

WP. 32
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (vi): Case studies

**SYNTHETIC DATA STRUCTURE FILES: DEVELOPMENT AND
DISCLOSURE CONTROL**

Invited Paper

Prepared by Hans-Peter Hafner (Statistical Office of Hessen) and Rainer Lenz (University of Applied
Sciences Mainz), Germany

Synthetic Data Structure Files: Development and Disclosure Control

Hans-Peter Hafner* and Rainer Lenz**

* Research Data Centre, Statistical Office of Hessen, Rheinstraße 35/37, 65175 Wiesbaden, Germany, e-mail hhafner@statistik-hessen.de

** School of Technology, University of Applied Sciences Mainz, Holzstraße 36, 55116 Mainz, Germany, e-mail rainer.lenz@fh-mainz.de

Abstract: In the last 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. Therefore, one aim of the project InfnitE (“An informational infrastructure for the e-science age”) - started in June 2009 – is to develop anonymised data structure files which can be utilised to specify econometric models and to formulate syntactical error-free codes. A way of providing such files is to produce synthetic data sets 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. The paper contains first approaches to generate what is called partial synthetic datasets of the monthly report for local units of the manufacturing sector for the year 2001, regarding their analytical potential by means of comparative analyses with the original data and their protection effect applying statistical matching.

1 The German Project InfnitE

The producers of data relating to surveys of economic statistics in Germany have observed a fundamental change in the demand for their products. In early 2000, providing the scientific community with so-called scientific use files (SUFs) – which researchers can use at their own workplaces outside the statistical offices - was considered a way, if not even the “gold standard”, towards giving empirical social and economic research adequate access to official microdata in Germany. Such SUFs are available for selected and strongly demanded statistics. However SUFs of economic surveys are not very well accepted. One reason for this are the new data perturbation methods which are necessary to ensure the confidentiality of the data, but they destroy parts of the statistic inference. In addition there is too much delay between the collection of the data and the generation of the associated SUF. In the last 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. Therefore, the aim of the project InfnitE (“An informational infrastructure for the e-science age”) - started in June 2009 - is, on the one hand, to develop anonymised data structure files which

can be utilised to specify econometric models and to formulate syntactical error-free codes. Furthermore, the checking of the output, which is very time-consuming for the employees of the research data centres (RDC), shall be automated as far as possible.¹

In this paper, we focus on the first issue, the development of data structure files and here especially on the generation of such files by means of (multiple) imputation. In chapter 2, we give a short introduction to the different methods to create synthetic datasets and to the imputation software IVEware which we apply in our project so far. The first data source we use to develop and to test our methods, is the monthly report for local units of the manufacturing sector. We give a short description of the data in chapter 3. A particular challenge consists in the imputation of categorical attributes; we discuss this in detail in chapter 4. One aspect of anonymisation 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 year 2001 which we present in chapter 5. 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. We give an overview of the theory behind these experiments and first results for the monthly report 2001 in chapter 6.

2 Development of Data Structure Files

So far, the data structure files often consist of a sample of the original material, which has been subjected to additional anonymisation measures, or of values generated at random within the value range of the data set. Although the variables are maintained in both approaches, their attributes and the dependence structure (filter, variance-covariance matrix) regarding other variables are 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.

2.1 Synthetic Data Structure Files

A way of providing data structure files of a significantly better quality is to produce synthetic data sets based on the idea of a multiple imputation of missing values. The decisive advantage of this method is the universality of its approach. Any restrictions

¹ For further information about the project see the paper of Brandt and Gürke for this work session

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 variables. Due to its high flexibility and also applicability to very complex and linked panel data sets, this innovative approach has been increasingly used at the international level in the past few years.²

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, Reiter und Rubin (2003). The basic principle is to produce in each case several synthetic datasets which are analysed individually. The actual result of the analysis follows by the application of simple combining rules (Raghunathan et al. (2003)).

In principle fully synthetic and partial synthetic datasets can be distinguished. For fully synthetic datasets, all units of the population which don't belong to the sample, are treated as missing values. For these "missing" units additional information is required (for example from the business register or from the employment statistics of the Federal Employment Agency) which is included in the model for the imputation. In contrast, for partial synthetic datasets all attributes or only sensitive attributes of the units contained in the survey, are replaced by synthetic values.³

2.2 IVEware

The project has the goal to develop standardised anonymisation procedures that can easily be applied and adapted to all kinds of surveys by every member of the RDC staff and other employees of the statistical offices. This requires that the software used is easy learnable (not only for computer scientists) and not too expensive. The imputation software IVEware fulfils these two conditions. IVEware was developed by Raghunathan, Solenberger and Van Hoewyk and is free available by download.⁴

The program uses the technique of sequential regression: Let X_1, \dots, X_i be the variables of the dataset without missing values, Y_1, \dots, Y_j the variables containing missing values. Thereby let the order of the Y – variables be ascending with respect to the number of missing values. In the first step, a model for the conditional distribution of Y_1 given the observed values of the X – variables is estimated. Afterwards, from this distribution the values for Y_1 are drawn. In the next step a model for the conditional distribution of Y_2 given the observed X – values and the previously imputed Y_1 – values is estimated and from this distribution the missing values of Y_2 are imputed and so on.⁵

² Cf. Lenz 2009

³ Cf. Reiter 2003

⁴ <http://www.isr.umich.edu/src/smp/ive/>

⁵ Cf. Raghunathan, Lepkowski, Van Hoewyk, Solenberger 2001

IVEware distinguishes four kinds of variables: Continuous, categorical, mixed (0 as categorical value, otherwise continuous) and count variables (for instance, the number of local units of an enterprise). For continuous variables, the ordinary linear regression model is used for the estimation, while for categorical variables a logistic or a generalized logistic model is applied. Mixed variables are imputed in two stages: At first zero or non-zero is estimated by means of a logistic regression; afterwards the values for the units with non-zero estimate are imputed using a linear regression model. Count variables are usually treated with a Poisson regression.

IVEware offers several possibilities to maintain the structure of the original data and to preserve dependencies between the variables. Lower and upper limits can be declared via the `bounds – statement`, while the `restrict – statement` determines that values are only imputed if a certain condition is satisfied, for example the number of birth shall only be imputed for women.⁶

IVEware can be launched from SAS or it can be executed independently. Unfortunately, the program doesn't work if SAS-codes are activated using the more comfortable Enterprise Guide interface.

3 The monthly Report on local Units of the Manufacturing Sector

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 1998 to 2001. 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 are all 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 possesses at least 20 employees.

Among the attributes reported are the sector of economic activity, the location, the number of employees, the (export) turnover, the wages and salaries paid and the number of working hours carried out.

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

⁶ For further details see the IVEware User Guide (ftp://ftp.isr.umich.edu/pub/src/smp/ive/ive_user.pdf)

4 Generation of synthetic Data Structure Files

To gain experience in imputation and the usage of the program IVEware, we examined at first only one year of the survey, namely 2001. In this year 50.347 local units were contained in the data.

4.1 Prerequisites

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 with business surveys.

Only one variable will be anonymised at a time. Therefore the dataset is reduplicated: After the original data the same dataset is added once again, but this time the values of the variable which shall be synthesised are replaced by missings.

4.2 Imputation of categorical variables

It turned out that the imputation of categorical variables with more than 4 to 6 distinct values poses a particular challenge. Three alternatives were tested. We discuss the results using the example of the location of the local unit coded by the 16 federal states of Germany.

Alternative 1: The attribute "federal state" is taken into the model as categorical variable with 16 values.

The program aborts repeatedly after 10 or more hours of running time. The message "Abnormal Termination" appears on the screen, the log-file contains no further information about possible sources of error. We suppose, that the abortion is a consequence of instabilities of the network at night time due to scheduled backups and updates. This alternative requires further testing on a stand-alone PC.

Alternative 2: For every federal state a dummy variable is created.

The imputation of the dummy variables is carried out according to the order of the local units located in the federal state. At first the values for the federal state having the most local units (Nordrhein-Westfalen) are estimated, then those for the federal state in which the second highest number of local units is located (Baden-Württemberg) and so on. If for a unit for the first time a value of 1 is imputed for a dummy, all other proceeding dummies are automatically set to 0. The computing time is acceptable for this alternative: 30 – 40 minutes are needed per synthetic data set, if the number of iterations for the regression procedure is set to 10. This alternative works quite well (cf. Appendix, Tables 1 - 3).

Three models were tested. In models 2 and 3 supplemental explanatory variables were added. For all three tested models, the both smallest federal states, Bremen and Saarland, have high percental deviations between synthetic and original data with regard to the fraction of local units located in the federal state. These deviations even increase by adding more explanatory variables. Sachsen has a very low deviation of about one percent in model 1 (Table 1); this value increases in model 2 (Table 2) to over 13 percent. The differences between the models for individual federal states suggest that it could make sense to estimate separate models for the federal states. This question needs further investigations.

Alternative 3: Three-stage imputation: At first old or new federal states, then regions in the old states and finally the federal states within the regions.

Germany is segmented into the four regions “old federal states north” (Schleswig-Holstein, Hamburg, Niedersachsen, Bremen), “old federal states middle” (Nordrhein-Westfalen, Hessen, Rheinland-Pfalz), “old federal states south” (Baden-Württemberg, Bayern, Saarland) and “new federal states”. At the first stage, it is estimated whether a local unit is located in the old or in the new federal states. At the second stage, the local units of Western Germany are attached to one of the regions north, middle or south. Finally a federal state within the allocated region is imputed.

The assignment to the old / new federal states and to the regions is computed by means of dummy variables, the assignment to the single federal states by means of categorical variables. The idea behind this alternative was to reduce the number of steps of the imputation process compared to alternative 2. Here altogether 7 runs are needed: One for the assignment to the old or the new federal states, two runs for the assignment of the regions in the old federal states (south, middle; units that aren't assigned in these two runs are automatically assigned to north) and 4 runs for the assignment of the concrete federal state (one run per region).

It appears that the computing time for multinomial regressions is considerably higher compared to logistic regressions for binary dummy variables, even if the number of categories is only 3. In particular, the run for the new federal states (6 categories) takes 3 – 4 hours. The percental deviations for the regions are all below one percent, however the deviations for the individual federal states are in part very extreme, up to 414 percent for Hessen (cf. Appendix, Table 4).

To summarise the three alternatives mentioned above, we conclude that alternative 2 is the most promising. Alternative 1 drops out because of the long associated computing time, alternative 3 because of the bad results regarding multinomial regressions.

5 Comparison between synthetic and original data

To get a first impression, to which extent the analytical potential of the data is preserved in the synthetic datasets, we take a look at some results of comparative analyses between the synthetic and the original data. We have generated five different synthetic datasets of the monthly report 2001 and we restrict our analyses to the core variables sector of economic activity, location of the local unit, turnover and number of persons employed. More comprehensive analyses will be carried out by the staff of the Institute for Applied Economic Research (IAW) which is one of the partners in our project.

The univariate distribution of the number of employed persons is well preserved in all of the five synthetic datasets. The mean fluctuates between 127,3 and 127,9 while the true value is 128,1. The standard deviation is in general slightly higher compared to the original data (synthetic data 585,8 to 598,1; original value 575,9). The results concerning the turnover aren't that excellent. Both the mean and the standard deviation are located considerably below the respective original characteristics: While the deviation for the mean is around 8 to 9 percent, it reaches even around 20 percent for the standard deviation.

Dataset	Number of persons employed		Turnover	
	Mean	Standard Deviation	Mean	Standard Deviation
Synthetic 1	127,3	585,8	24562062	216851902
Synthetic 2	127,9	598,1	24807294	216127156
Synthetic 3	127,3	587,9	24780019	221423558
Synthetic 4	127,9	587,5	24716646	226316867
Synthetic 5	127,3	586,0	24447647	211408230
Original	128,2	575,9	26911974	274285595

Table 1: Mean and standard deviation for employed persons and turnover

The correlation coefficients between number of employed persons and turnover are considerably higher for the synthetic datasets. While the original coefficient is 0.80, the coefficients for the synthetic data are all around 0.91.

The economic sector having in average the highest number of employees per local unit, is the automobile industry. The original mean for the number of employees is 651, while the corresponding means for the synthetic data differ between 370 and 471. This is in three of the five synthetic datasets the highest average value for an economic sector while in the other two cases the average for the mining industry exceeds the value of the automobile industry. In the original data mining takes place two with an average of 502 employees. In all datasets the sector “quarrying for stone and earth” takes the last place with an average between 25,8 and 28,1 while the true value is 21,0.

Especially the results for the turnover need some inspection to be able to explain why all synthetic values are too low in average. Further tests will show which other improvements to the synthetic data are necessary.

6 Confidentiality of micro data

Regarding confidential micro data, we always have to follow two objectives. On the one hand, as mentioned already in the previous chapter, the analytical validity of the data has to be widely preserved. On the other hand, the risk of disclosure of confidential information by a potential data intruder should be minimised to greatest possible extent. The latter can be tested by carrying out realistic matching scenarios, as described below.

6.1 Mathematical modelling

In a database cross match, see Elliott and Dale (1999), the data intruder matches an external database A with the whole confidential database B . For this, he uses variables which the external data have in common with the confidential data, the so-called key variables.

To be a candidate for a possible assignment, it is necessary for a record pair $(a,b) \in A \times B$ that both records coincide in their values of some specified variables. In the following these variables are called blocking variables, since they divide the whole data into disjoint blocks. The aim of blocking data is on the one hand to reduce the complexity of the subsequent assignment procedure and the allocated main storage and on the other hand the number of mismatches. Detailed empirical investigations on blocking data in the context of disclosure control can be found in Lenz and Vorgrimler (2005).

In a non-technical way, the concept of matching may be introduced as bringing together pairwise information from two records $a \in A$ and $b \in B$, taken from

different data sources, that are believed to refer to the same individual. The records a and b are then said to be matched. In the following it is tried to minimise the number of mismatches. If we presume that a potential data intruder had knowledge about the participation of the searched units in the target survey, the problem of matching might be formulated in mathematical terms as follows: Find an injective mapping $\varphi: A \rightarrow B$, based on some distance measure $d: A \times B \rightarrow [0,1]$ (or alternatively based on some similarity measure $w: A \times B \rightarrow [0,1]$), which maps every record of A onto a near (or similar) record of B .

More precisely, the mapping can be defined by the following single objective assignment problem:

$$\text{Minimise } \sum_{i=1}^n \sum_{j=1}^m d(a_i, b_j) x_{ij}, \quad (\text{AP})$$

$$\begin{aligned} \text{subject to } \quad & x_{ij} \in \{0, 1\} \quad \text{for } i = 1, \dots, n; j = 1, \dots, m, \\ & \sum_{j=1}^m x_{ij} = 1 \quad \text{for } i = 1, \dots, n \quad \text{and} \\ & \sum_{i=1}^n x_{ij} \leq 1 \quad \text{for } j = 1, \dots, m. \end{aligned}$$

The constraints of (AP) ensure that each record a of the external data A is one-to-one assigned to some record b of the target data B . That is, $x_{ij}=1$ if and only if a_i is connected with b_j . Therefore, it seems to be reasonable to assume that A possesses a smaller or equal number of records than B .

Once the coefficients $d(a_i, b_j)$ are calculated, we can solve the linear assignment problem using classical established methods such as the simplex method. For larger data blocks (typically generated when dealing with tax statistics) it is recommendable for reasons of efficiency that approximation heuristics should be used. Fortunately, the usage of appropriate heuristics yields results near the optimum solution of the assignment problem, see Lenz (2006).

6.2 Empirical results

We carried out first matching scenarios for the about 36.000 single-site enterprises contained in the monthly report for the year 2001. From a commercial database we have additional knowledge for about 9.000 of these units. As blocking variables we used the location of the local unit (only old or new federal states), the branch of economic activity (two-digit NACE code) and the number of employees summarised

to 6 size categories. The key variables for the assignment were the number of employees and the turnover. Until now we conducted this scenario only for one synthetic dataset. The result: Only negligible 18 of the about 9.000 units were matched correctly, that is 0,2%.

To check whether this finding is reliable, we calculate the fraction of corresponding values for the blocking variables in the original and in the synthetic data. The two-digit NACE code is the same in both sources for only 23,6% of the units, the size category of the employees for 80,2% and the location (old / new federal states) is identical for 77,5% of the cases. The combination of all three variable remains unchanged for only 14,7% of the local units. This is a strong indication that the generation of synthetic data has a huge protective effect regarding the confidentiality of data.

References

- Elliot, M., Dale, A. (1999). *Scenarios of attack: the data intruder's perspective on statistical disclosure risk*, Netherlands Official Statistics, 6-10.
- Lenz, R., Vorgrimler, D. (2005). *Matching German turnover statistics*, RDC working paper No. 4. Wiesbaden.
- Lenz, R. (2006). *Measuring the disclosure protection of micro aggregated business microdata - an analysis taking as an example the German Structure of Costs Survey*. *Journal of Official Statistics* **22**, 681-710.
- Lenz, R. (2009). *Défis méthodiques de la réalisation de l'accès aux données économiques allemandes par la téléinformatique automatisée*. 41èmes Journées de Statistique, SFdS. Bordeaux.
- 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**, 85-95.
- Raghunathan, T.E., Reiter, J.P. & Rubin, D.B. (2003). *Multiple Imputation for Statistical Disclosure Limitation*. *Journal of Official Statistics* **19**, 1-16.
- Reiter, J.P. (2003). *Inference for partially synthetic, public use microdata sets*. *Survey Methodology* **29**, 181-188.
- Rubin, D.B. (1993). *Discussion. Statistical Disclosure Limitation*. *Journal of Official Statistics* **9**, 461-468.

Appendix: Comparison local units by federal state – original and synthetic data

Federal State	Number of local units – original data	Number of local units – synthetic data	Percental deviation
Schleswig-Holstein	1517	1582	4,11
Hamburg	589	622	5,31
Niedersachsen	4262	4341	1,82
Bremen	358	407	12,04
NRW	11179	10962	-1,98
Hessen	3352	3435	2,42
Rheinland-Pfalz	2464	2419	-1,86
Baden-Württemberg	8931	9093	1,78
Bayern	8111	8171	0,73
Saarland	543	632	14,08
Berlin	959	995	3,62
Brandenburg	1217	1212	-0,41
Mecklenburg-Vorpommern	698	687	-1,60
Sachsen	2893	2862	-1,08
Sachsen-Anhalt	1376	1243	-10,70
Thüringen	1898	1684	-12,71

Table 1: Comparison local units by federal state – Alternative 2 / Model 1

Model 1:

The following explanatory variables are included (continuous and mixed variables all transformed by extracting the cubic root):

Categorical variables:

Kind of unit (single-site company, multi-site company), entry in the register of craftsmen, branch of economic activity

Count variable: Number of functional sections of the local unit

Continuous variable;

Number of persons employed in the functional sectors of the local unit

Mixed variables:

Domestic turnover, export turnover, incoming orders, wages, salaries

Federal State	Number of local units – original data	Number of local units – synthetic data	Percental deviation
Schleswig-Holstein	1517	1529	0,78
Hamburg	589	667	11,69
Niedersachsen	4262	4286	0,56
Bremen	358	439	18,45
NRW	11179	11245	0,59
Hessen	3352	3310	-1,27
Rheinland-Pfalz	2464	2518	2,14
Baden-Württemberg	8931	8927	-0,04
Bayern	8111	8219	1,31
Saarland	543	644	15,68
Berlin	959	1044	8,14
Brandenburg	1217	1243	2,09
Mecklenburg-Vorpommern	698	736	5,16
Sachsen	2893	2553	-13,32
Sachsen-Anhalt	1376	1217	-13,06
Thüringen	1898	1770	-7,23

Table 2: Comparison local units by federal state – Alternative 2 / Model 2

Model 2:

Additionally to model 1 the following explanatory variables are included:

Mixed variables: Number of persons employed in the building and construction sectors of the local unit, number of persons employed in other sectors of the local unit

Federal State	Number of local units – original data	Number of local units – synthetic data	Percental deviation
Schleswig-Holstein	1517	1591	4,65
Hamburg	589	643	8,40
Niedersachsen	4262	4223	-0,92
Bremen	358	471	23,99
NRW	11179	11094	-0,77
Hessen	3352	3318	-1,02
Rheinland-Pfalz	2464	2462	-0,08
Baden-Württemberg	8931	8859	-0,81
Bayern	8111	8357	2,94
Saarland	543	661	17,85
Berlin	959	1045	8,23
Brandenburg	1217	1274	4,47
Mecklenburg-Vorpommern	698	674	-3,56
Sachsen	2893	2630	-10,00
Sachsen-Anhalt	1376	1278	-7,67
Thüringen	1898	1767	-7,41

Table 3: Comparison local units by federal state – Alternative 2 / Model 3

Model 3:

Additionally to model 2 the following explanatory variables are included:

Mixed variables: Number of blue-collar workers in the technical sectors, number of blue-collar workers in the building and construction sectors, number of blue-collar workers in the other sectors, domestic turnover differentiated by technical, building and construction and other sectors instead of domestic turnover, export turnover differentiated by technical and other sectors instead of export turnover, domestic and foreign incoming orders instead of incoming orders.

Federal State	Number of local units – original data	Number of local units – synthetic data	Percental deviation
Schleswig-Holstein	537	1517	182,50
Hamburg	339	589	73,75
Niedersachsen	5361	4262	-20,50
Bremen	431	358	-16,94
North	6668	6726	0,87
NRW	13424	11179	-16,72
Hessen	652	3352	414,11
Rheinland-Pfalz	3001	2464	-17,89
Middle	17077	16995	-0,48
Baden-Württemberg	10554	8931	-15,38
Bayern	6486	8111	25,05
Saarland	456	543	19,08
South	17496	17585	0,51
Berlin	486	959	97,33
Brandenburg	598	1217	103,51
Mecklenburg-Vorpommern	953	698	-26,76
Sachsen	5031	2893	-42,50
Sachsen-Anhalt	907	1376	51,71
Thüringen	1131	1898	67,82
New federal states	9106	9041	-0,71

Table 4: Comparison local units by federal state – Alternative 3 / Model 3 (explanatory variables in model cf. annotations to table 3)