Modeling Projections in Microaggregation

Jordi Nin and Vicenç Torra

IIIA, Artificial Intelligence Research Institute
CSIC, Spanish National Research Council
Campus UAB s/n,
08193 Bellaterra,
Catalonia, Spain
{jnin,vtorra}@iiia.csic.es

Abstract

Microaggregation is a method used by statistical agencies to limit the disclosure of sensitive microdata. It has been proven that microaggregation is an NP-hard problem when more than one variable is microaggregated at the same time. To solve this problem in a heuristic way, a few methods based on projections have been introduced in the literature. The main drawback of such methods is that the projected axis is computed maximizing a statistical property (*e.g.*, the global variance of the data), disregarding the fact that the aim of microaggregation is to keep the disclosure risk as low as possible for all records.

In this paper we present some preliminary results on the application of aggregation functions for computing the projected axis. We show that, using the Sugeno integral to calculate the projected axis, we can reduce in some cases the disclosure risk of the protected data (when projected microaggregation is used).

Keywords: Microaggregation, Sugeno Integral, Statistical Disclosure Control.

1 Introduction

It is a common practice in all organizations to manage large volumes of confidential data. In many cases, data need to be transferred to third parties to be analyzed. In this case, privacy becomes an essential issue. Data has to be transferred but while statistics have to be preserved, confidential information has to be kept private. This is a typical problem, for instance, in national statistics offices.

Special efforts have been made to develop a wide range of protection methods [4]. These methods aim at guaranteeing an acceptable level of protection of the confidential data. Specific areas such as Privacy in Statistical Databases (PSD) tackle the problem of protecting confidential data in order to publicly release it, without revealing confidential information that could be linked to an specific individual or entity. Normally, Record linkage methods [15] are used to find these linkages.

Recently, microaggregation has emerged as one of the most promising protection methods. Microaggregation works as follows: given a data set with V variables, it builds small clusters of at least k elements and replaces the original values by the centroid of the cluster to which the element belongs. A certain level of privacy is ensured because k elements have an identical protected value.

Microaggregation techniques can be classified as *univariate microaggregation* (when the number V of variables is 1) or *multivariate microaggregation* (when $V > 1$). In the case of univariate microaggregation, there ex-

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 1384–1391
Torremolinos (Málaga), June 22–27, 2008

ist polynomial-time methods to obtain an optimal solution, for instance, the algorithm presented in [10] which works over a graph built from the sorted original data. On the other hand, optimal multivariate microaggregation was proved to be NP-hard [12]. Due to this, heuristic [5, 6] approaches have been proposed. In this paper, we focus on projected microaggregation, as in [5].

Projected microaggregation simplifies the multivariate microaggregation problem translating it into the univariate case. To do this, V variables are summarized / represented into a single value in a projected axis. Normally, this summarization is done using the Principal Component Analysis or the sum of Zscores (both methods are described in detail later on). The aim of both methods is to establish an order among records to apply an optimal univariate microaggregation algorithm.

In order to summarize several variables into a single value, aggregation functions [16] can be used. In this paper we propose replacing the use of projection methods in microaggregation by the use of methods based on aggregation functions. We will show that the trade-off between privacy and statistical utility achieved by microaggregation using the Sugeno integral [14] to summarize the variables is equal, better in many cases, than the traditional projected microaggregation methods.

The rest of the paper is organized as follows. In Section 2, we explain some preliminary concepts about projections and projected microaggregation. Then, in Section 3, we present our new approach to microaggregation, in Section 4 some preliminary results are described. Finally, Section 5 draws some conclusions and presents some future work.

2 Preliminaries

In this section, we explain some basic concepts about projections and their application to multivariate microaggregation. We also explain the score, a well-known measure to analyze (evaluate) protection methods. In relation to notation, we will assume in the rest of this paper that the values of the V variables for the n individuals (records) are stored in a matrix X of dimension $n \times v$.

2.1 PCP Projection

Formally Principal Component Projection (PCP) works as follows: let us assume that values of v attributes for n individuals are stored in a matrix X of dimension $n \times v$, where columns contain attributes and rows contain individuals. For the sake of simplicity, we will assume here that data is standardized (*i.e.*, the data has $\mu = 0$ and $\sigma = 1$, and so the covariance matrix is $S = 1/nX^TX$).

The first principal component is defined as the linear combination of the attributes which has the maximum variance. Therefore, this first principal component will be represented using a vector $z_1 = Xv_1$, for some vector v_1 with v components, to be found. Since the original values have $\mu = 0$, we have that z_1 also has $\mu = 0$, and its variance is

$$\frac{1}{n}z_1^Tz_1 = \frac{1}{n}v_1^TX^TXv_1 = v_1^TSv_1 \qquad (1)$$

Since S is positive-definite, the variance increases when the module of the vector v_1 does. For this reason, to find a concrete solution for the maximization of the expression (1), some constraint on the module of v_1 is needed. In this case, the search is limited to vectors v_1 with module 1 (i.e. $v_1^Tv_1 = 1$). This is equivalent to maximize the following expression, where a Lagrange multiplier has been added to the variance:

$$M = v^TSv_1 - \lambda(v_1^Tv_1 - 1) \qquad (2)$$

To maximize expression (2), the derivative with respect to the v_1 components must be made equal to 0.

$$\frac{\partial M}{\partial v_1} = 2Sv_1 - 2\lambda v_1 = 0 \qquad (3)$$

The solution for such equation is $Sv_1 = \lambda v_1$, which implies that v_1 is an eigenvector of the matrix S, and λ is its corresponding eigenvalue. To determine which eigenvalue of S is the right solution, Equation (3) is left-multiplied with v_1^T, leading to $v_1^T S v_1 = \lambda v_1^T v_1 = \lambda$.

Summing up, λ is the variance of z_1. Since the goal is to maximize the variance, λ is the largest eigenvalue of the matrix S, and its associate eigenvector v_1 defines the coefficients of the projection (PCP). Therefore, the final projected value is $PCP = \sum_{i=1}^{v} v_i x_i$.

The rationale of this process is to preserve, as maximum as possible, the total variance of the original variables in the projected one.

2.2 Zscores Projection

Given a record $(x_1, x_2, \ldots, x_v)$ in X, the sum of Zscores Projection is defined as

$$Z = \sum_{i=1}^{v} \frac{x_i - \mu_i}{\sigma_i}$$

where μ_i is the average and σ_i is the variance of the i-th variable, computed by taking into consideration all the records in X.

The rationale of this process is to sort the records taking into account the variance of all the variables.

2.3 Projected Microaggregation

As we have explained in the introduction, the main problem for extending optimal univariate microaggregation to the multivariate case is the sorting of multivariate data. One approach is to reduce the dimensionality of the problem. That is, to move from the case of several variables into 1 variable, by applying the projection methods explained before. Formally, the algorithms work as follows:

- Split the data set X into r sub-data sets $\{X_i\}_{1 \le i \le r}$, each one with v_i of the V attributes of the n records (V_i should define a partition of the V; i.e., $V_i \cap V_j = \emptyset$ for $i \neq j$ and $\cup_{i=1}^{r} V_i = V$, and $v_i = |V_i|$).
- For each sub-data set X_i:
    1. Apply a projection algorithm to the variables V_i in X_i, which results in an univariate vector p_i with n components (one for each record).
    2. Sort the components of p_i in increasing order.
    3. Apply to the sorted vector p_i the univariate optimal microaggregation.
    4. For each cluster resulting from the previous step, compute the v_i-dimensional centroid and replace all the records in the cluster by the centroid.

Depending on the projection method, we will obtain different methods of multivariate microaggregation. In this work we will use *PCP microaggregation* and *Zscores microaggregation.*

2.4 Protection Methods Evaluation

A protection method have to ensure a certain level of privacy (low disclosure risk). At the same time, since the goal is to allow third parties to perform reliable statistical computations over the protected information, a protection method must ensure that the protected data is still useful for statistical analysis (low information loss).

We thus have two inversely related values for the evaluation of a protection method: the *disclosure risk* (DR), which is the risk that an intruder obtains correct relations between the protected and the original data; and the *information loss* (IL) caused by the protection method. In the standard case, if one of these measures increases, the other one decreases.

Record linkage [15] has been used [3, 17] as a way to measure disclosure risk. Such methods try to model the situation where an intruder tries to link the protected data set with some records (original data) he/she has obtained from other sources. Of course, the more records that can be linked by means of record linkage methods, the more disclosure risk has the employed protection method. Some examples of record linkage methods are distance based and probabilistic record linkage.

Many approaches exist for computing information loss. We will use the measures defined in [3], where the authors calculate the average difference between some statistics computed on both the original and the protected microdata.

There are different ways to evaluate the quality of a data protection method, by taking into account these two values (DR and IL). We will use one of the most popular ones, the *score* [3]. This measure has been used in many works [11, 17]. The score is a simple and natural way to evaluate the trade-off between the information loss and the disclosure risk because it is defined as the average of these two values. Namely, score $= 0.5IL + 0.5DR$.

We use the definitions of IL and DR provided in [3]:

- **Information Loss (IL).** The overall IL is computed as $IL = 100\big(0.2\,IL_1 + 0.2\,IL_2 + 0.2\,IL_3 + 0.2\,IL_4 + 0.2\,IL_5\big)$, where

  (i) IL_1 is the mean absolute error of the original microdata X with respect to the protected data X'.

  (ii) IL_2 is the mean variation of the attribute average vectors.

  (iii) IL_3 is the mean variation of the attribute covariance matrices.

  (iv) IL_4 is the mean variation of the attribute variance vectors.

  (v) IL_5 is the mean variation of the attribute correlation matrices.

- **Disclosure Risk (DR).** Three alternatives are considered for measuring the disclosure risk. They are the two variations of record linkage explained before and the interval disclosure.

  Half weight is given to record linkage and the remaining half weight is given to interval disclosure. For record linkage, the average of the two methods is computed. Formally, this corresponds to $DR = 0.25 \cdot DLD + 0.25 \cdot PLD + 0.5 \cdot ID$, where

  (i) DLD, the *Distance based Linkage Disclosure risk*, is the average percentage of correctly linked records using distance based record linkage [13];

  (ii) PLD, the *Probabilistic Linkage Disclosure risk*, is the average percentage of correctly linked records using probabilistic record linkage [9]; and

  (iii) ID, the *Interval Disclosure risk*, is computed as the average percentage of original values falling into an interval defined around the corresponding masked value. The interval is defined as a percentage, between 1 per cent and 10 per cent, of the values.

  These three values are normally computed over the number of attributes that the intruder is assumed to know. In this paper, as we are comparing different microaggregation methods, for the sake of simplicity, we assume that the intruder knows different groups of attributes, instead of different groups of variables. Specifically, we assume that the intruder knows from two groups to all.

3 Modeling Projections

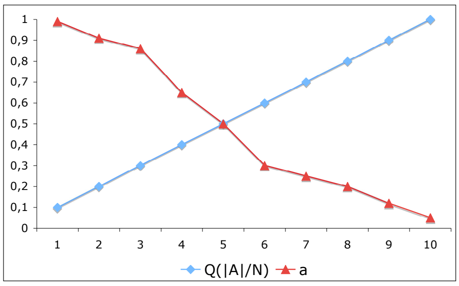
In this section, we explain some basics about the Sugeno integral and the quantifier used in the experiments. We also explain in detail the modifications of the standard projected microaggregation to include our modeling projection method.

3.1 Sugeno Integral

Now, we review a few definitions that are needed latter on. We start with the definition of the Sugeno integral [14] (see also [16]) in terms of a fuzzy quantifier.

Definition 1 *A function* $Q : [0, 1] \to [0, 1]$ *is a* regular monotonically non-decreasing fuzzy quantifier *(non-decreasing fuzzy quantifiers for short) if it satisfies: (i)* $Q(0) = 0$*; (ii)* $Q(1) = 1$*; (iii)* $x > y$ *implies* $Q(x) \geq Q(y)$*.*

Figure 1: Quantifier $Q(A)$



Definition 2 *Let $X := \{x_1, \ldots, x_N\}$ be a set of information sources, the Sugeno integral with respect to the measure $\mu(A) = Q(|A|/N)$ for $A \subseteq X$ is defined by:*

$$SI_Q(a_1, \ldots, a_N) = \max_{i=1}^{N} min(Q(i/N), a_{\sigma(i)})$$

where σ is a permutation such that $a_{\sigma(i)} \geq a_{\sigma(i+1)}$.

Figure 1 is a graphical representation of this integral with the quantifier $Q(x) = x$, the one used in the experiments.

3.2 Algorithm Description

As we have explained before, projected microaggregation defines a sorting criterion over the multivariate data. Traditional projected microaggregation methods build a projected axis to establish an order among records. We will do that using aggregation functions instead of building a projected axis.

The rationale of the approach is as follows. Decision makers use aggregation functions to evaluate a large number of projects. In this case, different criteria or expert opinions are considered in the aggregation process. Then, the decision maker decides which is the best project to invest his money using the summarized value provided by the aggregation function. In this kind of processes, decision maker are, in some way, sorting the projects using a summarized criteria / opinion. Following this idea, we propose to use aggregation functions over the records to be protected in order to compute a representative summarized value, and then, using such value, sort the records in the data set. Naturally, at that time, optimal univariate microaggregation methods can be applied.

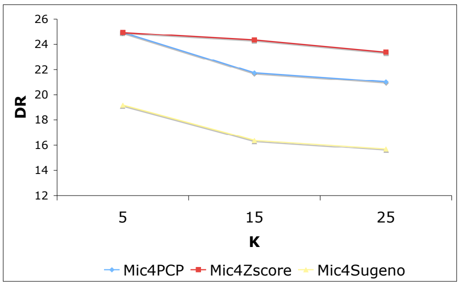
This new approach has many advantages with respect to the traditional projected approach. We underline the following ones.

- In projected methods, we need to compute some parameters. For instance, the sum of Zscores calculates the average and the variance of all the variables, PCP needs to solve an optimization problem. This is unnecessary using aggregation functions. Therefore, our new approach save execution time.
- Projected methods are not parametrizable. Using aggregation functions, one can define how data is sorted and, in some sense, protected.
- It is often the case that the projected values returned by a projection method are difficult to understand. Using aggregation functions one is able to understand the final summarized value for a concrete record.

In detail, the projected microaggregation algorithm works as follows.

- Split the data set X into r sub-data sets $\{X_i\}_{1 \leq i \leq r}$, each one with v_i attributes of the n records and according to a partition $\{V_i\}_i$ of the variables V (as before).
- For each sub-data set X_i:
  1. Compute an aggregation function with the variables V_i in X_i, which results in an univariate summarized vector p_i with n components (one for each record).
  2. Sort the components of p_i in increasing order.
  3. Apply to the sorted vector p_i the univariate optimal microaggregation.
  4. For each cluster resulting from the previous step, compute the v_i-dimensional centroid and replace all

Figure 2: Graphical representation of the DR of (PCP, Zscores and Sugeno) microaggregation using $v = 4$ and $k = 5, 15, 25$.



the records in the cluster by the centroid.

Depending on the aggregation function used, we will obtain different methods of modeling projection microaggregation. In this work we use the *Sugeno microaggregation*.

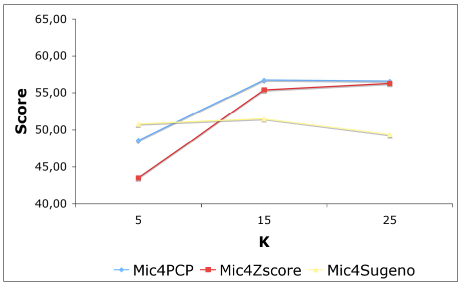
4 Experiments

We have implemented the projected microaggregation methods described in Section 2: PCP, Zscores and the new Sugeno-based ones. We have protected two different data sets with different instances of the three methods to compare them. These data sets were proposed in the CASC project [1] as the reference files to compare protection methods. The first microdata file, called Census, was extracted using the Data Extraction System (DES) from the U. S. Census Bureau [2]. The data set contains 1080 records with 13 attributes each (i.e., 14040 values to be protected). The second microdata file, called EIA, was extracted from the U.S. Energy Information Authority [8]. It contains 4092 records consisting of 10 attributes (i.e., 40920 values to be protected).

Figures 2 and 3 present in a graphical way disclosure risk (DR) and *score* for the microaggregation of the Census data set with $v = 4$ (the most protected configuration). We can observe that the Sugeno microaggregation algorithm obtains always the lowest DR and the best scores for $k = 15, 25$.

In Table 1 we present the scores as well as the unaggregated components. It can be seen that for some cases, Sugeno microaggregation leads to the lowest score (and the same for its components). E.g., the score obtained by the Sugeno microaggregation method is 49.39 in the Census data set with $v = 4$ and $k = 25$, while using PCP and Zscores microaggregation, the values are around 56. It is similar for the IL and DR components (IL, DLD, PLD and ID values). The values for Sugeno microaggregation are 83.08, 0.60, 0.37 and 30.90, respectively better than for PCP and Zscores microaggregation (89.02, 4.40, 3.38 and 42.86 for Zscores microaggregation; 92.17, 2.92, 1.71, 39.72 and 39.72 for PCP microaggregation).

Figure 3: Graphical representation of the Scores of (PCP, Zscores and Sugeno) microaggregation using $v = 4$ and $K = 5, 15, 25$.



Another interesting result can be observed analyzing Table 1: our approach never obtains the worst results (neither score values nor its components) in any case. This fact indicates that the results of our new approach are more independent of the data set than projected microaggregation methods.

5 Conclusions and Future Work

In this paper, we have presented a new family of microaggregation methods, which use aggregation functions as the sorting criteria. We have also shown that our method obtains better results than classical projected microaggregation methods, when we use the score from [3] to compare them.

As future work, we plan to further study the application of other aggregation functions and quantifiers to microaggregation and to other protection methods.

Acknowledgements

Partial support by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02) and by the Government of Catalonia (grant 2005-SGR-00093) is acknowledged. Jordi Nin wants to thank the Spanish Council for Scientific Research (CSIC) for his I3P grant.

References

[1] CASC: Computational Aspects of Statistical Confidentiality, European Project IST-2000-25069, http://neon.vb.cbs.nl/casc.

[2] Data Extraction System, U.S. Census Bureau, http://www.census.gov/

[3] Domingo-Ferrer, J., Torra, V., (2001), Disclosure control methods and information loss for microdata, 91-110 of [7].

[4] Domingo-Ferrer, J., Torra, V., (2001), A quantitative comparison of disclosure control methods for microdata, 111-133 of [7].

[5] Domingo-Ferrer, J., Mateo-Sanz, J. M. (2002) Practical data-oriented microaggregation for statistical disclosure control, IEEE TKDE, 14 189-201.

[6] Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J.M., Sebé, F. (2006) Efficient multivariate data-oriented microaggregation, The VLDB Journal, 15, 355-369.

[7] Doyle, P., Lane, J., Theeuwes, J., Zayatz, L., eds. (2001), Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies, Elsevier Science.

[8] U.S. Energy Information Authority, http://www.eia.doe.gov/

[9] Jaro, M. A. (1989) Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Society, 84:406, 414-420.

[10] Hansen, S., Mukherjee, S. (2003) A Polynomial Algorithm for Optimal Univariate Microaggregation. IEEE TKDE, 15:4 1043-1044.

[11] Medrano, P., Pont, J., Nin, J., Muntés, V., (2007), Ordered Data Set Vectorization for Linear Regression on Data Privacy, LNAI, Springer, 4617, 361-372.

[12] Oganian, A., Domingo-Ferrer, J. (2000) On the Complexity of Optimal Microaggregation for Statistical Disclosure Control, Statistical Journal of UNECE, 18, 4, 345-354.

[13] Pagliuca, D., Seri, G., (1999), Some results of individual ranking method on the system of enterprise accounts annual survey, Esprit SDC Project, MI-3/D2.

[14] Sugeno, M., (1974), Theory of fuzzy integrals and its application, Thesis, Tokyo Institute of Technology.

[15] Torra, V., Domingo-Ferrer, J., (2003), Record linkage methods for multidatabase data mining, Information Fusion in Data Mining, Springer, 101-132.

[16] Torra, V., Narukawa, Y. (2007) Modeling decisions: Information Fusion and Aggregation Operators, Springer.

[17] Yancey, W., Winkler, W., Creecy, R., (2002), Disclosure risk assessment in perturbative microdata protection, LNCS, Springer, 2316, 135-152.

A Score Computations

	EIA data set					Census data set				
Method	IL	DLD	PLD	ID	Score	IL	DLD	PLD	ID	Score
Mic2PCP05	13.90	2.94	6.91	70.04	25.69	80.96	12.93	5.70	42.60	53.46
Mic2PCP15	17.24	1.72	2.37	67.67	26.05	92.94	8.46	2.94	35.64	56.81
Mic2PCP25	19.98	1.42	1.58	67.21	27.17	84.77	6.61	1.94	32.93	51.69
Mic2Zscores05	4.27	25.36	36.08	89.26	32.13	81.57	16.78	7.85	48.27	55.93
Mic2Zscores15	5.08	21.92	34.37	87.85	31.54	98.05	12.96	6.19	44.33	62.50
Mic2Zscores25	5.52	21.05	35.06	87.33	31.61	100.92	12.85	4.83	42.90	63.40
Mic2Sugeno05	5.25	17.79	23.90	86.65	29.50	73.44	9.63	6.00	40.75	48.86
Mic2Sugeno15	6.24	14.14	19.11	85.08	28.55	79.39	4.85	4.63	34.02	49.39
Mic2Sugeno25	6.49	12.81	18.29	84.51	28.26	73.43	3.72	4.76	32.96	46.02
Mic3PCP05	16.08	2.47	2.69	62.79	24.38	57.72	10.15	5.71	43.48	41.71
Mic3PCP15	17.76	1.49	1.21	59.41	24.07	71.28	4.35	3.49	37.36	45.96
Mic3PCP25	18.49	1.31	0.90	58.49	24.14	72.49	4.07	2.65	35.51	45.96
Mic3Zscores05	13.24	6.40	9.26	72.31	26.66	60.98	14.44	13.67	50.63	46.66
Mic3Zscores15	15.30	3.79	5.50	69.35	26.15	75.21	9.38	10.46	45.71	51.51
Mic3Zscores25	15.73	3.21	5.02	68.65	26.06	79.38	7.47	9.04	44.20	52.80
Mic3Sugeno05	17.22	3.78	6.89	65.52	26.32	83.93	7.47	7.25	44.50	54.93
Mic3Sugeno15	21.31	1.74	3.21	61.22	26.58	122.52	3.55	5.52	39.47	72.26
Mic3Sugeno25	20.08	1.47	2.65	60.83	25.76	129.37	3.30	4.57	39.08	75.44
Mic4PCP05	18.25	4.23	4.81	73.22	28.56	72.23	6.48	3.06	45.12	48.59
Mic4PCP15	16.39	1.96	2.13	70.48	26.33	91.74	3.43	2.04	40.73	56.74
Mic4PCP25	17.27	1.93	1.91	69.66	26.53	92.17	2.92	1.71	39.72	56.59
Mic4Zscores05	13.91	5.21	8.50	78.73	28.35	62.04	11.71	7.04	40.50	43.49
Mic4Zscores15	21.79	2.71	4.40	77.41	31.14	86.47	5.60	4.21	43.77	55.40
Mic4Zscores25	21.66	2.35	3.89	76.76	30.80	89.20	4.40	3.38	42.86	56.29
Mic4Sugeno05	28.78	2.20	3.30	73.48	33.45	82.43	3.15	0.51	36.51	50.80
Mic4Sugeno15	35.97	0.71	1.03	70.86	35.92	86.64	0.83	0.28	32.19	51.51
Mic4Sugeno25	45.27	0.42	0.71	70.31	40.35	83.08	0.60	0.37	30.90	49.39

Table 1: Score of different microaggregation methods and parameterizations. Mic$i$var$j$ corresponds to microaggregation using variation var (either PCP, Zscores of Sugeno) with $v = i$ and $k = j$