On the way to remote access to German official microdata: a glimpse of work in progress

Rainer Lenz

Saarbruecken University of Applied Sciences
Faculty of engineering
Goebenstrasse 40
66117 Saarbruecken
Germany
Rainer.Lenz@htw-saarland.de
http://www.htw-saarland.de

Abstract. A number of methodological and technical preconditions have to be met in order to provide (automated) remote data access. This paper outlines the German approach to achieving this goal with a strong focus on the scientific and methodological challenges, in particular, regarding the access to cross-sectional and longitudinal economic statistics. The paper concentrates mainly on the generation of anonymised data with a structure similar to that of the real data, which are made available to data users. The data are provided in the form of what are called data structure files, which can be produced by using specific variants of microaggregation, multiplicative stochastic noise and multiple imputation. These data structure files are sent to the researcher after he has submitted a request for remote access. They allow checking a program code for syntactic and semantic errors before it can be finally applied to the original data by remote execution.

1 Introduction

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 the “royal road”, towards giving empirical social and economic research adequate access to official microdata in Germany. Such SUFs have been available for what is called the off-site use of selected and strongly demanded statistics. As regards SUFs of economic statistics, however, these data stocks have not been very well received. Reasons to be mentioned in this respect are the new data perturbating anonymisation methods which are not yet familiar to researchers, too long waiting periods and the excessive effort required to compile the SUFs. Due to the necessary and partly drastic interference in the information structure of the data in the course of anonymisation and, in particular, the reservations still existing with respect to the data perturbating anonymisation methods used, SUFs of microdata of economic statistics for off-site use have by far not achieved the same prominence as, for instance, SUFs used for off-site purposes in the area of individual or household-related statistics. As has become apparent 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 frequently used forms of accessing microdata of economic statistics. For this reason, the current long-term objective is, on the one hand, to improve access through the controlled remote data execution procedure (still manual) in a way to provide researchers with direct insight into so-called data structure files (anonymised form of microdata). The other goal is to facilitate both automated checks of the outcome produced, such as tables or individual values of estimating functions, and hence an early data transmission to researchers.

This paper introduces a project on “An informational infrastructure for the e-science age” which was launched recently and constitutes an essential milestone towards remote data access in Germany. The institutions involved in the project promoted by the Federal Ministry of Education and Research (BMBF) include the research data centre of the Federal Employment Agency in its Institute for Employment Research, the Mainz University of Applied Sciences, the Institute for Applied Economic Research, and the research data centres of the statistical offices of the Federation and the Länder.

Under ideal conditions, an empirical scientist would obtain authorised round-the-clock access to official statistical data from any computer terminal. The results would be delivered to the researchers in real time upon immediate and fully automated confidentiality checks. To ensure such a form of data access for independent scientific research, however, a number of methodological, technical and legal issues must be settled. The focus of this project is on tackling the methodological challenges of remote data access. During the three-year period of the project, solutions are to be developed for a direct (one-to-one) application of the analysis programs of scientists by compiling what are called data structure files. The latter are used to apply, without any further interference or adjustments by the staff in the research data centres, the analysis programs to the data of official statistics. Furthermore, solutions must be developed with respect to output confidentiality. This refers to confidentiality of both outputs in tables and outputs of estimation. While taking into account the legal situation, too, a technical solution can be implemented only after the methodological aspects of remote data access have been tackled. The project uses a sustainable approach in this respect, which means that it will be possible to implement the results developed, like data structure files and output confidentiality concepts, in the context of future technical solutions. In the following chapters, the paper will outline the overall objective of the project, the anonymisation methods to be used in generating so-called data structure files and the approach to be applied in measuring the protective effect of the methods used.

The paper is organised as follows. Section 2 contains a short summary of the main objectives of the research project. Afterwards, in section 3, the idea of data structure files is illustrated together with some potential anonymisation strategies to generate those data files. A recommendation how to measure the protection effect of these strategies is then made in section 4. Finally, in section 5, the previous concepts of disclosure control are applied to the German monthly reports on local units of the manufacturing sector for the years 1998 to 2001.

2 Scientific objectives of the research project

Manifold technical, legal and methodological issues (the latter are a focus of this project) must be settled before a pure remote access application can be completely implemented. Although first applications are available - Lissy in Luxembourg, Coder et al. (2003), and the methods of the Dutch (Hundepool and De Wolf, 2005) and the Danish (Borchsenius, 2005) national statistical institutes -, none of them provide fully automated access routines and some of them – e.g. Lissy – are restricted to specific applications. In Germany, SAM (Heitzig, 2006) is a first technical solution. JoSua[1], too, could possibly be advanced to become an application of this kind. In particular, „safe communication' through lines would be an issue in this context. However, the project deals only with the statistical disclosure control (SDC) side of remote access, not with the aspects of information technology (IT). Hence, it aims to provide the basis for the following three methodological approaches:

1. developing anonymous data structure files which can be used to specify analysis models and must therefore be suitable for semantic analysis, and which allow developing analysis programs that are error-free in terms of syntax.

2. developing and assessing standardised and completely automated output checking procedures.

3. simultaneous consideration of microdata anonymisation and output checks.

In the following we concentrate on the methods of anonymisation used to develop data structure files and on the investigations assuring protection of those data and hence confidentiality regarding the underlying individuals.

3 Development of data structure files

A first goal of the project is the standardisation of data in the form of so-called data structure files. The anonymised data sets, which have the same structure as the original data sets, are sent to the researcher after he/she has submitted a request for use. As a next step, the researcher develops an analysis program code and sends it to the competent research data centre (RDC). Members of the RDC staff then apply the program code to the original data and, upon data security and confidentiality checks, send the output back to the researcher. 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

[1] The data centre of the Institute for the Study of Labor in Bonn (IdZA) has developed an application allowing external researchers to start microdata analyses via the internet. On the one hand, the JoSuA application is user-friendly because researchers can monitor the status of their orders from their workstations and, on the other, it makes IdZA activities easier because the programs must no longer be started manually.

executable, though he does not obtain any information on whether the actual issue 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.

Basic strategies are to be developed for the production of anonymised data structure files which will allow checking a program run for syntactic and semantic errors. To produce such data structure files, data perturbation methods like multiplicative stochastic noise, multidimensional microaggregation and multiple imputation are particularly suitable. Against a modified background of goals, the results of the projects on "De facto anonymisation of economic microdata" and "Economic statistics panel data and de facto anonymisation" can be built on in this context. Both projects set the methodological basis for producing de facto anonymised data sets for enterprises and local units. In particular, data perturbating anonymisation methods were developed or adjusted to the requirements of economic statistics of the German statistical offices and the Federal Employment Agency (cf. Ronning et al., 2005).

3.1 Multiplicative stochastic noise

One main challenge regarding additive noise with constant variance is that on one hand small values are strongly perturbed and on the other large values are weakly perturbed. For instance, in a business microdata set the large enterprises -- which are much easier to re-identify than the smaller ones -- remain still high at risk after noise addition, as illustrated in Table 1 for the turnover variable.

Number of local unit	Region of residence	NACE 2	Number of employees	Turnover anonymised	Turnover original
1	A	34	10	1.460.705	1.903.425
2	B	28	6	3.355.041	3.313.910
3	A	17	2	918.941	810.729
4	C	28	4	743.497	794.211
5	A	29	4	1.355.569	906.228
6	C	15	43	19.938.428	20.195.146
7	A	15	15	6.953.886	6.645.567
8	A	29	7	3.610.630	3.054.587
9	B	24	7	2.292.144	2.398.257
10	A	29	11	3.990.655	3.777.735
11	B	15	81	28.119.621	28.056.704
12	A	29	6	3.673.967	4.073.463

TAB. 1 – *Small example for noise addition*

For instance, the total turnover of the local units nr. 6 and 11 has only slightly changed, whereas the corresponding values of the units nr. 1 and 5 are strongly modified after noise addition.

A possible way out is given by the multiplicative noise approach explained as follows. Let X be the matrix of the original data and W the matrix of continuous perturbation variables with expectation 1 and variance $\sigma^2_w > 0$. The corresponding anonymised data X^a is then obtained as $(X^a)_{ij} := w_{ij} X_{ij}$ for each pair (i, j). The following approach has been suggested by Höhne (2004). In a first step, for each record it is randomly decided whether its values are increased or decreased, each with 0.5-probability. This is done using the main factors $1-f$ and $1+f$. In order to avoid that all values of some record are perturbed with the same noise, these main factors are themselves perturbed with some additive noise s (where $s < f/2$). As an example, the factor f could be set as 0.1 (that is, the relative deviation of the anonymised values from their corresponding original values is about 10 percent) with an additional noise with variance $s=0.03$. In this case, the vector of the original variable values is elementwise multiplied by some vector Z_i like

$$Z_1=(1.09266,1.09102,1.07073,1.11946,1.08521,1.07056,...)$$

or

$$Z_2=(0.90734,0.90898,0.92927,0.88054,0.91479,0.92944,...).$$

Note that in this example all values are perturbed in the same direction. Multiplying by Z_1, they increase, using Z_2 they decrease about 10 percent.

Particularly if the original data follow a strongly skewed distribution, the deviations using this method may strongly depend on the configuration of the noise factors for some few, but large values. That is, despite consistency, means and sums might be unsatisfactorily reproduced. For this reason, Höhne (2008) suggests a slight modification of the method.

The experience gained in the context of the project on “Economic statistics panel data and de facto anonymisation” can be built on in using the relevant methods for the production of data structure files. As regards the application to integrated data material, however, some methodological advancements are still required as high-quality data structure files cannot be produced by a simple combination of sampling and the subsequent application of stochastic noise. Only if the existing procedures are extended/adjusted, the characteristics of the overall stock of original data can be maintained in a sub-stock of the data. If data structure files must be produced as fully anonymous public use files, the selected parameters will have to ensure a by far stronger distortion than has been the case with scientific use files. In particular, the methods including additive noise with mixed distributions (see Roque, 2000 and Yancey, 2002) are anonymisation methods that can be used to meet the increased protection requirements by appropriate parameter selection. However, detailed tests are required in this respect in order to determine the specific type of additive noise and the parameter constellations that are best suited to fulfil the relevant requirements.

3.2 Uni- and multivariate microaggregation

The rationale behind microaggregation (Domingo-Ferrer and Mateo-Sanz, 2002) is that confidentiality rules in use allow publication of microdata sets if records correspond to groups of k or more individuals, where no individual dominates (*i.e.* contributes too much to) the group and k is a threshold value. Strict application of such confidentiality rules leads to replacing individual values with values computed on small aggregates (microaggregates) prior to publication. This is the basic principle of microaggregation. To obtain microaggregates in a microdata set with *n* records, these are combined to form *g* groups of size at least *k*. For each variable, the average value over each group is computed and is used to replace each of the original averaged values. Groups are formed using a criterion of maximal similarity. Once the procedure has been completed, the resulting (modified) records can be published.

Table 2 contains a small example for univariate microaggregation. Here, at first the data are sorted increasingly by the variable "Turnover original". Afterwards, neighboured triples of values are replaced by their average value (that is, we set *k=3*), defining the anonymised variable "Turnover anonymised".

Number of local unit	Region of residence	Number of employees anonymised	Number of employees original	Turnover anonymised	Turnover original
1	A	8	10	2.452.089	1.903.425
2	B	7,67	6	3.721.703	3.313.910
3	A	3,33	2	837.056	810.729
4	C	3,33	4	837.056	794.211
5	A	3,33	4	837.056	906.228
6	C	46,67	43	18.299.139	20.195.146
7	A	46,67	15	18.299.139	6.645.567
8	A	8	7	2.452.089	3.054.587
9	B	8	7	2.452.089	2.398.257
10	A	7,67	11	3.721.703	3.777.735
11	B	46,67	81	18.299.139	28.056.704
12	A	7,67	6	3.721.703	4.073.463

TAB. 2 – *Small example for univariate microaggregation*

Comparing Tables 1 and 2 one may observe that univariate microaggregation provides a stronger protection to outsiders than additive noise. Note that in the current example the method is applied to all numerical variables simultaneously. That is, after anonymisation three grouped together records may differ only in their categorical information. As some further variant of univariate microaggregation one can apply the procedure to each numerical variable separately in order to preserve the analytical validity of the data to a better extent. Much stronger protected and for scientific analyses and in general not appropriate data are obtained applying variants of multivariate microaggregation, where the data are not sorted by some existing numerical variable, but by some additional variable, e.g., a weighted sum over the whole of numerical variables (so-called Z-score).

The advantage of microaggregation methods is that they produce anonymous values in the form of a linear combination of real values. Thus they automatically make sure that content-related linear dependencies are maintained in variables that are treated together. In the context of the projects on "Anonymisation of economic microdata" (Ronning et al., 2005 and Lenz et al., 2005) and "Economic statistics panel data and de facto anonymisation" (Brandt et al., 2008), the multidimensional microaggregation procedures convinced through their high protective effect (particularly in including categorical variables). With respect to the analysis quality achieved, however, they were greatly inferior to other methods (also the one-dimensional microaggregation procedures). Since data structure files have to fulfil other requirements than scientific use files, however, microaggregation procedures can serve as an approach to meeting the relevant requirements. Compared, in particular, to multiple imputation, the advantage of this approach would be that the data user would have to be provided with only one data structure file. Unfortunately, first project results testing microaggregated data for this purpose are less promising.

3.3 Generation of synthetic data sets

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 and filter structures can be taken into account in the production of the relevant sets. In addition, the approach can be applied to continuous variables in the same way as to categorical variables. Since, in case of continuous variables, an attempt is made to maintain the variance-covariance matrix of the complete data set, the results a scientist will get in using the data structure files will usually differ from the original data-based results to a minor extent only. 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. Note that there exist further anonymisation procedures preserving approximately the variance-covariance matrix of the original data (cf. Domingo-Ferrer and Gonzales-Nicolas, 2010).

The proposal to generate synthetic datasets for the scientific community by means of “multiple” imputation was made 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 which are analysed individually. The actual result of the analysis follows by the application of simple combining rules. In principle, fully synthetic and partially 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 German business register or from another official survey) which is included in the model for imputation. In contrast, regarding partially synthetic datasets all attributes or only sensitive attributes of the units contained in the survey, are replaced by synthetic values (cf. Reiter, 2003).

Since the project has the goal to develop standardised anonymisation procedures that can easy be applied and adapted to all kinds of surveys by every member of the RDC staff and other employees of the statistical offices, the software used should be 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.[2] 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

[1] http://www.isr.umich.edu/src/smp/ive/

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 (Raghunathan et al., 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 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 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 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.

4 Measuring the data protection

Of course, the risk of disclosure of confidential information by a potential data intruder should be minimised to greatest possible extent. In the case of data structure files, the German law requires that the data are absolutely anonymised. That is, the remaining risk of re-identification after perturbation has to be zero. The degree of confidentiality can be tested by carrying out realistic matching scenarios, as modelled below. A more detailed description can be found in Lenz (2006). An adaptation of the approach meeting the specific requirements of longitudinal data is described in Lenz (2008).

4.1 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.

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:

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

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

The constraints of the above *(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. A detailed empirical study is presented in Lenz (2006).

5 Application to real world data

The members of the 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 possesses a straightforward questionnaire containing about 30 variables.

Several test simulations carried out so far with microaggregated data and data perturbed by multiplicative noise have been carried out. As a result, it turned out that these anonymisation strategies led to weakly anonymised data. “Weak” in the sense that it seems difficult to use those methods in order to produce absolutely anonymous data structure files (cf. Lenz, 2010). For this reason, in this section we focus on synthetical data generated using the software IVEware.

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 also be possible. However, until now only the aggregated annual data are available for scientists at the RDC.

5.1 Anonymisation

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, about 50.300 local units were contained in the data.

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. Moreover, only one variable will be anonymised at a time. Therefore the dataset is reduplicated: Next to the original data the same dataset is added once again, but this time the values which shall be later imputed are replaced by missings.

In practice, it turned out that the imputation of categorical variables with more than 4 to 6 distinct values poses a particular challenge. Therefore, for technical reasons it was necessary to introduce binary dummy variables. Hence, considering the example of the location of the local unit coded by the 16 federal states of Germany, two alternatives are tested.

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 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 dummy value of 1 is imputed, all other proceeding dummies are automatically set to 0. The computing time is acceptable for this alternative: between 30 and 40 minutes are needed each synthetic data set, if the number of iterations for the regression procedure was previously set to 10. This alternative works quite well, see Table 3 below.

Federal State	Number of local units, original	Number of local Units, synthetic	Percentual deviation
Schleswig-Holstein	1517	1582	4,11
Hamburg	589	622	5,31
Niedersachsen	4262	4341	1,82

Federal State	Number of local units, original	Number of local Units, synthetic	Percentual deviation
Bremen	358	407	12,04
Nordrhein-Westfalen	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-Vorp.	698	687	-1,60
Sachsen	2893	2862	-1,08
Sachsen-Anhalt	1376	1243	-10,70
Thüringen	1898	1684	-12,71

TAB. 3 – *Comparison of local units by federal state – original versus synthetic data*

In Hafner and Lenz (2010) further models are tested. As a whole, it is observed that alternative 2 described above is the most promising one, while the other alternatives tested so far drop out due to their corresponding expenditure in computing time or non-satisfying imputation results concerning multinomial regression, respectively.

5.2 Measuring the data protection

In the following, as confidential target data about 36.000 single-site enterprises, taken from the monthly report for the year 2001, are used. In order to simulate matching scenarios it is necessary to create additional knowledge of some potential data intruder. For this purpose, a commercial database was built up containing about 9.000 common with the target records. As blocking variables we used the location of the local unit (summarised to old and new federal states), the branch of economic activity (NACE 2 code) and the number of employees (summarised to 6 employee size classes). As numerical key variables for the assignment the two variables number of employees and total turnover were available. The result was somewhat surprising: just negligible 18 of the about 9.000 units could be assigned correctly, that is about 0,2%.

To check whether this finding is reliable, we also calculated the fraction of corresponding values for the blocking variables in the original and in the synthetic data. In total, the two-digit NACE code was the same in both sources for only 23,6% of the units, the employee size class for 80,2% and the location (old / new federal state) for 77,5% of all cases. The combination of these three categorical variables just remained unchanged for 14,7% of the local units. This observation strongly indicates that the generation of synthetic data has a huge protective effect regarding the confidentiality of data. Future work will show which parameter settings and other improvements regarding the synthetic data generating process are necessary to obtain absolutely anonymised micro data. However, the recent results are very promising.

6 Prospects

The project outlined in this paper serves as an important link between the developments that have been made in the past few years to improve the ways of data access for the scientific community and the concepts the research data centres are currently preparing for the future. Hence, the project constitutes an essential milestone towards remote data access. In the long run, this way of data access seems to be the only practicable solution at both the national and international level, all the more since a method, once developed, can be applied to other surveys without delay and can hence ensure a just-in-time provision of data. The technical developments have reached a level which allows online access from anywhere and will allow online access to an adequate range of data soon. Remote access allows scientists to process data in a flexible manner which is independent of time and location. Also, it has the advantage that the real data remain in the protected rooms (and on the protected servers) of official statistics. Furthermore, this form of data access increases both networking among researchers and scientific transparency, because in future any scientist may access the data and replicate the results at any time.

Acknowledgement

This work has been partially supported by the German Ministry of Education and Research (BMBF). The author also greatfully acknowledges the three anonymous referees whose suggestions have improved the article.

References

Abowd, J.M., Stinson, M. and Benedetto, G. (2006) *Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project,* (www.sipp.census.gov/sipp/SSAfinal.pdf).

Borchsenius, L. (2005) New Developments in the Danish system for access to microdata, *Proceedings of the Joint Eurostat UN/ECE Worksession on Statistical data confidentiality,* Geneva, November 2005.

Brandt, M., Lenz, R. and Rosemann, M. (2008) Anonymisation of panel enterprise microdata - Survey of a German Project, in: Domingo-Ferrer, J., Saygin, Y. (Eds.): Privacy in Statistical Databases, *Lecture Notes in Computer Science*, vol. 5262, Springer, Heidelberg, 139-151.

Coder, J. and Cigrang, M. (2003) LISSY Remote Access System*, Proceedings of the Joint Eurostat UN/ECE Worksession on Statistical data confidentiality,* Luxembourg, April 2003.

Domingo-Ferrer, J. and Mateo-Sanz, J.M. (2002) Practical Data-oriented Microaggregation for Statistical Disclosure Control. *IEEE Transactions on Knowledge and Data Engineering*, 39, 189–201.

Domingo-Ferrer, J. and Gonzales-Nicolas, U. (2010) Hybrid microdata using microaggregation. *Information Sciences*, 180(15), 2834–2844.

Drechsler, J., Bender, S. and Rässler, S. (2007) Comparing Fully and Partially Synthetic Data Sets. for Statistical Disclosure Control in the German IAB Establishment Panel, *Proceedings of the Joint Eurostat UN/ECE Worksession on Statistical data confidentiality,* Manchester, December 2007.

Elliot, M. and Dale, A. (1999) Scenarios of attack: the data intruder's perspective on statistical disclosure risk, *Netherlands Official Statistics,* 6-10.

Graham, P., Young J. and Penny, R. (Eds) (2008) Methods for constructing synthetic data, *Official Statistics Research Series*, vol. 3, Wellington, New Zealand, Statistics New Zealand.

Hafner, H.-P. and Lenz, R. (2010) Synthetic data structure files: development and disclosure control. Appears in: Proceedings of the UNECE work session on statistical data confidentiality, Bilbao.

Heitzig, J. (2006) Wissenschaftsserver zur Auswertung von Mikrodaten, Federal Statistical Office of Germany, unpublished paper.

Höhne, J. (2003) Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten. *Forum der Bundesstatistik*, vol. 42, Federal Statistical Office Germany,Wiesbaden.

Höhne, J. (2008) Anonymisierungsverfahren für Paneldaten, *Journal of the German Statistical Society* (Wirtschafts- und Sozialstatistisches Archiv), vol. 2 (3), 259-276.

Hundepool, A. and de Wolf, P. (2005) *OnSite@Home: Remote Access at Statistics Netherlands,* Monographs of Official Statistics, Luxembourg.

Kennickell, A.B. (1999) Multiple Imputation and Disclosure Control: The Case of the 1995 Survey of Consumer Finances. In: *Record Linkage Techniques*, Washington D. C., 248-267.

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 (4), Sweden, 681-710.

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. (2010) Methoden der Geheimhaltung wirtschaftsstatistischer Einzeldaten und ihre Schutzwirkung, appears in: *Statistik und Wissenschaft*, Statistisches Bundesamt, Wiesbaden, Germany.

Lenz, R., Vorgrimler, D. and Scheffler, M. (2005) A standard for the release of business microdata, *Monographs of Official Statistics - Research in Official Statistics,* (presented at the UN-ECE/Eurostat work session on statistical data confidentiality,) 197-206.

Rubin, D.B. (1993) Discussion: Statistical Disclosure Limitation, *Journal of Official Statistics*, 462-468.

Raghunathan, T.E., Lepkowski, J. M., Van Hoewyk, J. and Solenberger, P. (2001) A

Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models*, Survey Methodology 27, 85-95.*

Raghunathan, T., Reiter, J. and Rubin, D. (2003) Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* 19 (1), Sweden, 1-16.

Reiter, J.P. (2003) Inference for partial synthetic, public use microdata sets, *Survey Methodology*, 181-189.

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M. and Vorgrimler, D. (2005) Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten, *Statistik und Wissenschaft*, vol. 4., Statistisches Bundesamt, Wiesbaden.

Roque, G.M. (2000) Masking Microdata Files with Mixtures of Multivariate Normal Distributions, Dissertation, University of California Riverside.

Yancey, W.E. (2002) Working Papers for Mixture Model Additive Noise for Microdata Masking, *Research Report Series* (Statistics #2002-03), Statistical Research Division U.S. Bureau of the Census, Washington D.C. 20233.

Résumé

Pour établir la téléinformatique automatisée, il faut partir des préliminaires méthodiques et techniques. L'étude suivante montre le chemin que les Allemands prennent pour atteindre ce but tout en se concentrant sur les défis méthodiques qui se posent avec l'accès à des enquêtes confidentielles du patronat (observées pour une année ou pour une période de plusieurs années). Un pirate informatique tente, à l'aide d'informations complémentaires extérieures, de réidentifier des ensembles de données d'entreprises dans un fichier protégé. Les défis consistent d'une part à produire des fichiers rendus anonymes, transparents et comprehensibles pour les scientifiques qui exercent leur métier chez eux ou à l'institut, des fichiers dont les structures ressemblent à ceux qui sont authentiques et originaux. Ces fichiers dits fichiers de structure sont produits à l'aide des procédés spéciaux garantissant l'anonymat (p. ex. microagrégation, recouvrement stochastique ou imputation multiple). Et d'autre part, les défis consistent à développer des procédés concernant le contrôle standardisé et automatisé de résultats d'analyse.