Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 449–461

Masking Microdata Using Micro-Aggregation

D. Defays and M.N. Anwar[1]

A family of disclosure avoidance methods based on micro-aggregation are introduced and the nature of confidentiality protection provided by these methods, the likely costs of such procedures to data providers and the likely benefits to data users are investigated.

Key words: Confidentiality; data transformation; data protection.

1. Introduction

Micro-aggregation is a technique for protecting individual data by aggregation (see Defays and Nanopoulos 1993). In its most basic form the idea is influenced by the work of Strudler, Lock, and Scheuren on the Tax Model at the U.S. Internal Revenue Service. However, in order to apply the method to a specific case we had to adapt and generalise it. The original idea could only be applied to a specific case, but in this article we show how it can equally be applied, with suitable modification, to other situations. A sample application is presented in Section 4. The advantages and weaknesses of the different variants of the generic method are discussed. We conclude by showing the links between this method and the more traditional techniques such as data suppression, the recoding of variables or their disturbance.

2. The Basic Method

The principle of the method is simple. In a statistical table, it is common practice to denote as confidential either cells with fewer than three units, or cells which are dominated by one observation – two in some countries – which covers an extremely large part of this whole. When applied to individual data this rule has immediate consequences: any observation with a frequency of less than three is deemed confidential. Data protection thus involves the re-grouping of micro-data (using automatic classification or value re-allocation techniques) in aggregates of three (or more if the threshold of three is considered too risky), while taking care that these aggregates are not dominated by a single observation. The method of micro-aggregates involves a straightforward application of this principle. In order to minimise data losses, it is proposed that the different unidimensional variables

[1] Research and development, and statistical methods unit, Eurostat – The Statistical Office of the European Communities, L-2920 Luxembourg, Luxembourg.
Acknowledgments: The authors would like to thank Messrs Ph. Nanopoulos, L. Kioussis, and W. Grünewald for their advice and comments.

be aggregated separately, by sorting the values according to their ranks, and by an aggregation in fixed size groups of contiguous values. To illustrate, we take data on three variables from a survey on technological innovation in Europe analysed by Eurostat and aggregate into groups of 3, 5, and 10 to show the likely structural changes under different group sizes[2].

As a first step, the units are sorted in ascending (or descending) order of Variable 1 and grouped k by k (where k in our case was 3, 5 and 10). The original Variable 1 value for each unit is then replaced by the average for Variable 1 of the corresponding group. In the next step, the units are again sorted, but by Variable 2 this time. Groups of k are formed and the original Variable 2 values are replaced by the averages of the corresponding groups. This procedure – sorting, grouping, replacement with average values – is repeated for the third variable, and a new file is created consisting of surrogate observations equal in number to the original file. The method acts more on outlying observations while leaving the majority of the data structure intact; this property is both interesting and useful from a statistical and confidentiality viewpoint. The masking of extreme values is a prerequisite of any method purporting to safeguard confidentiality, yet a method which destroys data structure also destroys the statistical properties of the data. The method proposed both decreases the risk of disclosure and maintains relationships (see Section 5). A ''grid'' structure is imposed by the method which becomes more pronounced as k is increased. This grid structure provides a guarantee of confidentiality by creating observations with identical values on a single given dimension (see Figure 1). The mesh of the grid becomes finer the more components there are, as illustrated in the figure below (case of numerical variables treated separately and replaced by averages).

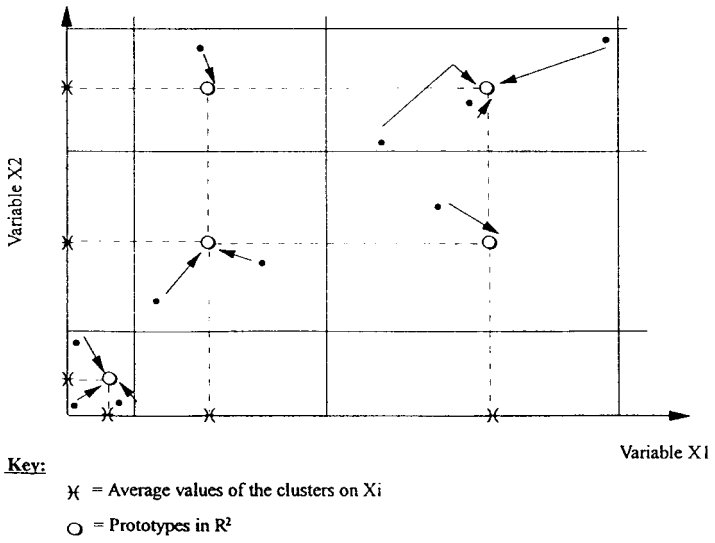

Fig. 1. Structural changes in the data

[2] The observations were ranked in ascending order in the example but the descending order can also be used.

Table 1. Summary statistics

Original Statistics				
Variable	Mean	Std Dev	Minimum	Maximum
VAR 1	272	2,383	0	177,183
VAR 2	29,175	374,863	0	25,867,102
VAR 3	10,358	169,553	0	14,057,249
Modified statistics ($k = 3$)				
Variable	Mean	Std Dev	Minimum	Maximum
VAR 1	272	2,369	0	159,207
VAR 2	29,175	373,858	0	24,345,544
VAR 3	10,358	167,873	0	12,252,076
Modified statistics ($k = 5$)				
Variable	Mean	Std Dev	Minimum	Maximum
VAR 1	272	2,331	0	125,056
VAR 2	29,175	370,123	0	20,056,637
VAR 3	10,358	163,383	0	8,967,085
Modified statistics ($k = 10$)				
Variable	Mean	Std Dev	Minimum	Maximum
VAR 1	272	2,274	0	103,871
VAR 2	29,175	358,465	0	16,251,070
VAR 3	10,358	159,649	0	7,548,644

This characteristic is very interesting; it demonstrates that the technique adjusts spontaneously to the distribution of variables and, where there is a high concentration, disturbs the data very little (more is not necessary since many units have similar values), whereas when the values are more dispersed, it superimposes a stronger noise on the original data.

However, as it can be seen from this figure, and as shown in Table 1, the variances of the variables are affected by the micro-aggregation. Their reduction increases with k (see Section 5.4). Note that the method leaves the means of the variables unchanged.

The survey to which we wanted to apply the method combined both quantitative and qualitative data, key data on the enterprise which might permit indirect identification and more neutral subjective assessments, simple questions or questions with a more complex structure. The method had first to be generalised so that both quantitative and non-quantitative variables could be considered. How, for example, should an assessment of the type

> ''*Crucial – Of major significance – Of average significance – Of minor significance – Of no significance*''

be dealt with? Again, how should a variable used to code a sector of activity be dealt with? Similarly, specific problems were caused by some items, such as:

> ''*Evaluate the effectiveness of the various methods outlined below in protecting your product innovations:*
> *– patents,*
> *– industrial secrecy,*
> *– product sophistication, etc.*''

where each method has to be evaluated on an ordinal scale of five levels. Dealing with each of the headings in the item separately would basically mean disturbing only a very small number of records, thereby providing poor protection of confidentiality. Asymmetry in the distribution of certain variables also presented problems: how was it possible to avoid very large enterprises being easily identifiable following micro-aggregation?

The generic method presented in the following section is an attempt to solve some of these problems.

3. A More Generic Version

Let $(X_1, X_2, \ldots, X_i, \ldots, X_p)$ be a vector of p variables, for example employment, turnover, investment. Let P denote a population of N units, for example the enterprises in a given country. Let L_x denote a method which replaces the original set of X values for a given population by a new set of values. Generally the new values are taken so as to respect as far as possible the original distribution; for example, L_x could ensure correspondence between a distribution and its average value or mode. Let H be a measure of the homogeneity of a group of units.

The definition of a micro-aggregation method involves a certain number of conventions which are outlined below. In what follows, we will suppose that we have at the outset a population P of N units on which a multivariate variable X has been defined.

3.1. Segmentation of the set of variables

The micro-aggregation method presented in Section 2 has been referred to as the ''micro-aggregation method by individual ranking''. This term underlines the necessarily separate treatment of the different individual variables which results in separate classifications and aggregations of units of P, as illustrated in Section 2. But what is regarded as an individual variable in this context? A set of p variables can be treated as a single multivariate variable, resulting in a single grouping, or as separate p variables, resulting in p groupings. More generally, X can be segmented into s variables, s varying from 1 to p: $X = (X_1, X_2, \ldots, X_i, \ldots, X_s)$. With this notation, the X_i are thus multivariate or univariate variables which we will call segments.

Each segment will be regarded as homogeneous, i.e., composed of the same type of variables. We will distinguish four types of variables: *quantitative*, *ordinal*, *nominal-hierarchical*, and *nominal-flat*. The two first types are well-known and require no explanation. The latter two are less well-known. A nominal variable is called *hierarchical* when the values it takes belong to a set on which a total hierarchy has been defined (generally represented by a tree), as in the well-known classifications of official statistics. This hierarchical structuring of values is frequent and permits certain operations, like taking a maximum. The nominal-flat variables are other qualitative variables.

3.2. Characterisation of groups

For each segment formed, units of population P will be regrouped and within each group, the X values of the units will be replaced by a central value which sums up the values included in the group, or by other values which will respect as far as possible the original

distribution in the groups. Let L_x refer to the method selected to associate new values (generally a central value) with X. There are different ways of attributing a central value to a distribution, corresponding to different variants of the method. If X is a quantitative variable, it can, for example, be associated with an average ($L_x = E(X)$), a median ($L_x = \text{med}(X)$), a mode ($L_x = \text{mod}(X)$), or an interval of variation.

As suggested by a referee, one could also in the case of each unit add white noise (with a suitable chosen variance) to the group averages, or replace each value of the group by a rounded value chosen so that the overall mean and the variance are preserved.

If X is a nominal-hierarchical variable, it can be associated with a mode, or a maximum (the order being defined by the hierarchy of values). In this case, each group will be associated with the node of the tree directly above all the values observed in the group. In effect, this involves changing the coding and, in the case of geographical coding for example, moving from the commune to the district and even to the province or the country if the values are very widely dispersed. When X is nominal, it can be summed up by a mode or a set of values (another form of recoding). In the case of multivariate nominal variables, it is possible to take either the multivariate mode or the modes for each of the variables separately.

3.3. Constraints on size

Respect for confidentiality involves avoiding conspicuous values by grouping them in sufficiently large groups. The minimum size of such groups can depend on several factors: the procedures or rules adopted in some countries, the degree of confidentiality of a segment of variables, or even the values taken by the variables in a group which may be so atypical that larger aggregations above certain thresholds are desirable; thus, it is not unreasonable to conclude that, for company statistics, small enterprises should be grouped three by three whereas larger enterprises should be aggregated in larger groups.

In the following, C represents the size constraints imposed. We write $C = \text{fix } 3$ if the size of all the groups is equal to three, or $C = \min 3$ if the size of all the groups is to be larger than or equal to three. Of course C can be a multiple constraint, e.g., consisting of a size constraint and a dominance constraint.

3.4. Measures of group homogeneity

Groups are formed for each segment of variables by maximising the homogeneity of the groups with respect to these variables. Homogeneity may be measured in different ways. Let H denote the measure of homogeneity for a given segment of variables. Obviously, different types of variables require different types of measurement. For the quantitative variables, the Euclidean distance or the variance ($H = \text{var}$) will determine the formation of groups, while for the ordinal variable this formation will be determined by the absolute value of the difference in rank ($H = \text{abs(dif } r)$) or more general measurements of homogeneity, such as that based on a generalisation of the concept of entropy ($H = \text{ent}$) proposed by Vogel et al. (1982).

In the case of nominal-hierarchical variables, the hierarchy of possible values introduces a natural distance which is the induced ultra-metric distance (it being assumed that the hierarchy is valued by its aggregation levels). With this arrangement,

two components belonging to two communes of the same district, for instance, are closer than two communes belonging to two different districts. In the case of the nominal-flat variables, a traditional entropy measurement or a chi square ($H = \text{chi}^2$) can be used (see, for example, Benzecri 1973).

3.5. *Classification techniques*

When a measure of similarity or homogeneity has been fixed for a segment of variables, the problem of aggregation is largely reduced to a standard problem of automatic classification (see Sneath 1973). Groups of maximum homogeneity have to be formed while respecting constraints (see Hannani 1979). If the variable is quantitative and univariate and if the constraint is to form groups of a fixed size k, it is obvious that the groups are formed by ranking the individuals and then grouping k by k. This will basically minimise the within-group variance. If the variable is quantitative and multivariate the problem is more complicated: when the distance used is Euclidean, it has been found that the optimum groups are defined by hyper-planes in R^p (Defays and Nanopoulos 1993). Iterative methods exist to approximate the optimum partition. The case of a univariate ordinal variable is also straightforward, involving a ranking of values while trying to minimise rank differences expressed in absolute values. This can also be applied for the multivariate case by summing the rank differences over the variables. However, this may not take adequately into account the semantic structure of the data. What happens if the multivariate case is more complex and one would like to incorporate the semantic structure of the multivariate variable? If it is decided to denote the dissimilarity between values using a distance, such as the city block distance, based on ranks, the problem involves the regrouping of similar units. To give an approximation of the optimal solution a more general use of the ranking method used in the previous case is possible. As the objective is to regroup units which are similar, it is sufficient to define a path of minimum length linking all the units, then regroup them k by k *along the path* as above. This is unfortunately the well-known and difficult salesman problem. However, when all the combinations of values are possible, there is a pragmatic approach, which is easy to implement and which makes it possible to find the path of minimum length. The values (in effect, the ranks) can be represented as nodes of a grid with p dimensions: as illustrated in the following graph ($p = 2$), it is simple to define a path on this grid, passing via adjacent peaks (distance one) and via all the peaks. This has been referred to as the snake method. Alternatively entropy can be used as a

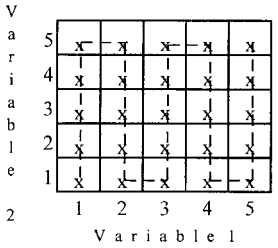

Fig. 2. ''The snake''

measure of the homogeneity of groups, in which case the algorithms used should be iterative (we will call it proxentropy): as with the snake method, they do not always guarantee the best solution.

Our snake takes the route $(1,1),\ldots,(1,5),\ (2,5),\ldots,(2,1),\ldots$ as shown in the figure above. Of course, the choice is somewhat arbitrary and other paths are possible.

Thus, by specifying the various components outlined above, namely segmentation, characterisation of groups, size constraints and definition of the measure of group homogeneity, we get a specific method of micro-aggregation. The classification techniques provide the solution or approximate solution to the problem thus specified. A particular micro-aggregation method will thus be presented as an s-couple (because s segments of variables were defined).

$$(< X_1, L_X, C, H >, < X_2, L_X, C, H >, \ldots, < X_s, L_X, C, H >)$$

The figure below represents the recommended range of options for our s-couple, with the arrows indicating the hierarchical nature of data values. A metric variable can be transformed into an ordinal or nominal variable so that the ''snake'' or entropy can be used as a measure of homogeneity. The exact mix chosen will depend on the quality of data needed to allow data users to draw adequate conclusions, and the degree of data protection required by data providers.

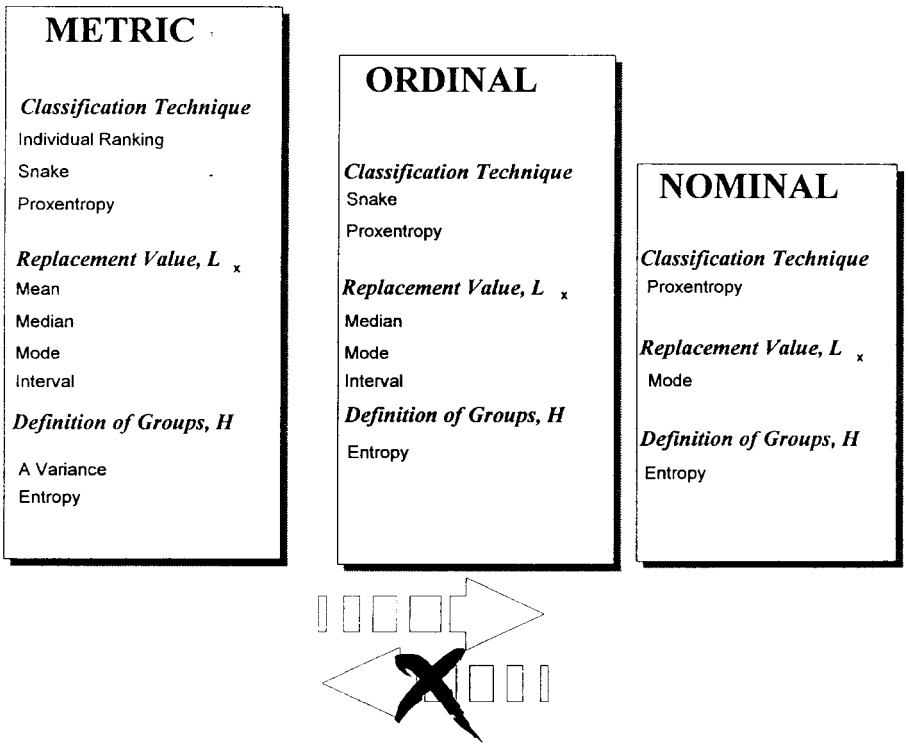

Fig. 3. Possible options for micro-aggregation.

Table 2a. Original data on nine companies

				What is more important?			
Company	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	12	1,000	2	1	1	N	Y
2	21	1,500	6	1	2	N	Y
3	39	2,000	5	2	5	Y	Y
4	40	3,000	3	2	4	N	N
5	42	1,000	4	2	4	N	Y
6	47	2,000	10	3	3	N	Y
7	53	1,500	11	4	3	N	N
8	58	1,500	10	4	2	Y	N
9	60	3,000	14	5	5	Y	Y
	METRIC			ORDINAL		NOMINAL	

4. A Sample Application

Table 2a presents an example of data adapted from a survey on technological innovation in Europe currently being coordinated by Eurostat. These are three metric variables, an ordinal question with two sub-questions, and two nominal variables which will be treated together as one segment. The particular method we shall use for our example is, ''individual ranking'' for the metric variables, ''individual ranking with snake'' for the ordinal question and ''proxentropy'' for the nominal variables, with k, the number of observations in each grouping, set at 3. The method can of course be varied according to the particular circumstances of each data set and the required degree of perturbation. The method can thus be described by the following:

$$(< X_1, \mathsf{E}(X_1)\text{fix } 3, \text{var} >,\ < X_2, \mathsf{E}(X_2), \text{fix } 3, \text{var} >,\ < X_3, \mathsf{E}(X_3), \text{fix } 3,\ \text{var} >,$$
$$< (X_4, X_5), \text{med}(X_4) - \text{med}(X_5), \text{fix } 3,\ \text{abs(dif } r) >,$$
$$< (X_6, X_7), \text{mod}(X_6) - \text{mod}(X_7), \text{fix } 3, \text{ent} >)$$

Step 1 – Metric variables

Each variable is taken to define one segment, and k is set at three. Each variable is independently ranked and for each grouping the original values are replaced by the arithmetic mean of the three observations comprising the cluster (see the output Table 2b below)

Step 2 – Ordinal variables

Since we have two variables in our segment and five classes, the snake method is applied. After ranking the observations accordingly and placing them in groupings of 3, one replaces the original values with an appropriate measure of central tendency such as the one in our example, the median (see output Table 2b below).

Table 2b. Surrogate data on nine companies following micro-aggregation

				What is more important?			
Company	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	24	1,167	3	1	2	N	Y
2	24	1,167	7	1	2	N	Y
3	24	1,667	7	1	2	Y	Y
4	43	2,667	3	2	4	N	N
5	43	1,167	3	2	4	N	Y
6	43	2,667	7	2	4	N	N
7	57	1,667	12	4	3	N	N
8	57	1,667	12	4	3	Y	Y
9	57	2,667	12	4	3	Y	Y

Step 3 – Nominal variables

The last two nominal variables are treated as one segment. The group homogeneity is measured by an entropy calculated as follows:

$$H = \left(-\sum_{i=1}^{L} p_i \; ld \; p_i \right) \Big/ ld \; L$$

where p_i is the frequency of observations belonging to category i in the group, ld is the logarithm to the base 2 and L is the number of different categories, that is to say the number of values taken by the multivariate variables $(X_6, X_7) - 4$ in our example. Once all groups have been identified, one selects an appropriate measure of central tendency, such as the mode, and the original values are replaced by surrogates.

5. Characteristics of the Method

The characteristics of the initial method have been outlined in detail in various documents. Some additional information may be added in the light of the more general application developed in this article.

5.1. Fine-tuning the degree of protection

The definition of segments on the X set of variables makes it possible to achieve a fine balance between the demands of data protection on the one hand and data quality on the other. At the extreme, consideration of only a single segment (with all the variables) provides an absolute guarantee of confidentiality if the minimum size of the groups is set at more than two and if no observation dominates in the groups. In other words, irrespective of the variables which are crossed, no table based on micro-aggregated data will contain confidential data.

As we have already indicated, it is possible to consider imposing different constraints according to the range of values involved in each segment. For example, above a certain threshold (or in a given sub-region of an area) the data could be more aggregated. Enterprises with a turnover higher than a given threshold could be grouped 5 by 5 and not 3 by 3, thereby strengthening the protection of marginal individuals in the population. Thus the method can be fine-tuned by modifying the segments, by changing k, or by choosing a different replacement value.

5.2. Protection obtained from the method

In principle the level of protection enjoyed by micro-aggregating is the same as for tabular data. Indirect identification of some outliers is possible but in most cases only perturbed data is available to the intruder. If the micro-aggregation is applied to continuous data (with replacement values equal to the mean of each k grouping) an intruder may be able to deduce what group size (k) has been used by grouping records on the basis of equal values of the variables. And since micro-aggregation does not alter the *order* of the contributions to the group total, the top k contributors remain the same after micro-aggregation. If the $k-1$ of the top k contributors pool their data they can deduce the value of the kth contributor. But this is also true for tabular data, and any predominance rules applying to tabular data should also be applied to micro-aggregated data. Sensitive variables could be aggregated above a certain threshold by choosing a higher value for k, or replacing it with a size class interval.

5.3. Relationships between variables

An important characteristic of a method for protecting individual data is to preserve the moments of the initial distributions. Indeed, some people reject techniques such as ''data swapping'' on the basis of such arguments. If the variant chosen is individual ranking it is easy to show that the second order moments will, as a rule, be slightly disturbed by the method (see Section 5.4. for a more theoretical approach). Qualitatively, this is easily explained. Each individual variable is disturbed as little as possible since it is aggregated with data which are as similar to it as possible. Thus, a unit characterised by a couple (X_1, X_2) of high values, where X_1 and X_2 belong to different segments, will still be characterised by high values following micro-aggregation, since the unit's relative position in R^2 changes little, as X_1 and X_2 are only slightly disturbed.

5.4. Reduction of variance

We have mentioned several times that micro-aggregation by individual ranking will reduce the variances of the variables. To study the bias introduced, some further notations are needed. We will concentrate in this section on the effect on one unidimensional variable X. Let $X_{(1)}, X_{(2)}, \ldots, X_{(N)}$ be the values taken by X, arranged in increasing order. When we apply the individual ranking method, the original values of X are replaced by micro-aggregated values denoted Y:

$$X_{(i)} \text{ is replaced by } Y_i = \sum_{j=1}^{k} X_{(lk+j)}/k$$

if, for some integer $l, i \in]lk, (l+1)k]$

This substitution can be seen as a perturbation of the original value $X_{(i)}$

$$Y_i = X_{(i)} + e_i$$

where

$$e_i = \frac{1}{k}\sum_{j=1}^{k}(X_{(lk+j)} - X_{(i)})$$

From the usual analysis of variance formula, decomposing the total sum of squares by the sums of squares between and within groups, we have

$$\text{var}(X) = \text{var}(Y) + \text{var}(e)$$

where var denotes the variance across the N finite population values.

This shows that $\text{var}(Y) \leq \text{var}(X)$.

A study of the reduction of the variance of X necessitates a more in-depth analysis of the error e.

This study is possible but fairly complex: in fact, the e_i can be expressed as a linear function of the spacings $D_j = X_{(j)} - X_{(j-1)}$ between contiguous X values.

It can be shown that

$$e_i = \left[\sum_{r=i+1}^{(l+1)k} \sum_{j=i+1}^{r} D_j - \sum_{r=lk+2}^{i} \sum_{j=r}^{i} D_j \right] \Big/ k$$

As the distribution of the D_i (when the N observations are considered as a random sample) depends on the distribution of X and has been studied (see for instance Pyke 1965), this formula makes it possible, in theory, to derive the bias introduced on the variance by the method under different distributional assumptions.

Furthermore, as

$$|e_i| \leq X_{(l+1)k} - X_{lk+1} \leq \sum_{j=lk+2}^{(l+1)k} D_j$$

and since

$$\text{var}(e) = \frac{1}{N} \sum_{i=1}^{N} e_1^2$$

upper bounds on the bias can be calculated.

Let us suppose for instance that X is uniform on (0, 1) and that the N values have been observed on a random sample. Under this assumption the distribution of the D_is is well-known; they all have the same distribution and

$$E(D_i^2) = \frac{2}{(N+1)(N+2)}$$

This means that

$$\begin{aligned} E\,\text{var}(e) &= \frac{1}{N} \sum_{i=1}^{N} E(e_i^2) \leq \frac{1}{N} \sum_{i=1}^{N} E\left(\sum_{j=2}^{k} D_j \right)^2 \\ &\leq \frac{1}{N} \sum_{i=1}^{N} (k-1)^2 \frac{2}{(N+1)(N+2)} \text{(by applying the Cauchy-Schwartz inequality)} \\ &\leq \frac{2(k-1)^2}{(N+1)(N+2)} \end{aligned}$$

Table 3. Upper bound of the relative bias for a uniform distribution (%)

N	$k = 3$	$k = 5$
50	3,5	14
100	1	3,7
1,000	0,01	0,04

This makes it possible to calculate, for different values of k, an upper bound on the relative bias since

$$\frac{E\,\mathrm{var}(e)}{E\,\mathrm{var}(X)} \leq \frac{24(k-1)^2(N-1)}{N(N+1)(N+2)}$$

Some values are given in Table 3.

6. Summary and Discussion

The preceding sections have provided a theoretical framework for the use of micro-aggregation techniques. In fact, the proposed generalisation encompasses some well-known procedures for protecting the confidentiality of individual data.

The simplest technique, involving the creation of classes of three individuals of minimum variance, using the average as a replacement value, corresponds to the method $(< X, E(X), \text{fix } 3, \text{var} >)$. The individual ranking presented at the outset as a paradigm of the method corresponds to $(< X_1, E(X_1), \text{fix } 3, \text{var} >, < X_2, E(X_2), \text{fix } 3, \text{var} >, \ldots, < X_p, E(X_p), \text{fix } 3, \text{var} >)$.

Micro-aggregation is also one way to recode data or to replace them by missing values. As stated in the paragraph on the characterisation of groups, values may be replaced by an interval, by a set of recorded values, by a broader set presented as a higher aggregation level in a hierarchy, or by a missing value.

Certain authors (cf. Bragard et al. 1988) have proposed that the initial population P be replaced by a set of prototypes representative of the initial units. The method of micro-aggregates is strongly influenced by this idea in its initial design. The units in the micro-aggregated data are real units but the values of these units are modified (by L_X methods) while taking account of the initial distributions, which are departed from as little as possible (the homogeneity measurements H are maximised).

When applied to numerical variables, micro-aggregation can be seen as a disturbance method. However, it does present some specificity: generally, if X is the initial variable and Y the disturbed variable, there is equality such that: $Y = X + e$, where e is a noise orthogonal to the variable X. With micro-aggregation the noise e is orthogonal to Y. The disturbance method thus amounts to extracting a given factor Y which is as close as possible to the initial variable X (thus Y would be some kind of the ''true score,'' to use measurement theory terminology).

It has been shown that the micro-aggregation method has several interesting features. First, it is simple and flexible in its approach. Second, it offers a compromise between data protection – the transformed data can always be constructed so that the units do

not correspond to any of the units in the original dataset (if they are unique) – and quality of the information. It has also pointed to areas of additional research, namely: (1) developing a theoretical model to estimate the errors resulting from the use of micro-aggregation data, and the change in relationship between observations; (2) expanding the approach to handle longitudinal data; (3) looking into the possibility of on-line access to micro-aggregated data for researchers; and (4) development of micro-aggregation software.

Privacy has become a real concern in our modern societies, but so has the need to access detailed statistical information. Micro-aggregation is only a first step towards satisfying both these conflicting demands; unfortunately it is an empirical approach based on empirical rules which have proved useful, rather than on pure statistical theory, and much more needs to be done both on the theoretical front and on development of more powerful techniques.

7. References

Benzecri, J.P. (1973). L'Analyse des données. I La Taxinomie. Paris: Dunod. (In French).

Bragard, L., Roubens, M., Libert, J., and Gailly, B. (1988). Examen d'une méthode d'échantillonnage par sélection de prototypes. Report produced by CEPS, Walferdange, Luxembourg. (In French).

Defays, D. and Nanopoulos, Ph. (1993). The Small Aggregates Method. Proceedings of the 92 Symposium on ''Design and Analysis of Longitudinal Survey.'' Ottawa: Statistics Canada.

Hannani, U. (1979). Multicriteria Dynamic Clustering. Rapport de recherche Laboria, No 358. Rocquencourt, France.

Pyke, R. (1965). Spacings. Journal of the Royal Statistical Society, Series B, 27, 395–449.

Sneath, P.H. and Sokal, R.R. (1973). Numerical Taxonomy. San Francisco: W.H. Freeman and company.

Strudler, M., Lock, O.H., and Scheuren, F. (1982). Protection of Taxpayer Confidentiality with Respect to the Tax Model. Washington: U.S. Internal Revenue Service.

Vogel, F., Dobbener, R., and Grünewald, W. (1982). Iterative Klassifikation von Merkmalsträgern – Programmpaket Komixi. Internal report, Bamberg: Universität Bamberg. (In German).

Received August 1995
Revised August 1997