WP. 25
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (iv): Balancing data quality and data confidentiality

Some aspects concerning analytical validity and disclosure risk of CART generated synthetic data

Prepared by Hans-Peter Hafner, Statistical Office of Hesse and
Rainer Lenz, Saarland State University of Applied Sciences, Germany

Some aspects concerning analytical validity and disclosure risk of CART generated synthetic data

Hans-Peter Hafner[*] and Rainer Lenz[**]

[*] Research Data Centre of the Statistical Offices of the Länder, Statistical Office of Hesse, Rheinstr. 35/37, 65185 Wiesbaden, Germany, e-mail hhafner@statistik-hessen.de

[**] Saarland State University of Applied Sciences, Goebenstrasse 40, 66117 Saarbrücken, Germany, e-mail rainer.lenz@htw-saarland.de

Abstract: Reiter (2005) proposed the generation of partially synthetic microdata from a series of CART models. However there are some open questions concerning this approach. These questions include the following: a) May the synthesis order of the variables be randomly chosen or has it an impact on the analytical validity of the synthetic data? b) In principle two procedures are discussed: The first is to build trees by conditioning on all variables that are already synthesized or that should not be synthesized at all. The second one is to include for each tree all variables in the corresponding CART model and to iterate between the draws of the synthetic values. Is one of these two methods superior regarding the analytical potential of the data? c) Can the setting of the complexity parameter (used to determine the size of a tree) be automated to achieve an optimal balance between analytical validity and disclosure risk? Which requirements to the disclosure risk assessment are necessary for those data? d) Is an adaption of record linkage technology, commonly used for non-synthetic perturbed data, possible? These issues are analyzed on the basis of a sample of the longitudinal section of the German monthly report on local units of the manufacturing sector for the years 1999 to 2002. We present first results of our work in progress which is part of the German project InfinitE ('An informational infrastructure for the e-science age').

1 The German national Project InfinitE

In recent years, in the research data centres of the statistical offices of the Federation and the Länder, controlled remote data execution and safe centres have become the most frequent used access ways to microdata of economic statistics. Controlled remote data execution means that the researcher send their program code to the research data centre (RDC), the RDC staff runs the code on the original data, checks the output with respect to the confidentiality rules and sends the checked output to the researcher. To develop their program code, the researcher receive so called data structure files (DSF). These are anonymized files which have the same structure as the underlying original data. Until now, the DSF only serve to check whether the code is syntactically correct; it can't be tested whether a model is correct since the relations between the variables are mostly destroyed. Therefore, the aim of the project InfinitE (`An informational infrastructure for the e-science age', see Brandt and Zwick (2009), Lenz (2009) and Lenz and Zwick (2009)), being started in June 2009 is, on the one hand, to develop anonymized DSFs which can be utilised for instance to specify econometric models and to formulate syntactical error-free codes.

Furthermore, the checking of the analysis output, which is at this time very time-consuming for the employees of the RDCs, shall be automated as far as possible in order to spare personal ressources. In this paper, we focus on the first issue, the development of DSFs and herein especially on the generation of such files by means of multiple imputation. In chapter 2, we give a short introduction to the development of (synthetic) DSFs. The following chapter deals with CART models and their application to the generation of DSFs. The related software is the R package `rpart'. The first data source we use to develop and to test our methods, is the German monthly report for local units of the manufacturing sector. We give a short description of the data in section 4.1. One essential aspect of anonymization is to maintain the analytical potential of the data; first hints, how well this is achieved with our methods, show some comparative results between the original and the synthetic data of the monthly report for the years 1999 to 2002 in section 4.2. But what is at least of the same importance, is that the confidentiality of the data is preserved. The protective effect of this special variant of synthetically generated data is shown by appropriate matching experiments. In section 4.3, we give an overview on the theory behind these experiments and first results for the German monthly report survey. We conclude with some remarks on potential future work.

2 Development of Data Structure Files

So far, DSFs often consist of a sample of the original material, which has been subjected to additional anonymization measures, or of values generated at random uniformly with respect to the range of the corresponding variables. Although the whole set of variables is maintained in both approaches, their attributes and dependence structure (filter, variance-covariance matrix) between variables are in most cases completely destroyed. Hence, a researcher can check whether his/her program is executable, though he/she does not obtain any information on whether or not the actual question has been adequately implemented.

For this reason, the analysis programs of scientists can often not be used in an unchanged form for the subsequent application to the original data. Instead, additional adjustments have to be made by the scientists and the RDC staff. A promising way of providing DSFs of a significantly better quality is to produce synthetic datasets based on the idea of multiple imputation of missing values. The decisive advantage of this method is its universality. Any restrictions and filter structures can be taken into account. In addition, the approach can be applied to continuous variables in the same way as to categorical ones. Due to its high exibility and also applicability to very complex and linked panel data sets, this approach has been increasingly used at the international level in the past few years. The proposal to generate synthetic datasets for the scientific community by means of multiple imputation was submitted first in Rubin(1993) and it was further expanded in Raghunathan et al. (2003). The basic principle is to produce in each case several

synthetic datasets being analysed individually. The actual analytical results follow by application of simple combining rules.
In principle, it is distinguished between fully and partially synthetic datasets. Regarding fully synthetic datasets, all units of the population which do not belong to the sample, are treated as missing values. For these `missing' units additional information is required (for example from the German official business register or from the employment survey of the Federal Employment Agency) which is included in the imputation model. In contrast, for partially synthetic datasets all attributes or only sensitive attributes of the units contained in the survey are replaced by synthetic ones.

3 Classification and Regression Trees

3.1 Description of the method

The CART methodology (Classifcation and Regression Trees) was introduced by Breiman et al. (1984). The idea is to partition the predictor space of some variable X with covariates $Y_1,\ldots, Y_n$ by a series of recursive binary splits. These splits are represented by a tree structure, where the units inside the emerging subsets (so called nodes or leaves) should be relatively homogeneous and those of different nodes should be as heterogeneous as possible. Having this in mind, a splitting criterion is calculated for every covariate and each potential splitting point. The covariate and splitting point which maximize the splitting criterion are choosen. If X is categorical with range {1,…, k}, a commonly used splitting criterion is the well-known Gini index: Let p(i | t) be the proportion of units in node t for which Y = i holds. Then the Gini index i(t) is defined as

$$i(t) = \sum_{i \neq j} p(i \mid t)\, p(j \mid t).$$

If X is continuous, a squared residiuals algorithm is applied minimising the variance within the two resulting nodes. Besides this splitting criterion, a stopping criterion is needed in order to decide when to stop the growing of the tree. In general, it is not recommendable to let the tree grow further, if the number of observations in the resulting nodes is too low or if there is no significant reduction of heterogeneity. If a tree is very complex, its interpretation often appears to be difficult or impossible. However, in the synthetic data context it is the other way around. Here the aim is not to interpret the trees. If the tree is too small, some important relationships might be lost. Thus, the use of stopping criteria that allow large trees, is approriate in our context. In recent years, CART models have become very popular, because the method is nonparametric, the covariates may be correlated and one can use different

types of variables. Last but not least, the results are easy to interpret because of the possibility of graphical visualisation. For further details see Breiman et al. (1984).
The first application of CART models in the context of the generation of synthetic data was suggested in Reiter(2005), sketched as follows. In the beginning, a tree T_1 is grown for the first variable Y_1 that we want to synthesize. In every node N of T_1 then a bayesian bootstrap is performed as follows: Let Y^N be the range of Y_1 restricted to node N with cardinality $|Y^N| = n_N$.
1. Draw n_N - 1 uniform random numbers. Sort these numbers in ascending order and label them as $a_o = 0$, a1, a2,…, a_{nN-1}, $a_{nN} = 1$.
2. Draw n_N uniform random numbers $u_1, u_{2,\ldots,}u_{nN}$. For each of these u_k and
$j \in \{1 ,\ldots, n_N\}$impute Y_j^N when $a_{j-1} < u_k \leq a_j$.
We replace the original values of Y_1 by bootstrapped values and proceed with variable Y_2. Now, in the tree T_2 - according to the synthetic values of Y_1 – there might be combinations of values of the covariates that do not exist in the original data. In this case, we have no observations in the leaf to draw the synthetic values. So we have to go up the tree until we find a valid combination.

3.2 Software

Meanwhile, CART models are implemented in many commercial software systems. Besides this there are at least two packages written for R, namely rpart and tree. We use the first one being developed by Terry M. Therneau and Beth Atkinson (cf. Therneau et al. (2010)). This package is easy to apply: In the majority of cases it suffices to declare dependent variable, covariates and the dataset in use. Additionally, there is a choice between two splitting criteria (Gini and information) and there are some parameters for the setting of appropriate stopping criteria. For example, if one sets minsplit = n, then a node with less than n observations won't be split further. If cp = x, a split is not executed when the overall lack of fit is not decreased by a factor of cp. A smaller value of cp results in more nodes.

3.3 Choices which may influence analytical validity and disclosure risk

There exist several choices for the generation of the CART models which might have an impact as well on the analytical validity as on the disclosure risk of the synthetic data. These choices can be divided into such for the synthesis order, such for the model specification and such for the stopping rules.

Synthesis Order

Reiter(2005) suggests to examine the depth of the variables in the trees, that is the depth of the first split at which one variable appears and to arrange the variables with respect to the minimum depth over all trees from lowest to highest. Then the synthesis of the variables is carried out according to this order. In practice this is an extreme laborious procedure and in most cases there are some variables that have the same depth.

Model Specification

In principle there are two ways to define the CART models. The first one is always to include all variables. In this case one has to iterate between the different draws of the synthetic values; that means to store synthetic values approximately every 20th to 50th draw. This ensures the independence of the multiple synthetic values. The second method is to include as dependent variables only such ones that are already synthesized or such ones that should not be synthesized at all. This is analogue to the approach of the parametric sequential regression (Ragunathan et al. (2001)).

Stopping Rules

Typically stopping rules relate to the number of observations in an existing node (no further splitting if the number of observations is below a certain threshold), to the number of observations in a resulting node (no further splitting if a new generated node would have too few observations) or to the amount of additional model fit declared by a further split (cf. the cp parameter in rpart).

4 Empirical results

4.1 The data used

The members of the InfinitE project agreed to develop and to compare different anonymisation strategies on the basis of the monthly report on local units of the manufacturing sector for the years 1999 to 2002. On the one hand, this survey is strongly demanded by scientists, on the other hand it has a straightforward questionnaire with about 30 variables. Subject to report is the whole of local units focussing mainly on economic activity in the manufacturing sector and occupying at least 20 employees. Also included are smaller local units, if the enterprise to which they belong occupies at least 20 employees. Among the attributes reported are the sector of economic activity, the unit's location, the number of employees, the total (export) turnover, the wages and salaries paid and the number of working hours carried out.

In principle, analyses on a monthly basis would also be possible. However, until now only the aggregated annual data are available for scientists at the RDCs.

4.2 Analytical validity

For reasons of the runtime of the programs, we use a 15% sample of the longitudinal section of the monthly report to test different options for the generation of the CART models. So 6483 local units are included in the file.

We coarsen the 4 digit NACE code into a 2 digit one and the code for the location of the local unit to 16 categories (federal states of Germany). Since there are only very few local units changing their branch of activity or location during the four years, we synthesize these variables only once for the first year. The continuous variables are transformed by extracting the cubic root. This function doesn't ascend as strong as the usually used natural logarithm. This is an advantage, if the data contains outliers as it is mostly the case in business surveys. We divide the data into five subsets built by classes of turnover and we compute separate CART trees for every subset.

We synthesize only the NACE code, the region, the number of employees and the turnover for all 4 years. We always generate five synthetic datasets.

At first we generate trees for different complexity parameters and the following two alternatives:

(i) We use the absolute numerical values for all variables and all years.

(ii) We use the absolute value for the year 1999 and the rates of change for the following years. (For an attribute X the rate of change from year t-1 to year t is defined as $r_t = (X_t - X_{t-1}) / X_t$.)

Besides the arithmetic means of the number of employees and the turnover we examine two job flow characteristics:

a) Net employment change (NEC): This is the difference between the rate of job creation and the rate of job destruction.
b) Job turnover (JT): This is the sum of the rate of job creation and the rate of job destruction.

The pooled NEC respective pooled JT is the average of the characteristics over the different years weighted by the number of employees.

	Original	cp = 0.01	cp = 0.001	cp = 0.0001	cp = 0.00001
Mean Turnover 1999	1936189.85	1847448.79	1883230.39	1934018.48	1926245.85
Mean Turnover 2002	2169133.1	2208620.66	2150722.43	2154219.84	2167506.11
Turnover Index 2002 (1999 = 100)	112.031013	119.679831	114.438163	111.560841	112.70672
NEC Pooled	-0.5321251	1.27624955	-0.2603989	-0.0055279	0.90603273
JT Pooled	7.2900441	31.8256513	20.8003609	18.1333015	17.5818771

Table 1. Characteristics for alternative (i) – absolute values for all years

The values for the job turnover are by far too high independent of the value of the complexity parameter. The turnover index is significantly better for smaller cp values.

	Original	cp = 0.01	cp = 0.001	cp = 0.0001	cp = 0.00001
Mean Turnover 1999	1936189.85	1925153.74	1921748.68	1968514.29	1911603.23
Mean Turnover 2002	2169133.1	2167223.29	2105423.7	2201567.62	2118043.63
Turnover Index 2002 (1999 = 100)	112.031013	112.600209	109.587064	111.822827	110.786968
NEC Pooled	-0.5321251	-0.3753351	-0.1881142	-0.395563	-0.4545543
JT Pooled	7.2900441	7.93833441	7.81700098	7.49026939	7.63130304

Table 2. Characteristics for alternative (ii) – rates of change for the years 2000 to 2002

The synthetic job turnover values for this alternative are close to the original value. However the turnover index is now nearer to the original value for the largest cp. Thus we conclude that the use of rates of change leads to better results for some characteristics while the impact of the cp parameter needs further investigation.

NACE Code	Original data Turnover Index 2002	cp = 0.1 Turnover Index 2002		cp = 0.0001 Turnover Index 2002	
		Mean	Range	Mean	Range
		Over 5 synthetic data sets		Over 5 synthetic data sets	
10	173.2	104.8	84.3 - 116.3	165.1	129.7 - 231.9
14	92.1	99.1	93.4 - 107.4	96.6	91.0 - 100.6
15	119.9	114.9	101.4 - 121.3	113.3	108.4 - 119.3
17	94.4	109.6	96.4 - 121.2	104.1	91.1 - 113.9
18	90.5	106.4	94.5 - 118.6	96.8	93.1 - 105.2
19	109.0	117.3	101.8 - 151.0	102.5	90.8 - 112.8
20	94.6	112.2	99.6 - 132.4	98.9	93.3 - 107.0
21	128.0	117.2	107.9 - 122.0	110.5	100.0 - 117.7
22	98.4	112.7	103.2 - 119.8	106.5	97.4 - 121.6
24	121.9	116.5	109.1 - 124.2	113.6	107.6 - 117.1
25	111.4	113.3	106.9 - 121.0	107.9	103.8 - 117.9
26	90.1	106.2	99.9 - 110.1	96.1	91.7 - 98.6
27	116.0	119.0	99.9 - 141.2	121.6	109.6 - 136.3
28	109.6	109.1	104.0 - 113.3	107.0	100.4 - 112.6
29	103.1	116.1	109.1 - 128.1	111.6	107.8 - 119.6
30	91.4	114.4	93.7 - 134.8	93.9	85.4 - 105.0

Table 3. Turnover Index 2002 by 2-digit NACE Code

NACE	Original data		cp = 0.1		cp = 0.00001	
Code	NEC pooled	JT pooled	NEC pooled	JT pooled	NEC pooled	JT pooled
10	0.31171745	11.9754249	-3.7144774	11.1255142	1.30541806	9.59327307
14	-1.7497556	7.82532614	0.08905479	8.87030376	8.4628104	20.4053842
15	-0.2030299	8.61607427	0.48534536	9.52861032	-0.3834236	7.46388163
17	-1.8735322	7.35598322	0.37102058	6.79085572	-1.186748	8.46139291
18	-3.6944797	7.25011794	0.1711825	10.5618033	-2.9289133	5.96712464
19	-2.1478618	8.76247606	0.81166447	7.86742969	-4.3069551	9.61020507
20	-2.5775355	8.47084155	0.30993357	7.72387462	-0.3768272	8.36781237
21	0.83667378	5.45816673	-0.066811	6.63163992	0.73325438	6.01821538
22	-0.096421	7.8442202	-1.0186887	8.11495442	-2.1205747	9.20886819
24	-1.0809645	5.62499728	-0.1769504	6.42928496	-0.8503427	5.3786609
25	0.17896174	7.66830703	-0.2985845	8.02291092	0.60403962	7.17278479
26	-3.1110766	7.6484509	-0.3396862	8.61460059	0.11626143	11.2480476
27	-2.9563047	8.15361279	-0.7689096	7.47226845	-0.4674697	5.91891298
28	0.09495236	7.3826078	-0.5480605	8.38193956	-0.9305251	6.17950251
29	-0.4198332	6.54224754	-0.3550072	7.64120872	-0.4976521	7.95096963
30	-2.3809798	6.75609294	2.07334828	7.04864868	-9.0250062	13.300375

Table 4. NEC and JT by 2-digit NACE Code

Now we look at some results by NACE code. For lack of space we present in tables 3 and 4 our findings only for 16 out of 23 groups of economic activity and only for alternative (ii). In table 3 the red marked values are those synthetic turnover indices which are nearest to the original value. The majority of these values is located in the column for the smaller cp value but for instance for NACE code 21 the average index for cp = 0.1 (117.2) is much closer to the original value (128.0) than the average index for cp = 0.00001 (110.5) and all synthetic indices are far below the original value (highest synthetic index is 122.0 for a dataset generated with cp = 0.1). Related to the net employment change the sign is preserved for the smaller cp value more often (table 4). But the absolute differences are sometimes better for cp = 0.1 (for example for NACE code 30).

A general problem is that the results between the synthetic datasets spread very wide. This fact is due to the presence of outliers. Since the CART trees not split by the separate NACE codes (there is nearly almost more than one NACE code in a node) an outlier can be assigned to local units from different NACE codes in the different synthetic datasets which leads to great differences between the results for the distinct synthetic datasets. Since in the end only one synthetic data structure file should be transferred to the researcher a procedure to reduce the variance between the datasets or to choose the best one out of the different synthetic datasets has to be to developed.

4.3 Simulation of data intrusion scenarios

A rational framework for generating confidential microdata has always to review two objectives. On the one hand, the analytical potential of the data has to be maintained

to a greatest possible extent. On the other hand, confidential information on the individuals (in our case enterprises or local units) behind the data has to be protected to a certain degree. The latter can be estimated by linking records of some external dataset A with the confidential target data B. As external data we use the original formally anonymised monthly report (that is, direct identifiers like name, address are removed from the dataset). The project has decided for this ‘worst-case’ scenario, since it is followed the objective to produce absolutely anonymised DSF’s. Nevertheless, within the linkage procedure we use only those variables as key variables which the confidential monthly report data have in common with commercially available databases.
We presume that a potential data intruder had knowledge about the participation of the searched units in the target survey, that is, the re-identification problem might be formulated in mathematical terms as follows: Find an injective mapping $\Psi: A \rightarrow B$, based on some distance measure $d: A \times B \rightarrow [0,1]$ mapping every record a_i of A onto a near (or alternatively similar) record b_j of B. We calculated the distances $d(a_i, b_j)$ are calculated for each pair $(a,b) \in A \times B$ and solved the corresponding parametric linear assignment problem using the LP-solver *linprog* in SAS/IML. For this, it was necessary to adopt the complementary conditions of the problem appropriately. For more details see Lenz (2008).

We carried out first matching scenarios for the above described alternative (ii) with complexity parameter cp = 0.00001. We used size classes of turnover and number of employees as blocking variables. This implies that only those units will be compared which belong to the same combination of size classes in both sources. The assignment to a specific size class was made by calculating the average of the turnover respective the number of employees for the years 1999 to 2002. The key variables to calculate the distances were the number of employees and the turnover for the four years. Until now the scenario was conducted only for one synthetic dataset. For one block 27% of the local units could be assigned correctly. For all other blocks not more than 15% of the units were assigned correctly. So the synthetic data are at least de facto anonymous; that means that the expense of a correct assignment is higher than the benefit from the assignment.
The major objection to this approach is that the results depend on the choice of the size classes. Furthermore other strategies could be imaginable which are more adequate to synthetic data.

5 Prospects

The implementation of CART models for the generation of synthetic data seems to be promising but time consuming. Our first attempts show that it is useful to build different CART trees for subsets of the original data and to draw the synthetic values from this distinct trees. Further in the case of longitudinal data it is evident that the use of rates of change from the second year onwards is superior to the use of the absolute values for all years. Smaller values of the complexity parameter cp tend to maintain more of the analytical potential; but this still needs further investigation. Due to time constraints not all of the in section 3.3 mentioned influencing factors could be examined until now. This will be on our agenda for the next months as well as the revision of the risk assignment.

Acknowledgement: This work was supported by the Federal Ministry of Education and Research (BMBF).

References

Brandt, M., Zwick, M. (2009). *An information infrastructure for the E-Science Age – On the way to remote data access*. Conference '*New Technics and Technologies for Statistics*' (NTTS). Brussels.

Breiman, L., Friedman, J.H., Olshen, R.H. & Stone, C.J. (1984). *Classification and Regression Trees.* Boca Raton.

Lenz, R. (2008). *Risk Assessment Methodology for Longitudinal Business Micro Data*. *Journal of the German Statistical Society* (Wirtschafts- und Sozialstatistisches Archiv), vol. **2** (3), 241-258.

Lenz, R. (2009). *Défis méthodiques lors de la réalisation de l'accès aux données économiques allemandes par la téléinformatique automatisée*. 41$^{\text{ème}}$ *Journées de Statistique de la Societé Francaise de Statistique* (SFdS). Bordeaux.

Lenz, R., Zwick, M. (2009). *Methodological aspects assuring remote access to German business microdaa. Bulletin of the 60th International Statistical Institute* (ISI). Durban.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. & Solenberger, P. (2001). *A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology* **27**(1), 85-95.

Raghunathan, T.E., Reiter, J. & Rubin, D.B. (2003). *Multiple Imputation for Statistical Disclosure Control. Journal of Official Statistics* **19**(1), 1-16.

Reiter, J.P. (2005). *Using CART to Generate Partially Synthetic Public Use Microdata. Journal of Official Statistics* **21**(3), 441-462.

Rubin, D.B. (1993). *Statistical Disclosure Limitation. Journal of Official Statistics* **9**(2), 461-468.

Terneau, T.M., Atkinson, B.(2010). *Rpart. R package version 3.1-46. http://CRAN.R-project.org/package=rpart.*