Extending Microaggregation Procedures for Time Series Protection

Jordi Nin and Vicenç Torra

IIIA-CSIC
Campus UAB s/n
08193 Bellaterra (Catalonia, Spain)
jnin@iiia.csic.es
vtorra@iiia.csic.es

Abstract. Privacy is becoming a pervasive issue, as nowadays information is gathered by all kind of information systems.

In this paper we introduce a method for database protection in the case that the data under consideration is expressed in terms of time series. We propose the use of microaggregation for this purpose and extend standard microaggregation so that it works for this kind of data.

Keywords: privacy, masking methods, time series, microaggregation, clustering, time series distances.

1 Introduction

In the last years, the need for tools to ensure data privacy is increasing as people is more and more concerned with privacy issues. At the same time, there is an increasing demand of data by researchers and decision makers. Privacy preserving data mining and inference control [13] are to develop tools for data protection with the aim that protected data can be released for further study without compromising the privacy of data respondents.

Masking methods [13] are the specific tools that are used for data protection. Among all masking methods, perturbative ones are those that modify the original data so that the perturbed data avoids disclosure. Nevertheless, data perturbation might cause data to lose their utility. This is so because a *maximum* perturbation makes disclosure impossible but at the same time data is useless for any analysis. Instead, when no perturbation is applied, we have maximum data utility (only original data is published) but data permit disclosure. To measure all these aspects, some measures have been defined. They are the so-called measures for information loss (to evaluate in what extent data has lost its utility), and measures for disclosure risk. Besides, there are scores and other functions to combine or visualize these measures to evaluate the tradeoff between data utility and disclosure risk.

Most masking methods have been developed for standard databases. That is, databases in which records take values on a set of variables. Information loss and disclosure risk measures have been defined for such kind of records.

S. Greco et al. (Eds.): RSCTC 2006, LNAI 4259, pp. 899–908, 2006.

Due to the increasing amount of information currently available, and due to the increasing rate on data storage, data is no longer a static object but it has a temporal component. Therefore, it is of interest the study of masking methods for temporal data protection. That is, the protection of time series. Some research has been done in this line. See *e.g.* [1].

In this paper we develop a new method for time series protection. The method is based on microaggregation [2], a tool for data protection that has a good performance in standard numerical data with respect to information loss and disclosure risk measures as shown in [4]. Microaggregation is one of the standard tools for database protection commonly in use in National Statistical Offices (see *e.g.* [6]).

Microaggregation requires the definition of a distance on the data. In microaggregation for standard data such distance is usually the Euclidean distance. In the case of time series, several distance on time series can be considered. In this paper we propose and use two different distances: the short time series distance and Euclidean distance. We can see in section 2.3 the formal definition of these distances and in section 3 we will study how the choice of the distance affects microaggregation results.

The structure of the paper is as follows. In Section 2 we describe some preliminaries required in the rest of the paper. In particular, this section describes standard microaggregation and some distance functions for time series. Then, in Section 3, we propose our method for time series protection. In Section 4 we describe the experiments done. The paper finishes with some conclusions and some research lines for future work.

2 Preliminaries

This section presents standard microaggregation and some results on time series that are needed in the rest of this work.

2.1 Microaggregation

Microaggregation is a masking method for database protection. From the procedural point of view, it works as follows:

1. Clusters are built from the original data. Each cluster should contain at least k records.
2. A representative is built for each cluster.
3. Original records are replaced by the corresponding representatives.

The fact of having clusters containing at least k records is to ensure data privacy. Note that after microaggregation is applied, we will have at least k records indistinguishable for each cluster (with respect to the variables clustered).

This method was originally defined on numerical data. Then, it was extended to categorical data [11]. The method can be formally defined in terms of an optimization problem. Nevertheless, it was proven that finding the optimal solution

of such optimization problem is an NP-problem [9]. Therefore, some research has been done to find heuristic approaches. One of the methods is the so-called MDAV (Maximum Distance Average Vector)-generic algorithm [5]. This method is described in the next section.

2.2 MDAV-Generic Algorithm

The MDAV-generic algorithm is an heuristic algorithm for clustering records in a dataset R. Each cluster is constrained to have at least k records. The algorithm is as follows:

Algorithm (MDAV-generic) (R: dataset, k: integer) is
1. while ($|R| > k$) do
   - **(1.a)** Compute the average record $\tilde{x}$ of all records in R
   - **(1.b)** Consider the most distant record x_r to the average record $\tilde{x}$ using and appropriate distance
   - **(1.c)** Form a cluster around x_r. The cluster contains x_r together with the $k-1$ closest records to x_r
   - **(1.d)** Remove these records from dataset R
2. if ($|R| > k$) then
   - **(2.a)** Find the most distant record x_s from record x_r (from step 1.b)
   - **(2.b)** Form a cluster around x_s. The cluster contains x_s together with the $k-1$ closest records to x_s
   - **(2.c)** Remove these records from dataset R
3. end if
4. end while
5. form a cluster with the remaining records

This algorithm is generic and it can be applied to different kind of data using appropriate definitions of distance and average. That is, we need to formulate what the *most distant record* means, and which are the *closest records* of a given record. Additionally, we need to define *the average record* of a set of records. This average record is needed in step (1.a) and later to mask the original data. Recall that we need to build a representative for each cluster and then replace each original record by the corresponding representative.

In [2] this method is applied to numerical data, using the Euclidean distance for computing the distance between records and the arithmetic mean to compute the average. In [5] this method was extended to categorical data using appropriate functions.

2.3 Time Series

Now we turn into the problem of defining distances for time series. We focus on numerical time series. Formally speaking, a time series is defined by pairs $\{(v_k, t_k)\}$ for $k = 1, \ldots, N$ where t_k corresponds to the temporal variable and v_k is the variable that depends on time (dependent variable). Naturally, $t_{k+1} > t_k$. Stock prices are examples of time series, as they depend on time.

There are different methods to compute the distances between time series. We can use distance based on raw values of equal or unequal length, cross-correlation matrix, vectors of feature-value pairs, probability-based functions and so on. See [12] for more details.

For our experiments we have implemented two different distance functions, Euclidean distance and the Short Time Series distance proposed by Möller-Levet et al. in [7].

Let x and v be two N-dimensional time series. Let x be defined by the pairs $\{(x_k, t_k)\}$ for $k = 1, \ldots, N$ and let v defined by the pairs $\{(x_k, t_k)\}$ for $k = 1, \ldots, N$. Note that we assume that the two time series under consideration have exactly the same length and that they are aligned. That is, the temporal component in both series is exactly the same. Also, note that we use x and v for both denoting the time serie and the dependent variable.

Then, the Euclidean distance is defined by:

$$d_{EU}(x, v) = \sqrt[2]{\sum_{k=1}^{N} (x_k - v_k)^2}$$

The short time series distance, STS distance in short, is defined as

$$d_{STS}(x, v) = \sqrt[2]{\sum_{k=1}^{N} \left(\frac{v_{k+1} - v_k}{t_{k+1} - t_k} - \frac{x_{k+1} - x_k}{t_{k+1} - t_k} \right)^2}$$

3 Time Series Microaggregation

To specialize the MDAV-generic algorithm for time series we need to make the distances concrete, and then consider the particular average functions. We have implemented the algorithm described in Section 2.2 with the following parameterizations:

Distance functions: We have used Euclidean and STS distances: d_{EU} and d_{STS} as defined in Section 2.3.

Average: We have used a kind of arithmetic mean. The mean has been defined component-wise. That is, given the set $V = \{v^j\}_{j=1,\ldots,J}$ with time series v^j for $j = 1, \ldots, J$, each one with v_k^j, we define $\tilde{v}_k = (1/J) \sum_{j=1,\ldots,J} v_k^j$.

Therefore, we have applied the MDAV-generic algorithm where $\tilde{x}$ is the average of all records (time series) in R. These distance functions have been used to determine the most distant records as well as the closest records to a given record r.

Different distance functions cause the microaggregation algorithm to compute different clusters. While Euclidean distance makes clusters based on the distance between data components, the STS distance makes clusters based on the shape of the time series. This is illustrated in the following example.

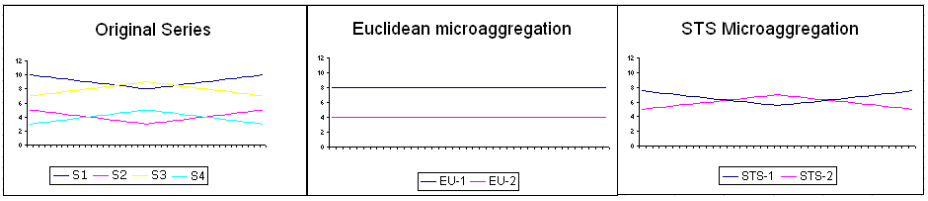

Fig. 1. Graphical representation of distance function selection

Example 1. Figure 1 (left) illustrates this problem: 4 series are to be microaggregated. The results of microaggregating these 4 series into 2 clusters using Euclidean and STS distances are given, respectively, in middle and right figures of Figure 1. It can be seen that the Euclidean distance gathers together the most near series although they have different shapes (and, thus, the outcomes are just lines but that mainly keep the original values). Instead, the STS distance gathers series according to shapes (and, thus, the outcomes keep such shapes but not the original position of the series).

In this example, we have used point-wise average for computing the representative of each cluster.

According to this, in the step of selecting the distance function, we have the opportunity to model how the microaggregation procedure makes the clusters and decide which information is the most important to be kept in the final protected model.

4 Experiments

We have applied our method to a data set consisting on several time series. We describe below the data considered. We have applied our algorithm to these data, testing different values of k. In particular, we have used: $k \in \{2, 3, 6, 9, 12\}$.

4.1 Data

The data under consideration correspond to the Stock Exchange information of the thirty five most important Spanish companies. These companies are ranked in the so-called Ibex35 stock market. We have got historical information about company prices in the Ibex35 stock market for about a year from [10]. This information is publicly available.

We have obtained thirty five files, one for each company ranked in the Ibex35 stock market. These files have been processed to obtain a new file with the opening prices of the companies. In this way, we have 35 time series. Tables 1 and 2 give details on the 35 series considered for applying microaggregation.

The selection of economic data was done for two main reasons. First, data can be obtained easily and free in electronic markets. Second, economic information clearly corresponds to a time series structure, so it is a good example for our method.

Table 1. Original data for one company (Abertis) in Ibex35. Thirty five files with this structure were downloaded.

Abertis	opening value	maximum value	minimum value	closing value	volume
06-21-2005	19.94	19.94	19.81	19.89	841
06-22-2005	19.95	20.10	19.88	19.99	799
06-23-2005	19.90	20.05	19.90	19.95	708
...	...	...	...	...	...
04-27-2006	20.93	20.97	20.67	20.95	2507
04-28-2006	21.00	21.00	20.65	20.92	2442

Table 2. 35 time series corresponding to the opening prices of all companies in the Ibex35 stock market

Company name	business	05-03-2005	05-04-2005	05-05-2005	...	04-27-2006	04-28-2006
Abertis	Building firm	19,94	19,95	19,90	...	20,93	21,00
Acciona	Building firm	76,85	79,00	81,40	...	139,65	134,65
Acerinox	Steel firm	11,75	11,58	11,55	...	13,72	13,29
ACS	Building firm	22,05	22,38	22,41	...	33,14	32,87
Altadis	Tobacco firm	33,85	34,28	34,68	...	37,75	37,25
Antena3 TV	Private television	17,25	17,11	16,92	...	21,84	21,67
Arcelor	Steel firm	16,64	16,47	16,22	...	34,03	33,13
Banco Popular	Bank	10,00	9,98	9,94	...	12,12	12,06
Banco Sabadell	Bank	21,12	21,14	21,29	...	28,47	28,87
Bankinter	Bank	42,63	42,65	42,70	...	55,80	54,90
BBVA	Bank	12,87	12,77	12,79	...	17,44	17,32
Cintra	Building firm	8,97	9,00	9,05	...	10,88	10,85
ENAGAS	Energy firm	13,63	13,73	13,90	...	17,93	17,57
Endesa	Energy firm	18,18	18,37	18,45	...	26,51	26,35
F.C. Contratas	Building firm	46,67	46,32	45,93	...	64,10	64,25
Ferrovial	Building firm	50,60	52,00	51,95	...	65,15	65,00
Gamesa	Aeronautics industry	11,27	11,35	11,42	...	17,41	16,93
Gas Natural	Energy firm	23,30	23,57	23,45	...	24,40	24,31
Iberdrola	Energy firm	21,40	21,63	21,59	...	25,89	25,80
Iberia	AirLine	2,47	2,49	2,50	...	2,24	2,22
Inditex	Textile firm	21,97	21,94	21,70	...	32,85	32,41
Indra	New Technology firm	15,60	15,65	15,85	...	16,73	16,41
Metrovacesa	Building firm	49,57	50,75	52,65	...	72,95	71,90
NH Hoteles	Hotel firm	10,89	11,06	11,03	...	14,28	14,35
Prisa	Press firm	16,06	16,19	16,12	...	14,65	14,69
R. E. Española	Energy firm	20,90	21,24	21,15	...	27,92	27,87
Repsol YPF	Energy firm	21,35	21,26	21,28	...	24,02	23,77
SCH	Bank	9,52	9,58	9,59	...	12,15	12,19
Sogecable	Private television	30,23	30,24	30,55	...	30,78	30,20
Telecinco	Private television	19,42	19,48	19,30	...	20,70	20,63
Telefónica	Telecom firm	13,54	13,51	13,51	...	12,77	12,74
Telefónica Móvil	Telecom firm	8,76	8,75	8,80	...	10,52	10,46
TPI	Telecom firm	6,90	7,09	7,05	...	8,90	8,82
Unión Fenosa	Energy firm	24,04	24,40	24,59	...	31,00	30,70
Vallehermoso	Building firm	19,50	19,93	19,37	...	27,51	27,44

Before applying microaggregation, time series were standardized to avoid any scale problems. This standardization step consists in normalizing the data values using the mean and the standard deviation. We have calculated the mean and the standard deviation for all values in all time series.

4.2 Microaggregation and Results

In Figure 2 we can see all time serials of the Ibex35 stock market and in Figures 3 and 4 we can see the cluster centroids when we use, respectively, Euclidean and STS distances. In these figures the cluster size is fixed to six series per cluster.

The clusters obtained with the Euclidean distance and cluster size fixed to six are as follows:

Cluster 1: Abertis, Antena 3 TV, Enagas, Endesa, Indra and Telecinco
Cluster 2: Acciona, Altadis, Bankinter, F.C. Contratas, Ferrovial and Metrovacesa
Cluster 3: ACS, Inditex, Red Eléctrica Española, Repsol YPF, Sogecable and Unión Fenosa
Cluster 4: Arcelor, Banco Sabadell, Gas Natural, Iberdrola and Vallehermoso
Cluster 5: Banco Popular, Cintra, Iberia, SCH, Telefónica Móvil and TPI
Cluster 6: BBVA, Gamesa, NH Hoteles, Prisa and Telefónica

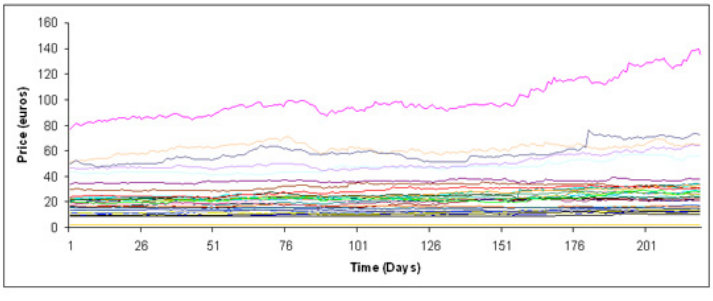

Fig. 2. Graphical representation of Ibex35 time series

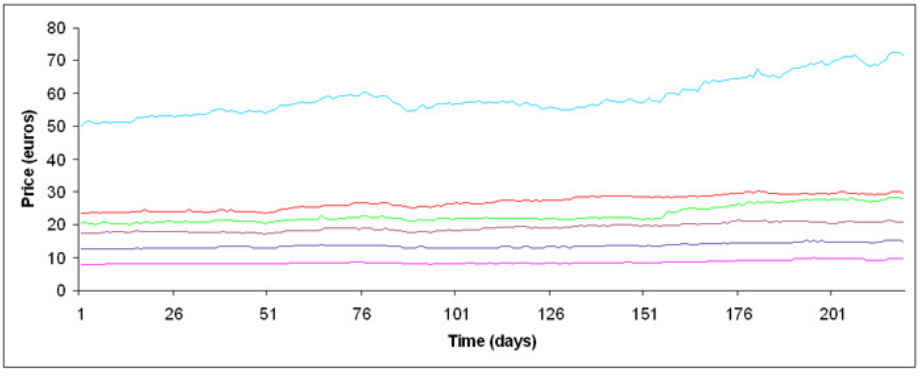

Fig. 3. Graphical representation of Euclidean distance clustering

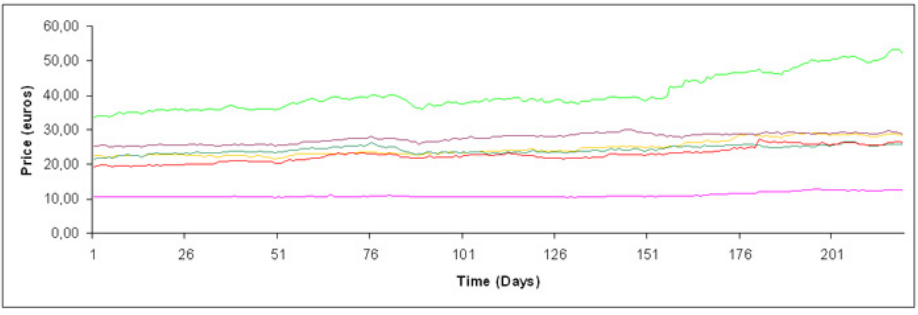

Fig. 4. Graphical representation of STS distance clustering

If we compare these results with Table 2, we can see that the companies in the first cluster have the lowest opening prices, the prices of these companies are around seventeen during June 2005 and around twenty-one during April 2006. In the second cluster, on the other side, we have companies with the highest opening prices, around fifty three during June 2005 and seventy-one during April 2006. The remaining clusters are between these opening values.

The clusters obtained using the STS distance and cluster size fixed to six are as follows:

Cluster 1: Abertis, ACS, Arcelor, F.C. Contratas and Vallehermoso
Cluster 2: Acerinox, Enagas, Ferrovial, Indra, Prisa and Red Eléctrica Española
Cluster 3: Altadis, Inditex, Repsol YPF, Sogecable and Telecinco
Cluster 4: Antena 3 TV, Bankinter, Endesa, Gamesa, Gas Natural and Iberdrola
Cluster 5: Banco Popular, Banco Sabadell, Iberia, Telefónica, Telefónica Móvil and TPI
Cluster 6: BBVA, Cintra, Metrovacesa, NH Hoteles, SCH and Unión Fenosa

In this case (see Table 2), clusters are not based on the opening prices but on the business type. If we observe the first cluster we notice that five of the six companies in the cluster are construction firms and if we check the fourth cluster we take into account that three companies are energy firms or in the fifth cluster three companies are telecommunications firms. On the remaining clusters we can find the same effect with two or more companies.

This effect in STS distance is possible because stock markets have been affected for social or external conditions like the price of money or fuel and all companies with a similar business have similar trends during a certain time.

From these results we can say that Euclidean distance measures differences between time serial values, and this distance benefits time series with closer sample values. Meanwhile STS distance measures difference between trends, this measure clusters time series with respect to their closer *shape*.

5 Conclusions and Future Work

In this paper, we have introduced a new method for protection of time series based on microaggregation. We have applied our approach using two different distances for time series. In particular, a distance based on the Euclidean distance and the STS distance. We have applied our approach to a data set defined in terms of time series.

The comparison of the two distances shows that the Euclidean distance gathers time series with similar values, while the STS one focuses on the *shape* of the series instead of the values themselves. This corresponds to the effect illustrated in Figure 1 and described in Example 1.

As future work we include the analysis of the method with respect to information loss and disclosure risk measures (some preliminary results can be found in [8]). These measures are required to properly evaluate the performance of the new methods and compare different approaches.

Although not analysed in this paper, the procedure for computing the representative of a cluster is also a relevant point. Further work is also needed in this direction.

Acknowledgements

This work was partly funded by the Spanish Ministry of Education and Science under project SEG2004-04352-C04-02 "PROPRIETAS".

References

1. Abowd, J. M., Woodcock, S. D., (2001), Disclosure Limitation in Longitudinal Linked Data, in P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. V. Zayatz (Eds), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland, 215-277.
2. Domingo-Ferrer, J., Mateo-Sanz, J. M., (2002), Practical data-oriented microaggregation for statistical disclosure control, IEEE Trans. on Knowledge and Data Engineering, 14 189-201.
3. Domingo-Ferrer, J., Torra, V., (2001), Disclosure methods and information loss for microdata, in P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. V. Zayatz (Eds), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland, 91-110.
4. Domingo-Ferrer, J., Torra, V., (2001), A quantitative comparison of disclosure methods for microdata, in P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. V. Zayatz (Eds), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland, 111-133.
5. Domingo-Ferrer, J., Torra, V., (2005), Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation, Data Mining and Knowledge Discovery, 11 195-212
6. Felso, F., Theeuwes, J., Wagner, G. G., (2001), Disclosure Limitation Methods in Use: Results of a Survey, in P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. V. Zayatz (Eds), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland, 17-42.

7. Möller-Levet, C. S., Klawonn, F., Cho, K.-H., Wolkenhauer, O., (2003), Fuzzy clustering of short time series and unevenly distributed sampling points, Proceedings of the 5th International Symposium on Intelligent Data Analysis, Berlin, Germany, August 28-30, 2003.
8. Nin, J., Torra, V.,(2006), Distance based re-identication for time series, Analysis of distances, Proc. Privacy in Statistical Databases (PSD 2006), in press
9. Oganian, A., Domingo-Ferrer, J., (2001), On the complexity of microaggregation, in Second Eurostat-UN/ECE Joint Work Session on Statistical Data Confidentiality, Skopje, Macedonia, March.
10. Stock Exchange web, Sabadell Bank, http://www.bsmarkets.com/
11. Torra, V., (2004), Microaggregation for categorical variables: a median based approach, Proc. Privacy in Statistical Databases (PSD 2004), Lecture Notes in Computer Science, 3050 162-174.
12. Warren Liao, T., (2005), Clustering of time series data - a survey, Pattern Recognition, 38 1857-1874.
13. Willenborg, L., de Waal, T., (2001), Elements of Statistical Disclosure Control, Lecture Notes in Statistics, Springer-Verlag.