Quantile-based Bootstrap Methods to Generate Continuous Synthetic Data

Daniela Ichim
Istituto Nazionale di Statistica
via Cesare Balbo, 00184
Rome, Italy
ichim@istat.it

ABSTRACT

To face the increasing demand from users, National Statistical Institutes (NSI) release different information products. The dissemination of this information should be performed in full compliance with the regulations pertaining to the privacy of respondents. One product that could belong to a dissemination portfolio is represented by synthetic data. In this paper a very brief review of several methods to generate synthetic data is given. The emphasis is put on bootstrap methods that might be used in complex surveys. A quantile-based bootstrap method is proposed, avoiding any model assumption. Different bootstrap strategies were empirically compared from the point of view of some univariate statistics and in a linear regression framework. The Italian Structure of Earnings Survey 2006 data were used in these preliminary experiments.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications–statistical databases; K.4.1 [**Computers and Society**]: Public Policy Issues–privacy

General Terms

Privacy

Keywords

statistical disclosure control, microdata dissemination, data utility, synthetic data, bootstrap

1. INTRODUCTION

The National Statistical Institutes (NSI) mission is to produce reliable, impartial, transparent, accessible and pertinent information. The dissemination of this information should be performed in full compliance with the regulations pertaining to the privacy of respondents.

After the removal of direct identifiers, e.g. name and address, other indirect identifiers, called key variables, could still allow the disclosure of some confidential information on respondents. In business microdata files, both categorical and continuous key variables are commonly registered. Due to the latter, each record is almost a unique case; risk measures based on rareness concepts are not appropriate. In the statistical disclosure control literature, the generation of synthetic data is considered as a strategy to minimize the disclosure risk, especially when the synthetic values are different from the original values. There are several approaches to the disclosure concept. The first one, the inferential disclosure, was introduced in [3]. More precisely, a disclosure is said to occur when the intruder, using the released microdata file, is able to increase his/her information about a target unit. The implementation of this scenario in practical situations requires distributional and modelling assumptions about what an intruder might a priori know. Even if it is not explicitly stated, the inclusion of the target unit in the released sample is also assumed. This leads to another definition of disclosure, namely the identity disclosure, see [16]. An identity disclosure is said to occur when the intruder is able to identify a unit in the microdata file. There is a strong relationship between attribute and inferential disclosure. An exact disclosure is a particular type of identity disclosure, when a respondent is identified using only the exact values of the key variables. Anyway, issues related to inferential disclosure should be taken into consideration, too.

Generally, synthetic data are viewed as an alternative to masked data. The usefulness of synthetic data is usually evaluated from the point of view of the analytical validity. Anyway, some other requirements, e.g. data structure and ease of use and implementation, should also be considered.

In Section 2 a very brief review of some methodologies that generate synthetic data is given. In Section 3 several bootstrap methods that might be used to create synthetic data are discussed. The emphasis is put on issues related to the complex surveys and a quantile based method is proposed. In Section 4, the results of the application of different bootstrap strategies to the Italian Structure of Earnings Survey 2006 are compared from the point of view of some univariate statistics and in a linear regression framework.

2. MODEL-BASED APPROACHES TO THE GENERATION OF SYNTHETIC DATA

In this section several methods for synthetic data generation are discussed. There are two general classes of methods. The first class includes the methods based on the distribution function estimation. An example is the Latin Hyper-

PAIS '10, March 22, 2010, Lausanne, Switzerland.

cube Sampling. The second category of methods includes the ones that model the relationships between survey variables. Well-known examples are the multiple imputation and IPSO methods.

Probability Distribution The oldest proposal, in [12], is to simulate a data set by randomly sampling from the distribution function of the original data set. This is a very simple method, easy to explain and implement. It has also the advantage that it works for both categorical and continuous variables. In spite of its simplicity, the method heavily relies on the identification (and estimation) of the cumulative distribution function. This is probably the reason why in many practical situations the NSIs do not generally adopt this approach. Moreover, its adaptation to data stemming from complex surveys was not yet discussed.

Latin Hypercube Sampling In [4], another method for generating multivariate synthetic data sets was proposed. The method uses a functional of the empirical distribution function to generate synthetic data. The protection is achieved by both resampling (Latin hypercube sampling) and a data transformation which reduces the informative content. The method preserves the univariate statistics as well the rank correlations. Consequently, the pairwise associations between variables are preserved. Nonetheless, the method is very time consuming. Different aspects of this methodology are discussed in [11].

IPSO Information Preserving Statistical Obfuscation (IPSO) was originally proposed in [2]. Suppose X and Y should be released, where the X are the confidential outcome attributes and Y are the key variables. A multiple regression of Y on X is performed and the fitted $\hat{Y}$ values are computed. Finally, attributes X and $\hat{Y}$ are released by IPSO in place of X and Y. Obviously, IPSO is efficient against any type of exact disclosure based on the key variables Y. The IPSO variant proposed in [13] reproduces, in a single computation step, the means and covariance matrix of the original data set. In [13], other IPSO-based methods are discussed, too.

Multiple Imputation In [18], a method for generating fully synthetic data sets by multiple imputation was proposed. All the non sampled observations from the population/sampling frame are considered as missing data. Subsequently, they are imputed according to the multiple imputation framework. Afterwards, several simple random samples from these fully imputed data sets could be released. The efficiency of this method strongly depends on the robustness and accuracy of the model used to impute the "missing" values. Since the released data sets could be random samples from the population, the sampling design complexity is no more an issue to deal with. From a data utility/validity point of view, the release of multiple data sets might not be so easily accepted by the users. This might happen because a multiple data set is not an "object" similar to the original survey data.

Data Shuffling The data shuffling was introduced in [15]. It might be interpreted as a generalization of the data swapping, see [20]. By using the data shuffling, values are indeed shuffled and consequently, a value of the i-th record could possibly be assigned to the j-th record, that of the j-th record to the k-th record, etc. Thus, the shuffled values are the original values, just shuffled in such a manner as to maintain the univariate statistics. Other quality criteria were assessed by means of some empirical experiments. The authors of the data shuffling claim that the disclosure risk is reduced to the minimum possible risk. The data shuffling is guided by a model used to preserve some association statistics.

Even if the released data are synthetic, they do not completely solve the problem if disclosure risk. From the disclosure risk point of view, if the released synthetic values are different from the original values, the risk of *exact disclosure* is considered very low. Probably the risk of *inferential disclosure* might not be negligible, depending on the adopted disclosure scenario. This statement holds especially for establishment survey data for variables having very skewed distributions. Indeed, in a business data framework, the (extreme) outliers are generally very hard to mask.

For each of the five discussed methods, the model assumptions should be rigorously tested. Moreover, the role played by the sampling design features is not always clear. It should also be noted that except for the multiple imputation and IPSO, the illustrated methods for generating synthetic data were designed for univariate variables.

In the next section, a class model-free methods to generate synthetic data will be analyzed. It is based on a well-known resampling procedure, i.e. the bootstrap.

3. QUANTILE-BASED BOOTSTRAP

3.1 Standard Bootstrap

Suppose we have a finite population of N elements labeled as $1, \ldots, N$. Let y_i be the value of a characteristic (or possibly a vector of characteristics) of the i-th unit of the finite population ($i = 1, \ldots, N$). In general, we are interested in a variety of (nonlinear) functions of y_i, ($i = 1, \ldots, N$), e.g. means or quantiles. Since it is not possible to observe the population values, the population characteristics are estimated by means of samples of size n, say $y_1, \ldots, y_n$.

The (non-parametric) bootstrap method is an application of the plug-in principle. By non-parametric, we mean that only y is known (observed) and no prior knowledge on the population density function (or the cumulative distribution function) is available. Originally, the bootstrap was introduced in [6] to estimate the standard error of an arbitrary estimator θ and to-date the basic idea remains the same.

The bootstrap methods discussed in this section involve the following common steps:

1. Using a suitable probability sampling scheme, generate a resample (or bootstrap sample, bootstrap replicate) of size n from the original sample.

2. Calculate the (nonlinear) statistic, named $\hat{\theta}$, using the resample or a suitable rescaled version of the resample. Denote it by $\hat{\theta}^*$.

3. Calculate $\hat{\theta}$ for a large number (say, B) of independent resamples. Let $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ be the estimates from the B independent bootstrap replicates. These $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ form the basis of inference for $\hat{\theta}$.

Even if it was initially introduced to compute the standard error of an estimator whose exact distribution could not be easily derived, this resampling methodology found its utility in many other statistical applications. In [9], different bootstrap methods and their potential are illustrated. In the field of statistical disclosure control, some bootstrap methods were proposed to preserve confidentiality in contingency tables, see [10] or [14].

3.2 Complex Surveys

The bootstrap is probably the most flexible and efficient method of analyzing survey data since it can be used to solve a variety of challenging statistical problems (e.g., variance estimation, imputation, small-area estimation, etc.) for complex surveys involving both smooth and nonsmooth statistics.

Various federal and private agencies routinely conduct large-scale sample surveys. In a typical sample survey, a suitable probability sampling scheme is used to collect data from a finite population. Population stratification and selection of ultimate sampling units in several stages are generally involved. A survey weight is usually attached to each unit in the sample to account for various factors such as unequal selection probability, non-response, post-stratification and calibration. The incorporation of these important survey weights in the bootstrap method has received considerable attention in the survey from finite populations literature. Bootstrap methods in finite populations are usually justified from the randomization approach to survey sampling.

The finiteness of the survey population, the complexity of the survey design and the complex weighting scheme all contribute to the challenging task of finding a valid bootstrap procedure.

A first simple bootstrap in finite population would be the resampling with replacement, but using probabilities directly related to the sampling weights.
In the literature, the important problem of recovering the f.p.c. (finite population correction, $f = n/N$) resulted in mainly two different approaches: (i) without-replacement bootstrap (BWO) and (ii) with-replacement bootstrap (BWR). The latter approach (BWR) attempts to adapt Efron's original bootstrap by carefully choosing the bootstrap sample size or by rescaling the generated bootstrap replicate. On the other hand, the first approach (BWO) tries to mimic the original SRSWOR (simple random sampling without replacement).

3.2.1 Without-replacement bootstrap - BWO

Obviously, drawing a bootstrap sample of size n without replacement does not make sense since it would provide the original sample. One might think of a naive bootstrap without-replacement sample of size $n^{'} = fn$ in order to capture the f.p.c. But this bootstrap sampling yields an overestimation of the true variance. In [19] the following method was proposed: if $N = kn$, create an artificial population of N units simply by copying each of the n elements in the original sample k times and then take a SRSWOR bootstrap sample of size n from this artificial population. When k is not an integer a randomization is needed. Asymptotically, when both n and N increase, with f fixed, the method provides a consistent variance estimator.

3.2.2 Rescaling - BWR

In finite populations literature, it was noted that it is not essential to mimic the original sampling design to capture the f.p.c, f. One could simply adapt Efron's original bootstrap by taking a larger bootstrap sample than is required if the original sampling plan were SRSWR (simple random sampling with replacement). A sample of size $n^{'} = (1 - f)^{-1}(n - 1)$ yielding the customary variance estimator was suggested. One problem with this method is that $n^{'}$ could be non-integer; some randomization might be needed in most practical situations.

In [17] a method which rescales the bootstrap replicate so as to recover the f.p.c. in the usual SRSWOR variance formula was proposed. This procedure selects samples of size m, say $\mathbf{y}^* = (y_1, \dots, y_m)$, from the original sample and then rescales the bootstrap sample by

$$\tilde{y}_i = \bar{y} + \frac{\sqrt{m}}{\sqrt{n-1}} \sqrt{1-f}(y_i^* - \bar{y}) \qquad (1)$$

In [17], a discussion on the possible choices of m is given, too.

3.3 Synthetic data

For the generation of synthetic data, the bootstrap was initially discussed in [8]. Given an original sample set $y = (y_1, \dots, y_n)$, the empirical cumulative distribution function (F) is computed. Next, F is smoothed and then it is sampled with replacement to obtain a synthetic microdata set Z. Needless to say, the smoothing choice plays an important role in this approach. Anyway, as remarked in [8], the bootstrap theory relies on the repeated sampling, not on a single sample. A similar approach based on a single bootstrap replicate may be found in [1]. Here, the synthetic value is a linear combination of the original value and its unique bootstrap replicate. The bootstrapping method is the one introduced in [6]. The parameter α defining the linear combination (or the synthetic percentage) might be used to control some data utility criteria. None of these approaches takes into account the sampling design and in particular the sampling weights.

In this work, an univariate synthetic data generator is proposed. Let $y = (y_1, \dots, y_n)$ be the original sample. Here it is proposed to generate the synthetic data set by using the n quantiles that could be calculated from a sample of size n. The idea is that all sample quantiles are defined as averages of consecutive order statistics. In other words, for $h = 1, \dots, n$, the statistic θ of interest introduced in section 3.1 is set equal to the h-th quantile. For example, let h be equal to 0.15. The bootstrap procedure would be as follows: 1) generate a bootstrap replicate (using standard bootstrap, BWO or BWR), 2) estimate $\hat{q}^*_{0.15}$, the 0.15-th quantile of the bootstrap sample, 3) repeat steps 1 and 2 a large number of times, say B and finally, 4) compute the quantile estimate $\hat{q}_{0.15} = 1/B \sum_1^B \hat{q}^*_{0.15}$. The steps 1-4 are then repeated for each of the n quantiles. Finally, the univariate synthetic data would be $\hat{q}_{1/n}, \hat{q}_{2/n}, \dots, \hat{q}_{n/n}$. Several general considerations hold. First, such a procedure would avoid any model-assumption. Second, it would be based on the generation of a large number of bootstrap samples.

It is important to stress here that any of the previously discussed bootstrap strategies may be applied for selecting the bootstrap replicates.

4. APPLICATION

In this section several simulations of the bootstrap methodologies in a real case-study are illustrated. First the data stemming from the Italian Structure of Earnings Survey are briefly described. Then, the results of the application of the bootstrap methodologies are discussed. The different bootstrap methodologies are compared from the point of view of some univariate statistics and in a linear regression framework.

4.1 Italian Structure of Earnings Survey

The Structure of Earnings Survey (SES) provides detailed information on the level and structure of remuneration of employees, their individual characteristics and the enterprise or local unit to which they belong to. The SES outcome represents an uniquely rich data source on gross earnings in Europe which is increasingly important for evidence-based policy making, in particular for monitoring economic growth and social cohesion. Furthermore, the SES data are indispensable for employers and employees as regards the demand and supply of labour.

In Italy, SES data were collected by means of a sampling survey. A two stage sampling scheme was followed. The enterprises sampling frame was the most up-to-date version of Archivio Statistico delle Imprese Attive (statistical business register of active enterprises). The employees were sampled through a proportional to size sampling scheme. The observed variables are indicated in the Commission Regulation 1916/2000. The optional variables were generally not observed in the Italian survey. At enterprise level, some structural economic variables were registered: principal economic activity ($Nace$), number of employees, geographical location ($Nuts$), form of economic and financial control and existence of collective pay agreements, etc. On employees, besides gender and age, variables related to education, profession and contractual position were surveyed. Annual and monthly earnings corresponding to the reference month (October 2006) were registered together with their various components. The working time was accounted for through variables like number of paid hours.

The sampling weights were computed in order to indicate each respondent representativeness, see [7]. Generally, the weights should satisfy some restrictions, for example the preservation of some known population totals. The enterprise sampling weights were derived by means of a multivariate calibration procedure, see [5]. To compute the employees final weights, the procedure indicated in [7] was followed.

4.2 Bootstrap Experiments

45 categories of the variable principal economic activity ($Nace$) were considered, spanning a wide range of economic sectors. Summary statistics on the number of observations in the sample and population are shown in table 1. The summary statistics were computed over the 45 categories of the principal economic activity variable. The population totals were computed using the sample weights. From the SES 2006 data set, five variables were selected, i.e. *Total gross earnings for a representative month* (ME), *Earnings related to overtime* (MO), *Special payment for shift work* (MS), *Total gross annual earnings in the reference year* (AE) and *Total annual bonuses* (AB). Summary statistics on these five variables are shown in table 2. It may be observed that each variable has a skewed distribution. Moreover, extreme outliers are always present. Additionally, variables like MS or MO are extremely sparse, probably because this kind of remuneration is not included in the working contract. In table 2, the summary statistics were computed considering a single category of the principal economic activity. Anyway, the same qualitative conclusions hold for the other categories.

By principal economic activity category, the standard deviations of the employees survey weights vary from a minimum equal to 0 up to a maximum value of 650. This suggests that there is a significant variability of these survey weights which were derived from the first order inclusion probabilities.

Table 1: Summary statistics of the number of units in the sample and population. Q_1 = the first quantile, Q_2 = the second quantile (median) and Q_3 = the third quantile

Statistic	Sample	Population
Min	49	1643
Q_1	442	14829
Q_2	952	38820
Mean	1397	74273
Q_3	1676	89528
Max	6109	491850

Table 2: Summary statistics of the five variables ME, MO, MS, AE and AB, respectively. Q_1 = the first quantile, Q_2 = the second quantile (median) and Q_3 = the third quantile

Statistic	ME	MO	MS	AE	AB
Min	421	0	0	602	0
Q_1	1775	0	0	20727	0
Q_2	2211	0	0	26746	1943
Mean	2516	146	9.92	29990	2632
Q_3	2959	220.5	0	35444	3908
Max	12300	2231	663	173852	31972

The quantile-based approach was applied to generate synthetic data. As discussed in the previous section, this approach has the advantage of generating a single synthetic data set, that is, an object similar to the original data. In the core of the quantile-based synthetic data generator, the BWO, BWR and the standard bootstraps were used to select the bootstrap replicate. A stratified bootstrap was applied, using the principal economic activity as stratification variable. That is, for each stratum, an independent bootstrapping procedure was applied. In all cases, B, the number of bootstrap replications was set equal to 1000. For the BWR method, the parameter m was set equal to $n - 3$, as suggested in [17]. For the BWR method, no randomization was applied when k was not an integer number. In a first run, for the three bootstrap versions, BWR, BWO and standard, for each unit the probabilities of inclusion in the bootstrap replicate were set equal to $1/n$. In a second experiment, those probabilities of inclusion in the bootstrap replicate were set equal to the inverse of their weight. This approach might be interpreted as a tentative to mimic their probability of inclusion in the sample.

4.2.1 Descriptive statistics - data utility and risk of disclosure

In table 3, several summary statistics on the percentage relative variation between the synthetic and original values are shown. Let T be the statistic of interest; denote by T^* the statistic computed from the synthetic values and by T_o the statistic computed from the original values. Then the percentage relative variation is defined as $\frac{T_o - T^*}{T_o} * 100$. As for T, the minimum, mean, maximum, variance and weighted means were considered. T = minimum and T = maximum are related to the disclosure risk, while T = (weighted) mean and T = variance might be considered data utility issues. For each stratum defined by the categories of the principal

Table 3: Summary statistics of the percentage variations of T (minimum, mean, maximum, variances and weighted means) for ME and AE. Q_1 = the first quantile, Q_2 = the second quantile (median) and Q_3 = the third quantile. s = standard bootstrap, bwo = BWO bootstrap method, bwr = BWR bootstrap method, ".w" = unequal inclusion probabilities in the bootstrap replicate.

	ME						**AE**					
Type	Min	Q_1	Q_2	Mean	Q_3	Max	Min	Q_1	Q_2	Mean	Q_3	Max
T= **Min**												
s	-975	-103	-46	-111	-21	-2.57	-1598	-154	-50	-187.50	-24.51	-3.60
s.w	-647	-145	-57	-137	-19	-1.74	-2885	-211	-58	-230.40	-27.26	-5.12
bwo	-9607	-95	-49	-110	-21	-2.51	-1584	-155	-50	-187.90	-26.24	-3.96
bwo.w	-668	-150	-50	-152	-20	-9.52	-2759	-119	-39	-209.40	-21.24	-5.11
bwr	-1046	-132	-83	-133	-28	-3.79	-1842	-214	-91	-245.30	-47.47	-13.86
bwr.w	-748	-162	-85	-161	-26	-2.70	-3184	-240	-81	-290.50	-41.89	-11.62
T= **Mean**												
s	-1.07	-0.20	-0.11	-0.17	-0.06	0.04	-1.44	-0.25	-0.11	-0.21	-0.05	0.02
s.w	-12.69	-1.53	0.51	0.69	2.60	21.52	-15.77	-2.07	0.22	0.64	2.81	21.80
bwo	-1.14	-0.17	-0.09	-0.16	-0.05	0.03	-1.24	-0.18	-0.10	-0.19	-0.05	0.03
bwo.w	-5.73	-1.37	0.50	1.28	2.74	21.25	-4.03	-2.07	0.46	1.28	3.61	21.49
bwr	-1.10	-0.19	-0.09	-0.16	-0.02	0.05	-1.17	-0.22	-0.09	-0.19	-0.03	0.07
bwr.w	-12.45	-1.48	0.53	0.69	2.54	21.21	-15.60	-1.92	0.21	0.63	2.70	21.39
T= **Max**												
s	0.61	3.64	5.56	6.38	8.84	19.71	0.50	3.46	6.18	6.94	10.12	16.73
s.w	0.55	3.87	6.63	9.05	12.15	30.02	0.81	3.60	5.87	9.67	11.88	36.04
bwo	0.47	3.74	5.37	6.34	8.36	19.30	0.40	3.55	5.83	6.96	10.11	18.16
bwo.w	2.69	5.11	7.18	8.61	11.66	19.53	0.98	3.16	5.91	7.73	9.54	24.76
bwr	1.69	4.72	7.16	7.68	9.85	20.99	1.84	4.78	7.52	8.27	11.31	18.37
bwr.w	0.81	5.32	7.74	10.28	13.25	30.31	1.08	4.54	7.85	10.81	12.94	35.22
T= **Var**												
s	0.04	1.31	2.10	3.14	4.06	15.05	-0.15	1.53	2.47	3.47	4.89	16.48
s.w	-31.32	-0.37	8.67	7.5	15.77	39.77	-19.45	-1.50	7.933	7.88	16.25	39.63
bwo	0.32	1.39	2.27	3.15	3.96	14.04	0.50	1.55	2.50	3.50	4.58	15.52
bwo.w	-13.29	0.28	9.53	8.59	14.21	38.84	-17.34	-0.34	9.49	8.74	16.10	38.48
bwr	2.26	3.97	5.43	6.46	7.13	21.75	2.12	4.11	5.321	6.72	7.55	23.24
bwr.w	-29.21	2.84	12.29	10.72	18.24	41.33	-17.4	3.42	10.77	11.12	19.78	40.75
T= **WMean**												
s	-75.89	-6.23	-0.32	-1.56	6.76	30.11	-1.44	-0.20	-0.10	-0.20	-0.02	0.15
s.w	-79.44	-6.93	0.05	-0.99	9.16	36.77	-15.46	-1.89	0.41	0.80	3.41	22.94
bwo	-75.99	-6.21	-0.38	-1.55	6.69	30.09	-1.24	-0.19	-0.08	-0.19	-0.02	0.31
bwo.w	-37.16	-7.01	-0.56	0.86	6.60	36.60	-3.93	-1.91	0.59	1.36	3.06	22.59
bwr	-74.72	-6.05	-0.37	-1.49	6.78	29.87	-1.17	-0.22	-0.09	-0.18	0.02	0.26
bwr.w	-78.20	-6.75	0.04	-0.91	9.21	36.42	-15.31	-1.81	0.36	0.79	3.30	22.66

economic activities, the percentage relative variations of T were computed. In order to summarize this information, the minimum, median, mean etc. indicators were computed over the strata. For brevity, only the variables ME and AE are presented in table 3. Similar results were obtained for the other three variables, even those are more sparse. The other summary descriptive statistics T (quartiles, median, etc) have an intermediate behaviour. From table 3, it may be observed that the minimum and the maximum values are always modified. During the simulations, it was observed that the values on the tails are generally modified in a significant manner. That is, the values with the highest risk of disclosure, i.e. the values on the tails, receive a greater amount of modification, especially for the bwr.w method. The values in intervals with high density of values might be considered safe because the possible intruder would have a greater amount of uncertainty in identifying a unit or its corresponding values.

4.2.2 Linear models - data utility

To assess the quality of the bootstrap approaches in a regression framework, two very simple linear models were considered. *Total gross annual earnings in the reference year* (AE) was considered as response variable, while the explanatory variables were *Total gross earnings for a representative month* (ME) and *Total annual bonuses* (AB). The only difference between the tested regression models was the consideration of an intercept. That is, in Model1, the intercept was included, while in Model2 the intercept was not considered. Consequently, the two used models may be easily written as:

$$\begin{array}{lll} \text{Model1}: & \text{AE} & = \gamma_1 + \alpha_1 \text{ME} + \beta_1 \text{AB} \\ \text{Model2}: & \text{AE} & = \alpha_2 \text{ME} + \beta_2 \text{AB} \end{array} \qquad (2)$$

The ordinary least squares method was used to compute the estimates $\hat{\gamma}_1, \hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2, \hat{\beta}_2$ of $\gamma_1, \alpha_1, \beta_1, \alpha_2, \beta_2$, respectively. To evaluate the synthetic data usage in a hypothesis testing framework, the standard errors of the estimated parameters were estimated as well.

For each stratum, the goodness of fit R^2 statistics were computed. Denote by R^{2*} and R^2_o the corresponding statistic estimated using the synthetic data and original data, respectively. The similarity between the original and synthetic data was evaluated by means of the summary statistics of $\frac{R^{2*}}{R^2_o}$. These summary statistics are presented in table 4, together with the summary statistics of the original R^2_o values. It may be observed that there is no much variability among the bootstrap versions. Moreover, the R^2 statistics of Model2 seems better preserved. Anyway, it should be noted that even for the original data, from this goodness of fit criteria point view, the Model1 seems less adequate, see the last row of table 4. Both the R^2_o and $\frac{R^{2*}}{R^2_o}$ values could be considered sufficiently high to support the usage of Model2 in practice. Using the synthetic data and original data, for both Model1 and Model2, generally, the p-values of the fitted coefficients were always sufficiently small to reject the null hypothesis of their elimination from the model. This statement holds for all types of bootstrap methods. It was observed that the estimates $\hat{\gamma}_1, \hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2, \hat{\beta}_2$ computed using the original data do not belong to the 95% or 99% confidence intervals of the estimates $\hat{\gamma_1}^*, \hat{\alpha_1}^*, \hat{\beta_1}^*, \hat{\alpha_2}^*, \hat{\beta_2}^*$ computed using any of the bootstrap synthetic data. This feature is probably due to the exaggerate simplicity of the tested models, i.e. Model1 and Model2. Further investigation on the impact of bootstrap methods on (linear) regression models should be performed.

4.2.3 Confidence Intervals

Data quality

The quality of the synthetic data set might be measured by the differences between the original data and the bootstrapped data, too. Since the quantile based bootstrap synthetic data set generation has the same dimension as the original data set, the element-wise differences can be computed. For each of the five variables, ME, MO, MS, AE and AB, it was assessed whether the original value v belongs to one of the corresponding intervals $v^* \pm se_{v^*}$ and $v^* \pm 2se_{v^*}$. Here se_{v^*} denotes the standard error of the estimate v^*; its estimation was possible using the B replicates of the quantiles, in a classical bootstrap manner. For each principal economic activity category, the number of original values belonging to the intervals derived from the corresponding bootstrap estimate was computed. In table 5, the summary statistics of these counts are illustrated. The influence of the sparse variables, e.g. MO or MS, may be easily observed. Indeed, the percentages of original values falling in the corresponding bootstrap interval decreases significantly for the sparse variables, almost independently on the bootstrap variant.

Disclosure risk

For each of the five variables, for each bootstrap confidence interval BCI (computed as $v^* \pm 2se_{v^*}$), the number of original values belonging to the BCI interval was calculated. Not only the original values corresponding to the bootstrap values were considered. In table 6, the summary statistics of the number BCI intervals containing no original values is shown for the variables MO, MS and AB. The variables ME and AE were not considered since their relative statistics were always equal to 0. This difference between ME, AE and MO, MS and AB is again due to the sparsity of the latter variables. It may be observed that achieved protection depends on the degree of sparsity rather than on the bootstrap method. In table 7, some summary statistics of the percentages of BCI intervals containing only one or two original values (not necessarily the corresponding values) is shown. The other summary statistics (min, Q_1 and Q_2) were always equal to 0. As usually in this work, the summary statistics were computed over the categories of the principal economic activity variable. In table 7, it may be observed that the number of BCI intervals containing less than 3 original values is very reduced. It follows that the risk of inferential disclosure is kept under control since the synthetic values do not contain really isolated values.

Disclosure risk

In table 8 the summary statistics of the number of records having a bootstrap value different from the original value are shown. Since the statistics of interest, θ, in the proposed bootstrap procedure, see Section 3.3, were the quantiles, it was possible to match the bootstrapped values and the original values by means of their corresponding ranks. The summary statistics were computed over the principal economic activity categories. It may be observed that the presented percentages are generally quite high. For the extremely sparse variables, e.g. MS, those percentages are not so high. This is surely explained by the zero value of a large

Table 4: Summary statistics for ratios $\frac{R^{2*}}{R_o^2}$. Both Model1 (with intercept) and Model2 (without intercept). Q_1 = the first quantile, Q_2 = the second quantile (median) and Q_3 = the third quantile. s = standard bootstrap, bwo = BWO bootstrap method, bwr = BWR bootstrap method, ".w" = unequal inclusion probabilities in the bootstrap replicate.

	Model1						**Model2**					
Type	Min	Q_1	Q_2	Mean	Q_3	Max	Min	Q_1	Q_2	Mean	Q_3	Max
s	0.23	0.38	0.51	0.51	0.61	0.82	0.69	0.83	0.87	0.86	0.91	0.96
s.w	0.23	0.39	0.50	0.51	0.62	0.82	0.69	0.83	0.87	0.87	0.92	0.96
bwo	0.23	0.38	0.51	0.50	0.61	0.83	0.69	0.83	0.87	0.86	0.91	0.96
bwo.w	0.25	0.42	0.50	0.51	0.62	0.82	0.76	0.84	0.86	0.86	0.90	0.95
bwr	0.23	0.38	0.51	0.50	0.61	0.83	0.71	0.83	0.88	0.87	0.92	0.96
bwr.w	0.23	0.39	0.51	0.51	0.62	0.82	0.71	0.84	0.88	0.87	0.92	0.96
R_o^2	0.43	0.83	0.87	0.86	0.92	0.96	0.89	0.96	0.97	0.97	0.98	0.99

Table 5: Summary statistics of the percentage of original values that are within a quantity r (either se^* or $2se^*$) far from the corresponding bootstrap value v^*. Q_1 = the first quantile, Q_2 = the second quantile (median) and Q_3 = the third quantile. s = standard bootstrap, bwo = BWO bootstrap method, bwr = BWR bootstrap method, ".w" = unequal inclusion probabilities in the bootstrap replicate.

		$r = se^*$						$r = 2se^*$					
Type	Var	Min	Q_1	Q_2	Mean	Q_3	Max	Min	Q_1	Q_2	Mean	Q_3	Max
s	ME	0.94	1.84	2.52	3.81	4.24	15.28	1.91	3.87	5.06	7.75	8.13	31.58
s	MO	23.30	40.67	49.00	50.29	56.56	87.56	23.30	40.67	49.00	50.30	56.56	87.56
s	MS	0.00	14.84	20.39	23.86	38.18	49.30	0.00	14.84	20.39	23.86	38.18	49.30
s	AE	95.92	99.65	99.88	99.62	99.96	100.00	100.00	100.00	100.00	100.00	100.00	100.00
s	AB	65.79	87.35	94.51	91.54	97.86	100.00	65.79	87.35	94.51	91.72	98.65	100.00
s.w	ME	0.71	1.86	2.52	3.74	4.04	15.28	1.86	4.00	5.17	7.66	7.84	31.58
s.w	MO	0.00	16.52	32.51	30.27	44.08	79.17	0.09	32.60	46.46	41.73	49.88	86.11
s.w	MS	0.00	2.08	10.98	12.51	17.59	39.91	0.00	3.62	15.52	16.48	23.60	43.69
s.w	AE	2.70	13.20	35.57	46.66	82.87	99.80	6.97	34.15	78.74	66.49	99.72	100.00
s.w	AB	0.11	12.34	32.90	37.83	55.79	100.00	2.84	31.30	66.07	59.06	84.57	100.00
bwr	ME	0.84	1.76	2.47	3.68	4.29	15.28	1.96	3.94	5.02	7.62	7.95	32.24
bwr	MO	23.21	39.07	46.84	47.44	52.19	82.13	23.78	40.93	47.89	49.03	54.64	83.48
bwr	MS	0.00	13.80	19.08	23.04	35.89	47.38	0.00	14.01	19.23	23.54	36.55	48.04
bwr	AE	82.93	95.36	98.40	96.71	99.35	99.89	94.88	100.00	100.00	99.83	100.00	100.00
bwr	AB	61.84	83.55	88.98	87.74	94.63	98.77	67.11	86.02	91.87	90.61	97.13	100.00
brw.w	ME	0.84	1.76	2.47	3.68	4.29	15.28	1.96	3.94	5.02	7.62	7.95	32.24
brw.w	MO	23.21	39.07	46.84	47.44	52.19	82.13	23.78	40.93	47.89	49.03	54.64	83.48
brw.w	MS	0.00	13.80	19.08	23.04	35.89	47.38	0.00	14.01	19.23	23.54	36.55	48.04
brw.w	AE	82.93	95.36	98.40	96.71	99.35	99.89	94.88	100.00	100.00	99.83	100.00	100.00
brw.w	AB	61.84	83.55	88.98	87.74	94.63	98.77	67.11	86.02	91.87	90.61	97.13	100.00
bwo	ME	0.84	1.76	2.47	3.68	4.29	15.28	1.88	3.75	5.00	7.61	8.23	30.92
bwo	MO	23.21	39.07	46.84	47.44	52.19	82.13	24.27	40.92	48.86	50.49	56.03	87.50
bwo	MS	0.00	13.80	19.08	23.04	35.89	47.38	0.00	14.58	21.05	23.85	37.33	49.74
bwo	AE	82.93	95.36	98.40	96.71	99.35	99.89	100.00	100.00	100.00	100.00	100.00	100.00
bwo	AB	61.84	83.55	88.98	87.74	94.63	98.77	71.05	86.69	94.24	91.79	98.73	100.00
bwo.w	ME	0.84	1.76	2.47	3.68	4.29	15.28	1.90	3.94	5.17	7.50	7.67	31.58
bwo.w	MO	23.21	39.07	46.84	47.44	52.19	82.13	0.19	33.20	43.84	41.56	49.81	88.89
bwo.w	MS	0.00	13.80	19.08	23.04	35.89	47.38	0.00	3.32	15.52	16.16	22.11	44.75
bwo.w	AE	82.93	95.36	98.40	96.71	99.35	99.89	6.78	34.00	81.10	66.49	99.58	100.00
bwo.w	AB	61.84	83.55	88.98	87.74	94.63	98.77	3.26	29.97	64.77	58.53	82.48	100.00

Table 6: Summary statistics of the percentages bootstrap confidence intervals containing no original values. Q_1 = the first quantile, Q_2 = the second quantile (median) and Q_3 = the third quantile. s = standard bootstrap, bwr = BWR bootstrap method, w = unequal inclusion probabilities.

Type	Var	Min	Q_1	Q_2	Mean	Q_3	Max
s	MO	12.44	43.44	51.00	49.70	59.33	76.70
s.w	MO	8.15	44.02	50.35	49.08	57.83	72.73
bwr	MO	14.03	45.36	52.11	50.62	59.07	76.22
bwr.w	MO	7.92	45.59	52.22	50.35	58.42	73.78
bwo	MO	12.50	43.97	51.14	49.51	59.08	75.73
bwo.w	MO	7.69	44.08	51.14	49.18	56.98	72.89
s	MS	50.70	61.82	79.61	76.14	85.16	100.00
s.w	MS	40.95	63.55	78.36	76.73	86.46	100.00
bwr	MS	51.52	62.83	80.77	76.33	85.99	100.00
bwr.w	MS	42.94	63.89	79.61	77.12	87.13	100.00
bwo	MS	50.26	62.67	78.95	76.15	85.42	100.00
bwo.w	MS	43.53	64.26	79.10	76.98	86.54	100.00
s	AB	0.00	1.35	5.49	8.28	12.65	34.21
s.w	AB	0.00	0.94	6.69	7.79	12.43	25.11
bwr	AB	0.00	2.15	5.73	8.58	13.17	32.89
bwr.w	AB	0.00	1.45	6.83	8.34	12.15	26.11
bwo	AB	0.00	1.27	5.76	8.21	13.31	28.95
bwo.w	AB	0.00	0.72	6.72	7.96	12.60	26.16

Table 7: Summary statistics of the percentages of bootstrap confidence intervals containing only one or two original values. Q_3 = the third quantile. s = standard bootstrap, bwr = BWR bootstrap method, ".w" = unequal inclusion probabilities in the bootstrap replicate.

		Uniques			Doubles		
Type	Var	Mean	Q_3	Max	Mean	Q_3	Max
s	ME	0.00	0.00	0.00	0.02	0.00	0.68
s.w	ME	0.00	0.00	0.12	0.04	0.00	0.68
bwr	ME	0.00	0.00	0.00	0.02	0.00	0.68
bwr.w	ME	0.00	0.00	0.12	0.00	0.00	0.12
bwo	ME	0.00	0.00	0.00	0.03	0.00	0.68
bwo.w	ME	0.00	0.00	0.12	0.00	0.00	0.12
s	MO	0.00	0.00	0.00	0.03	0.00	0.49
s.w	MO	0.00	0.00	0.19	0.07	0.00	1.39
bwr	MO	0.04	0.00	1.39	0.09	0.00	2.78
bwr.w	MO	0.00	0.00	0.15	0.11	0.00	2.78
bwo	MO	0.00	0.00	0.00	0.07	0.04	1.39
bwo.w	MO	0.00	0.00	0.19	0.08	0.03	1.39
s	MS	0.00	0.00	0.00	0.05	0.00	2.04
s.w	MS	0.02	0.00	0.71	0.08	0.00	2.04
bwr	MS	0.09	0.00	1.24	0.06	0.00	2.04
bwr.w	MS	0.10	0.00	1.06	0.09	0.00	2.04
bwo	MS	0.00	0.00	0.00	0.07	0.00	2.04
bwo.w	MS	0.02	0.00	0.71	0.07	0.00	2.04
s	AE	0.00	0.00	0.00	0.07	0.00	2.04
s.w	AE	0.00	0.00	0.04	0.09	0.02	2.04
bwr	AE	0.00	0.00	0.00	0.01	0.00	0.49
bwr.w	AE	0.01	0.00	0.19	0.07	0.00	2.04
bwo	AE	0.00	0.00	0.00	0.08	0.00	2.04
bwo.w	AE	0.00	0.00	0.04	0.09	0.00	2.04
s	AB	0.00	0.00	0.00	0.02	0.00	0.66
s.w	AB	0.00	0.00	0.12	0.04	0.00	0.66
bwr	AB	0.15	0.14	1.26	0.05	0.00	0.66
bwr.w	AB	0.19	0.22	1.70	0.08	0.08	0.66
bwo	AB	0.00	0.00	0.00	0.03	0.00	0.66
bwo.w	AB	0.00	0.00	0.12	0.04	0.01	0.66

Table 8: Summary statistics of the percentages of units for which the bootstrap value is different from the original value. Q_1 = the first quantile, Q_2 = the second quantile (median) and Q_3 = the third quantile. s = standard bootstrap, bwr = BWR bootstrap method, ".w" = unequal inclusion probabilities in the bootstrap replicate.

Type	Var	Min	Q_1	Q_2	Mean	Q_3	Max
s	ME	100.00	100.00	100.00	100.00	100.00	100.00
s	MO	23.70	41.32	48.76	50.66	56.92	87.50
s	MS	0.00	14.00	21.83	23.98	38.90	51.07
s	AE	100.00	100.00	100.00	100.00	100.00	100.00
s	AB	67.11	87.11	94.38	91.61	97.96	100.00
s.w	ME	100.00	100.00	100.00	100.00	100.00	100.00
s.w	MO	26.46	43.66	49.57	51.23	55.82	92.31
s.w	MS	0.00	14.36	22.09	23.68	37.14	57.21
s.w	AE	100.00	100.00	100.00	100.00	100.00	100.00
s.w	AB	75.57	87.57	93.39	92.19	98.85	100.00
bwr	ME	100.00	100.00	100.00	100.00	100.00	100.00
bwr	MO	100.00	100.00	100.00	100.00	100.00	100.00
bwr	MS	0.00	100.00	100.00	92.86	100.00	100.00
bwr	AE	100.00	100.00	100.00	100.00	100.00	100.00
bwr	AB	100.00	100.00	100.00	100.00	100.00	100.00
bwr.w	ME	100.00	100.00	100.00	100.00	100.00	100.00
bwr.w	MO	100.00	100.00	100.00	100.00	100.00	100.00
bwr.w	MS	0.00	100.00	100.00	92.86	100.00	100.00
bwr.w	AE	100.00	100.00	100.00	100.00	100.00	100.00
bwr.w	AB	100.00	100.00	100.00	100.00	100.00	100.00
bwo	ME	100.00	100.00	100.00	100.00	100.00	100.00
bwo	MO	24.27	41.49	48.92	50.72	56.84	87.50
bwo	MS	0.00	14.19	21.61	24.05	37.53	49.74
bwo	AE	100.00	100.00	100.00	100.00	100.00	100.00
bwo	AB	71.05	86.18	94.14	91.60	97.92	100.00
bwo.w	ME	100.00	100.00	100.00	100.00	100.00	100.00
bwo.w	MO	27.11	43.58	48.90	51.08	56.67	92.31
bwo.w	MS	0.00	13.29	21.26	23.38	36.15	56.47
bwo.w	AE	100.00	100.00	100.00	100.00	100.00	100.00
bwo.w	AB	75.93	87.34	93.35	92.14	98.81	100.00

number of quantiles of a vector having the majority of elements equal to zero. The high percentages show bootstrap capacity to protect against exact disclosure, especially for the bwr.w variant.

5. CONCLUSIONS

In this paper the problem of synthetic data generation was addressed. First, a brief review of the existing methods was presented. The approaches to synthetic data generation are based on some model assumption which in practical situations may be difficult to validate. A model free quantile based bootstrap strategy was then proposed. Complex survey data issues were also considered, by illustrating different bootstrap variants. The bootstrap methods for synthetic data generation were tested in a real situation using the Italian Structure of Earnings Survey 2006 data. In these preliminary experiments, the results show that the quantile based bootstrap methodology might be a valid alternative for producing synthetic data. Different forms of bootstrapping might also be used. Further studies will be performed aiming at the preservation of the most important survey characteristics. Extensions to the multivariate bootstrapping will also be considered. Experiments will be performed taking into account disclosure scenarios based on multivariate key variables.

Acknowledgments.

Istat is not responsible for any views or results presented. The author was supported by the Network of Excellence in the European Statistical System in the field of SDC (ESSnet-SDC) and the 7^{th} Framework Programme Blu-ets.

6. REFERENCES

[1] R. J. Bowden, A. B. Sim. The Privacy Bootstrap. *Journal of Business and Economic Statistics*, 10(3):337–345, July 1992.

[2] J. Burridge. Information preserving statistical obfuscation. *Statistics and Computing*, 13:321–327, 2003.

[3] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistisk Tidskrift*, 85:429–444, 1977.

[4] R. Dandekar, M. Cohen, N. Kirkendall. Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique. In *Inference Control in Statistical Databases*. Lecture Notes in Computer Science, vol. 2316, Springer, Berlin Heidelberg, 2002, pp. 245–253.

[5] J.C. Deville, C.E. Sarndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382, 1992.

[6] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, 15(7):1–26, 1979.

[7] Eurostat *Structure of Earnings Survey 2002 - Eurostat arrangements for implementing the Council Regulation 530/1999 and the Commission Regulation 1916/2000*, Eurostat, 6 April 2004.

[8] S. E. Fienberg. A radical proposal for the provision of micro-data samples and the preservation of confidentiality. *Carnegie Mellon University Department of Statistics, Technical Report 611*, 1994.

[9] D. V. Hinkley. Bootstrap Methods. *Journal of the Royal Statistical Society, Series B*, 50(3):321–337, 1988.

[10] G. R. Heer. A bootstrap procedure to preserve statistical confidentiality in contingency tables. In *Proc. of the International Seminar on Statistical Confidentiality* (ed. D. Lievesley). Eurostat, Luxembourg, 1993, pp. 261–271.

[11] D. E. Huntington, C. S. Lyrintzis. Improvements to and Limitations of Latin Hypercube Sampling. *Probabilistic Engineering Mechanics*, 13(4):245–253, 1998.

[12] C.K. Liew, U.J. Choi, C.J. Liew. A data distortion by probability distribution. *ACM Transactions on Database Systems*, 10:395–411, 1985.

[13] J. M. Mateo-Sanz, A. Martinez-Balleste, J. Domingo-Ferrer. Fast generation of accurate synthetic microdata. In *Privacy in Statistical Databases*. Lecture Notes in Computer Science, vol. 3050, Springer, Berlin Heidelberg, 2004, pp. 298–306.

[14] J. Domingo-Ferrer, J. M. Mateo-Sanz. On resampling for statistical confidentiality in contingency tables. *Computers and Mathematics with Applications*, 38:13–32, 1999.

[15] K. Muralidhar, R. Sarathy. Data Shuffling: a New Masking Appraoch for Numerical Data. *Management Science*, 52(5):658–670, 2006.

[16] G. Paass, G. Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics*, 6:487–500, 1988.

[17] J.N.K. Rao, C.F.J. Wu. Resampling Inference With Complex Survey Data. *Journal of the American Statistical Association*, 83(401):231–241, March 1988.

[18] D. R. Rubin. Discussion of Statistical Disclosure Limitation. *Journal of Official Statistics*, 9(2):461–468, 1973.

[19] R.R. Sitter. A Resampling Procedure for Complex Survey Data. *Journal of the American Statistical Association*, 87(419):755–765, September 1992.

[20] Willenborg, L. & De Waal, T. (2001). *Elements of statistical disclosure control.* Lecture Notes in Statistics 155. New York: Springer.