Supervised Learning Using Mahalanobis Distance for Record Linkage

Daniel Abril
IIIA-CSIC
Campus UAB, Bellaterra (Spain)
dabril@iiia.csic.es

Vicenç Torra
IIIA-CSIC
Campus UAB, Bellaterra (Spain)
vtorra@iiia.csic.es

Guillermo Navarro-Arribas
DEIC-UAB
Campus UAB, Bellaterra (Spain)
guillermo.navarro@uab.cat

Summary

In data privacy, record linkage is a well known technique used to evaluate the disclosure risk of protected data. Mainly, the idea is the linkage between records of different databases, which make reference to the same individuals. In this paper we introduce a new parametrized variation of record linkage relying on the Mahalanobis distance, and a supervised learning method to determine the optimum simulated covariance matrix for the linkage process. We evaluate and compare our proposal with other studied parametrized and not parametrized variations of record linkage, such as weighted mean or the Choquet integral, which determines the optimal fuzzy measure.

Keywords: data privacy, record linkage, disclosure risk, Mahalanobis distance , fuzzy measure, Choquet integral.

1 Introduction

Record linkage is the process of finding quickly and accurately two or more records distributed in different databases (or data sources in general) that make reference to the same entity or individual. This term was initially introduced in the public health area by [9], when files of individual patients were brought together using name, date-of-birth and other information. In the following years, this idea was developed in [17, 16, 11], and nowadays it is a popular technique used by statistical agencies, research communities and corporations. Record linkage is one of the existing preprocessing techniques used for data cleaning [15, 24], and it is also used to control the quality of the data [2]. For example, data sources could be analyzed to deal with dirty data like duplicate records [10], data entry mistakes, transcription errors, lack of standards for recording data fields, etc. Moreover, it is nowadays a popular technique employed to integrate different data sets that provide information regarding to the same entities [6, 4].

In the last years, record linkage techniques have also emerged in the data privacy context. Many governments agencies and companies need to collect and analyze sensitive data about individuals. So, it is fundamental to provide security to statistical databases against disclosure of confidential information. Privacy preserving data mining [1] and Statistical Disclosure Control [23] research on methods and tools for ensuring the privacy of this data. Record linkage permits the evaluation of disclosure risk of protected data [19, 25]. By identifying links between the protected data set and the original one, we can evaluate the re-identification risk of the data by an intruder. For example [7], it defines a score using the combination of disclosure risk techniques, to evaluate the risk of re-identification, and another method, which readily quantified the information loss of a protected data set using analytical measures (either generic or data-use-specific).

In this paper we introduce a new distance based record linkage for data privacy based on the Mahalanobis distance [14]. It calculates distances taking into account the covariance among the variables. Moreover, we present a supervised learning approach adapted to this distance. It learns the covariance matrix of the distance, so that the linkage between the two data sets is maximized. The approach also gives us the relevance of single variables and each pair of them in the linkage process. In this paper we do a comparison between the proposed method and others non-supervised, such as arithmetic mean and the Mahalanobis distance [19] and, also with other supervised variations based on weighted mean [22] and the Choquet integral.

The outline of this paper is as follows. In section 2, we review some concepts needed in the rest of the paper. In section 3, we describe the supervised learning approach for distance based record linkage. The evaluation of the method is introduced in section 4. Finally, Section 5 presents the conclusions of the paper.

2 Preliminaries

In this section we review some ideas and definitions that are needed to understand the rest of the paper. We explain some ideas of the record linkage in the data privacy area and how the data sets are.

A dataset X can be viewed as a matrix with n rows (*records*) and V columns (*attributes*), where each row refers to a single individual. The attributes in a dataset can be classified in two different categories:

- *Identifiers*: attributes that can identify an individual unambiguously, e.g., the passport number.
- *Quasi-identifiers*: attributes that are not able to identify a single individual when they are used alone. However, when combining several of them, they can unequivocally identify it. Among the quasi-identifier attributes, we distinguish between confidential (X_c) and non-confidential (X_{nc}), depending on the kind of information that they contain. An example of non-confidential quasi-identifier attribute would be the zip code, while a confidential quasi-identifier might be the salary.

Before releasing the data, a protection method ρ is applied, leading to a protected dataset X'. This protection method will protect the non-confidential quasi-identifiers, $X'_{nc} = \rho(X_{nc})$. To ensure the privacy the identifiers are eiher removed or encrypted and the confidential quasi-identifiers are not modified because are the interesting for third parties. Then, everybody can see the protected data set, $X' = X'_{nc}||X_c$. This scenario, first used in [7] to compare several protection methods and then, adopted in other works like [25].

In data privacy, record linkage can be used to reidentify individuals between the protected dataset and a part or the whole original dataset as an evaluator of disclosure risk. There are two extensively used approaches of record linkage to evaluate the disclosure risk of protected data. The **Probabilistic record linkage (PRL)** [12] and the **Distance based record linkage (DBRL)** [18], which links each record a to the *closest* record in b, by means of a distance function.

The work in this paper is focused on distance based record linkage, which is further described below.

2.1 Distance-based record linkage

The main point in distance-based record linkage is in the definition of a distance. Nevertheless, different distances can be defined, each obtaining different results. Different distances have been considered and tested in the literature. We review two of them that are used in this work.

We will use $V_1^X, \ldots, V_n^X$ and $V_1^Y, \ldots, V_n^Y$ to denote the set of variables of file X and Y, respectively. Using this notation, we express the values of each variable of a record a in X as $a = (V_1^X(a), \ldots, V_n^X(a))$ and of a record b in Y as $b = (V_1^Y(b), \ldots, V_n^Y(b))$. $\overline{V_i^X}$ corresponds to the mean of the values of variable V_i^X.

DBRL: The Euclidean distance is used for attribute-standardized data. Accordingly, the distance between two records a and b is defined by:

$$d(a,b)^2 = \sum_{i=1}^{n} \left(\frac{V_i^X(a) - \overline{V_i^X}}{\sigma(V_i^X)} - \frac{V_i^Y(b) - \overline{V_i^Y}}{\sigma(V_i^Y)} \right)^2$$

DBRLM: Distance based record linkage using the Mahalanobis distance is as follows:

$$d(a,b)^2 = (a-b)'\Sigma^{-1}(a-b)$$

where, $\Sigma = [Var(V^X) + Var(V^Y) - 2Cov(V^X, V^Y]$ and $Var(V^X)$ is the variance of attributes V^X, $Var(V^Y)$ is the variance of attributes V^Y and $Cov(V^X, V^Y)$ is the covariance between attributes V^X and V^Y. Note that if the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean.

3 Supervised learning for record linkage

In this paper we focus on the utilization of parametrized distances, which used together the supervised learning they give us the best combination of the parameters to obtain the best reidentification between records of original and protected data. To that end, we first introduce a parametrized version of the Mahalanobis distance, an then, we present the supervised method that we use to determining the best weights, which in the Mahalanobis distance is a matrix that simulates a covariance matrix.

3.1 A parametric distance for record linkage

It is well known that the multiplication of the Euclidean distance by a constant will not change the results of any record linkage algorithm. Due to this, we can express the distance DBRL given in Section 2.1 as a weighted mean of the distances for the attributes.

In a formal way, we redefine DBRL as follows:

$$d(a,b)^2 = \sum_{i=1}^{n} \frac{1}{n} \left(\frac{V_i^X(a) - \overline{V^X}_i(a)}{\sigma(V_i^X)} - \frac{V_i^Y(b) - \overline{V^Y}_i(b)}{\sigma(V_i^Y)} \right)^2$$

Now, defining

$$d_i(a,b)^2 = \left(\frac{V_i^X(a) - \overline{V^X}_i(a)}{\sigma(V_i^X)} - \frac{V_i^Y(b) - \overline{V^Y}_i(b)}{\sigma(V_i^Y)} \right)^2$$

we can rewrite this expression as

$$d(a,b)^2 = AM(d_1(a,b)^2, \ldots, d_n(a,b)^2),$$

where AM is the arithmetic mean $AM(c_1, \ldots, c_n) = \sum_i c_i/n$.

In general, any aggregation operator $\mathbb{C}$ [20] might be used:

$$d(a,b)^2 = \mathbb{C}(d_1(a,b)^2, \ldots, d_n(a,b)^2).$$

From this definition, it is straightforward to consider weighted versions of the DBRL. That is as follows.

Definition 1 *Let $p = (p_1, \ldots, p_n)$ be a weighting vector (i.e., $p_i \geq 0$ and $\sum_i p_i = 1$). Then, the weighted distance is defined as:*

$$d^2WM_p(a,b) = WM_p(d_1(a,b)^2, \ldots, d_n(a,b)^2),$$

where $WM_p = (c_1, \ldots, c_n) = \sum_i p_i \cdot c_i$.

Another aggregation operator used is the Choquet integral (Definition 2). From a definitional point of view, its main difference with the previous tool is its use of fuzzy measures. In this way, this last operator expresses new information like redundancy, complementariness, and interactions among the variables, which are not reflected in weighted mean. Therefore, tools that use fuzzy measures to represent background knowledge permit the consideration of variables that are not independent.

Definition 2 *Let μ be an unconstrained fuzzy measure on the set of variables V, i.e. $\mu(\emptyset) = 0$, $\mu(V) = 1$, and $\mu(A) \leq \mu(B)$ when $A \subseteq B$ for $A \subseteq V$, and $B \subseteq V$. Then, the Choquet integral distance is defined as:*

$$d^2CI_\mu(a,b) = CI_\mu(d_1(a,b)^2, \ldots, d_n(a,b)^2),$$

where $CI_\mu(c_1, \ldots, c_n) = \sum_{i=1}^{n} (c_{s(i)} - c_{s(i-1)})\mu(A_{s(i)})$, given that $c_{s(i)}$ indicates a permutation of the indexes so that $0 \leq c_{s(1)} \leq \ldots \leq c_{s(i-1)}$, $c_{s(0)} = 0$, and $A_{s(i)} = \{c_{s(i)}, \ldots, c_{s(n)}\}$.

Now that we have briefly explained two existing parametrized distances, we present a novel approach relying on the Mahalanobis distance. To do so, firstly, we have to compute the normalized difference between two records $a \in X$ and $b \in Y$, in the following way:

$$d_i(a,b) = \left(\frac{V_i^X(a) - \overline{V_i^X}}{\sigma(V_i^X)} - \frac{V_i^Y(b) - \overline{V_i^Y}}{\sigma(V_i^Y)} \right)$$

Then, using it as an aggregation operator:

Definition 3 *Let Σ be an nxn weighting matrix, which simulates a covariance matrix. Then, the Mahalanobis distance is defined as:*

$$d^2MD(a,b) = MD_\Sigma(d_1(a,b), ..., d_n(a,b))$$

where $MD_\Sigma(c_1, ..., c_n) = (c_1, ..., c_n)^T \Sigma^{-1} (c_1, ..., c_n)$.

Comment that Σ, is a symmetric matrix. Then, the diagonal of the matrix expresses the relevance of each single variable in the reidentification process, whereas the up or down triangle values of the matrix are the weights that evaluates the interactions between each pair of variables.

The interest of these variations is that we do not need to assume that all the attributes are equally important in the re-identification. This would be the case if one of the attributes is a key-attribute, e.g. an attribute where $V_i^X = V_i^Y$. In this case, the corresponding weight would be assigned to one, and all the others to zero. Such an approach would lead to 100% of reidentifications. Note that in Definition 2 and 3 the interaction of different variables is taken into account by the fuzzy measure in contrast to Definition 1 which it can only weight the variables individually.

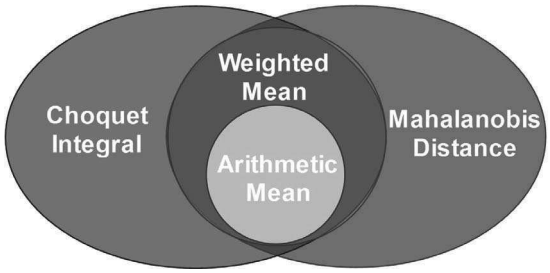

Figure 1: Distances classifications

Figure 1 shows the classification of the different distances that we have explained in this paper. As you can see arithmetic mean is a special case of weighted mean and at the same time this last is also a shared special case between the Choquet integral and the Mahalanobis distance, more details in [21].

3.2 Determining the optimal weights

For the sake of simplicity, we presume that each record of X, $a_i = (V_1^X(a_i), \ldots, V_N^X(a_i))$, is the protected record of Y, $b_i = (V_1^Y(b_i), \ldots, V_N^Y(b_i))$. That is, files are aligned. Then, if $V_k(a_i)$ represents the value of the kth variable of the $i$threcord, we will consider the

sets of values $d(V_k(a_i), V_k(b_j))$ for all pairs of records a_i and b_j.

Then, record i is correctly linked using an aggregation operator $\mathbb{C}$ when the aggregation of the values $d(V_k(a_i), V_k(b_i))$ for all k is smaller than $d(V_k(a_i), V_k(b_j))$ for all $i \neq j$. That is,

$$\begin{aligned}\mathbb{C}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) < \\ \mathbb{C}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j)))\end{aligned} \quad (1)$$

for all $i \neq j$. Then, the optimal performance of record linkage is achieved when this equation holds for all records i.

To formalize the optimization problem and permit that the solution violates some equations we consider the equation in blocks. We consider a block as the set of equations concerning record i. Therefore, we define a block as the set of all the distances between one record of the original data and all the records of the protected data. Therefore, we have as many K as the number of rows of our original file. Besides, we need a constant C that multiplies K to avoid the inconsistencies and satisfy the constraint.

The rationale of this approach is as follows. The variable K indicates, for each block, if all the corresponding constraints are accomplished ($K = 0$) or not ($K = 1$). Then, we want to minimize the number of blocks non compliant with the constraints. This way, we can find the best weights that minimize the number of violations, or in other words, we can find the weights that maximize the number of re-identifications between the original and protected data. Therefore, we have as many K as the number of rows of our original file. Besides, we need a constant C that multiplies K to avoid the inconsistencies and satisfy the constraint.

Note that if for a record i, Equation (1) is violated for a certain record j, then, it does not matter that other records j also violate the same Equation for the same record i. This is so because record i will not be re-identified.

Using these variables, K_i and the constant C are defined as follows:

$$\begin{aligned}\mathbb{C}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) - \\ -\mathbb{C}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) + CK_i > 0\end{aligned}$$

for all $i \neq j$. The constant C is used to express the *minimum distance* we require between the correct link and the other incorrect links. The larger it is, the more correct links are distinguished from incorrect links.

Using these constraints we can define the optimization problem for a given aggregation operator $\mathbb{C}$ as:

$$Minimize \sum_{i=1}^{N} K_i \quad (2)$$

$$\begin{aligned}&Subject\ to: \\ &\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{C}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) - \\ &\quad - \mathbb{C}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) + \\ &\quad + CK_i > 0\end{aligned} \quad (3)$$

$$K_i \in \{0, 1\} \quad (4)$$

where N is the number of records, and n the number of variables. This problem is a linear optimization problem with linear constraints and the (global) optimum solution can be found with an optimization algorithm.

If N is the number of records, and n the number of variables of the two data sets X and Y. We have N terms of K_i in the objective function, that is N variables for Equation (2). The total number of constraints in the optimization problem is $N^2 + N$. There are N^2 constraints from Equation (3), and N for Equation (4). Note that depending on the aggregation operator $\mathbb{C}$ used, there will be more constraints in the problem.

3.3 Learning the optimal weights using the Mahalanobis distance

Once we have seen the generalized constraint problem in the last section, we define the problem for the Mahalanobis distance d^2MD introduced in Section 3.1. The minimization problem can be expressed as:

$$Minimize \sum_{i=1}^{N} K_i \quad (5)$$

$$\begin{aligned}&Subject\ to: \\ &\sum_{i=1}^{N}\sum_{j=1}^{N} MD_{\Sigma}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) - \\ &\quad - MD_{\Sigma}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) + \\ &\quad + CK_i > 0\end{aligned} \quad (6)$$

$$MD_{\Sigma}(c_1, \ldots, c_n) \geq 0 \quad (7)$$

$$K_i \in \{0, 1\} \quad (8)$$

where N is the number of records, and n the number of variables. Comment that due to the weighing matrix is symmetric we only take into account $n(n+1)/2$ weights, instead of n^2 that has the whole matrix.

The number of constraints is: N^2 for Equation (6); N^2 for Equation (7) and N for Equation (8). While the number constraints for the Choquet integral problem is $N^2 + N + \sum_{k=2}^{n} \binom{n}{k} k + 1$ and $N^2 + N + n + 1$ is the number of constraints for weighted mean problem.

4 Evaluation

We have evaluated our proposal with different protected files using *microaggregation*[6], a well-known microdata protection method, which broadly speaking,

provides privacy by means of clustering the data into small clusters of size k, and then replacing the original data by the centroid of their corresponding clusters. This parameter k determines the protection level: the greater the k, the greater the protection and at the same time the greater the information loss.

We have considered files with the following protection parameters:

- *M4-33*: 4 variables microaggregated in groups of 2 with $k = 3$.
- *M4-28*: 4 variables, first 2 variables with $k = 2$, and last 2 with $k = 8$.
- *M4-82*: 4 variables, first 2 variables with $k = 8$, and last 2 with $k = 2$.
- *M5-38*: 5 variables, first 3 variables with $k = 3$, and last 2 with $k = 8$.
- *M6-385*: 6 variables, first 2 variables with $k = 3$, next 2 variables with $k = 8$, and last 2 with $k = 5$.
- *M6-853*: 6 variables, first 2 variables with $k = 8$, next 2 variables with $k = 5$, and last 2 with $k = 3$.

For each case, we have protected 400 records randomly selected from the Census dataset [5] from the European CASC project [3], which contains 1080 records and 13 variables, and has been extensively used in other works [13, 8, 26].

Note that in our experiments we apply different protection degrees to different variables of the same file. The values used vary between 2 to 8, i.e., values between the lowest protection value and a good protection degree in accordance with [7]. This is especially interesting when variables have different sensitivity.

	d^2AM	d^2MD	d^2WM	d^2CI	d^2MD^*
M4-33	0.84	0.94	0.955	0.9575	0.9675
M4-28	0.685	0.9	0.93	0.9375	0.9425
M4-82	0.71	0.9275	0.9425	0.9425	0.9525
M5-38	0.3975	0.8825	0.905	0.9125	0.9225
M6-385	0.78	0.985	0.9925	0.9975	0.9975
M6-853	0.8475	0.98	0.9875	0.9925	0.995

Table 1: Improvement in the linkage ratio.

Table 1 shows the linkage ratio between the standard record linkage method (d^2AM); the Mahalanobis distance (d^2MD); two currently existing supervised learning approaches: the weighted mean (d^2WM) and the Choquet integral (d^2CI), which were described in Section 3.2; and, finally, the new supervised approach presented in this paper based on the Mahalanobis distance (d^2MD*). The values in the table are the ratio determining the correctly identified records from the total, so a ratio of 1 means a 100% re-identification.

As it can be appreciated, our proposed method achieves an important improvement with respect to the standard distance based record linkage. However, the improvement with respect to the d^2MD and the two other supervised approaches is relatively small, especially with d^2CI. Although the difference between methods d^2CI and d^2MD* is small, it is important to bear in mind that the Choquet integral approach is computationally more expensive and complex. This is due to the number of constraints required in the optimization problem. This makes the proposed use of the Mahalanobis distance more effective than the one using the Choquet integral.

Moreover, we compare the covariance matrix used in d^2MD and the inverse matrix obtained by the supervised approach using Mahalanobis (d^2MD*), which it is supposed to be the same than the covariance matrix or a scaled variation of it. However, when we compare both matrices after their normalization, by means of the Frobenius matrix normalization, the results obtained shows that both matrices are different.

5 Conclusions

In data privacy and statistical disclosure control, record linkage is used as a disclosure risk estimation of the protected data. This estimation is based on the links between records of the original and the protected data.

In this paper we have introduced a distance based record linkage. Our proposal uses a supervised learning approach relying on the Mahalanobis distance to determine the optimal weighting matrix for the linkage, which also provides information about the interaction between each pairs of variables. Furthermore, we have evaluated this supervised learning with other supervised and no supervised methods and we have achieved the bests results, even when we have compared with the Choquet integral approach using a fuzzy measure.

Acknowledgements

Partial support by the Spanish MICINN (projects TSI2007-65406-C03-02, ARES- CONSOLIDER INGENIO 2010 CSD2007-00004) is acknowledged.

Some of the results described in this paper have been obtained using the Centro de Supercomputación de Galicia (CESGA). This partial support is gratefully acknowledged.

*This is the supervised learning approach using the Mahalanobis distance.

References

[1] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 439–450. ACM Press, 2000.

[2] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*. Springer-Verlag New York, Inc., 2006.

[3] R. Brand, J. Domingo-Ferrer, and J. Mateo-Sanz. Reference datasets to test and compare sdc methods for protection of numerical microdata. *Technical report, European Project IST-2000-25069 CASC*, 2002.

[4] Canada. Record linkage at statistics canada, 2010.

[5] U. Census Bureau. Data extraction system.

[6] D. Defays and P. Nanopoulos. Panels of enterprises and confidentiality: The small aggregates method. In *Proc. of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, pages 195–204. Statistics Canada, 1993.

[7] J. Domingo-Ferrer and V. Torra. *A quantitative comparison of disclosure control methods for microdata*, pages 111–133. Elsevier, 2001.

[8] J. Domingo-Ferrer and V. Torra. Ordinal, continous and heterogeneous anonymity through microaggregation. *Data Mining and Knowledge Discovery.*, 11(2):195 – 212, 2005.

[9] H. Dunn. Record linkage. *American Journal of Public Health*, 36(12):1412–1416, 1946.

[10] A. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16, 2007.

[11] I. Fellegi and A. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[12] M. A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.

[13] M. Laszlo and S. Mukherjee. Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Trans. on Knowl. and Data Eng.*, 17(7):902–911, 2005.

[14] P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings National Institute of Science, India*, volume 2, pages 49–55, Apr. 1936.

[15] A. Mccallum and B. Wellner. Object consolodation by graph partitioning with a conditionally-trained distance metric. *In Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pages 19–24, 2003.

[16] H. B. Newcombe and J. M. Kennedy. Record linkage: making maximum use of the discriminating power of identifying information. *Commun. ACM*, 5(11):563–566, 1962.

[17] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records. *Science*, 130:954–959, 1959.

[18] D. Pagliuca and G. Seri. Some results of individual ranking method on the system of enterprise acounts annual survey. *Esprit SDC Project, Delivrable MI-3/D2*, 1999.

[19] V. Torra, J. Abowd, and J. Domingo-Ferrer. Using mahalanobis distance-based record linkage for disclosure risk assessment. *Lecture Notes in Computer Science*, (4302):233–242, 2006.

[20] V. Torra and Y. Narukawa. *Modeling Decisions: Information Fusion and Aggregation Operators*. Springer, 2007.

[21] V. Torra and Y. Narukawa. On independece, expectation and distances: Choquet integrals and mahalanobis distance. *Modeling Decisions for Artificial Intelligence*, 2010.

[22] V. Torra, G. Navarro-Arribas, and D. Abril. Supervised learning for record linkage through weighted means and owa operators. *Control and Cybernetics*, 39(4):1011–1026, 2010.

[23] L. Willenborg and T. Waal. Elements of statistical disclosure control. Springer-Verlag, 2001.

[24] W. E. Winkler. Data cleaning methods. *Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[25] W. E. Winkler. Re-identification methods for masked microdata. volume 3050, pages 216–230, Heidelberg, Berlin, 2004. Springer.

[26] W. E. Yancey, W. E. Winkler, and R. H. Creecy. Disclosure risk assessment in perturbative microdata protection. In *Inference Control in Statistical Databases, From Theory to Practice*, volume 2316, pages 135–152, London, UK, 2002. Springer-Verlag.