Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment

Vicenç Torra[1], John M. Abowd[2] and Josep Domingo-Ferrer[3]

[1] IIIA-CSIC, Campus UAB, E-08193 Bellaterra, Catalonia. E-mail vtorra@iiia.csic.es
[2] Edmund Ezra Day Professor of Industrial and Labor Relations, Director, Cornell Institute for Social and Economic Research (CISER), 391 Pine Tree Road, Ithaca, NY 14850, USA. E-mail john.abowd@cornell.edu
[3] Universitat Rovira i Virgili, Dept. of Computer Engineering and Maths, Av. Països Catalans 26, E-43007 Tarragona, Catalonia. E-mail josep.domingo@urv.cat

Abstract. Distance-based record linkage (DBRL) is a common approach to empirically assessing the disclosure risk in SDC-protected microdata. Usually, the Euclidean distance is used. In this paper, we explore the potential advantages of using the Mahalanobis distance for DBRL. We illustrate our point for partially synthetic microdata and show that, in some cases, Mahalanobis DBRL can yield a very high re-identification percentage, far superior to the one offered by other record linkage methods.
Keywords: Microdata protection, Distance-based record linkage, Mahalanobis distance.

1 Introduction

A microdata set V can be viewed as a file with n records, where each record contains p attributes on an individual respondent. The attributes in the original unprotected dataset can be classified in four categories which are not necessarily disjoint:

- *Identifiers.* These are attributes that *unambiguously* identify the respondent. Examples are passport number, social security number, full name, etc. Since our objective is to prevent confidential information from being linked to specific respondents, we will assume in what follows that, in a pre-processing step, identifiers in V have been removed/encrypted.
- *Quasi-identifiers.* Borrowing the definition from [3, 13], a quasi-identifier is a set of attributes in V that, in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in V refer. Examples of quasi-identifier attributes are birth date, gender, job, zipcode, etc. Unlike identifiers, quasi-identifiers cannot be removed from V. The reason is that any

attribute in V potentially belongs to a quasi-identifier (depending on the external data sources available to the user of V). Thus one would need to remove all attributes (!) to make sure that the dataset no longer contains quasi-identifiers.
- *Confidential outcome attributes.* These are attributes which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.
- *Non-confidential outcome attributes.* Those are attributes which contain non-sensitive information on the respondent. Note that attributes of this kind cannot be neglected when protecting a dataset, because they can be part of a quasi-identifier. For instance, "Job" and "Town of residence" may reasonably be considered non-confidential outcome attributes, but their combination can be a quasi-identifier, because everyone knows who is the doctor in a small village.

Disclosure risk assessment is needed to measure the safety in a masked microdata being considered for release. The standard procedure is to use quasi-identifier attributes to perform record linkage between the masked dataset and an external identified data source. Each correctly linked pair yields a re-identification. To be more specific, the disclosure model considered in this paper is depicted in Figure 1 and is described next:

- We assume that the released microdata set (on the right-hand side in Figure 1) contains records with quasi-identifier attributes Y' and confidential outcome attributes X. Attributes Y' are masked, synthetic or partially synthetic versions of original quasi-identifier attributes.
- A snooper has obtained an external identified microdata set (on the left-hand side in Figure 1) which consists of one or several identifier attributes Id and several quasi-identifier attributes Y. Attributes Y are original (unmasked) versions of attributes Y' in the released dataset.
- The snooper attempts to link records in the external identified dataset with records in the released masked dataset. Linkage is done by matching quasi-identifier attributes Y and Y'. The snooper's goal is to pair identifier values with confidential attribute values (*e.g.* to pair citizens' names with health conditions).

1.1 Contribution and plan of this paper

We offer here an empirical comparison of various record linkage methods for re-identification. The masked datasets have been generated using the IPSO family of partially synthetic data generators [2] in the same way

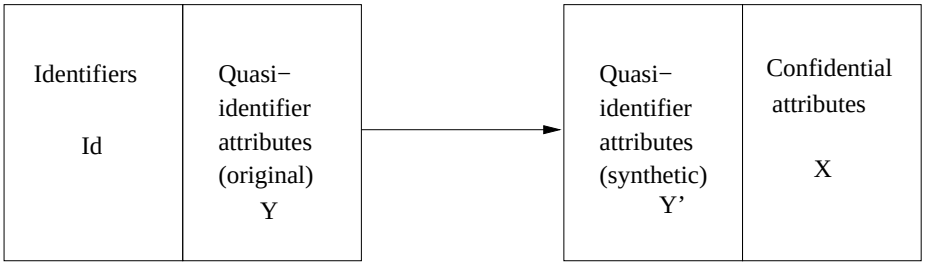

Fig. 1. Re-identification scenario

described in [9]. The range of record linkage methods tried is broader than in [9] and includes distance-based record linkage (DBRL) based on the Mahalanobis distance. This latter method yields surprising good results when there are strongly correlated attributes among in the quasi-identifiers.

Section 2 describes the record linkage methods used. The IPSO synthetic data generators are briefly recalled in Section 3. Section 4 specifies the two test datasets used as original datasets in the empirical study. Section 5 describes the experiments that were carried out. Conclusions are drawn in Section 6

2 Record linkage methods used

We list below the record linkage methods implemented. For additional details and notation see [9, 8]. In what follows, when the distance between pairs of records (a, b) where $a \in A$ and $b \in B$ is considered, we assume that files A and B are defined, respectively, on attributes $V_1^A, \ldots, V_n^A$ and $V_1^B, \ldots, V_n^B$. Accordingly, the actual values of a and b are, respectively, $a = (V_1^A(a), \ldots, V_n^A(a))$ and $b = (V_1^B(b), \ldots, V_n^B(b))$. The following record linkage methods were considered:

DBRL1: Attribute-standardizing implementation of distance-based record linkage. The Euclidean distance was used. Accordingly, given the notation for a and b given above, the distance between a and b is defined by:

$$d(a,b)^2 = \sum_{i=1}^{n} \left(\frac{V_i^A(a) - \bar{V_i^A}}{\sigma(V_i^A)} - \frac{V_i^B(b) - \bar{V_i^B}}{\sigma(V_i^B)}\right)^2$$

DBRL2: Distance-standardizing implementation of distance-based record linkage. The Euclidean distance was used. Therefore, the distance between a and b is defined by:

$$d(a,b)^2 = \sum_{i=1}^{n} (\frac{V_i^A(a) - V_i^B(b)}{\sigma(V_i^A - V_i^B)})^2$$

DBRLM: Distance-based record linkage using the Mahalanobis distance. That is:

$$d(a,b)^2 = (a-b)'[Var(V^A) + Var(V^B) - 2Cov(V^A, V^B)]^{-1}(a-b)$$

where $Var(V^A)$ is the variance of attributes V^A, $Var(V^B)$ is the variance of attributes V^B and $Cov(V^A, V^B)$ is the covariance between attributes V^A and V^B.
The computation of $Cov(V^A, V^B)$ poses one difficulty: how records in A are lined up with records in B to compute the covariances. Two approaches can be considered:

- In a worst case scenario, it would be possible to know the correct links (a, b). Therefore, the covariance of attributes might be computed with the correct alignment between records.
- It is not possible to know a priori which are the correct matches between pairs of records. Therefore, any pair of records (a, b) are feasible. If any pair of records (a, b) are considered, the covariance is zero.

The re-identification using Mahalanobis distance with the first approach for computing the covariance will be denoted by DBRLM-COV. The second approach will be denoted by DBRLM-COV0.

KDBRL: Distance-based record linkage using a Kernel distance. That is, instead of computing distances between records (a, b) in the original n dimensional space, records are compared in a higher dimensional space H. Thus, let $\Phi(x)$ be the mapping of x into the higher space. Then, the distance between records a and b in H is defined as follows:

$$d(a,b)^2 = ||\Phi(a) - \Phi(b)||^2 = (\Phi(a) - \Phi(b))^2 =$$

$$= \Phi(a)\cdot\Phi(a) - 2\Phi(a)\cdot\Phi(b) + \Phi(b)\cdot\Phi(b) = K(a,a) - 2K(a,b) + K(b,b)$$

where K is a kernel function (*i.e.*, $K(a,b) = \Phi(a)\cdot\Phi(b)$).
We have considered polynomial kernels $K(x,y) = (1 + x\cdot y)^d$ for $d > 1$. With $d = 1$, the kernel record-linkage corresponds to the distance-based record linkage with the Euclidean distance.

Taking all this into account, the distance between a and b is defined as:

$$d(a,b)^2 = K(a,a) - 2K(a,b) + K(b,b)$$

with a kernel function K.

PRL: Probabilistic record linkage. The method is based on [10] and [11]. Our implementation follows [14].

3 The IPSO synthetic data generators

Three variants of a procedure called Information Preserving Statistical Obfuscation (IPSO) are proposed in [2]. The basic form of IPSO will be called here Method A. Informally, suppose two sets of attributes X and Y, where the former are the confidential outcome attributes and the latter are quasi-identifier attributes. Then X are taken as independent and Y as dependent attributes. A multiple regression of Y on X is computed and fitted Y'_A attributes are computed. Finally, attributes X and Y'_A are released in place of X and Y.

In the above setting, conditional on the specific confidential attributes x_i, the quasi-identifier attributes Y_i are assumed to follow a multivariate normal distribution with covariance matrix $\Sigma = \{\sigma_{jk}\}$ and a mean vector $x_i B$, where B is the matrix of regression coefficients. Let $\hat{B}$ and $\hat{\Sigma}$ be the maximum likelihood estimates of B and Σ derived from the complete dataset (y, x). If a user fits a multiple regression model to (y'_A, x), she will get estimates $\hat{B}_A$ and $\hat{\Sigma}_A$ which, in general, are different from the estimates $\hat{B}$ and $\hat{\Sigma}$ obtained when fitting the model to the original data (y, x). IPSO Method B, IPSO-B, modifies y'_A into y'_B in such a way that the estimate $\hat{B}_B$ obtained by multiple linear regression from (y'_B, x) satisfies $\hat{B}_B = \hat{B}$.

A more ambitious goal is to come up with a data matrix y'_C such that, when a multivariate multiple regression model is fitted to (y'_C, x), *both* sufficient statistics $\hat{B}$ and $\hat{\Sigma}$ obtained on the original data (y, x) are preserved. This is done by the third IPSO method, IPSO-C.

4 The test datasets

We have used two reference datasets [1] used in the European project CASC:

1. The "Census" dataset contains 1080 records with 13 numerical attributes labeled $v1$ to $v13$. This dataset was used in CASC and in several other papers. [5, 4, 15, 12, 7, 6, 9].

Table 1. Re-identification experiments using dataset "Census" and methods IPSO-A, IPSO-B and IPSO-C

Quasi-identifier in external **A**	Quasi-identifier in released **B**
$v7, v12$	$v7_A^{S1}, v12_A^{S1}$
$v4, v7, v11, v12$	$v4_A^{S1}, v7_A^{S1}, v11_A^{S1}, v12_A^{S1}$
$v4, v7, v12, v13$	$v4_A^{S1}, v7_A^{S1}, v12_A^{S1}, v13_A^{S1}$
$v4, v7, v11, v12, v13$	$v4_A^{S1}, v7_A^{S1}, v11_A^{S1}, v12_A^{S1}, v13_A^{S1}$
$v1, v3, v4, v6, v7, v9, v11, v12, v13$	$v9_A^{S1}, v11_A^{S1}, v12_A^{S1}, v13_A^{S1}, v1_A^{S1}, v3_A^{S1}, v4_A^{S1}, v6_A^{S1}, v7_A^{S1}$
$v7, v12$	$v7_A^{S2}, v12_A^{S2}$
$v4, v13$	$v4_A^{S2}, v13_A^{S2}$
$v7, v12, v13$	$v7_A^{S2}, v12_A^{S2}, v13_A^{S2}$
$v4, v7, v12, v13$	$v4_A^{S2}, v7_A^{S2}, v12_A^{S2}, v13_A^{S2}$

2. The "EIA" dataset contains 4092 records with 15 attributes. The first five attributes are categorical and will not be used. We restrict attention to the last 10 numerical attributes, which will be labeled $v1$ to $v10$. This dataset was used in CASC, in [4, 6, 9] and partially in [12] (an undocumented subset of 1080 records from "EIA", called "Creta" dataset, was used in the latter paper).

5 Experiments

We have considered the datafiles "Census" and "EIA", with the same scenarios and the same re-identification experiments we used in [9]. In short, re-identification experiments are applied to pairs of external and released files using subsets of quasi-identifiers. In scenario $S1$ for the dataset "Census" there are nine quasi-identifiers; in scenario $S2$ for "Census" there are four quasi-identifiers. For "EIA" there is a single scenario with five quasi-identifier attributes highly correlated with the rest of attributes. Released files (see [9, 8] for details) were generated using the synthetic data generators IPSO-A, IPSO-B and IPSO-C. Table 1 lists the sets of quasi-identifiers considered for the "Census" data in the case of data generated using IPSO-A. Analogous sets of quasi-identifiers (vi_B^{S1} and vi_C^{S1} instead of vi_A^{S1}) were considered for the other IPSO-B and IPSO-C methods. Table 2 contains similar information corresponding to "EIA" datasets.

Note that in this paper only experiments with files sharing attributes have been considered.

Table 2. Re-identification experiments using dataset "EIA" and methods IPSO-A, IPSO-B and IPSO-C

Quasi-identifier in external **A**	Quasi-identifier in released **B**
$v1$	$v1_A$
$v1, v7, v8$	$v1_A, v7_A, v8_A$
$v1, v2, v7, v8, v9$	$v1_A, v2_A, v7_A, v8_A, v9_A$
$v1$	$v1_B$
$v1, v7, v8$	$v1_B, v7_B, v8_B$
$v1, v2, v7, v8, v9$	$v1_B, v2_B, v7_B, v8_B, v9_B$
$v1$	$v1_C$
$v1, v7, v8$	$v1_C, v7_C, v8_C$
$v1, v2, v7, v8, v9$	$v1_C, v2_C, v7_C, v8_C, v9_C$

The results of the experiments considered for the "Census" data for methods IPSO-A, IPSO-B and IPSO-C are given in Tables 3, 4 and 5. The results of the experiments using the file "EIA" are given in Table 6.

6 Conclusions

Conclusions in [9] with respect to distance-based and probabilistic record linkage are also applicable here. In relation to the additional methods considered here we should point out that:

- Distance-based record linkage based on Mahalanobis distance achieves the highest number of re-identifications (3206 over 4092 records) in the case of the EIA datafile when the synthetic data generator is IPSO-A and all quasi-identifiers are considered. This corresponds to the re-identification of 78.3% of the records. Similarly, 3194 (78.05%) re-identifications are obtained for IPSO-B data. In the case of IPSO-C, the best performance is 773 re-identifications (which corresponds to 18.9% of the records).
- With respect to distance-based record linkage based on Mahalanobis distance, DBRLM-COV0 (*i.e.*, covariances between attributes V^A and V^B are set to zero) has a better performance than DBRLM-COV.
- The distance-based record linkage based on the kernel distance leads to results equivalents to the other distance-based methods. Only in one experiment does this method outperform the other ones. This experiment corresponds to "Census" data with synthetic data generated with IPSO-A (first experiment with two variables). In this case, 146 records are re-identified.

One possible explanation for the different behaviour of DBRLM-COV0 in "Census" and "EIA" is that quasi-identifiers in the latter dataset are more highly correlated.

In the experiments performed here, re-identification consists of finding the links between the original and the synthetic data. This corresponds to the assumption that the snooper knows a subset of the original data and tries to link such data with the synthetic data in order to disclose sensitive attributes. This re-identification is directed following the scheme in Figure 1. This re-identification scheme differs from the scheme considered in [9]. There, synthetic data was re-identified back to the original source data. The change in the scheme does not reveal any substantial differences among the methods already considered in [9]. The following results illustrate the minor differences:

- DBRL1 for "Census" data in scenario $S1$ on the data generated with IPSO-A leads to 144, 85, 104, 79 and 36 records re-identified when using the scheme in [9]. Instead, the current scheme leads to 145, 91, 95, 98 and 23, respectively.
- DBRL1 for "EIA" data on the data generated with IPSO-A, the previous scheme leads to 10, 23 and 65 re-identifications while the new one yields 14, 16 and 65 re-identifications, respectively.

Acknowledgments

We acknowledge partial support by the Government of Catalonia under grant 2005 SGR 00446, by the Spanish Ministry of Science and Education under project SEG2004-04352-C04-01/02 "PROPRIETAS" and by Cornell University under contracts no. 47632-10042 and 10043. Abowd acknowledges support from NSF-ITR grant SES-0427889 to Cornell University.

References

1. R. Brand, J. Domingo-Ferrer, and J. M. Mateo-Sanz. Reference data sets to test and compare sdc methods for protection of numerical microdata, 2002. European Project IST-2000-25069 CASC, http://neon.vb.cbs.nl/casc.
2. J. Burridge. Information preserving statistical obfuscation. *Statistics and Computing*, 13:321–327, 2003.
3. T. Dalenius. Finding a needle in a haystack - or identifying anonymous census records. *Journal of Official Statistics*, 2(3):329–336, 1986.

Table 3. Re-identification experiments using dataset "Census" and method IPSO-A. Results in number of correct re-identifications over an overall number of 1080 records. DBRL1: attribute-standardizing implementation of distance-based record linkage (DBRL); DBRL2: distance-standardizing implementation of DBRL; DBRLM-COV and DBRLM-COV0: distance-based record linkage using Mahalanobis distance (covariances computed using the appropriate alingment or covariances set to zero); KDBRL: distance-based record linkage with kernel distance (polynomic kernel with d=2); PRL: probabilistic record linkage

DBRL1	DBRL2	DBRLM-COV0	DBRLM-COV	KDBRL	PRL
145	133	135	123	146	133
91	75	126	60	89	82
95	87	137	66	94	103
98	87	129	62	97	86
23	40	123	67	24	97
104	92	93	84	100	92
59	65	63	57	61	65
94	85	89	68	91	86
109	104	106	44	106	103

4. R. Dandekar, J. Domingo-Ferrer, and F. Sebé. Lhs-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 153–162, Berlin Heidelberg, 2002. Springer.
5. J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra. Comparing sdc methods for microdata on the basis of information loss and disclosure risk. In *Pre-proceedings of ETK-NTTS'2001 (vol. 2)*, pages 807–826, Luxemburg, 2001. Eurostat.
6. J. Domingo-Ferrer, F. Sebé, and A. Solanas. A polynomial-time approximation to optimal multivariate microaggregation. *Manuscript*, 2005.
7. J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogenerous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.
8. J. Domingo-Ferrer, V. Torra, J. M. Mateo-Sanz, and F. Sebé. Research data center-based confidentiality research: Systematic measures of re-identification risk based on the probabilistic links of the partially synthetic data back to the original microdata, final report. Technical report, Rovira i Virgili University and IIIA-CSIC, 2005.
9. J. Domingo-Ferrer, V. Torra, J. M. Mateo-Sanz, and F. Sebé. Empirical disclosure risk assessment of the ipso synthetic data generators. In *Monographs in Official Statistics-Work Session On Statistical Data Confidentiality*, pages 227–238, Luxemburg, 2006. Eurostat.
10. I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.
11. 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.

Table 4. Re-identification experiments using dataset "Census" and method IPSO-B. Results in number of correct re-identifications over an overall number of 1080 records. DBRL1: attribute-standardizing implementation of distance-based record linkage (DBRL); DBRL2: distance-standardizing implementation of DBRL; DBRLM-COV and DBRLM-COV0: distance-based record linkage using Mahalanobis distance (covariances computed using the appropriate alingment or covariances set to zero); KDBRL: distance-based record linkage with kernel distance (polynomic kernel with d=2); PRL: probabilistic record linkage

DBRL1	DBRL2	DBRLM-COV0	DBRLM-COV	KDBRL	PRL
146	133	135	123	133	133
89	75	126	61	73	81
95	86	138	66	87	103
97	85	130	62	86	86
23	40	123	63	5	94
104	92	93	83	92	92
59	65	63	57	65	65
94	85	89	68	85	86
109	104	106	44	103	103

12. M. Laszlo and S. Mukherjee. Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 17(7):902–911, 2005.
13. P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.
14. V. Torra and J. Domingo-Ferrer. Record linkage methods for multidatabase data mining. In V. Torra, editor, *Information Fusion in Data Mining*, pages 101–132, Berlin, 2003. Springer.
15. W. E. Yancey, W. E. Winkler, and R. H. Creecy. Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 135–152, Berlin Heidelberg, 2002. Springer.

Table 5. Re-identification experiments using dataset "Census" and method IPSO-C. Results in number of correct re-identifications over an overall number of 1080 records. DBRL1: attribute-standardizing implementation of distance-based record linkage (DBRL); DBRL2: distance-standardizing implementation of DBRL; DBRLM-COV and DBRLM-COV0: distance-based record linkage using Mahalanobis distance (covariances computed using the appropriate alingment or covariances set to zero); KDBRL: distance-based record linkage with kernel distance (polynomic kernel with d=2); PRL: probabilistic record linkage

DBRL1	DBRL2	DBRLM-COV0	DBRLM-COV	KDBRL	PRL
34	34	34	34	33	34
37	37	42	19	39	32
24	24	24	11	23	23
39	39	44	17	40	36
24	24	50	11	25	43
47	47	47	44	49	48
19	19	20	20	19	18
40	40	34	34	41	37
35	35	41	41	32	33

Table 6. Re-identification experiments using dataset "EIA" and methods IPSO-A, IPSO-B and IPSO-C. Results in number of correct re-identifications over an overall number of 4092 records. DBRL1: attribute-standardizing implementation of distance-based record linkage (DBRL); DBRL2: distance-standardizing implementation of DBRL; DBRLM-COV and DBRLM-COV0: distance-based record linkage using Mahalanobis distance (covariances computed using the appropriate alingment or covariances set to zero); KDBRL: distance-based record linkage with kernel distance (polynomic kernel with d=2); PRL: probabilistic record linkage

DBRL1	DBRL2	DBRLM-COV0	DBRLM-COV	KDBRL	PRL
14	9	9	9	14	8
16	15	18	9	16	16
65	121	3206	143	63	159
14	9	9	9	14	8
17	15	18	8	17	16
65	120	3194	135	62	159
11	11	11	11	11	10
6	6	14	8	6	5
53	53	773	46	54	93