Artificial Continuous Data for SDC

Flavio Foschi[1], Brunero Liseo[2†]

[1] Istat, Division for Information Technology and Methodology, via Cesare Balbo 16, 00184 Rome, Italy, foschi@istat.it
[2]Sapienza Università di Roma, Dept. Of geoeconomics and statistics, viale del Castro Laurenziano 9, 00161 Rome, Italy, brunero.liseo@uniroma1.it

Abstract. The principal aim of this paper is to investigate the use of some theoretical tools to generate artificial data under some empirical constraints. This problem is at the core of *Statistical Disclosure Control (SDC*, henceforth) procedure where one is faced with the problem of providing official statistical information preserving protection of statistical units. In particular we propose the use of Dirichlet processes and empirical copulas to generate values from continuous variables with potentially high skewness and kurtosis. We present a brief methodological outline of the procedure and discuss, in some detail, the result of an application related to Small and Medium Enterprises Survey data released by the Italian National Institute of Statistics (Istat).

1 Outline

SDC procedures, when applied to microdata aim at providing statistical information without jeopardising the confidentiality of single responders. This problem is particularly important in national offices of Statistics (*NIS*, henceforth), and several different techniques have been proposed in the literature.

The main *SDC* techniques used to anonymise data can be classified into three categories:

- pertubative masking; a series of different methods can be used to modify the original observed data,
- non perturbative-masking methods which do not alter data,
- production of synthetic or artificial data files which reproduces some specific features of the original data

We will concentrate on the third approach. The synthetic data approach models the joint distribution of the variables in the dataset; then, synthetic values are generated from their predictive distributions to replace actual sensitive values in the data. In this way simulated values replace part or even all of the dataset, thus limiting the disclosure risk. Usually, this procedure is repeated several times to generate multiple synthetic datasets. These datasets are then released by the agency to the public. In general, researchers perform their analyses on each of the synthetic datasets separately and then, via simple combining rules, are able to obtain interval estimates for a variety of

† All computations were performed by Flavio Foschi, who also wrote Sections 2.2, 3 and 4, while Brunero Liseo wrote the rest of Section 2, and Sections 1 and 5.

quantities of interest. The very important goal is being able to reflect statistical properties of the original data in the synthetic datasets in order to increase the utility of the released data to researchers.

After a brief introduction to the problem of generating artificial continuous data in the *SDC* framework, we introduce some technical tools we consider relevant for the task. In particular we refer to the use of Dirichlet Processes (*DP*) as a flexible nonparametric model to model marginal distributions and the use of empirical copulas in order to maintain the observed structure of dependence. Dirichlet processes as a nonparametric Bayesian tool were introduced by Ferguson (1973, 1974). For a modern perspective on *DP* see, for example, Hjort et. Al (2010). We also explore how to combine *DP* and empirical copulas in order to generate artificial data which preserve some given specific values of the observed data, i.e. multivariate cumulants. Small and Medium Enterprise data released by the Italian National Institute of Statistics are described in section three; they are also used to illustrate the procedure. The obvious trade-off between accuracy and protection ability is investigated and quantified in section four via simulation studies. Section five contains a summary and outlook on future research.

2 SDC Methods for Continuous Data

The most used and popular approaches to deal with continuous variables in *SDC* literature are:

- Data Swapping (Fienberg and McIntyre, 2005),
- General Additive Data Perturbation (Muralidhar, Parsa, Sarathy, 1999),
- Information Preserving Statistical Obfuscation (Burridge, 2003),
- Data Shuffling (Muralidhar, Sarathy, 2006),
- Micro-aggregation (Domingo-Ferrer, Mateo-Sanz, 2002).

When dealing with accuracy, it should be pointed out that the preserved multivariate features are prevalently those related to covariance/correlation matrices, while assumptions about the availability of non sensitive variables (acting as covariates w.r.t. the other variables) are sometimes needed. In terms of data protection, some popular measures are described by Ferrer, Sanz, and Sebé (2006):

- Distance-based Linkage Disclosure (*DLD*), which considers the difficulty in identify statistical unites via record linkage techniques;
- Rank Interval Disclosure (*RID*);
- Standard Deviation Interval Disclosure (*SDID*).

In this paper we will adopt a model free approach: in this respect we need not make any prior assumption on dependence or independence among variables; also the availability of non sensitive variables is not deemed necessary. For a given data set, the irrelevance of some conditioning variable subsets with respect to other variables induces robustness against the lack of model specifications: it is not always possible to get satisfactory fits for all of them. Another important feature of our proposal is the gain in terms of computational effort; although multivariate linear and non linear relationships are preserved, the computational burden increases by only a linear factor with the number of variables. A proposal which is similar in spirit is described in Foschi (2009), where the use of calibration techniques is proposed for the correction of the information loss due to the use of compressed univariate supports. In this work we do not use a support compression approach in order to enhance protection offered by simulated data; this approach is better suited for very large data set, in that it avoids calibration steps and related computational costs. Since we do not rely in any form of rank swapping, a somehow different (with respect to the usual *SDC* literature proposals) definition of data protection is deemed necessary. Under the most pessimistic assumption that an intruder is aware of the presence of a given statistical unit in the sample, the amount of uncertainty about the values that variables can assume is critical to minimize the intrusion utility.

2.1 Artificial Data from Dirichlet Processes

The definition of the Dirichlet process and a review of its features are presented in Ferguson (1973, 1974). In simple word, a Dirichlet process is a probability distribution over the set of all cumulative distribution functions. To our aims it will be sufficient to summarize some basic concepts. Let (P, A) a measurable space and let α be a not null finite measure ($\alpha(P) < \infty$) on (P, A). We say that F is a Dirichlet process with parameter α if, for each partition $A=(A_1,\ldots,A_n)$, of P the random distribution function $(F(A_1),\ldots,F(A_n))$ is a Dirichlet random variable $D(\alpha(A_1),\ldots,\alpha(A_n))$:

$$\forall\{A_i\}_{i=1}^{n}\left(F(A_1),F(A_2),\cdots,F(A_n)\right)\sim D\left(\alpha(A_1),\alpha(A_2),\cdots,\alpha(A_n)\right)$$

Ferguson (1973) shows that if $\mathbf{x}$ is a sample of size n from $F\sim DP(\alpha)$, then the *posterior* distribution of F is also a Dirichlet process, whose parameter is updated in terms of the observed values of $\mathbf{x}$:

$$\left(F(A_1)\cdots,F(A_n)|\mathbf{x}\right)\sim D\left(\alpha(A_1)+\sum\delta_x(A_1),\cdots,\alpha(A_n)+\sum\delta_x(A_n)\right)$$

where $\Sigma\delta_x(A_i)\equiv F_n(A_i)$ is the number of sample observations falling in A_i. Marginal posterior expected values $\pi(x^+\in A_i|\ \mathbf{x},\alpha)$ are convex combinations of prior means and observed frequencies: in particular,

$$E_F\left(F(A_i)|\mathbf{x},\alpha\right)=\frac{\alpha(A_i)+\sum\delta_x(A_i)}{\alpha(P)+n}=\frac{\alpha(A_i)}{\alpha(P)}\frac{\alpha(P)}{\alpha(P)+n}+F_n(A_i)\frac{n}{\alpha(P)+n}$$

It is also easy to see that the predictive distribution is equal to the posterior mean of the process. If x^+ is a "new" observation from F, then,

$$\pi(x^+|\mathbf{x},\alpha) = E\left(F|\mathbf{x},\alpha\right) = \frac{\alpha + F_n}{\alpha(P)+n}$$

It is often useful to specify $\alpha(\cdot)=kF_0(\cdot)$ where k is called the concentration parameter, while F_0 is the base cumulative distribution function. This way k describes the concentration of F around F_0. Conditionally on $F\sim DP(kF_0)$, the random variables $\{X_1, X_2,\ldots,X_n\}$ are independent and identically distributed with cumulative distribution function (*CDF*) F and the following theorem holds:

$$\forall n, B \quad \pi\left(x_{n+1}\in B|\mathbf{x}\right) = \frac{k}{n+k}F_0(B) + \frac{n}{n+k}F_n(B) \Leftrightarrow F \sim DP\left(kF_0\right)$$

When $k/n \to 0$, the role of F_0 becomes negligible and the empirical *CDF* F_n coincides with the posterior distribution. The Rubin's Bayesian bootstrap is obtained when k=0. Muliere and Secchi (1994, 1996) propose to use a proper prior F_0 in order to overcome some logical pitfalls, already noticed in Rubin (1981). This way they are able to achieve some advantages in terms of estimator efficiency. In our framework consistency is not a primary concern: base distributions allow the distortion necessary to achieve the preferred trade-off between accuracy and protection. From this perspective, any distribution which penalizes tail intensities w.r.t. the Empirical Cumulative Distribution Function (*ECDF*) could be used.

2.2 Empirical copulas

The study of copula functions has received growing attention in the recent years: a comprehensive review is presented in Nelsen (2006). The Sklar's theorem (Rueschendorf, 2009) establishes that, given an m-dimensional distribution function F, with given marginals F_1, F_2,…, F_m, there exists a correspondent m-dimensional distribution function C (named the "copula" function) on $[0,1]^m$ with uniform marginals and such that

$$F\left(x_1, x_2, \cdots, x_m\right) = C\left[F_1(x_1), \cdots, F_m\left(x_m\right)\right].$$

Focusing on continuous variables, since $F(X) \equiv p(X \le x)$=p[$\cap(X_j \le x_j)$], setting $U\sim U(0,1)$, and $X=F^{-1}(U)$,

$$p\left[\cap\left(X_j \le x_j\right)\right] = p\left\{\cap\left[F_j^{-1}\left(U_j\right) \le x_j\right]\right\} = p\left\{\cap\left[U_j \le F_j\left(x_j\right)\right]\right\} \equiv C\left[F_1(x_1), \cdots, F_m\left(x_m\right)\right]$$

If the marginal distributions F_j's are not known, their sample counterparts, namely the empirical *CDF's*, could be used, and the inverse distributional transform produces vectors which are proportional to the observed ranks:

$$n\hat{U}_j = n\hat{F}_j\left(X_j\right) = R_j \qquad \left(j = 1, \cdots, m\right)$$

Hence, the Empirical Copula Function is defined as

$$\hat{C}(\mathbf{t}) = n^{-1}\sum I\left[\bigcap\left(\hat{U}_j \le t_j\right)\right] \;=\; n^{-1}\sum I\left[\bigcap\left(R_j \le nt_j\right)\right] \qquad \mathbf{t} \in [0,1]^m$$

The last expression clarifies the key role of the observed rank matrix in conveying the dependence structures observed in the real data. If $\{x_{(1)j}, x_{(2)j}, \ldots, x_{(n)j}\}$ and $\{r_{1j}, r_{2j}, \ldots, r_{nj}\}$ represent, respectively, the order statistics and the ranks for the j^{th}- variable, the preservation of the dependence structure is approximately accomplished by setting

$$x_{ij} \equiv x_{(r_{ij})j} \qquad \left(i = 1,\cdots,n \qquad j = 1,\cdots,m\right)$$

As a result, no rank swapping is necessary and the Empirical Copula Function (*ECF*) does not play a specific role in terms of data perturbation.

2.3 A synthesis of the procedure

In this section we describe the proposed procedure as sequence of steps.

a) Each observed variable $\mathbf{y}_j$ (j =1,…, m) is sorted and then decomposed as:

$$\mathbf{y}_j \equiv f\left(\mathbf{r}_j\right) + \mathbf{x}_j$$

The choice of ranks as covariates would avoid to convey information not predictable from ordinal values. In the above formula the $\mathbf{x}_j$ plays the role of errors. The structural components of the above equation is then modelled through a spline regression

$$f\left(\mathbf{r}_j\right) = \beta_0 + \beta_1\mathbf{r}_j + \sum_s \beta_{2+s}\left(\mathbf{r}_j - \kappa_s\right)_+$$

b) For each marginal distribution, a Dirichlet process prior is defined: the base distribution F_0 is determined by fixing a percentage ξ of extreme observations in the sample and considering the (1- ξ) % observations in the centre of the empirical distribution; then a uniform distribution that covers the central (1−ξ)% part of the support can be used as F_0;

c) for a selected concentration parameter k, sorted samples of length equal to the original number of records are drawn, by inverting each posterior univariate distribution for $\mathbf{x}_j$; in other terms, $\mathbf{x}_{jh}$ is the h^{th} Dirichlet process outcome for the j^{th} variable);

d) the $\mathbf{x}_{jh}$ realizations are added to the correspondent $f(\mathbf{r}_j)$ values and new samples $\mathbf{y}_{jh}$'s are then constructed. All of the new values are centred on the observed mean values in order to alleviate the bias induced by the penalization of the tails of the marginal distributions;

e) each single $\mathbf{y}_{jh}$ generated in (d) is ordered according to the observed ranks; this step is performed one variable at-a-time, in order to maintain univariate relations

between statistical units; implicitly, this allows to recover multivariate links between variables included in the data set; as stressed above, no rank swapping is necessary at this stage.

It should also be stressed that our procedure does not aim at any inferential purpose about the "true" probability laws. Our main goal is then to reproduce, in the synthetic data, much of the features observed in the real data set. The "gaps" between observed and simulated values are then used to evaluate the accuracy protection trade-off. According our framework:

a) for each variable, only the values which fall in distribution tails are to be protected; however, all records are simulated,

b) Suppose l_j and u_j define the lower and the upper critical values selected to detect observations "far" from the body of the marginal distribution of the j^{th} variable: then $z_{ijh}=|y_{ijh}/y_{ij} -1|$, evaluated for the i^{th} record and the j^{th} variable, provides a measure of the bias achieved in the h^{th} simulation, that is the value on which relies the utility intrusion decrement;

c) given a (subjective) threshold $q \in [0, \infty]$, the percentage of $\{y_{ij}\}$ values secured through the artificial data generation can be assessed by considering the fraction of $\{z_{ij}\}$ beyond q. A synthetic evaluation of the protection implied by simulated values should be based on some functions that summarize the z_{ijh}'s quantities; a very simple one could be the pooled average over observations exceeding critical values from (b), that is:

$$T_h = \left[\sum_{i=1}^{n} \sum_{j=1}^{m} I\left(y_{ij} \notin \left[l_j, u_j \right] \right) \right]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ijh} I\left(y_{ij} \notin \left[l_j, u_j \right] \right)$$

A standardised indicator (which takes values in [0, 1]) can then be achieved using a different definition of z_{ijh}:

$$z_{ijh} \equiv |min(y_{ijh}, y_{ij}) / max(y_{ijh}, y_{ij}) - 1|.$$

Then, the correspondent T_h represents a Symmetrised Mean Absolute Percentage Error (*SMAPE*).

3 Data used

Some economic account data about italian enterprises with no more than 99 workers are gathered by the Small and Medium Enterprise (*SME*) survey. Statistical units are sampled from the Statistical Archive of Active Enterprises according to strata identified by economic activity, region (*NUTS2*), and size class. All economic activities are surveyed, excluding agriculture, zootechnics, hunting and fishing, financial activities (except for the financial intermediation and insurance auxiliary

ones), public administration and associative organization activities as well activities carried out by families and cohabitations. From the 2004 *SME* survey, a subset of 12 balance sheet quantities in 56,080 records is considered:

Labels	Variables
T	Turnover
G	Variation in stock of finished and semi-finished goods
J	Variation on construction contracts
K	Work performed by the undertaking for its own purposes and capitalized
I	Other operating income
M	Cost of materials, power consumptions and goods to resale
S	Cost for services
R	Cost for leased and rental assets
L	Staff costs
D	Depreciation and value adjustments on non financial assets
V	Variation in stock of materials and goods to resale
A	Allowance accounts and other operating charges

Table 3.1 Subset of variables from *SME* survey data

var.	m	sd	sk	ku	var.	m	sd	sk	ku	var.	m	sd	sk	ku
T	3430.8	16808.3	32.6	1753.7	I	92.5	1151.7	60.5	4699.7	L	401.4	829.7	21.7	1897.4
G	12.5	1680.2	-147.8	24164.5	M	1901.0	12889.4	38.7	2392.2	D	121.5	1187.8	114.1	18231.9
J	2.3	636.4	-38.6	4831.8	S	764.1	5029.1	47.2	3489.3	V	-5.5	609.0	-74.6	12513.3
K	7.0	101.2	37.5	2132.5	R	91.8	1229.8	196.0	43016.7	A	97.2	1545.7	183.9	38840.2

Table 3.2 First four cumulants for observed data

Data are represented in thousands of euro. Some cumulants are showed in table 3.2, while divergence measures between simulated and observed order statistics (for avoiding occasional disclosures) are illustrated in section 4.1. About multivariate dependence relationships, the pairwise linear ones are summarized in table 3.3. Defining product and cost of product respectively as P≡T+G+J+K+I and C≡M+S+R+L+D+V+A, the reproduction ability of mixed moments and functions (including P and C) of variables grouped according to known strata (49 class of economic activity and 5 size class in term of workers) will be discussed in section 4.2.

var.	ρ	var.	ρ	var.	ρ	var.	ρ	var.	ρ	var.	ρ
T, G	-0.11	G, J	-0.02	J, I	0.07	K, R	0.02	I, A	0.12	S, A	0.10
T, J	-0.03	G, K	0.00	J, M	0.00	K, L	0.16	M, S	0.38	R, L	0.10
T, K	0.05	G, I	-0.02	J, S	0.00	K, D	0.10	M, R	0.05	R, D	0.04
T, I	0.33	G, M	0.01	J, R	0.00	K, V	0.00	M, L	0.23	R, V	-0.01
T, M	0.92	G, S	-0.02	J, L	0.00	K, A	0.03	M, D	0.09	R, A	0.06
T, S	0.64	G, R	0.00	J, D	0.00	I, M	0.34	M, V	-0.14	L, D	0.26
T, R	0.16	G, L	0.01	J, V	0.01	I, S	0.28	M, A	0.21	L, V	-0.03
T, L	0.33	G, D	0.00	J, A	-0.01	I, R	0.05	S, R	0.11	L, A	0.12
T, D	0.23	G, V	0.02	K, I	0.06	I, L	0.29	S, L	0.28	D, V	-0.02
T, V	-0.11	G, A	-0.01	K, M	0.02	I, D	0.18	S, D	0.16	D, A	0.12
T, A	0.30	J, K	-0.04	K, S	0.06	I, V	-0.15	S, V	-0.10	V, A	-0.02

Table 3.3 Correlations for observed data

4 Simulation results

Using the R software (R Development Core Team, 2009), simulations were performed in 1600 replications for each combination of parameters related to tails protection and mixture weights, respectively ξ={0.1, 0.2} and $\delta \equiv k/(k+n)$={0.3,0.5,0.7}. Synthetic measures of gaps between artificial and observed values assumed to be secured can be obtained from the empirical distribution of $T=\{T_h\}$ as defined in section 2.3; given thresholds l_j=0.05 and u_j=0.95 for j=1,…,m, a summary of quantities T_h according to each simulation setting, is illustrated in table 4.1:

	MAPE						*SMAPE*					
ξ	0.1			0.2			0.1			0.2		
δ	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
p2.5	0.16	0.21	0.25	0.16	0.21	0.26	0.09	0.14	0.19	0.09	0.14	0.19
m	3.82	2.92	1.80	4.46	3.68	2.92	0.14	0.19	0.24	0.13	0.18	0.23
p97.5	13.14	10.52	6.68	14.08	11.88	8.73	0.18	0.23	0.27	0.18	0.22	0.27

Table 4.1 *MAPE* and *SMAPE* for data to protect

With respect to proposed δ and ξ values, the influence of ξ on simulation results from the accuracy point of view is weak and it will be sufficient to consider outcomes related to ξ=0.2 only. A brief discussion about univariate and multivariate characteristics of simulated data follows.

4.1 Univariate features

Empirical *CI*s and mean values for simulated 2nd, 3rd and 4th cumulants are collected in table 4.2: due to the centering of simulated univariate data on observed mean values according to section 2.3, 1st ones are perfectly fitted and hence omitted. It is possible to notice that in two instances only simulated *CIs* boundaries do not cover observed values; those happen for δ=0.7, about the standard deviation of T (gap of 274.87) and the kurtosis of L (gap of 7.85). About the effect of the concentration parameter on obtained artificial distributions, the higher δ the lower the standard deviation for all variables, while no general tendency appears w.r.t. skewness and kurtosis.

Gaps between simulated and observed order statistics are summarized in table 4.3, considering *CI*s and mean values of minima, maxima and quartiles. As a by-product of the cumulants reproduction ability, similarity between observed and artificial order statistics is achieved. Positive values in the columns related to minima describe the shrinking of each empirical distribution left tail towards the region where its probability laws is predominantly allocated, meaning the increase of measured quantities; that happens markedly only for variables G, J and V which can assume negative values. Symmetrically, negative values in the columns of maxima evaluate the shrinking of each right tail towards the respective central part of the distribution. The greater δ, the greater the discrepancy between observed and simulated extremes,

while simulated quartiles do not reflect heavy modifications. This fact is consistent with the choice ξ=0.2 which leaves unchanged the central 60% of each empirical distribution.

		sd			sk			ku		
var	δ	p2.5	m	p97.5	p2.5	m	p97.5	p2.5	m	p97.5
T	0.3	12665.9	15414.9	19514.7	17.9	28.7	40.2	571.2	1475.0	2633.4
	0.5	11794.1	14403.4	18304.5	12.8	25.5	39.9	342.5	1329.6	2705.1
	0.7	11561.1	13381.6	16533.4	9.7	20.6	37.8	167.7	1040.5	2653.6
G	0.3	394.0	1369.2	2941.1	-201.8	-105.9	21.4	907.1	18828.0	45363.4
	0.5	397.7	1181.7	2782.4	-201.7	-96.5	16.5	362.0	17893.3	45308.3
	0.7	418.7	903.3	2417.5	-199.2	-69.8	13.9	169.8	13540.6	44594.1
J	0.3	484.3	602.9	767.6	-57.0	-31.0	3.2	1928.6	4028.4	5854.2
	0.5	488.7	586.9	753.6	-55.0	-25.1	6.0	1424.3	3425.3	5571.0
	0.7	499.6	572.8	722.9	-50.1	-19.4	4.5	1126.9	2570.6	5021.8
K	0.3	92.8	100.4	114.1	29.3	34.1	42.3	1081.8	1629.1	2528.9
	0.5	93.7	100.1	114.1	29.3	33.1	42.0	1071.3	1501.9	2484.9
	0.7	95.0	99.4	111.6	29.3	31.6	40.3	1055.9	1320.8	2310.6
I	0.3	677.6	1026.9	1482.0	31.8	56.4	71.6	1691.8	4648.3	7614.6
	0.5	613.5	936.4	1387.2	18.7	51.5	70.6	579.8	4337.1	7616.1
	0.7	616.8	834.8	1238.0	13.4	42.7	68.7	215.7	3634.3	7515.9
M	0.3	9192.9	11914.7	15900.4	18.7	34.4	48.1	611.5	2060.0	3583.1
	0.5	8770.3	10965.8	14432.0	14.7	30.2	47.4	392.7	1817.5	3693.6
	0.7	8571.0	10222.4	13349.2	10.6	24.8	45.7	174.1	1457.7	3681.3
S	0.3	3386.4	4561.2	6203.4	21.0	41.1	57.2	764.8	2883.1	5270.1
	0.5	3110.7	4132.9	5768.2	14.7	36.5	57.8	444.7	2614.8	5548.0
	0.7	3042.0	3773.5	5274.8	10.3	29.5	56.9	167.5	2087.5	5550.6
R	0.3	322.0	843.9	1920.9	15.9	97.6	199.1	532.6	15784.2	44211.3
	0.5	325.8	716.5	1695.8	11.9	85.9	201.3	249.8	14178.0	44959.4
	0.7	337.1	575.4	1549.6	9.7	60.0	197.4	121.0	9462.7	43804.9
L	0.3	731.6	794.3	962.1	3.0	11.3	30.8	16.0	667.0	2234.7
	0.5	733.1	781.7	938.0	3.0	9.5	29.0	14.8	533.1	2197.5
	0.7	735.0	763.6	878.5	2.9	6.8	25.0	13.9	320.0	1889.5
D	0.3	526.6	976.1	1754.5	27.3	75.9	134.8	1424.2	9279.8	24891.8
	0.5	489.4	846.7	1590.7	15.7	66.4	134.1	539.0	8148.5	24924.0
	0.7	486.0	723.9	1418.6	10.1	51.1	136.4	139.4	6100.7	25716.0
V	0.3	315.6	548.6	983.4	-110.7	-46.0	34.8	472.9	8094.4	19673.6
	0.5	305.8	493.1	875.1	-113.9	-42.0	33.5	268.4	7564.2	20975.2
	0.7	309.2	425.6	788.3	-113.2	-27.2	31.6	140.8	5406.6	20936.0
A	0.3	399.1	1094.7	2410.4	20.9	102.4	195.2	745.5	16113.1	43101.7
	0.5	396.9	928.0	2194.8	14.9	90.7	196.5	391.0	14525.0	43325.1
	0.7	403.8	736.5	1970.6	10.9	67.2	197.1	158.8	10511.3	43690.6

Table 4.2 Simulated 2^{nd}, 3^{rd} and 4^{th} cumulants, ξ=0.2

The univariate tails contraction explains how it is possible to expect a decrease of the intrusion utility when original observations close to the highest density values are assumed riskless. That circumstance is evident for all variables and it is relevant for T, M, S, R, D, A maxima and G, J, V minima.

		min			p25			p50			p75			max		
var.	δ	p2.5	m	p97.5	p25	m	p97.5	p2.5	m	p97.5	p2.5	m	p97.5	p2.5	m	p97.5
T	0.3	0.0	20.9	93.6	-57.8	2.6	93.6	-142.5	-4.4	94.0	-142.3	-4.2	94.2	-775196.5	-316370.6	-6231.3
	0.5	0.0	18.3	83.7	-57.8	1.6	83.8	-124.4	-3.3	84.2	-123.9	-2.8	84.7	-860299.9	-392323.8	-9682.6
	0.7	0.0	14.2	60.4	-57.8	0.1	60.6	-95.5	-2.0	61.1	-95.0	-1.4	61.6	-993544.7	-508628.5	-16538.5
G	0.3	-2.8	89258.0	282725.0	-9.4	1.6	23.1	-9.4	1.6	23.1	-9.4	1.6	23.1	-22377.4	-6881.2	-160.3
	0.5	-0.6	115016.2	285570.9	-7.0	1.5	21.4	-7.0	1.5	21.4	-7.0	1.5	21.4	-24163.3	-8959.0	-230.2
	0.7	2.0	167827.6	285956.6	-5.1	0.8	17.2	-5.1	0.8	17.2	-5.1	0.8	17.2	-25684.1	-12422.5	-354.2
J	0.3	-2.0	11556.7	42419.5	-3.5	0.1	4.4	-3.5	0.1	4.4	-3.5	0.1	4.4	-15513.6	-5585.7	-113.4
	0.5	-1.1	15332.7	42591.8	-3.0	0.0	3.9	-3.0	0.0	3.9	-3.0	0.0	3.9	-16020.0	-6841.9	-134.5
	0.7	-0.4	21928.4	43242.0	-2.3	0.1	3.1	-2.3	0.1	3.1	-2.3	0.1	3.1	-17884.6	-9477.5	-201.8
K	0.3	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.0	0.2	-4371.5	-2013.1	-52.4
	0.5	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.0	0.2	-4443.4	-2427.5	-60.5
	0.7	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	-4532.7	-3111.4	-85.1
I	0.3	0.0	1.4	6.4	0.0	1.4	6.4	-0.1	1.4	6.5	-12.2	-1.5	6.0	-60089.7	-12922.4	-117.8
	0.5	0.0	1.0	5.2	0.0	1.0	5.2	-0.1	1.0	5.4	-11.4	-1.9	4.7	-75962.5	-19985.9	-193.0
	0.7	0.0	0.5	2.8	0.0	0.5	2.8	-0.1	0.5	3.0	-9.7	-2.2	2.2	-94562.0	-32047.0	-263.8
M	0.3	0.0	15.5	78.1	-3.3	13.8	78.1	-58.6	-3.4	78.0	-134.8	-19.2	68.4	-648948.2	-221090.3	-4078.0
	0.5	0.0	13.0	63.0	-3.3	11.4	62.9	-58.6	-3.2	62.8	-107.6	-19.2	49.5	-719767.4	-301896.0	-5256.5
	0.7	0.0	8.8	43.0	-3.3	7.1	43.0	-58.6	-6.1	42.8	-100.2	-23.2	27.6	-834083.9	-403212.0	-8318.8
S	0.3	0.0	6.0	28.5	-6.4	3.2	28.5	-48.7	-3.4	28.5	-54.7	-9.6	22.1	-330530.4	-116684.6	-2340.3
	0.5	0.0	5.4	25.2	-6.4	2.7	25.2	-40.3	-2.4	25.2	-47.4	-9.7	17.9	-362566.7	-158703.5	-3527.3
	0.7	0.0	3.5	16.7	-6.4	0.7	16.7	-37.3	-3.2	16.6	-45.6	-11.5	8.4	-407396.2	-209758.1	-7881.7
R	0.3	0.0	1.7	5.6	0.0	1.7	5.6	-7.6	-0.1	5.6	-13.9	-0.7	5.4	-254642.5	-139457.3	-4101.4
	0.5	0.0	1.4	4.2	0.0	1.4	4.2	-7.6	0.0	4.2	-11.4	-0.5	3.9	-258147.5	-161959.5	-5891.8
	0.7	0.0	1.0	2.7	0.0	1.0	2.7	-7.6	-0.1	2.7	-9.9	-0.6	2.4	-263090.4	-196885.4	-11954.5
L	0.3	0.0	0.6	2.1	0.0	0.6	2.1	-4.5	-0.1	2.1	-5.0	-0.5	1.7	-72028.5	-37967.1	-1253.0
	0.5	0.0	0.5	1.7	0.0	0.5	1.7	-4.0	-0.1	1.7	-4.6	-0.8	1.0	-74261.0	-43281.6	-1570.8
	0.7	0.0	0.4	1.2	0.0	0.4	1.2	-2.8	0.0	1.2	-3.5	-0.7	0.5	-76622.0	-53729.2	-2473.2
D	0.3	0.0	1.6	6.7	-2.1	0.7	6.7	-11.8	-0.7	6.7	-14.8	-1.6	5.9	-172445.3	-82187.1	-1962.2
	0.5	0.0	1.3	5.3	-2.1	0.5	5.3	-11.8	-0.6	5.2	-12.9	-1.6	4.3	-178035.0	-103035.9	-2645.2
	0.7	0.0	0.9	3.6	-2.1	0.1	3.6	-9.7	-0.7	3.5	-10.8	-1.8	2.4	-194871.6	-127275.7	-6046.9
V	0.3	-1.2	31986.8	84551.7	-4.0	0.4	7.5	-4.0	0.4	7.5	-4.0	0.4	7.5	-34160.4	-14666.0	-355.6
	0.5	-0.3	39967.9	87381.8	-3.2	0.3	5.9	-3.2	0.3	5.9	-3.2	0.3	5.9	-34915.9	-18406.5	-572.2
	0.7	0.3	56032.8	89281.8	-2.6	0.1	4.9	-2.6	0.1	4.9	-2.6	0.1	4.9	-36478.6	-22765.8	-1023.9
A	0.3	0.0	1.8	7.2	-0.7	1.5	7.2	-8.9	-0.6	7.2	-19.8	-2.4	6.1	-309257.9	-160835.4	-4843.4
	0.5	0.0	1.3	4.8	-0.7	1.0	4.9	-8.9	-1.0	4.8	-16.9	-2.8	3.4	-314362.9	-188892.1	-5285.5
	0.7	0.0	0.7	2.7	-0.7	0.3	2.7	-8.9	-1.2	2.6	-15.9	-3.2	0.9	-321310.0	-230983.5	-11391.5

Table 4.3 Order statistics gaps, ξ=0.2

4.2 Multivariate features

CI correlation boundaries, as functions of δ are collected in table 4.4. Observed correlations not covered by *CIs* (four when δ=0.5, ten when δ=0.7) are often upwardly approximated. With δ=0.5 only simulated boundaries for T and G exceed the observed correlation more than 0.01. For δ=0.7, maximum discrepancy increase to 0.05 for T and L. Those exceeding 0.02 are on the whole 3 (T and G, T and L, M and L).

δ	0.3		0.5		0.7			0.3		0.5		0.7	
var.	p2.5	p97.5	p2.5	p97.5	p2.5	p97.5	var.	p2.5	p97.5	p2.5	p97.5	p2.5	p97.5
T, G	-0.11	0.01	-0.09	0.01	-0.07	0.01	K, R	0.01	0.07	0.01	0.07	0.02	0.08
T, J	-0.06	-0.03	-0.06	-0.03	-0.06	-0.03	K, L	0.14	0.17	0.14	0.17	0.14	0.17
T, K	0.05	0.06	0.05	0.07	0.05	0.07	K, D	0.09	0.18	0.09	0.19	0.10	0.20
T, I	0.25	0.40	0.26	0.43	0.27	0.44	K, V	-0.01	0.00	-0.01	0.00	-0.01	0.00
T, M	0.85	0.92	0.84	0.91	0.81	0.91	K, A	0.02	0.10	0.02	0.10	0.03	0.10
T, S	0.58	0.66	0.55	0.66	0.54	0.65	I, M	0.26	0.41	0.25	0.43	0.25	0.43
T, R	0.12	0.30	0.12	0.33	0.13	0.34	I, S	0.25	0.36	0.25	0.37	0.26	0.38
T, L	0.30	0.46	0.33	0.50	0.38	0.53	I, R	0.05	0.19	0.05	0.21	0.06	0.23
T, D	0.20	0.35	0.20	0.37	0.20	0.39	I, L	0.19	0.31	0.20	0.32	0.20	0.32
T, V	-0.21	0.01	-0.22	0.00	-0.22	-0.01	I, D	0.15	0.30	0.16	0.32	0.17	0.35
T, A	0.24	0.49	0.22	0.49	0.20	0.48	I, V	-0.27	-0.02	-0.28	-0.01	-0.26	0.00
G, J	-0.06	-0.01	-0.06	-0.01	-0.04	0.00	I, A	0.09	0.35	0.10	0.36	0.10	0.37
G, K	-0.01	0.00	-0.02	0.00	-0.03	0.00	M, S	0.32	0.40	0.31	0.40	0.29	0.39
G, I	-0.05	0.00	-0.06	0.00	-0.06	0.00	M, R	0.05	0.19	0.06	0.20	0.06	0.21
G, M	0.01	0.05	0.01	0.05	0.01	0.04	M, L	0.21	0.33	0.23	0.36	0.26	0.37
G, S	-0.03	0.04	-0.03	0.04	-0.04	0.04	M, D	0.08	0.19	0.08	0.21	0.09	0.23
G, R	0.00	0.02	0.00	0.02	0.00	0.02	M, V	-0.27	0.00	-0.28	-0.02	-0.27	-0.03
G, L	0.00	0.05	0.00	0.04	0.00	0.03	M, A	0.16	0.36	0.16	0.36	0.14	0.34
G, D	-0.02	0.01	-0.02	0.01	-0.03	0.01	S, R	0.08	0.28	0.10	0.32	0.11	0.36
G, V	0.00	0.09	0.00	0.08	0.00	0.08	S, L	0.24	0.42	0.27	0.46	0.30	0.49
G, A	-0.05	0.02	-0.06	0.03	-0.06	0.03	S, D	0.13	0.31	0.15	0.34	0.18	0.38
J, K	-0.05	-0.03	-0.05	-0.03	-0.05	-0.04	S, V	-0.18	-0.01	-0.18	-0.01	-0.15	-0.01
J, I	0.01	0.09	-0.01	0.09	-0.01	0.08	S, A	0.08	0.35	0.09	0.38	0.10	0.41
J, M	0.00	0.01	0.00	0.01	0.00	0.01	R, L	0.09	0.43	0.10	0.44	0.11	0.44
J, S	-0.01	0.02	-0.01	0.02	-0.02	0.02	R, D	0.04	0.26	0.05	0.30	0.06	0.32
J, R	0.00	0.02	0.00	0.02	0.00	0.02	R, V	-0.03	0.00	-0.04	0.00	-0.04	0.00
J, L	-0.02	0.02	-0.01	0.02	-0.01	0.02	R, A	0.04	0.25	0.04	0.27	0.05	0.29
J, D	-0.01	0.01	-0.01	0.01	-0.01	0.01	L, D	0.21	0.45	0.22	0.48	0.23	0.49
J, V	0.00	0.02	0.00	0.01	0.00	0.01	L, V	-0.05	-0.02	-0.06	-0.02	-0.06	-0.03
J, A	-0.03	0.00	-0.04	0.00	-0.04	-0.01	L, A	0.09	0.38	0.09	0.39	0.10	0.39
K, I	0.05	0.08	0.05	0.09	0.06	0.09	D, V	-0.05	-0.01	-0.05	-0.01	-0.06	-0.01
K, M	0.02	0.03	0.02	0.03	0.02	0.03	D, A	0.09	0.39	0.09	0.40	0.10	0.40
K, S	0.04	0.08	0.05	0.08	0.05	0.09	V, A	-0.09	0.01	-0.08	0.01	-0.07	0.01

Table 4.4 Simulated correlations *CIs*, ξ=0.2

Hints about the retention ability for more complex dependence relationships can be gathered considering mixed moments. In details, 500 of those have been selected raising each variable to a power given by a random integer in [0, 4].

order	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
frequency	2	5	6	15	36	44	54	56	73	61	52	40	27	16	6	5	1	1
δ=0.3	0.06	0.23	0.24	0.23	0.21	0.22	0.21	0.16	0.23	0.18	0.19	0.23	0.26	0.23	0.07	0.13	0.04	0.04
δ=0.5	0.06	0.29	0.27	0.25	0.23	0.25	0.23	0.21	0.24	0.21	0.21	0.24	0.26	0.24	0.09	0.14	0.05	0.06
δ=0.7	0.07	0.37	0.37	0.32	0.28	0.30	0.28	0.29	0.27	0.26	0.25	0.27	0.30	0.28	0.10	0.19	0.06	0.09

Table 4.5 *MAPEs* w.r.t. moment orders, ξ=0.2

By means of 1,600 replications of data simulation *MAPEs* are calculated for each mixed moment. In table 4.5 *MAPEs* are averaged w.r.t. moment orders according to each class size. Moreover, some linear and non linear functions of variables, when the outcomes are grouped according to the 2 digit *NACE* and enterprise size in term of workers (up to 3, left closed intervals 3 – 6, 6 – 10, 10 – 20, 20 – 100), were considered. For each stratum η and each function *g*, *MAPE*s are calculated over the *s* replications performed:

$$G_\eta = \frac{1}{s}\sum_h \left| \frac{g_{\eta h}\left(y_{j_1},\ldots,y_{j_u}\right)}{g_\eta\left(y_{j_1},\ldots,y_{j_u}\right)} - 1 \right|$$

Averaging those outcomes using strata size $\{\nu_\eta\}$ as weights, $G=(\Sigma\nu_\eta)^{-1}\Sigma G_\eta\nu_\eta$ is defined. Since for every δ, simulated mean (and consequently total) values are close to those observed in each stratum, increasing δ implies decreasing standard deviations and a growing concentration around stratum quantities as table 4.6 shows.

δ	T	P	M	S	L	C	T/P	M/C	S/C	L/C	P/C	T/L
0.3	0.17	0.17	0.47	0.23	0.11	0.15	0.04	0.26	0.19	0.16	0.19	0.19
0.5	0.15	0.15	0.41	0.21	0.09	0.13	0.04	0.24	0.17	0.14	0.17	0.17
0.7	0.13	0.13	0.33	0.17	0.07	0.11	0.03	0.21	0.15	0.13	0.14	0.14

Table 4.6 Averaged *MAPEs* for stratum quantities, ξ=0.2

5 Conclusions and Future Work

A data simulation method based on Dirichlet processes for spline regression residuals has been proposed. The choice of ranks as regressors is intended to avoid the systematic part of each simulated value conveys information not predictable from ordinal ones: from this perspective, operating on residuals makes it possible a control of those fractions of intensities which are crucial for the usefulness of disclosure attempts. Residuals are sampled inverting their posterior cumulative distribution function. The trade-off between accuracy and protection against disclosure is controlled through cumulative distribution function tail penalization; in turn, this is suitably accomplished by choosing the base distribution of the process in terms of the marginal empirical distribution functions, i.e. a uniform distribution that covers the central part of the support. Each simulated variable is centred on the observed mean, so that the distortion implied by the univariate distribution tail penalization is partially corrected. Multivariate dependence relationships are maintained ordering simulated samples according to the matrix of ranks. The irrelevance of explicit conditional independence assumptions regarding subsets of variables, often used in the literature about the matter, avoids model specification problems. Moreover, empirical copula representation of multivariate dependences reduces the computational burden ensuring a growth only linear w.r.t. the number of variables. Exercises conducted on Small and

Medium Enterprises survey data of year 2004, collected by the Italian National Institute of Statistics, have shown encouraging results. Univariate and multivariate features have been investigated. About the former, reproduction abilities are satisfactory, at least for first four cumulants as well order statistics do not belonging to tails (since in the proposed setting, values far from each univariate density core need to be changed for protection aims). For the latter linear and non linear dependence preserving has been verified. Considering all simulation settings, correlations are satisfactory maintained on the whole. Mean Absolute Percentage Errors for a sample of 500 mixed moments related to five variables randomly chosen and orders from 2 to 19 are meaningful of non linear multivariate dependence retention. Linear and non linear variable functions like ratios referred to subdomains totals (49 class of economic activity and 5 class of size in term of workers) have been considered too. Since simulated mean (and consequently total) values are close to those observed in each stratum, the greater the penalization of tails, the greater the concentration around observed stratum quantities. Although simulations do not take into account any information about strata, *MAPEs* are generally less than 0.2, with the exception of quantities involving the variable M. Due to the absence of any form of rank swapping the protection offered by simulated values relies on the idea the intrusion utility decrease when changes are achieved w.r.t. observed values. The Symmetrised Mean Absolute Percentage Error over values to be secured offers a measure bounded on the unit interval which summarizes that protection ability. Especially w.r.t. multivariate dependence relationships, increasing the shrinking towards any base distribution of each Dirichlet process improves the protection with some drawback for the accuracy. The possibility of extending the proposed method to *SDC* oriented simulations for discrete quantitative or qualitative data remains to be verified in further studies.

Acknowledgments. Istat is not responsible for any view or result presented. The authors were supported by the European Project ESSnet-SDC in the field of Statistical Disclosure Control.

References

Burridge, J. (2003). Information preserving statistical obfuscation. *Statistical Computing, 13*. 321–327.

Domingo-Ferrer, J., Mateo-Sanz, J.M. (2002). Practical Data-Oriented Micro-aggregation for Statistical Disclosure Control. *IEEE Transactions on Knowledge and Data Engineering, 14(1)*. 189-201.

Domingo-Ferrer, J., Mateo-Sanz, J.M., Sebé, F. (2006). Information Loss in Continuous Hybrid Microdata: Subdomain-Level Probabilistic Measures. *Studies in Fuzziness and Soft Computing, 197/2006.* 287-298. Springer Berlin / Heidelberg.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics, 1*. 209-230

Ferguson, T.S. (1974). Prior distributions on space of probability measures. *The Annals of Statistics, 2*. 615-629

Fienberg, S.E., McIntyre, J. (2005). Data Swapping: Variations on a Theme by Dalenius and Reiss. *Journal of Official Statistics, 21(2)*. 309-323

Foschi, F. (2009). Artificial data through calibration and empirical copulas. *Paper invited to the Joint UNECE/Eurostat work session on statistical data confidentiality*. Bilbao, Spain, 2-4 December 2009.

Hiort, N., Holmes, C., Mueller, P. and Walker, S. (2010) Bayesian Nonparametrics, Cambridge University Press.

Muliere, P., Secchi P. (1994). Una generalizzazione del bootstrap bayesiano. *Atti XXXVII Riunione Scientifica SIS, 2.* 527-534.

Muliere, P., Secchi P. (1996). Bayesian nonparametric predictive inference and bootstrap techniques. *Annals of the Institute of Statistical Mathematics, 48*. 663-673.

Muralidhar, K., Parsa, R., Sarathy, R. (1999). A General Additive Data Perturbation Method for Database Security. *Management Science, 45(10)*. 1399-1415.

Muralidhar, K., Sarathy, R. (2006). Data shuffling - A new masking approach for numerical data. *Management Science, 52(5)*. 658-670.

Nelsen, R.B. (2006). *An Introduction to Copula*. Springer-Verlag.

Polettini, S. (2003). Maximum entropy simulation for data protection, Statistics and Computing, 13(4), 307-320.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rubin, D.B. (1981), The Bayesian bootstrap. *The Annals of Statistics, 9*. 130-134.

Rueschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference, 139(11)*. 3921-3927