Chapter 5

Statistical Microdata Protection

In Section 2.3 we pointed out the need to protect statistical microdata when released to possibly dishonest users that may infer about individual respondents. In this chapter, we present our contributions to statistical disclosure control (SDC) for continuous microdata.

5.1 A modified score

In Subsections 2.3.2 and 2.3.3 above, measures to compute information loss and disclosure risk were shown. Those measures assume that the i-th masked record corresponds to the i-th original record.

Such one-to-one mapping cannot be assumed when the original and masked files have a different number of records or when masked records have been permuted. In this case, a new way to compute IL_1 must be defined. A natural way is to map each published masked record i to the nearest original record $c(i)$ using the d-dimensional Euclidean distance between standardized

records (where d is the number of variables in the data sets). Then a new IL'_1 is computed as the mean variation between masked records and the original records to which they are mapped. Denoting by n' the number of masked records, we have

$$IL'_1 = \frac{\sum_{j=1}^{d} \sum_{i=1}^{n'} \frac{|x_{c(i),j} - x'_{ij}|}{|x_{c(i),j}|}}{n'd}$$

where x_{ij}, x'_{ij} are the values taken by the j-th variable for the i-th record of the original and masked data sets, respectively.

Replacing IL_1 by IL'_1 leads to a modified information loss measure IL':

$$IL' = 100 \cdot \frac{(IL'_1 + IL_2 + IL_3 + IL_4 + IL_5)}{5}$$

Also, the lack of a one-to-one mapping between original and masked records forces a redefinition of disclosure risk measures DLD and PLD. As in the definition of IL'_1, we will say that a masked record is correctly linked to an original record if they are at the shortest possible d-dimensional Euclidean distance. Additionally, we redefine ID so that "corresponding values" mean values in records at shortest d-dimensional Euclidean distance. Distance is always measured over standardized records. Call DLD', PLD' and ID' the resulting redefined disclosure risk measures.

The new $Score'$ is computed by replacing IL, DLD and ID with IL', DLD' and ID' as well as dropping PLD for computational reasons. This yields:

$$Score' = 0.5 \cdot IL' + 0.25 \cdot DLD' + 0.25 \cdot ID' \tag{5.1}$$

This new $Score'$ was first used to evaluate the performance of a recently proposed method for synthetic microdata generation. We published our results in [DDS02].

5.2 Post-masking optimization

In the previous section, a measure $Score'$ has been proposed to measure how good is a masking method in terms of information loss and disclosure risk.

In this section, a post-masking optimization approach is presented which seeks to modify the masked data set to minimize information loss without increasing disclosure risk, leading to a better $Score'$ and thus to a better masking. Results presented here have been published in [SDMT02].

Once an original data set X has been masked as X', post-masking optimization aims at modifying X' into X'' so that the first and second-order moments of X are preserved as much as possible by X'' while keeping IL_1' around a prescribed value. Near preservation of first and second-order moments results in (constrained) minimization of IL_2, IL_3, IL_4 and IL_5, which implies near preservation of multivariate statistics. Regarding IL_1', a slight reduction is reasonable and desirable, whereas minimization is not; too small an IL_1' would most likely result in a dramatic disclosure risk increase, because post-masking optimized data would look too much like the original data.

5.2.1 Mathematical background

Next, we explain the mathematical background on which our post-masking procedure is based.

Preserving averages

Let X_1 and X_2 be two data sets with d common variables and with n_1 and n_2 records, respectively. Then, it is easy to see that, if

$$\frac{\sum_{i=1}^{n_1} x_{1ij}}{n_1} = \frac{\sum_{i=1}^{n_2} x_{2ij}}{n_2}, \quad \forall j \in (1 \cdots d)$$

where x_{1ij}, x_{2ij} are the values taken by the j-th variable for the i-th record of X_1 and X_2, respectively, then first-order moments of X_2 match those of X_1 (thus causing IL_2 between X_1 and X_2 to be 0).

Preserving variances

Let X_1 and X_2 be two data sets with d common variables and with n_1 and n_2 records, respectively. Let x_{1ij}, x_{2ij} be the values taken by the j-th variable for the i-th record and $\overline{x}_{1j}$, $\overline{x}_{2j}$ be the averages of the j-th variables of X_1 and X_2 respectively. Preserving the variances of the j-th variables in both data sets can be written as

$$\frac{\sum_{i=1}^{n_1}(x_{1ij} - \overline{x}_{1j})^2}{n_1} = \frac{\sum_{i=1}^{n_2}(x_{2ij} - \overline{x}_{2j})^2}{n_2}, \quad \forall j \in (1 \cdots d)$$

The above is equivalent to

$$\frac{\sum_{i=1}^{n_1}(x_{1ij}^2 - 2x_{1ij}\overline{x}_{1j} + \overline{x}_{1j}^2)}{n_1} = \frac{\sum_{i=1}^{n_2}(x_{2ij}^2 - 2x_{2ij}\overline{x}_{2j} + \overline{x}_{2j}^2)}{n_2}$$

and

$$\frac{\sum_{i=1}^{n_1} x_{1ij}^2}{n_1} - 2\overline{x}_{1j}\frac{\sum_{i=1}^{n_1} x_{1ij}}{n_1} + \overline{x}_{1j}^2 = \frac{\sum_{i=1}^{n_2} x_{2ij}^2}{n_2} - 2\overline{x}_{2j}\frac{\sum_{i=1}^{n_2} x_{2ij}}{n_2} + \overline{x}_{2j}^2$$

Finally, the previous expression is equivalent to

$$\frac{\sum_{i=1}^{n_1} {x_1}_{ij}^2}{n_1} - {\overline{x}_1}_j^2 = \frac{\sum_{i=1}^{n_2} {x_2}_{ij}^2}{n_2} - {\overline{x}_2}_j^2$$

From the expression above, we can see that, if $\overline{x}_{1j} = \overline{x}_{2j}$ (first-order moments are preserved), and $\frac{\sum_{i=1}^{n_1} {x_1}_{ij}^2}{n_1} = \frac{\sum_{i=1}^{n_2} {x_2}_{ij}^2}{n_2}$, $\forall 1 \leq j \leq d$, then the variance of corresponding j-th variables of X_1 and X_2 will be the same (which will result in IL_4 between X_1 and X_2 being 0).

Preserving covariances

In a similar way, let X_1 and X_2 be two data sets (with n_1 and n_2 registers respectively and d variables). Preserving the covariance between any pair $1 \leq j < k \leq d$ of variables can be written as:

$$\frac{\sum_{i=1}^{n_1}(x_{1ij} - \overline{x}_{1j})(x_{1ik} - \overline{x}_{1k})}{n_1} = \frac{\sum_{i=1}^{n_2}(x_{2ij} - \overline{x}_{2j})(x_{2ik} - \overline{x}_{2k})}{n_2}, \quad \forall j,k \in (1 \cdots d),\ j < k$$

This is equivalent to

$$\frac{\sum_{i=1}^{n_1}(x_{1ij}x_{1ik} - x_{1ij}\overline{x}_{1k} - \overline{x}_{1j}x_{1ik} + \overline{x}_{1j}\overline{x}_{1k})}{n_1} = \frac{\sum_{i=1}^{n_2}(x_{2ij}x_{2ik} - x_{2ij}\overline{x}_{2k} - \overline{x}_{2j}x_{2ik} + \overline{x}_{2j}\overline{x}_{2k})}{n_2}$$

Finally, the above is equivalent to

$$\frac{\sum_{i=1}^{n_1} x_{1ij}x_{1ik}}{n_1} - \overline{x}_{1j}\overline{x}_{1k} = \frac{\sum_{i=1}^{n_2} x_{2ij}x_{2ik}}{n_2} - \overline{x}_{2j}\overline{x}_{2k}$$

From the expression above, we can see that if $\overline{x}_{1j} = \overline{x}_{2j}, \quad \forall 1 \leq j \leq d$ (first-order moments are preserved) and $\frac{\sum_{i=1}^{n_1} x_{1ij}x_{1ik}}{n_1} = \frac{\sum_{i=1}^{n_2} x_{2ij}x_{2ik}}{n_2}, \quad \forall 1 \leq j < k \leq d$, then the covariance between the j-th and k-th variables of X_1 will match with the corresponding one of X_2 (if variances are also preserved,

this causes IL_3 between X_1 and X_2 to take a value of 0).

Preserving correlations

From the definition of correlation, it is trivial to see that, when two data sets X_1 and X_2 preserve variances and covariances, correlations are also preserved (causing IL_5 between X_1 and X_2 to be 0).

5.2.2 The model

As it has been shown in the previous section, the first-order moments of a data set X depend on the sums

$$\frac{\sum_{i=1}^{n} x_{ij}}{n} \text{ for } j = 1, \cdots, d$$

where x_{ij} is the value taken by the j-th variable for the i-th record. The second-order moments of X depend on the sums

$$\frac{\sum_{i=1}^{n} x_{ij}^2}{n} \text{ for } j = 1, \cdots, d$$

$$\frac{\sum_{i=1}^{n} x_{ij} x_{ik}}{n} \text{ for } j, k = 1, \cdots, d \text{ and } j < k$$

Therefore, our goal is to modify X' (masked data set) to obtain a X'' (optimized masked data set) so that the above $2d + d(d-1)/2$ sums are nearly preserved between X (original data set) and X'', IL_1' is reduced to a desired value and disclosure risk stays similar in X' and X''. First, let us compute IL_1' of X' vs X as

$$IL_1' := \frac{\sum_{i=1}^{n'} \sum_{j=1}^{d} \frac{|x'_{ij} - x_{c(i),j}|}{|x_{c(i),j}|}}{dn'} \tag{5.2}$$

where $c(i)$ is the original record nearest to the i-th masked record of X' (d-dimensional Euclidean distance [1] is used). Now let $0 < q \leq 1$ be a parameter and let M be the set formed by the $100q\%$ records of X' contributing most to IL_1' above. Then let us compute the values x_{ij}'' of X'' as follows. For $x_{ij}' \notin M$ then $x_{ij}'' := x_{ij}'$. For $x_{ij}' \in M$, the corresponding x_{ij}'' are solutions of the following minimization problem:

$$\min_{\{x_{ij}''|x_{ij}'\in M\}} \sum_{j=1}^{d}\left(\frac{\sum_{i=1}^{n'} x_{ij}''}{n'} - \frac{\sum_{i=1}^{n} x_{ij}}{n}\right)^2 + \sum_{j=1}^{d}\left(\frac{\sum_{i=1}^{n'} x_{ij}''^2}{n'} - \frac{\sum_{i=1}^{n} x_{ij}^2}{n}\right)^2$$

$$+ \sum_{1\leq j<k\leq d}\left(\frac{\sum_{i=1}^{n'} x_{ij}'' x_{ik}''}{n'} - \frac{\sum_{i=1}^{n} x_{ij}x_{ik}}{n}\right)^2 \qquad (5.3)$$

subject to

$$0.99 \cdot p \cdot IL_1' \leq \frac{\sum_{j=1}^{d}\sum_{i=1}^{n'} \frac{|x_{ij}''-x_{C(i),j}|}{|x_{C(i),j}|}}{dn'} \leq 1.01 \cdot p \cdot IL_1' \qquad (5.4)$$

where $p > 0$ is a parameter and $C(i)$ is the original record nearest to the i-th masked record of X'' after optimization. Note that, in general, $C(i) \neq c(i)$, because in general $X'' \neq X'$.

5.2.3 A heuristic optimization procedure

To solve the minimization problem (5.3) subject to constraint (5.4), the following hill-climbing heuristic procedure has been devised:

[1] Distance is always measured over standardized variables.

Algorithm 13 (PostMaskOptim(X,X',p,q,TargetE))

1. *Standardize all variables in X and X' by using for both data sets the averages and standard deviations of variables in X.*
2. *Compute IL_1' between X and X' according to expression (5.2).*
3. *Let $TargetIL_1' := p \cdot IL_1'$.*
4. *Let $X'' := X'$.*
5. *Rank records in X'' according to their contribution to IL_1'. Let M be the subset of the $100q\%$ records in X'' contributing most to IL_1.*
6. *For each record i in X'', determine its nearest record $C(i)$ in X (use d-dimensional Euclidean distance).*
7. *Compute E, where E denotes the objective function in Expression (5.3).*
8. *While $E \geq TargetE$*
   (a) *Randomly select one value v of a record i_v in $M \subset X''$ and randomly perturb it to get v'. Replace v with v' in record i_v.*
   (b) *Recompute the nearest record $C(i_v)$ in X nearest to the updated i_v.*
   (c) *Let $PreviousIL_1' := IL_1'$.*
   (d) *Compute IL_1' between X and X''. To do this, use Expression (5.2) while replacing x_{ij}' by x_{ij}'' and $c(i)$ by $C(i)$.*
   (e) *Let $PreviousE := E$.*
   (f) *Recompute E (X'' has been modified).*

(g) *If* $E \geq previousE$ *then* $undo := true$.

(h) *If* $IL'_1 \notin [0.99 \cdot TargetIL'_1, 1.01 \cdot TargetIL'_1]$ *and* $|IL'_1 - TargetIL'_1| \geq |PreviousIL'_1 - TargetIL'_1|$ *then* $undo := true$.

(i) *If* $undo = true$ *then restore the original value* v *of record* i_v *and recompute the nearest record* $C(i_v)$ *in* X *nearest to* i_v.

9. *Destandardize all variables in* X *and* X'' *by using the same averages and standard deviations used in Step 1.*

Note that, by minimizing E, the algorithm above attempts to minimize the information loss IL'. No direct action is taken to reduce or control disclosure risk measures DLD' and ID', beyond forcing that IL'_1 should be in a pre-specified interval to prevent the optimized data set from being dangerously close to the original one. The performance of Algorithm 13 is evaluated *a posteriori*: once E reaches $TargetE$, the algorithm stops and yields an optimized data set for which IL', DLD' and ID' must be measured.

5.2.4 Computational results

The test microdata set no. 1 of [DDS02] was used. This microdata set was constructed using the Data Extraction System (DES) of the U.S. Census Bureau (`http://www.census.gov/DES`). $d = 13$ continuous variables were chosen and 1080 records were selected so that there were not many repeated values for any of the attributes (in principle, one would not expect repeated values for a continuous attribute, but there were repetitions in the data set).

In the comparison of [DMT01, DT01a], two masking methods were singled out as particularly well-performing to protect numerical microdata: rank swapping [Moo96] and multivariate microaggregation [DM02]. For both methods, the number of masked records is the same as the number of

p	q	$Score'$	IL'	DLD'	ID'	E
None	None	25.66	23.83	14.74	40.23	0.419
0.5	0.5	24.45	14.73	20.30	48.03	0.04
0.5	0.3	22.15	13.65	16.30	44.98	0.04
0.5	0.1	21.71	15.26	14.81	41.51	0.09

Table 5.1: Rank-swapping with parameter 14. First row, best $Score'$ without optimization; next rows, scores after optimization.

original records ($n = n' = 1080$). Several experiments have been conducted to demonstrate the usefulness of post-masking optimization to improve on the best (lowest) scores reached by rank swapping and multivariate microaggregation.

The first row of Table 5.1 shows the lowest $Score'$ reached by rank swapping for the test microdata set: the $Score'$ is 25.66 and is reached for parameter value 14 (see [DDS02]). The next rows of the table show $Score'$ reached when Algorithm 13 is used with several different values of parameters p (proportion between target IL'_1 and initial IL'_1) and q (proportion of records in M). The last column shows the value of the objective function E reached (for all rows but the first one, this is the $TargetE$ parameter of Algorithm 13). The $Score'$ is computed using Expression (5.1) and the values of IL', DLD' and ID' reached are also given in Table 5.1.

The first row of Table 5.2 shows the lowest $Score'$ reached by multivariate microaggregation for the test data set: the score is 31.86 and is reached for parameter values 4 and 10, that is, when four variables are microaggregated at a time and a minimal group size of 10 is considered (see [DDS02]). The next rows of the table show $Score'$ reached when Algorithm 13 is used with several different values of parameters p and q.

When looking at the results on rankswapped data (Table 5.1), we can

p	q	$Score'$	IL'	DLD'	ID'	E
None	None	31.86	22.48	22.14	60.34	0.122
0.5	0.5	26.96	14.16	21.06	58.54	0.008
0.5	0.3	27.39	14.74	21.29	58.80	0.008
0.5	0.1	28.03	14.94	21.83	60.38	0.008

Table 5.2: Multivariate microaggregation with parameters 4 and 10. First row, best $Score'$ without optimization; next rows, scores after optimization.

observe the following:

- There is substantial improvement of the $Score'$: 21.71 for post-masking optimization with $p = 0.5$ and $q = 0.1$ in front of 25.66 for the initial rankswapped data set.

- The lower q (*i.e.* the smaller the number of records altered by post-masking optimization), the better is $Score'$. In fact, $Score'$ for $q = 0.1$ is lower than for $q = 0.3, 0.5$ even if the target E for $q = 0.1$ is less stringent (higher) than for the other values of q.

- Post-masking optimization improves the score by reducing information loss IL' and hoping that disclosure risks DLD' and ID' will not grow. In fact, Table 5.1 shows that DLD' and ID' increase in the optimized data set with respect to the rankswapped initial data set. The lower q, the lower is the impact on the rankswapped initial data set, which results in a smaller increase in the disclosure risk. This small increase in disclosure risk is dominated by the decrease in information loss, hence the improved $Score'$.

The results on microaggregated data (Table 5.2) are somewhat different. The following comments are in order:

- Like for rankswapping, there is substantial improvement of $Score'$: 26.96 for post-masking optimization with $p = 0.5$ and $q = 0.5$ in front of 31.86 for the initial microaggregated data set.

- The higher q, the better is $Score'$. This can be explained by looking at the variation of IL', DLD' and ID'. Microaggregated data are such that there is room for decreasing IL' while keeping DLD' and ID' at the same level they had in the initial microaggregated data set. In this respect, we could interpret that, multivariate microaggregation being "less optimal" than rank swapping, we should not be afraid of changing a substantial number of values because this can still lead to improvement.