

(5) Replace $G_i^{(r)}$ and $G_j^{(r)}$ by the two groups separated by the hyperplane.

(6) Set j=j+1

(7) If $j \le n$ go to (4) Else go to (8)

(8) Set i=i+1

(9) If $i \le n-1$ set j=j+1 and go to (4). Else set r=r+1 and go to (3)

(10) Stop when for the last $\binom{n}{2}$ iterations the classes have not been modified.

The present method compared to Hanani's algorithm has two advantages. Firstly, the solution will always obey to the equal size constraint. This is very important since in order to be in accordance with the confidentiality rules, clusters with less than k units are not allowed. Secondly, if one reaches at some stage the best solution the algorithm will be terminated. This is not the case in Hanani's algorithm which it will continue searching for a partition with a lower within-group variance breaking the equal size constraint. Finally, in both cases there is a danger of falling into local minima. It is therefore recommended to start the algorithm from different initial partitions.

2.4.2.2 Adaptation of Ward's Method

There are proposals in bibliography that suggest micro-aggregated groups of variable and not of fixed size (Domingo-Matteo 1997). The idea behind this is that groups need not to consist of exactly k data vectors but of at least k data vectors. Methods yielding variable sized groups are a bit more complex than methods obtaining fixed size groups. However, the former ones may take advantage of the distribution of the original data. The following figure illustrates this idea.

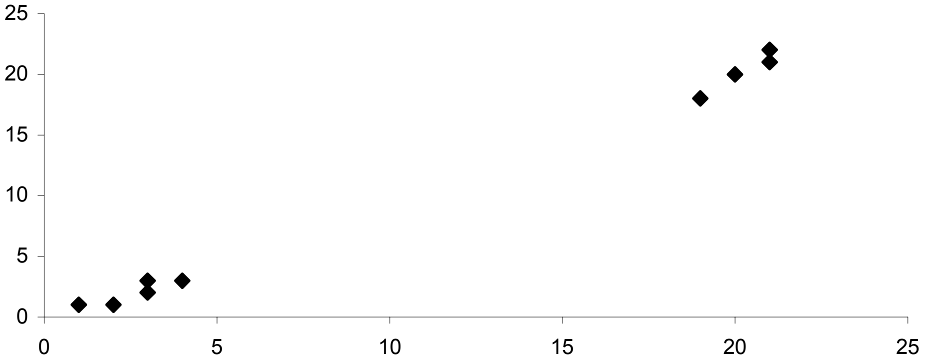

Fig. 2.2: Variable sized groups versus fixed sized groups.

If we want to form fixed sized groups e.g. $k=3$ we will force heterogeneous units belong to the same group. On the contrary if we allow variable sized groups, i.e. a group consisting of the 5 units on the left and another group containing the four units on the right we obtain a more natural partitioning. It is apparent that this type of partitioning, with variable sized groups, gives a smaller loss of information.

Heuristic micro-aggregation methods attempt to minimise the information loss. Since for a given data set the total sum of squares is fixed, each method is aiming at minimising the error sum of squares.

One alternative approach that gives variable sized groups is the modified Ward's algorithm. Generally speaking, the Ward's criterion is one of the various used in the hierarchical clustering. Clustering algorithms are aiming at obtaining groups of observations such that (1) each group is homogeneous with respect to certain characteristics and (2) each group should be different from other groups with respect to the same characteristics. The Ward's criterion forms clusters by maximising the within clusters homogeneity. More specifically clusters are formed at each step such that the resulting cluster solution has the minimum within cluster sum of squares. The mathematical expression for the within cluster sum of squares is given by

$$\mathrm{ESS} = \sum \left(x_i - \bar{x}_i\right)\left(x_i - \bar{x}_i\right)' .$$

The modified Ward's algorithm is a micro-aggregation method for quantitative or for qualitative data where a distance has been defined. In order to describe this algorithm the following definitions are needed.

Definition 1: For a given data set, a k-partition P is any partition of the data set such that each group consists of at least k elements.

Definition 2: For a given data set, a k-partition P is said to be finer than a k-partition P' if any group in P is contained by a group in P'

Definition 3: A k-partition P is said to be minimal if there is no $P' \neq P$ such that P' is finer than P.

Proposition 4: A k-partition P is said to be minimal if it consists of groups with sizes in the range $\{k,2k\}$.

Corollary 5: An optimal solution to the k-partition problem of a set of data exists that is minimal.

The modified Ward's algorithm was first applied in the univariate case. This algorithm is as follows:

Step 1: Form a group with the first (smallest) k elements of the data set and another group the last (largest) k elements of the data set.

Step 2: Use Ward's method until all elements in the data set belong to a group with k or more elements. In the process of forming the groups never join two groups which have size greater or equal to k.

Step 3: Apply the algorithm recursively for each group that in the final partition contains $2k$ or more elements.

By step1 each new recursion starts by splitting the initial data set into at least two groups. Step2 ensures that the groups formed by the smallest units and the groups formed by the largest units are not joined. In this way, at the end of a recursion step, the final k-partition consists of at least two groups and thus this partition is finer than the initial one (consisting of one group). If a group consists of more $2k$ elements then the algorithm is applied recursively to this group in order to obtain smaller groups. After a finite number of steps the maximal group size is less than $2k$.

The only modification that is required in order to define a multivariate modified Ward's algorithm is in step1. Basically, what is needed is a multivariate criterion for specifying the first k data vectors and the last k data vectors in step one. A solution in the problem of defining this multivariate criterion, can be derived by combining the Ward's algorithm with the existing micro-aggregation techniques. More specifically, instead of using the single axis or individual ranking techniques to perform the micro-aggregation, we can use these methods as the sorting criteria of the first step of the algorithm. In other words we can obtain the initial partition by applying the principal component analysis method or the sum of Z-scores method, and then apply the Ward's algorithm to this initial partition.

An additional sorting criterion is the maximum distance one. Under this criterion we have to calculate a distance matrix. Then we have to define the two extreme vectors i.e. those that are the most distant. For each of the extreme vectors take $k-1$ data vectors closest to them. In this way, a group with the first data vectors and another with the last are obtained.

The main drawback of this method is its storage complexity defined here as $S(MWA)$. This happens because Ward's method assumes a distance matrix containing the distances of each pair. Such a distance is symmetrical with zeros in the main diagonal. As a result $S(MWA) = (n-1)n/2$ which means that the storage requirements will increase with increasing sample size. A way of overcoming this difficulty is to partition the initial data set into subsets with smaller sample sizes.

2.4.3 Individual Methods

This kind of methods is no longer aims at partitioning all the data in groups of k elements but treats each variable in a separate way. The methods belong to this category is the individual ranking method and the moving averages method. These techniques are not basically different from the partitioning techniques. The set of variables can be considered as a set of n different variables or as a single multidimensional variable with n components.

2.4.3.1 Individual Ranking Method

Under the methods of single axis all observations are first ranked according to a sorting variable (e.g. first principal component, sum of Z-scores) and then aggregated into groups of k units. The original values are then replaced by a summary statistics. These methods provide satisfactory results if the underlying data are highly correlated, resulting in a high degree of explanation by the single axis. However, there are cases that where the population under study is not highly correlated. Using the single axis techniques we force heterogeneous values to become members of the same group and one of our goals is to have within groups homogeneity and between groups heterogeneity. An alternative to the single axis techniques is the individual ranking method. This method is based on sorting and aggregating the observations according to each variable separately. The following artificial example illustrates the individual ranking method.

Table2.4: Artificial example for the application of the individual ranking method

Company	Number of employees	Turnover	Number of sites
1	12	1000	2
2	21	1500	6
3	39	2000	5
4	40	3000	3
5	42	1000	4
6	47	2000	10
7	53	1500	11
8	58	1500	10
9	60	3000	14

Table2.4a: Observations sorted by the "number of employees " in descending order

Company	Number of employees	Turnover	Number of sites
9	60	3000	14
8	58	1500	10
7	53	1500	11

6	47	2000	10
5	42	1000	4
4	40	3000	3
3	39	2000	5
2	21	1500	6
1	12	1000	2

The observations are ranked according only to the first variable. Each individual value is replaced by the group means. For example the mean for the first group is $(60+58+53)/3=57$. The next step is to sort the data according to the variable "turnover".

Table2.4b: Observations sorted by "turnover" in descending order

Company	Number of employees	Turnover	Number of sites
9	57	3000	14
4	43	3000	3
6	43	2000	10
3	24	2000	5
8	57	1500	10
7	57	1500	11
2	24	1500	6
5	43	1000	4
1	24	1000	2

The observations are ranked according only to the variable "Turnover". Each individual is then replaced by the group means. In this case, the mean of the first group is for example $(3000+3000+2000)/3=2667$. The final step in the present example is to sort the observations according to the variable "number of sites".

Table2.4c: Observations sorted by "number of sites" in descending order

Company	Number of employees	Turnover	Number of sites
9	57	2667	14
7	57	1667	11
6	43	2667	10
8	57	1667	10
2	24	1667	6
3	24	1667	5
5	43	1667	4
4	43	2667	3
1	24	1667	2

The observations are ranked according only to the variable "number of sites". Each individual is then replaced by the group means. In this case, the mean of the first group is for example $(14+11+10)/3=11.67$. The final table containing the micro-aggregated data according to the individual ranking technique has as follows:

Table2.4d: Results from the application of the individual ranking method

Company	Number of employees	Turnover	Number of sites
1	24	1167	3
2	24	1167	7
3	24	1167	7
4	43	2667	3
5	43	1167	3
6	43	2667	7
7	57	1167	11.66
8	57	1167	11.66
9	57	2667	11.66

This method has a number of advantages if it is compared to ranking according to a single variable or a combination of variables. It does not depend on the correlation structure as the first principal component method does and it produces more homogeneous groupings by treating one variable at each time.

Estimation of Variance Loss Due to Micro-aggregation By The Individual Ranking Method

There are theoretical developments that make possible to measure the variance loss due to micro-aggregation by the individual ranking method.

Let $x_1, x_2, ..., x_n$ denote the data relating to the original sample and y_i the data related to the sorted sample. Let I_j be all the index numbers of group j i.e. we have $\bigcup_{1 \le j \le p} I_j = \{1, ..., n\}$ where p is the number of groups each one consisting of three units.

Let $\bar{y}_j$ be the mean of the sample in group j i.e. $\bar{y}_j = \frac{1}{3}\sum_{y_i \in j} y_i$. Using ANOVA notation we have that

$$\sum \left(y_i - \bar{y}_j \right)^2 = 3\sum_{j=1}^{p}\left(\bar{y}_j - \bar{y} \right)^2 + \sum_{j=1}^{p}\sum_{i \in I_j}\left(y_i - \bar{y}_j \right)^2 .$$

$$\text{Total sum of squares} = \text{sum of factorial squares} + \text{sum of residual squares}$$

The y_i's in this equation are order statistics, with $y_i = X_{(i)}$. We assume that the X_i's are independent and identically distributed according to a fully specified distribution. Rewriting the expected variance loss in order statistics notation we have

$$\text{Expected variance loss} = \frac{1}{n-1}E(SRS) = \frac{1}{n-1}E\left(\sum_{j=1}^{p}\sum_{i \in I_j}\left(X_{(i)} - \frac{\sum_{i \in I_j} X_{(i)}}{3} \right)^2 \right)$$

$$= \frac{2}{n-1}\sum_{j=1}^{p} E\left(\frac{1}{2}\sum_{i \in I_j}\left(X_{(i)} - \frac{\sum_{i \in I_j} X_{(i)}}{3} \right)^2 = \frac{2}{n-1}\sum_{j=1}^{p} E(\text{variance of group j}) \right)$$

Evaluating the expected variance loss in each group is sufficient to estimate the total variance due to micro-aggregation. Two methods are used for estimating analytically the expected variance of a group. The moments approach and the spacings approach.

Under the moments approach we have to calculate the expected variances of the different groups that consist of successive order statistics. Consequently, we have to estimate the following expression: $\mathrm{E(variance\,of\ group\ j)} = \mathrm{E}\left[\mathrm{var}\left\{X_{(k)}, X_{(k+1)}, X_{(k+2)}\right\}\right]$ where $k = 3*(j-1)+1$. We can show that

$$\mathrm{E}\left[\mathrm{var}\left\{X_{(k)}, X_{(k+1)}, X_{(k+2)}\right\}\right] = \frac{1}{3}\sum_{i=k}^{k+2}\mathrm{E}(X_{(i)})^2 - \frac{1}{3}\sum_{i<m}\mathrm{E}\left[(X_{(i)})(X_{(m)})\right]$$

The knowledge of $\mathrm{E}(X_{(k)}),^2\ \mathrm{E}(X_{(k)}X_{(k+1)}), \mathrm{E}(X_{(k)}X_{(k+2)})$ is sufficient to estimate the expected variance of a given group.

Under the spacings approach the expected variance of a group can be written in the following form $\mathrm{E}\left[\mathrm{var}\left\{X_{(k)}, X_{(k+1)}, X_{(k+2)}\right\}\right] = \frac{1}{3}\left(\mathrm{E}(D_k)^2 + \mathrm{E}(D_k D_{k+1})^2 + \mathrm{E}(D_{k+1})^2\right)$ with $D_k = X_{(k+1)} - X_{(k)}$.

For further developments we have to make assumptions about the distribution from which the sample comes. Assume that the sample comes from a uniform distribution in the interval [0,1]. Under the moments approach we have the following results:

$$\mathrm{E}(X_{(k)})^2 = \frac{k(k+1)}{(n+2)(n+1)}, \mathrm{E}(X_{(k)}X_{(k+1)}) = \frac{k(k+1)}{(n+2)(n+1)},$$
$$\mathrm{E}(X_{(k)}X_{(k+2)}) = \frac{k(k+3)}{(n+2)(n+1)}$$

Applying algebraic computations we find that the expected variance of a group is given by $\mathrm{E}\left[\mathrm{var}\left\{X_{(k)}, X_{(k+1)}, X_{(k+2)}\right\}\right] = \frac{5}{3(n+2)(n+1)}$.

Under the spacings approach we derive the same results by using the definition of the D_k quantity. More specifically,

$$\mathrm{E}(D_k)^2 = \frac{2}{(n+2)(n+1)}, \mathrm{E}(D_k X_{k+1}) = \frac{1}{(n+2)(n+1)}$$

$$\mathrm{E}\left[\mathrm{var}\left\{X_{(k)}, X_{(k+1)}, X_{(k+2)}\right\}\right] = \frac{5}{3(n+2)(n+1)}.$$

The expected total variance loss can be calculated as follows:

$$\text{Expected variance loss} = \frac{1}{n-1}E(SRS) = \frac{2}{n-1}\sum_{j=1}^{p} E(\text{variance of group j}) =$$

$$= \frac{2}{n-1}p\frac{5}{3(n+2)(n+1)} = \frac{10p}{3(n-1)(n+2)(n+1)}$$

This final result shows that the expected total variance loss is going to decrease with increasing sample size. In bibliography, similar results have been derived also for the exponential and the normal distributions. The interest of these results is primary to examine the impact of the micro-aggregation procedure on the variance loss and further to identify sample sizes for which aggregation is inappropriate.

2.4.3.2 Weighted Moving Average Method

This method requires the choice of weights i.e. a triplet (a, b, c) if the units that we wish to include in each group are three. The observations are first sorted according to each variable and then are replaced by new values consisted of a% of the previous observation, b% of the present observation and c% of the proceeding observation. The use of weights allows to control the emphasis that we want to place on the original data values and also allows us to obtain a data set that does not have an aggregate repeated k times in each group. Special attention should be paid to the first and the last observation. Since the first observation does not have any precedent value and the last any following, the first observation takes the value of the second and the last the value of the penultimate observation. The following artificial example illustrates this method.

Table2.5: Artificial example for the application of the weighting moving averages method

Company	Number of employees	Turnover	Number of sites
1	12	1000	2
2	21	1500	6
3	39	2000	5
4	40	3000	3
5	42	1000	4
6	47	2000	10
7	53	1500	11
8	58	1500	10
9	60	3000	14

The observations are first ranked according to the variable “number of employees”.

Table2.5a: Observations sorted by the “number of employees “ in descending order

Company	Number of employees	Turnover	Number of sites
9	60	3000	14
8	58	1500	10
7	53	1500	11
6	47	2000	10
5	42	1000	4
4	40	3000	3
3	39	2000	5
2	21	1500	6
1	12	1000	2

The values are replaced by moving averages with weights given by $(\mathrm{a} = 0.25, \mathrm{b} = 0.5, \mathrm{c} = 0.25)$. Consequently, in the above example the second value is replaced by $0.25 * 60 + 0.5 * 58 + 0.25 * 53 = 57$, the third value by $0.25 * 58 + 0.5 * 53 + 0.25 * 47 = 52.75$ and so on until the ultimate observation. The next step is to sort the data according to the variable "turnover" and to calculate moving averages according to this sorting.

Table2.5b: Observations sorted by "turnover" in descending order

Company	Number of employees	Turnover	Number of sites
9	57.25	3000	14
4	40.25	3000	3
6	47.25	2000	10
3	34.75	2000	5
8	57.25	1500	10
7	52.75	1500	11
2	23.25	1500	6
5	42.75	1000	4
1	23.25	1000	2

Replace the individual values by moving averages of the variable "turnover" i.e. the second value (i.e. fourth company) for example is replaced by $0.25*3000+0.5*3000+0.25*2000=2750$. The final step is to sort the companies according to the variable "number of sites".

Table2.5c: Observations sorted by "number of sites" in descending order

Company	Number of employees	Turnover	Number of sites
9	57.25	2750	14
7	52.75	1500	11
6	57.25	2167	10
8	47.25	2250	10
2	23.25	1375	6
3	34.75	1875	5
5	42.75	1125	4
4	40.25	2750	3
1	23.25	1125	2

In this case, the second observation (i.e. company 7) for example is replaced by $0.25*14+0.5*11+0.25*10=11.5$.

The final table containing the micro-aggregated data according to the weighted moving average technique has as follows:

Table2.5d: Results from the application of the weighting moving averages method

Company	Number of employees	Turnover	Number of sites
1	23.25	1125	3
2	23.25	1375	6.75
3	34.75	1875	5
4	40.25	2750	3
5	42.75	1125	4
6	47.25	2250	9
7	52.75	1500	11.5
8	57.25	2167	10.25
9	57.25	2750	11.5

Under this method we expect an increase in the loss of information resulting from the two missing weighted moving averages at the beginning and at the end. The replacement of these observations by using repetition of existed values maybe be far from the original data set. This is more likely at the tail of the distribution where extreme values occur.

2.5 Methods Applicable to Qualitative Variables

In practice it is rare to find a data set consisting only of metric variables. More common is to find all three types of variables, metric, ordinal and nominal ones coexisting in a single set of data. In the sequel we examine micro-aggregation techniques applied to qualitative variables. Such techniques are applied only to ordinal variables or only to nominal variables or simultaneously to ordinal and nominal variables under suitable definitions.

2.5.1 The Method of Snake

This technique concerns the multidimensional ordinal variables. Assume that we have an ordinal variable consisting of five levels and two (in our case) or more variables, and construct a 5×5 matrix (see fig.2). In case that you have more than two variables separate the variables in segments of two. The sorting of the observations is done on a basis of a starting point, $\{1,1\}$ in the present example, and by taking a relatively arbitrary route $\{1,1\},...\{1,5\},...,\{5,5\}$. More specifically, the snake defines a route on which the different values of the two dimensional space can be ordered.

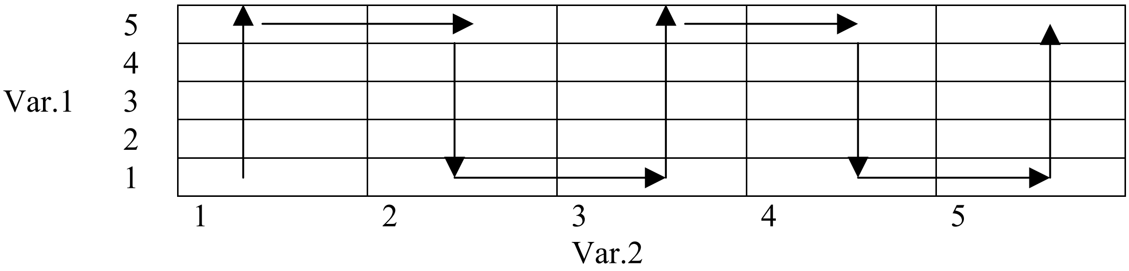

Fig.2.3: Graphical representation of the method of snake.

The ranking in the case of the method of snake is based on the computation of the so- called snake variable and on a recursive algorithm in which the results from the last level are transferred into the next level. More specifically, the snake variable is calculated in each step, the results are ranked in descending or ascending, and groups of k units are constructed. Finally, each individual value is replaced by the group mode or another statistic of central tendency. The formulas for calculating the snake variable in the case that the range of variables is between 2 and 5 and the recursive algorithm are as follows:

$$\begin{cases} v(l,m) = 6*1 + m^{(-1)^{l+1}} \\ w(k,l,m) = 6^2*1 + [v(l,m)]^{(-1)^{k+1}} \\ x(j,k,l,m) = 6^3*1 + [w(k,l,m)]^{(-1)^{j+1}} \\ y(i,j,k,l,m) = 6^4*1 + [x(j,k,l,m)]^{(-1)^{i+1}} \end{cases}$$

where y is the resulting snake variable on which the sorting of the observations is based, i is the index of the first variable, j the index of the second variable and m is the index of the fifth variable. In this method it is important to realise that the indexes i, j, m are defined by the arbitrary route that we choose to follow.

2.5.2 Calculation of Entropy

Entropy is a measure of homogeneity of the variables. It is used in the case of qualitative variables and adopts different definitions for ordinal or nominal ones.

The entropy, H, in the case of a nominal variable is given by $H_i = \frac{\left(-\sum_{i=1}^{L} p_i \text{ld} p_i\right)}{\text{ld}L}$ where p_i is the probability that an observation belongs to category i, ld is the logarithm base 2 and L is the number of categories. In the case of an ordinal variable the entropy is given by $H_i = \frac{\left(-\sum_{i=1}^{L-1}\left(\left(p_i' \text{ld} p_i'\right)+\left(1-p_i'\right)\text{ld}\left(1-p_i'\right)\right)\right)}{L-1}$ where p_i' is the probability that an observation belongs to category i or to a higher category, ld is the logarithm base 2 and L is the number of categories. Let Ω denote the population consisting of p variables, then the entropy of the whole data set both in the case of an ordinal and a nominal variable is given by $H = \sum_{i=1}^{p} H_i$.

Entropy can be used as a proximity measure that can take the form of a micro-aggregation process. Assume that we have a data set consisting of qualitative variables and we wish to create a micro-aggregated data set in which each group consists of three units. At first all entropies of all possible paired observations are calculated. These pairs are ranked according to their entropy in ascending order and the pair with the least entropy is matched with all other observations to create all possible triads given the initial two observations as the seed. The entropy is recalculated for all these combinations of triplets and the triplets are ordered in ascending order according to their entropy. The triad with the least entropy constitutes the first cluster of observations and the observations that constitute this triad are removed from the list of two observation combinations. In the sequel the next two available combination with the least entropy is taken to create all possible triads. The entropies of the triads are ordered and the triad with the least entropy is the second cluster of observations. Following the same logic, all possible combinations of groupings that minimise the entropy are taken and these groupings can be thought as observations with similar characteristics.

The variance of the set of corresponding variables can no longer be used to measure the loss of information since here we deal with qualitative variables. As a result, apart from its use as a micro-aggregation technique, entropy can be used in order to define a loss of information criterion in the case of ordinal and nominal variables. More specifically a loss of information criterion is calculated as

$$\text{Loss of Information Ratio} = \frac{\text{Original Entropy - New Entropy}}{\text{Original Entropy}} \times 100 .$$

2.5.3 Similarity of Distributions

This method concerns nominal variables for which a comparison between the values taken by the individuals is made. The nominal variables are transformed into binary variables and the individuals that have the closest distributions are regrouped.

In order to define the closeness of the distributions, a dissimilarity or equivalently a similarity index of the characteristics that two individuals have at the same time is calculated. The sorting of the observations is based on this dissimilarity (similarity) index. After forming the groups, each individual value is replaced by the mode or a weighted mode of the group that it belongs.

Another alternative, if the nominal variables are not so numerous, is to set up groups associated with all possible combinations of the values of these variables and to assign each observation to one of these groups and form groups of k units. In fact, the pre-established groups may not be so numerous because certain combinations of values will be impossible or very rare.

2.6 Evaluation Criteria

After applying the micro-aggregation techniques, we have to develop criteria for evaluating the derived results. The micro-aggregation procedure can be viewed as an effort to keep a balance between data confidentiality and data quality. This means that our main objectives are to crate confidential data that at the same time respect as much as possible the structure of the original data. The preservation of the structure of the original data ensures that the statistical analysis based on the modified set is valuable. Consequently, the evaluation criteria serve a dual purpose. On the one hand they measure the data quality and on the other hand they evaluate the confidentiality of the data set. For example, a high value of a perturbation indicator means that the data modification is high enough to ensure confidentiality. However, this also means that the micro-aggregated data have distribution very far from the original data.

2.6.1 Measuring Criteria of Data Confidentiality

According to the statistical law, data in form of tables or of aggregates must meet the following minimal constraints.

- Aggregates must not contain only one unit because it is treated as an individual piece of data.
- Aggregates must not contain only two units because it would be possible to find the value of the one by withdrawing the other unit.
- A cell is regarded as confidential if the n largest units contribute more than k percent to the cell total. This is the dominance rule. The values of n and k are determined by the provider of the original data who wants not to be identified (e.g. the Member States or a company).

2.6.1.1 Value of The Threshold 'k'

The threshold is the minimum number of units that a micro-aggregated cluster contains. The threshold can have different values that influence the result of micro-aggregation. The higher the threshold the more the data protection and consequently the higher the data perturbation. The minimal value of the threshold is derived by the first two rules of the statistical law. That means that the minimal confidentiality is achieved with three units in each micro-aggregated cluster. The grouping of units into clusters of three is sufficient when for small observations but not for large ones. This means that the threshold can be increased in the case that we have units larger than a fixed size but this implies also a higher loss of information. The variable nature of the threshold depends also on the fact one can form clusters by aggregating at least k units and not exactly k units. These is made in the case of automatic classification methods where the aggregation is made for elements that are close to each other and not because we have to satisfy an equal size constraint.

2.6.1.2 Concentration or Predominance Rule

According to this criterion a cluster is defined as confidential if the n largest elements of the cluster contribute more than k percent to the cluster total. This is also known with the name the (n,k) rule and is connected with the third point of the statistical law.

One of the means of determining the predominance level is by using a concentration coefficient of the form $C = \frac{\left(\sum x_i\right)^2}{\left(\sum x_i^2\right)}$ where x_i is the value of variable x for observation i. When this indicator is around one, the concentration relies on one value and if it is equal to n for n values then the distribution is uniform. In practice for a threshold of three, a cell is regarded as confidential if the two larger elements contribute more than 85% of the cell total.

2.6.1.3 Indicator of Data Perturbation Before and After Micro-aggregation

One can evaluate the number of units that are not sufficiently transformed by micro-aggregation, in order to have a measure of the maintenance of the confidentiality. This is achieved by using the indicator $d(j) = \text{abs}\left[\frac{x_{(j)}^m - x_{(j)}^o}{x_{(j)}^o}\right]$ where $x_{(j)}^m$ is the value of the variable x for observation j after micro-aggregation and $x_{(j)}^o$ is the value of the variable x for observation j before micro-aggregation. If the indicator is lower than a certain bound for a too large number of observations, we suppose that the method does not sufficiently maintain the confidentiality of the data.

2.6.2 Criteria for Evaluating The Maintenance of The Structure

As we have already stated, when using the micro-aggregation methods there is a trade off between confidentiality and data quality. The quality of the modified data set is very important because it determines the quality of the results obtained from the statistical processes with the modified data set. Criteria for evaluating the quality of the modified data set are mainly based on the proximity between the aggregated and the original data.

2.6.2.1 Use of Summary Statistics

One way to examine if the micro-aggregated data set preserves the characteristics of the original data set is by using summary statistics of position or dispresion. These statistics are different depending on whether the variables are numeric or ordinal. For quantitative variables we can use statistics such as the mean, the median the deciles, the variance and the Pearson correlation coefficient. For ordinal values we can use the median the mode and the entropy and for nominal values the mode and the entropy. More specifically, the calculation of such statistics aims at examining the closeness between the distribution of the original and the aggregated data. For example assume the following index. $\text{Ratio of deciles} = \frac{Q_i^m - Q_i^o}{Q_i^o}$ where Q_i^m is the ith decile of the micro-aggregated data and Q_i^o is the ith decile of the original data set. The lower the ratio the smaller the difference between the deciles of the original and the modified data. Moreover, apart from studying the univarite structure of the data we may wish to examine also the multivariate structure. In order to do so we use the Pearson correlation coefficient. If the correlation coefficient is more or less the same when it is computed for the original and the modified data, we can assume that the micro-aggregation process has not significantly changed the relationship that exists between the different metric variables.

2.6.2.2 Loss of Information Criteria

The loss of information criteria are usually described by means of analysis of variance. In case of metric variables the loss of information is defined as the ratio $\frac{\text{IntraVariance}}{\text{TotalVariance}}$. A low ratio, implies closeness between the original and the perturbed data set. In the case of qualitative variables, the measurement of the variability of the observations is the entropy. The loss of information is given by the ratio $\frac{\text{Original Entropy - New Entropy}}{\text{Original Entropy}} \times 100$. Again, the lower the ratio the closer the original and the modified data are.

2.6.3.3 Further Processing Ability

This indicator represents the ability of the transformed data to be used for making new analysis and for obtaining results sufficiently close to those that we would have obtained if instead we had used the original data. One will understand this criterion concretely by applying to the original data and to the micro-aggregated data identical analyses (e.g. regression models) and by comparing the results obtained for both data sets. The closer the results the better the preservation of the original data structure and the less the data perturbation.

2.7 Procedures in Eurostat and in European Countries

The micro-aggregation methods have been applied in Eurostat in the frame of the first and the second Community Innovation Survey (CIS).

The CIS data sent by the Member States consist of more than 40000 anonymised enterprise-level data on, among others, production, exports, R&D, innovation costs, innovation outputs and information on how companies promote and hinder innovation (a total of more than 200 variables some metric, some nominal and others ordinal). The database thus includes sensitive data and the respondents have the right to be protected from any disclosure of these data. At the same time the data available to the public have to be modified in such a way that they contain the maximum of the information.

Among other confidentiality techniques, Eurostat has chosen the micro-aggregation procedures for the CIS data. In the frame of the first Innovation Survey, quantitative variables micro-aggregated according to the individual ranking method using the mean as the aggregation statistic. Ordinal values modified according to the 'snake' method using the mode as the aggregation statistic and nominal variables were perturbed according to the similarities of distributions method using as an aggregation statistic the mode. In the case of nominal variables the first thought was to use entropy method. However, this method was rejected because it is time consuming. In the frame of the second Innovation Survey some changes have been made. More specifically, the ordinal variables are still micro-aggregated using the 'snake' method but the route followed is the 'snail' one. This happens because there is evidence that a route of a 'snail' type is more efficient i.e. in each step there is a change in the categories of the variables in an alternate way. Furthermore, the sorting order becomes descending so as not to find the largest elements in groups of four or five units. This happens because both in the first and in the second Innovation Survey the threshold used was three units per cluster. This means that if the number of records is not multiple of three the last cluster will contain 4 or more observations. Concerning the tests for the data quality of the modified data set the following methods had been used. The ratio of deciles, the ratio of variances, the examination of the marginal distributions, the correlation coefficient and the index of data perturbation. The following table summarises the procedures, concerning the micro-aggregation, in Eurostat, in Italy, in Spain and in Russia.

Table2.6: Procedures in Eurostat and in European countries concerning the micro-aggregation techniques

	Eurostat (CIS)[12]	Italy	Russia	Spain
Method for numeric variables	Individual ranking	Sum of z-scores	Individual ranking	Adaptation of Ward's method
Aggregation statistic	Mean	Mean	Mean	
Method for ordinal variables	-Snake	Not	Not	

[12] Community Innovation Survey

	-Snail	treated	treated	
Aggregation statistic	-Mode -Weighted mode			
Method for nominal variables	Similarities of distributions	Not treated	Not treated	
Aggregation statistic	-Mode -Weighted mode			
Threshold	3	3,4,5 (3 by default)		
Measure of quality and confidentiality	-Deciles, -variances, -marginal distributions, -correlation coefficient			

2.8 Concluding Remarks

The micro-aggregation process for creating confidential data is of high interest for both Eurostat and the Member States. Attempting to make some proposals concerning the micro-aggregation procedures that are followed by Eurostat we can say the following.

It is very important before starting the micro-aggregation procedure to examine always the correlation structure of the data. If there is a high correlation then we can resort to simple techniques like the Principal Component Analysis and obtain at the same time good results.

Moreover Eurostat should foresee the development of more complex methods. For example the application of Ward's method if it is combined with sorting criteria such as the first principal component, or the sum of Z-scores gives good results. Also the improvement of Hanani's algorithm, although its algorithmic complexity, must be considered. These methods must be included into the micro-aggregation software that is developed by Eurostat at this time point.

A lot of work is required in the domain of the assessment of the results obtained by the micro-aggregation procedures. The software that is under development must be enriched with more criteria for assessing the confidentiality aspect of the modified data set. For example the indicator of data perturbation must be included into the new software. Furthermore the software must be enriched also with practical criteria for assessing the closeness of the original and the modified data set. For example it is important to compare the correlation structure between the original and the modified data set for assessing the quality of the data. Moreover, it is very interesting to examine the further processing ability of the modified data set. This means that we have to apply a simple linear model to both data sets (original and modified) and see if there are significant differences in the results.

Generally speaking, in order to select an appropriate micro-aggregation technique, we have to adopt a model choice behaviour i.e. to apply several methods and several evaluation criteria, and then decide which method is the "optimal" one. Using the term 'optimal' solution we refer to a solution that is acceptable by the Member States and is valuable for the purposes of the statistical analysis.