Generation of Synthetic Data by means of Fuzzy c-Regression

Isaac Cano, Vicenç Torra

***Abstract*—Problems related to data privacy are studied in the areas of privacy preserving data mining (PPDM) and statistical disclosure control (SDC). Their goal is to avoid the disclosure of sensitive or proprietary information to third parties. In this paper a new synthetic data generation method is proposed and the information loss and disclosure risk are measured. The method is based on fuzzy techniques. Informally, a *fuzzy c-regression* method is applied to the original data set and synthetic data is released with an appropriate information loss and disclosure risk depending on *c*. As other data protection methods do, our synthetic data generation procedure allows third parties to do some statistical computations with a limited risk of disclosure. The trade-off between data utility and data safety of our proposed method will be assessed.**

I. Introduction

The digital age has enabled widespread access to collections of data. While there are several advantages to ubiquitous access to data, there is also the potential for breaching the privacy of individuals. Data perturbation [14] is a classical technique for solving the problem of simultaneously enabling access to data and preserving their privacy. As data usually contains sensitive information about the respondents, their release to third parties requires the application of mechanisms to ensure data privacy. Information loss measures evaluate in what extent the protected data is still valid for analysis (data utility), and disclosure risk measures evaluate in what extent data satisfy the privacy constraints (data safety) [1].

Protection methods, which are studied in the areas of privacy preserving data mining (PPDM) [1] and statistical disclosure control (SDC) [18], can be classified into three broad families of techniques depending on how they manipulate the original data in order to obtain a protected data set. This description is based on [17]

- **Perturbative methods.** Data perturbation involves modifying confidential variables introducing some kind of noise in them. E.g. noise is added to an attribute following a *N(0,a)* for a given *a*. This method perturbs the relationship between the variables from the original data and it may create new relationships in the protected data set. This obfuscation makes disclosure difficult to intruders.
- **Non-perturbative methods.** Instead of perturbing the original value it is replaced by another one that is not incorrect but less specific. E.g. replacing a real number by an interval. In general, non-perturbative methods reduce the level of detail of the data set, which means a higher information loss and less disclosure risk.
- **Synthetic Data Generators.** In this case, new artificial data is generated and used to substitute the original values. Formally, synthetic data generators build a data model from the original data set and, subsequently, a new (protected) data set is randomly generated constrained by the model. Although it is possible to publish the model, third parties usually prefer to receive the artificial data.

Current approaches on synthetic data generation does not permit to control the level of information loss or disclosure risk. While for most perturbative methods there is a parameter to select an appropriate perturbation, this is not the case for synthetic data generators.

In this paper we propose using the *fuzzy c-regression* technique [11] to build a data model from an original data set, when it consists of only continuous (numerical) attributes, and then generate synthetic data. This model can be controlled by *c* (number of classes) in order to get a desirable level of information loss and disclosure risk. In order to assess the risk, we used some of the most representative and well-known disclosure risk measures based on record linkage techniques, Distance Based Record Linkage (DBRL), Probabilistic Record Linkage (PRL) and Interval Disclosure (ID) [10]. In a similar way, to measure the information loss caused by our approach, we compare some statistics [13] computed on the original and the protected datasets.

The structure of this paper is as follows. In Section 2, we describe a family of methods for synthetic data generation named Information Preserving Statistical Obfuscation (IPSO) and the fuzzy c-regression technique. In addition, we explain the fuzzy c-means algorithm as it is used to bootstrap the fuzzy c-regression. In Section 3, we explain the method proposed to generate synthetic data by means of *fuzzy c-regression*. In Section 4, we describe the experiments we have performed and the results obtained. This paper finishes with conclusions and future work.

II. Preliminaries

The protection of datasets based on synthetic data generation is becoming a hot research topic. When protected data is synthetic, the privacy of the respondents seems to be protected because the published data is not ”real”. Nevertheless both disclosure risk and information loss have to be measured. On the one hand, it has been shown [16] that some reidentifications are possible, so disclosure risk has to be taken into account. On the other hand, the synthetic data must keep some sufficient statistics from the original values in order to minimize the information loss. Related

I. Cano and V. Torra are with the IIIA, Artificial Intelligence Research Institute; CSIC, Spanish National Research Council; Campus UAB; 08193 Bellaterra, Catalonia, Spain. E-mail: cano@iiia.csic.es, vtorra@iiia.csic.es

to information loss, synthetic data generators have to ensure that the same analysis applied both on the original data and on the synthetic data generated, obtain the same results. As synthetic data are generated from a data model built from the original data, those characteristics that are not explicitly included in the model are not usually included in the protected data.

A. The IPSO procedure

There have been proposed several methods for synthetic data generation. One of them is the family of methods named Information Preserving Statistical Obfuscation (IPSO) [4]. The aim of the IPSO procedure is to obscure the identity of the data while preserving certain statistics. This family comprises three methods IPSO-A, IPSO-B and IPSO-C, where IPSO-C is the method with less information loss and IPSO-A is the simplest of these three since it produces more information loss. IPSO methods are based on dividing the original data into sets of attributes X and Y. The attributes in the set X are considered independent and Y are considered as dependent. Then, a data model, e.g. multivariate normal multiple regression model, able to represent the information contained in Y is built and a protected set of attributes, say Y', is computed from the conditional distribution $Y|(T, x)$. Because the model is constructed from the data Y, the protection achieved is data dependent. Hence, the degree of protection is a direct function of the data and cannot be parametrized to obtain a balance between information loss and disclosure risk. We detail the differences among IPSO-A, IPSO-B and IPSO-C below:

- **IPSO-A.** In this method a multiple regression of Y on X is computed and fitted values Y'_A are used to replace the attributes in Y. Then, the released data is formed by attributes on X and Y'_A instead of the original set of attributes X and Y. The attributes in Y are supposed to follow a multivariate normal distribution with covariance matrix $\sum = \sigma_{jk}$ and a mean vector $x_i B$, where B is the matrix of regression coefficients. The disadvantage of IPSO-A is the following, if a multiple regression model is fitted to (y'_A, x) we will get estimates $\hat{B}_A$ and $\hat{\Sigma}_A$ which in general, are different from the estimates B and Σ obtained when fitting the model to the original data (y, x).
- **IPSO-B.** In the next IPSO method, IPSO-B, the prediction y'_B is fixed in such a way that the estimated mean vector $\hat{B}_B$ obtained from (y'_B, x) is equal to $\hat{B}$. Hence the new value y'_B can be used as a perturbed value for the public data preserving the sufficient statistic $\hat{B}$ so that the information loss is decreased.
- **IPSO-C.** In the last IPSO method, IPSO-C, the prediction y'_C is fixed to obtain $\hat{B}_C = \hat{B}$ and also $\hat{\Sigma}_C = \hat{\Sigma}$. This is obtained by fitting a multivariate multiple regression model to (y'_C, x).

The IPSO procedure is similar to the General Additive Data Perturbation (GADP) class of methods described by Muralidhar and Sarathy [14]. They both attempt to preserve features such as the means, variances and covariances of the original data. The principal difference between GADP and IPSO is that the latter preserves the values of the statistics of a sample even in the case of small to medium samples.

B. Fuzzy c-means

The fuzzy c-means (FCM) algorithm [2] is one of the most widely used methods in fuzzy clustering. It is based on the concept of fuzzy c-partition, introduced by Ruspini in 1969 [15]. The fuzzy c-means algorithm makes a fuzzy partition of a given set of elements. From a conceptual point of view, the underlying data categories are considered as fuzzy. Then, with a set of objects $X = \{x_1, x_2, \ldots, x_N\}$ evaluated in terms of attributes $A = \{A_1, A_2, \ldots, A_M\}$ fuzzy c-means makes a fuzzy partition of the objects X. Therefore, considering c categories ($C = \{C_1, \ldots, C_c\}$) the problem turns out to be the determination of c membership functions $\mu_1, \mu_2, \ldots, \mu_c$, where μ_i is the membership function corresponding to C_i. μ_i are such that for each object x their membership to all category C adds to one. In addition, it is required at least one element with a non zero membership for every category. Thus, the membership functions have to satisfy the following two conditions:

$$\sum_{i=1}^{c} \mu_i(x) = 1 \qquad \text{for all } x \in X$$

$$0 < \sum_{x \in X} \mu_i(x) < N \qquad \text{for all } C_i \in C$$

Once we have the membership's restrictions, the problem can be formulated as follows. The parameters to minimize are μ and P, where μ is as above and P are the centroids P_k of the clusters $k = 1 \ldots c$

$$\text{minimize} \qquad FO(\mu, P) = \sum_{k=1}^{c} \sum_{x} (\mu_k(x))^m \left\| A(x) - p_k \right\|^2$$

restricted to

$$\mu \in M_f = \left\{ (\mu_k(x)) \mid \mu_k(x) \in [0,1], \sum_{k=1}^{c} \mu_k(x) = 1, \forall x \in X \right\}$$

Where c is a constant value representing the number of fuzzy categories allowed. The other constant value m, which has to be grater than 1, is the fuzziness degree of the categories. The greater the value of m, the fuzzier the categories are. When $m \to \infty$, all categories cover all the points. In contrast, the smaller the m, the less fuzzy the categories are. When $m = 1$ it corresponds to the c-means algorithm.

The fuzzy c-means algorithm constructs a feasible solution of the above problem as follows:

1) Define an initial partition μ and compute the centroids P. This can be done randomly.
2) In this second step, for each element x_i update the membership of x_i to every category C_k as follows:

- If $\|x_i - p_k\|^2 > 0$ then for each category C_k:

$$\mu_k(x_i) = \left[\sum_{j=1..c} \left(\frac{\|x_i - p_k\|^2}{\|x_i - p_j\|^2} \right)^{\frac{1}{(m-1)}} \right]^{-1}$$

- If there is any category C_k for which $\|x_i - p_k\|^2 = 0$ it means that x_i has the same value as some centroid p_k. Hence, in this case, the membership of x_i must be randomly shared with all the centroids that match x_i.

3) The goal of next step is to update the centroid's value. So for each category C_k its centroid is defined as follows:

$$p_k = \frac{\sum_{i=1}^{N} (\mu_k(x_i))^m A(x_i)}{\sum_{i=1}^{N} (\mu_k(x_i))^m}$$

and for the jth component, this is defined as

$$A_j(p_k) = \frac{\sum_{i=1}^{N} (\mu_k(x_i))^m A_j(x_i)}{\sum_{i=1}^{N} (\mu_k(x_i))^m}$$

4) Stop when the convergence threshold is exceeded, otherwise go to step 2. The convergence threshold can be defined as the comparison of the membership functions of two consecutive iterations. Considering a threshold λ and also μ' and μ as the membership functions from two consecutive iterations, the stopping condition can be defined as the follows:

$$\lambda > \max_{k=1,\ldots,c} \max_{x \in X} |\mu'_k(x) - \mu_k(x)|$$

C. Fuzzy c-regression

Fuzzy c-regression models is a family of objective functions which can be used to fit switching regression models to numerical and continuous mixed data. For a given c (the number of clusters, $1 < c < n$), the fuzzy c-regression algorithm is able to get an estimation for the parameters of c regression models, together with a fuzzy c-partition of the data. Let us consider a set of object data of size n, $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where each feature vector (x_i, y_i) has a dependent observation $y_i \in \mathbb{R}^t$ corresponding to a certain independent observation $x_i \in \mathbb{R}^s$. The main difference between fuzzy c-regression models and the simplest data fitting problems is that the latter assume that a single functional relationship between x and y holds for all the data while the former assume the data to be drown from c models:

$$y = f_i(x; \beta_i) + \epsilon, \qquad 1 \leq i \leq c \tag{1}$$

each $\beta_i \in \Omega_i \subset \mathbb{R}^{k_i}$, and each ϵ_i is a random vector with mean vector $\mu_i = 0 \in \mathbb{R}^t$ and covariance matrix Σ_i. It must be told that S is unlabeled, so, for a given feature vector (x_i, y_i), it is not known which model from 1 applies. Hathaway and Bezdek published in [11] a feasible solution for this problem. Their approach is based on fuzzy clustering techniques and is able to produce good estimates of $\{\beta_1, \ldots, \beta_c\}$ while labeling with a fuzzy label vector each datum in S. The labeling problem is solved by means of fuzzy clustering assigning constrained label vectors representing the membership of each object (x_i, y_i) to each of the classes c.

The algorithm for building the Fuzzy c-Regression Models (FCRM) has similar steps to the ones in Fuzzy c-Means:

1) **Step 1.** Given a set of object data $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Set $m > 1$ (a reasonable choice is $m = 1.5$), specify regression models 1, and choose a measure of error $E = \{E_{ik}\}$ so that $E_{ik}(\beta_i) \geq 0$ for i and k and also satisfying the minimizer property [11]. Pick a termination threshold $\epsilon > 0$ (a choice for ϵ in the range 0.0001 to 0.00001 usually yields good estimates) and an initial partition $U^{(0)} \in M_f$. In our experiments, we used the fuzzy c-means algorithm to get such initial partition. Then set a threshold for r_{max}, the maximum number of iterations, so that $r = 1, \ldots, rmax$ in case FCRM does not converge (in our experiments a value of $r_{max} = 30$ was used).
2) **Step 2.** Update the values for the c model parameters $\beta_i = \beta_i^{(r)}$ and then the measure of error $E_{ik}(\beta_i)$ in $f_i(x_k; \beta_i)$ that globally minimize (over $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_c$) the restricted function:

$$\psi(\beta_1, \ldots, \beta_c) \equiv E_m(U^{(r)}, \beta_1, \ldots, \beta c)$$

The most common example for the measure of error $E_{ik}(\beta_i)$ is the squared vector norm $E_{ik}(\beta_i) = \|f_i(x_k; \beta_i) - y_k\|^2$ In our case this second step can be specified by fixing $\Omega_i = \mathbb{R}^s$, $f_i(x_k; \beta_i) = ((x_k)^T \beta i)$ and $1 \leq i \leq c$, so, the objective function $E_m(U^{(r)}, \beta_1, \ldots, \beta c)$ becomes a fuzzy multimodel extension of the least squares criterion for model fitting:

$$E_{ik}(\beta_i) = (y_k - (x_k)^T \beta i)^2.$$

In addition, the new values for the regression model parameters $\beta_i^{(r)}$, $1 \leq i \leq c$ can be computed using the following explicit formula if the columns of X are linearly independent and $U_{ik}^{(r)} > 0$ for $1 \leq k \leq n$:

$$\beta_i^{(r)} = \left[X^T D_i X \right]^{-1} X^T D_i Y \tag{2}$$

where X denotes the matrix in $\mathbb{R}^{n \times s}$ having x_k as its kth row. Y denotes the vector in $\mathbb{R}^n$ having y_k as its kth component, and D_i denotes the diagonal matrix in $\mathbb{R}^{n \times n}$ having $(U_{ik}^{(r)})^m$ as its kth diagonal element.
3) **Step 3.** The aim of this step is to update $U^{(r)} \rightarrow U^{(r+1)} \in M_f$, interpreting U_{ik} as the importance or weight attached to the extent to which the model value $f_i(x_k; \beta_i)$ matches y_k (fuzzy membership on all c models). The update is performed by the next formula:

$$U_{ik} = \left[\sum_{j=1}^{c} \left(\frac{E_{ik}}{E_{jk}} \right)^{\frac{1}{m-1}} \right]^{-1}, \text{ if } E_{ik} > 0 \text{ for } 1 \leq i \leq c$$

In case we encounter some $E_{ik} = 0$, its value can be replaced by adding a small positive number (we used

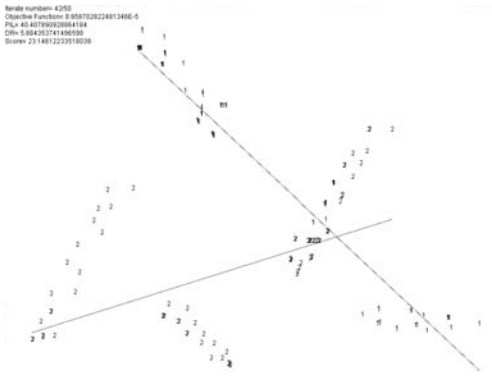

Fig. 1. Results of the models according to Equation 1 with $c = 2$ for the data in the Example.

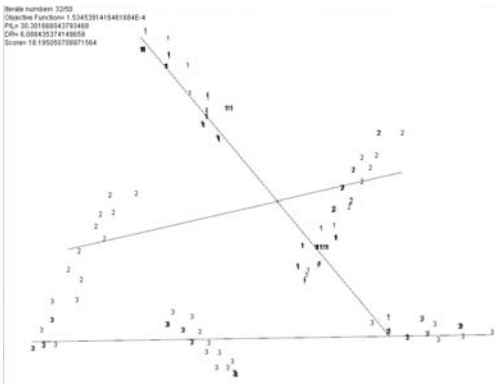

Fig. 2. Results of the models according to Equation 1 with $c = 3$ for the data in the Example.

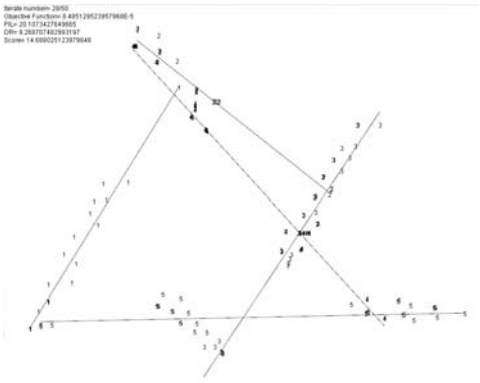

Fig. 3. Results of the models according to Equation 1 with $c = 5$ for the data in the Example.

10^{-100} in our experiments), so step 3 can be performed anyway.

4) **Step 4.** This step checks the termination of the algorithm. If the difference between U^r and U^{r+1} corresponding to two consecutive iterations is greater than the termination threshold, or r is less or equal to r_{max} then $r := r + 1$ and go to step 2. Otherwise stop.

III. Using fuzzy c-regression to generate synthetic data

Once we have introduced all the proper concepts relative to our work, the next step is to combine fuzzy clustering and switching regression models to generate synthetic data. In the previous section we have pointed out the formulas we use to implement the Fuzzy c-Means and the Fuzzy c-Regression model and now we present the basic steps needed to generate the synthetic data while preserving the privacy.

- The first step is to divide the dataset into independent X and dependent Y attributes.
- Next step is to run the Fuzzy c-regression algorithm bootstrapping it using the Fuzzy c-Means to compute the initial partition $U^{(0)} \in M_f$.
- Once FCRM is finished the synthetic data generation step comes up. For each feature vector $(x_s, y_s) \in S$, $1 \leq s \leq n$, select the ith cluster with maximal membership. I.e, select the $arg\ max_{i=1}^{c} U_{is}$. Now we know which centroid p_i best approximates each feature vector. Hence, we can use the regression model β_i corresponding to every centroid p_i to forecast the synthetic value $y_s^{'}$ in order to replace the original one y_s.

A. Example

Now we are going to show how to generate synthetic data by means of FCRM with a simple example. We consider two attributes, the first one is taken as independent and the second one as dependent. Hence we can plot into a two dimensional axis every feature vector labeled with the class number that has a higher fuzzy membership value, see Figures 1 to 3. As usual, the horizontal axis represents the X coordinate and the vertical one represents the Y coordinate. In this example we consider $2 \leq c \leq 10$ and the regression curve corresponding to each centroid is painted. The set of object data $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is formed by five independent clusters of feature vectors. For each c we computed the Probabilistic Information Loss (PIL) [13], the Disclosure Risk (DR) [10] obtained and the standard score [20], [10] (computed as $score = 0.5 * PIL + 0.5 * DR$). In Table I and Figure 4, the relationship between c and PIL/DR is shown: the largest the c, the smaller the information loss and the larger the risk. E.g., the lowest risk and maximum information loss is with $c = 2$ where $PILmax = 40.41\%$ and $DRmin = 5.88\%$. For a $c = 5$ we get a $PIL = 20.11\%, DR = 9.27\%$. Finally, the minimum information loss and maximum risk is with $c = 10$ where we get a $PILmin = 12.79\%$ and $DRmax = 22.48\%$ This example shows how using fuzzy c-regression models we can use c (number of centroids) in order to get a desirable balance between information loss and disclosure risk. This is a clear advantage between our approach respect to the IPSO procedure which generates synthetic data with a fixed information loss and disclosure risk.

IV. Experiments

The test dataset used is one out of two reference datasets [3] used in the European project CASC. We re-

C	O.F.	PIL	DR	SCORE
2	1.09E-04	40.41	5.88	23.15
3	1.62E-04	30.3	6.09	18.2
4	4.63E-03	23.97	8.03	16
5	1.46E-04	20.11	9.27	14.69
6	1.67E-04	18.89	9.98	14.43
7	1.07E-04	15.96	15.12	15.54
8	3.66E-03	15.24	16.17	15.7
9	1.12E-03	14.22	21.92	18.07
10	1.09E-01	12.79	22.48	17.64

TABLE I
EVALUATION IN THE EXAMPLE. C STANDS FOR NUMBER OF CLUSTERS, O.F. FOR OBJECTIVE FUNCTION, PIL FOR PROBABILISTIC INFORMATION LOSS, DR FOR DISCLOSURE RISK AND THE SCORE CORRESPONDS TO THE AVERAGE OF PIL AND DR.

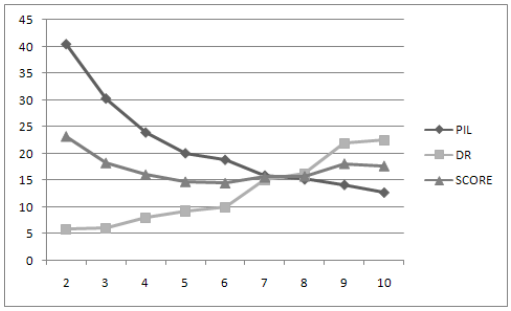

Fig. 4. Scatter plot showing the relationship between PIL and DR with respect to the number of clusters for the Example.

C	O.F.	PIL	DR	SCORE
2	0.184	44.677	9.583	27.13
3	0.005	32.614	12.294	22.454
4	0.078	26.668	14.791	20.73
5	0.195	21.797	16.9516	19.374
6	0.104	18.357	17.137	17.747
7	0.191	16.249	18.791	17.52
8	0.031	13.495	20.492	16.994
9	0.683	12.941	22.999	17.97
10	0.362	11.424	24.256	17.84
11	0.295	10.249	24.255	17.252
12	0.993	9.104	23.969	16.536
13	0.405	8.449	25.374	16.912
14	0.208	8.551	25.408	16.98
15	0.753	7.164	26.970	17.067
IPSO-A	-	49.163	10.044	29.603
IPSO-B	-	49.164	10.04	29.602
IPSO-C	-	9.522	6.392	7.957

TABLE II
EVALUATION IN SCENARIO S1: CENSUS DATA SET WITH 9 DEPENDENT VARIABLES.

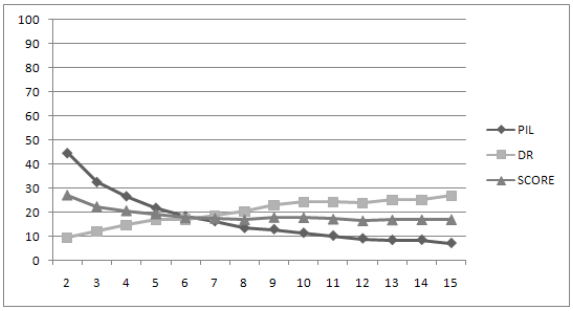

Fig. 5. Scatter plot showing the relationship between PIL and DR with respect to the number of clusters for scenario S1.

fer to the "Census" dataset which contains 1080 records with 13 numerical attributes labeled from $v1$ to $v13$. This dataset had been used in CASC and in several other papers [5], [6], [8], [9], [7], [12], [19]. We have considered two scenarios. In scenario $S1$ there are 9 dependent variables $v1, v3, v4, v6, v7, v9, v11, v12, v13$, and 3 independent variables, $v2, v8, v10$. The variable $v5$ is not considered because the attributes in the independent set are not linearly independent and this causes the singularity of $[X^T D_i X]$ in Equation 2. In scenario $S2$ there are 4 dependent variables $v4, v7, v12, v13$, and 9 independent variables, $v1, v2, v3, v5, v6, v8, v9, v10, v11$. For each scenario we have generated the synthetic data using different values for c and then we have computed the objective function (O.F.), Probabilistic Information Loss, Disclosure Risk and the standard score. As in the simple example, for small values of c the information loss is maximum while the disclosure risk is minimum. In contrast, for big values of c the information loss is minimum, hence, the disclosure risk is maximum. In addition there is a direct relationship between the number of centroids and the information loss because when c is incremented the PIL value decreases and there is an inverse relationship between the number of centroids and the disclosure risk because when c increases DR decreases. This is a property accomplished in both scenarios. I.e., in scenario $S1$ for a $c = 2$ we have a $PIL_{max} = 44.677\%, DR_{min} = 9.583\%$ and for a $c = 15$ we have a $PIL_{min} = 7.164\%, DR_{max} = 26.97\%$. Also in scenario $S2$ for a $c = 2$ we have a $PIL_{max} = 94.053\%, DR_{min} = 6.21\%$ and for a $c = 81$ we have a $PIL_{min} = 11.895\%, DR_{max} = 24.662\%$. In this experiment with the "Census" dataset we wanted to study in detail the evolution of the information loss and the disclosure risk so we have increased c until PIL values become smaller than DR values to point out the relationship of this two data measures. A good selection usually corresponds to the case with minimum score,i.e. in scenario $S1$ $SCORE_{min} = 16.912\%$ corresponding with $c = 13$. Note that in particular scenarios (with very sensitive data) other c with less risk (smaller DR) might be more adequate. That is why the availability of a parameter is specially meaningful.

To compare FCRM and IPSO procedures we have evaluated in both scenarios the information loss, disclosure risk and standard score when using the IPSO-A, IPSO-B or IPSO-C to generate the synthetic data. In scenario $S1$ we have for IPSO-A and IPSO-B a maximum information loss of 49.16% related with a disclosure risk of 10.04% while using FCRM with $c = 2$ the maximum information loss is 44.67% corresponding to a disclosure risk of 9.58%. However, using the IPSO-C procedure we obtained a best score of 7.95% because it obtains a $PIL = 9.52\%, DR = 6.39\%$. With respect to scenario $S2$, the IPSO-A and IPSO-B procedures got a information loss of 44.44% and a disclosure risk of 14.493%, in contrast when using FCRM to generate the

C	O.F.	PIL	DR	SCORE
2	0,999945	94,053	6,210	50,132
4	0,999999966	81,792	6,428	44,110
5	0,999999715	61,834	7,782	34,808
6	0,999999995	72,550	7,424	39,987
8	0,999995425	59,113	10,152	34,633
9	0,9999968	38,990	11,042	25,016
10	9,917E-06	49,157	9,261	29,209
14	0,999895297	42,246	11,710	26,978
18	0,999992791	43,083	15,407	29,245
20	0,999993721	31,422	11,454	21,438
22	0,999995385	35,233	13,895	24,564
24	0,999910478	29,362	16,184	22,773
26	0,999999944	24,750	15,186	19,968
28	0,999999785	23,098	17,857	20,478
30	0,999713871	25,222	17,376	21,299
34	0,967993197	21,397	16,296	18,847
45	4,70034E-25	17,606	19,612	18,609
56	0,999999909	13,541	22,085	17,813
77	0,999832048	12,751	25,257	19,004
81	0,999997699	11,895	24,662	18,279
IPSO-A	-	44,44	14,493	29,467
IPSO-B	-	44,441	14,493	29,467
IPSO-C	-	17,037	11,022	14,029

TABLE III
EVALUATION IN SCENARIO S2: CENSUS DATA SET WITH 4 DEPENDENT VARIABLES.

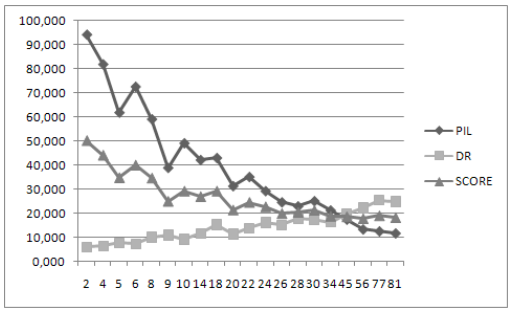

Fig. 6. Scatter plot showing the relationship between PIL and DR with respect to the number of clusters for scenario S2.

synthetic data we get for a $c = 26$ a similar disclosure risk value but almost half the information loss. Also in scenario $S2$, IPSO-C procedure obtains the best score because in FCRM for a similar disclosure value it introduces twice as much information loss.

V. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed using fuzzy c-regression models to generate synthetic data. We have assessed the information loss and disclosure risk when using different number of centroids as an input of the FCRM procedure. Our proposal has been compared to a well known family of methods named Information Preserving Statistical Obfuscation (IPSO) also used to generate synthetic data. We have presented the results of our approach that are better than the ones obtained when using IPSO-A and IPSO-B and worse when using IPSO-C. Also we have pointed out the advantage of the parameter c (number of centroids) when using our approach based on FCRM with respect to the IPSO procedures, which always preserves the same level of data privacy while in FCRM the aim of the parameter c is to obtain the desirable balance between information loss and disclosure risk. As future work we consider the extension of the approach to deal with degenerate datasets which causes singularity in FCRM and to use a Radial Basis Function (RBF) network.

ACKNOWLEDGMENTS

This work is partially supported by the Spanish MEC (CONSOLIDER INGENIO 2010 CSD2007-00004, and TSI2007-65406-C03-02).

REFERENCES

[1] Aggarwal, C.C, Yu, P.S., (2008) Privacy Preserving Data Mining: Models and Algorithms, Springer.
[2] Bezdek, J., (1981) Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York.
[3] Brand, R., Domingo-Ferrer, J., Mateo-Sanz, J.M., (2002) Reference data sets to test and compare SDC methods for protection of numerical microdata. European Project IST-2000-25069 CASC, http://neon.vb.cbs.nl/casc
[4] Burridge, J. (2003) Information Preserving Statistical Obfuscation. Statistics and Computing 13:321-327.
[5] Dandekar, R., Domingo-Ferrer, J., Sebé, F., (2002) Lhs-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, volume 2316:153-162 of Lecture Notes in Computer Science, Berlin Heidelberg, Springer.
[6] Domingo-Ferrer, J., Mateo-Sanz, J.M., Torra, V., (2001) Comparing SDC methods methods for microdata on the basis of information loss and disclosure risk. In Pre-proceedings of ETK-NTTS'2001, Vol. 2:807-826, Luxemburg. Eurostat.
[7] Domingo-Ferrer, J., Torra, V., Mateo-Sanz, J.M., Sebé, F., (2006) Empirical Disclosure risk assessment of the ipso synthetic data generators. In Monographs in Official Statistics-Work Session On Statistical Data Confidenciality, pages 227-238, Luxemburg. Eurostat.
[8] Domingo-Ferrer, J., Sebé, F., Solanas, A., (2005) A polynomial-time approximation to optimal multivariate microaggregation. Computers and Mathematics with Applications, Vol. 55(4):714-732.
[9] Domingo-Ferrer, J., Torra, V., (2005) Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery, Vol. 11(2):195-212.
[10] Domingo-Ferrer, J., Torra, V., (2001) A quantitative comparison of disclosure control methods for microdata, Confidentiality, disclosure, and data access : Theory and practical applications for statistical agencies. Doyle, P.; Lane, J.I.; Theeuwes, J.J.M.; Zayatz, L.V. eds., Elsevier, pp. 111-133.
[11] Hathaway, R.J., Bezdek, J.C., (1993) Switching Regression Models and Fuzzy Clustering, IEEE Transactions on Fuzzy Systems, Vol. 1(3):195-204.
[12] Laszlo, M., Mukherjee, S., (2005) Minimum spanning tree partitioning algorithm for microaggregation. IEEE Transactions on Knowledge and Data Engineering. Vol. 17(7):902-911.
[13] Mateo-Sanz, J.M., Domingo-Ferrer, J., Sebé, F. Probabilistic information loss measures in confidentiality protection of continuous microdata, Data Mining and Knowledge Discovery, Vol. 11, pp. 181-193. Sep 2005. ISSN: 1384-5810
[14] Muralidhar, K., Sarathy, R., (2003) A theoretical basis for perturbation methods. Statistics and Computing 13:329-335.
[15] Ruspini, E., (1969) A new approach to clustering, Information and Control, Vol. 15:22-32.
[16] Torra, V., Abowd, J., Domingo-Ferrer, J., (2006) Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment, Lecture Notes in Computer Science, Number 4302, p.233-242.
[17] Torra, V., (2009) Privacy in Data Mining, in Handbook of Data Mining, 2nd Edition, forthcoming.
[18] Willenborg, L., De Waal, T., (2001) Elements of Statistical Disclosure Control, New York: Springer-Verlag.
[19] Yancey, W.E., Winkler, W.E., Creecy, R.H., (2002) Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, Vol. 2316:195-152 of Lecture Notes in Computer Science, Berlin Heidelberg, Springer.
[20] http://ppdm.iiia.csic.es/