Working Paper
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
(Helsinki, Finland, 5 to 7 October 2015)

Topic (iii): Preserving Data Quality and Usability in Disclosure-Limited Data

Anonymization of longitudinal surveys in the presence of outliers

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

[*] Department of Engineering, Saarland State University of Applied Sciences, 66117 Saarbrücken, Germany, Hans-Peter.Hafner@htwsaar.de

[**] Department of Engineering, Saarland State University of Applied Sciences, 66117 Saarbrücken, Germany and Department of Statistics, Technical University of Dortmund, 44227 Dortmund, Germany, Rainer.Lenz@htwsaar.de

Abstract: Outliers are units deviating significantly from the rest of the sample for one or more attributes, or units showing a rare combination of some categorical attributes. The presence of such outliers, which is typical for business surveys, poses a great challenge for data holders to anonymize confidential micro data. The task is even more complex for longitudinal data. Using as an example four consecutive waves of the German cost structure survey, we compare several anonymization approaches based on the theory of linear mixed models and extensions thereof. For each of them, we discuss both the analysis potential and the associated disclosure risk of the anonymized data file. Recent results show that robust models are very promising, but future work is needed to improve the analytical potential and to reduce the computing time.

1 Introduction

There is a high demand for micro data of business surveys of official statistics by the scientific community, but due to confidentiality requirements and staff shortage, the statistical offices are not able to respond in adequate time to all requests of researchers. In recent years there were efforts to develop better anonymization methods for micro data stemming from business data by means of generating synthetic values from CART (Classification and Regression Trees) models Reiter, 2005; Hafner and Lenz, 2011). However especially for longitudinal data including extreme outliers,

the analytical validity of the synthetic data turns out to be not satisfactory. It seems to be more promising to use models that can handle correlation between data values, as it is the case when an enterprise reports its characteristics (like turnover, number of employees) for several years. Linear mixed models (LMM) define a promising class of models appropriate for this task. In Chapter 2, we introduce LMM and some extensions. Chapter 3 presents in brief the data we use for test purposes. Chapter 4 gives a detailed description of several examined simulation settings while the subsequent chapters evaluate the analytical potential and associated disclosure risk of the different anonymized data files. We finish with a conclusion and an outlook for further work.

2 Linear Mixed Models and Extensions

A linear mixed model, abbreviated LMM, is any model satisfying the following four conditions:

(1) $Y_i = X_i\beta + Z_i b_i + \varepsilon_i$,
(2) $b_i \sim N(0, D)$,
(3) $\varepsilon_i \sim N(0, \Sigma_i)$,
(4) $b_1,\ldots,b_N, \varepsilon_1,\ldots, \varepsilon_N$ are independent.

Y_i is the n_i-dimensional vector of responses for unit i, $1 \leq i \leq n$, X_i and Z_i are (n_i x p) respective (n_i x q) – dimensional matrices of known covariates. β is a p – dimensional vector containing the fixed effects (that is, effects being constant across units), b_i is a q – dimensional vector containing the random effects (that is, effects varying across units) and ε_i is a n_i – dimensional vector of residuals. Finally D is a (q x q) – covariance matrix and Σ_i is a (n_i x n_i) – dimensional covariance matrix which depends on i only through the dimension n_i; that means that the set of parameters of Σ_i does not depend on i.

For an introduction to LMM, see Verbeke and Molenberghs, 2009.

Koller, 2014 developed a robust estimation method for LMM. Robust techniques distinguish between observed values that originate from a certain supposed distribution and what are called *contaminated* values like measurement errors or outliers. Robust estimation equations use different weights, higher ones for observations that fit to the supposed distribution and lower ones for extreme values. Thus, robust methods are less vulnerable to the violation of distributional assumptions. The treatise of the foundations of robust statistics is beyond the scope of this paper, the interested reader is referred to Maronna et al., 2006.

Another possibility to circumvent the strong assumptions of normally distributed random effects and residuals is the use of Generalised Linear Mixed Models (GLMM). Here the outcomes of a single unit are supposed to be observations generated by a distribution belonging to the exponential family. For details, see Fitzmaurice et al. 2011.

3 The Cost Structure Survey

The Cost Structure Survey provides extensive information on enterprises of the manufacturing industry including attributes like branch of economic activity, location

of the head quarter (new and old federal states), number of employees, turnover, power consumption and raw material consumption. At most 18.000 enterprises occupying 20 or more employees are included in the survey each year. The sample is drawn every four years; however, enterprises with 500 and more employees have to participate every year at the survey. For further information on the sampling strategy, see Fritsch et al., 2004.

For the comparison of different anonymization methods, we use panel data of this survey consisting of 13.227 enterprises for the years 1999 to 2002.

4 Description of the models used for anonymization

We consider two approaches and for each of them several variations.

Approach 1: For the estimation of the linear mixed model and the prediction of the anonymized values, the same set of units is used.

Approach 2: We divide the whole dataset into two subsamples. For the estimation of the model we use the first sample (or a subset thereof), for the prediction of the anonymized values we use the second sample (or a subset thereof) and the other way around.

We describe the second approach in more detail:

(1) Divide the whole dataset D into layers according to the 17 economic groups and new and old federal states

(2) Let $Y_1,..,Y_t$ be the attributes to be anonymized; let $(Y_i)_{1999},\ldots, (Y_i)_{2002}$, $i = 1,\ldots, t$ be the values of the attribute Y_i for the years 1999 to 2002.
Compute $Y_iSQSum = ((Y_i)_{1999})^2 + ((Y_i)_{2000})^2 + ((Y_i)_{2001})^2 + ((Y_i)_{2002})^2$ and sort each layer by the values of this variable in descending order.

(3) Prior to the anonymization of attribute Y_i we construct two samples S_1 and S_2 of D: For each layer the units are distributed to one of the two samples alternating according to their value of Y_iSQSum (the unit having the highest value is assigned to S_1, the one with the second highest value to S_2 and so on). If a layer is made up of an odd number of enterprises, the record having the lowest Y_iSQSum value is excluded.

(4) S_1 and S_2 are sorted by layer and Y_iSQSum values. Let s be the number of records in S_1 respective S_2 (equal cardinality by construction); let $(E_{i,j})_l$ be the jth unit (j=1,…,s) in sample l (l=1,2) sorted by Y_iSQSum. Then we assign $(E_{i,j})_1$ to $(E_{i,j})_2$ and proceed as follows: We estimate a model $(M_i)_1$ for Y_i using a subset T_1 of S_1. Let T_2 be the subset of S_2 that contains all units being assigned to a unit of T_1. Then we predict the anonymized values of Y_i for the units of T_2 by usage of the parameters of model (M_i). Since the unit-specific random effects parameters b_i of this model are only applicable for the units of T_1, we exchange the unit identifier by the identifier of the assigned unit of T_1.

The variations for each approach are the following:

Variation 1: We estimate a common LMM for every variable to be anonymized for the whole dataset (approach 1) respective one LMM for each of the two samples (approach 2).

Variation 2: We estimate separate LMM for the data in the centre of the distribution and the outliers. We use the Hampel identifier with *t=2* as criterion to define outliers. Thereafter a value x_i of the variable *X* is an outlier if

$$| x_i - Median(X)| > t * MAD(X),$$

where the median absolute deviation MAD(X) is defined as $Median(|x_j - Median(X)|)$ over all values x_j, j=1,..,n of *X*.

Variation 3: We estimate separate LMM for each of the 17 economic groups.

Variation 4: Different models as for variation 3, but GLMM instead of LMM with Poisson distribution assumption.

Variation 5: Different models as for variation 3, but robust LMM instead of simple LMM.

We label the models by *ij* (*i* = 1,2; *j* = 1,..,5), where *i* denotes approach 1 or 2 and *j* the chosen variation.

For our simulations, we use only a subset of the variables of the survey consisting of the numerical attributes *number of employees* and *turnover* and the categorical attributes *branch of economic activity*, *legal form* and *new / old federal states*. We combine the branches of economic activity into 17 groups. We do this because the 4-digit NACE code that is included in the survey is a highly identifiable attribute. For now, we anonymize only the two numerical attributes. First, we build the model(s) for the turnover and calculate the anonymized turnover values as predictions from the model(s). Then we proceed with the number of employees analogous, but we use the anonymized turnover values to estimate the parameters of the model(s). For a better reproduction of the variations over the four years, we calculate the rate of change of the turnover

ch_turn = 0 for the year 1999 and *ch_turn* = $(\text{turnover}_t - \text{turnover}_{t-1}) / \text{turnover}_{t-1}$

for *t = 2000, 2001, 2002.*

Analogous, we define *ch_employ* as the rate of change of the number of employees.

A crucial part of the process is the model selection and there are many more or less sophisticated methods to support this. For an overview, see Müller et al., 2013. For reasons of simplicity, we used the widely applied Akaike Information Criterion (AIC) which is defined as

$$AIC = -2 * L(\hat{\theta} | y) + 2 * k,$$

where $\hat{\theta}$ is the maximum likelihood estimator for the parameter θ, *y* is the vector of observed values, *L* the Likelihood function and *k* is the number of free parameters. The model that fits the data best possible is the one showing the smallest AIC value.

At first, we determined the random factors. The number of random parameters may not exceed the number of observations. We tested models with a random intercept (that is, every unit has a different parameter for the intercept) against models with a random slope for every year. The random slope models show a much lower AIC value.

To specify the fixed effects we added the variables and 2-way interaction terms stepwise. Interestingly, the inclusion of terms containing the legal form increased the AIC value, while the addition of all other terms led to a decrease.

We performed our simulations with the software R (see R Core Team, 2014). For the estimation of the LMM and GLMM we used the package *lme4* (Bates et al., 2015), for the robust models *robustlmm* (Koller, 2013).

5 Analysis of the analytical potential

Since the dataset contains just two numeric variables, an analysis of the impact of the anonymization methods on multivariate models does not make sense. Therefore, we limit the study to the changes of appropriate descriptive measures, considering three groups of measures:

I) Deviations of single values:

Portion of deviations

(i) at most 5% (*d0*)

(ii) above 5 and at most 10% (*d5*)

(iii) above 10 and at most 20% (*d10*)

(iv) above 20% (*d20)*.

II) Deviations of aggregated values over the 17 economic groups

Portion of deviations regarding descriptive measures:

(i) correlation coefficients between waves > 0,1 (*r_c)*

(ii) rank correlations between waves > 0,05 (*r_s)*

(iii) arithmetic means > 10% (*r_mean)*

(iv) standard deviation > 10% (*r_sd*).

Portion of deviations regarding descriptive measures of the change rates of the variables between consecutive waves: Analogous to the measures for the variable values (*cr_r_c* for correlation coefficients, *cr_r_s* for rank correlations, *cr_r_mean* for arithmetic means and *cr_r_sd* for standard deviations).

III) Deviations of trends over the 17 economic groups

We examine whether the change rates between two consecutive waves have the same sign for each economic group and differentiate between

(i) portion of differing trends for the change rate of one variable (*t_s*)

(ii) portion of differing trends for the combination of change rates of occupation and turnover (*t_m*).

Until now, we computed nine of the proposed ten models. The robust method has very heavy memory requirements. Using a notebook with 4 GB main memory and a 2.3 GHz processor, the computation of model 15 (approach 1, robust estimation) was aborted because not enough memory could be allocated. The execution of model 25 took nearly 24 hours.

The deviation measures for the nine models are displayed in table 1.

Model									
Deviation Measure	11	12	13	14	21	22	23	24	25
d0	81,4%	82%	81,6%	78%	12,5%	14,4%	13,5%	17,9%	22%
d5	13,6%	13,3%	13,1%	15,8%	12%	13,6%	12,6%	15,9%	18,6%
d10	4,2%	3,9%	4,3%	5,1%	21,1%	22,8%	21,6%	24,1%	25,1%
d20	0,9%	0,8%	0,9%	1,2%	54,4%	49,2%	52,1%	42%	34,3%
r_c	0%	0%	0%	0%	21%	18,7%	21,6%	22,3%	18,1%
r_s	1,9%	0,4%	1,5%	0%	33,2%	28,8%	32,1%	36,1%	31,9%
r_mean	0%	0%	0,7%	0%	29,4%	28,7%	27,9%	23,5%	13,2%
r_sd	1,5%	1,5%	0,7%	0,7%	42,6%	51,5%	47,8%	42,6%	24,3%
cr_r_c	80%	77,6%	76%	80,4%	69%	69%	68,6%	71%	62,4%
cr_r_s	91,4%	89,8%	88,2%	89,4%	83,1%	82,4%	82%	83,1%	79,6%
cr_r_mean	88,2%	90,2%	90,2%	90,2%	86,3%	82,4%	85,3%	89,2%	86,3%
cr_r_sd	57,8%	50,0%	48%	64,7%	51%	52%	50%	61,8%	49,3%
t_s	12,7%	14,7%	17,6%	17,6%	18,6%	19,6%	18,6%	25,5%	17,6%
t_m	25,5%	29,4%	31,4%	33,3%	35,3%	35,3%	33,3%	41,2%	31,4%

Table 1. Deviation measures for the different anonymization models

Regarding the deviation of single values and descriptive measures of the anonymized attributes, approach 1 is very close to the original data. There are no significant differences between the four variations. On the other hand, the equivalent results for approach 2 are not satisfactory. Here, by far the best results are achieved for the robust variation (model 25).

The deviations for the descriptive measures of the change rates are all very high. One explanation for this is that most of these rates are below 10% such that an absolute deviation of 1% between original and anonymized data is already a relative deviation of more than 10%. Therefore, a more meaningful characteristic should be developed.

Surprisingly, the single trends are best preserved for model 11. We expected a better trend preservation in the case of different models for the distinct economic groups (models 13 and 14). The models for approach 2 behave slightly worse. By far the largest distortion is observed for the GLMM (model 24). The portion of different combined trends is nearly twice as high as for the single ones. For the future, an additional measure is planned to distinguish between large and small trend changes. Especially for approach 2, it is problematic when there are a few enterprises with very large change rates for a variable.

6 Analysis of the disclosure risk

We apply the scenario of the so-called database cross match to measure the disclosure risk for the different anonymization models. Within this scenario, a data intruder tries to assign as many records of an external database *A* (additional knowledge) as possible to the confidential target data *B*. For our simulations, we used the original data as additional knowledge. The reason for this is that the effort to generate a realistic external database is too laborious. For a former project, an external database for the data of the cost structure survey was built. Unfortunately, this database is not available to the authors. So, since external knowledge from commercial databases is in general of very low quality (see Hafner, 2008), we overestimate the real disclosure risk by far.

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. These variables are called blocking variables, since they divide the whole data into disjoint blocks. We use as blocking variables the 17 economic groups, new / old federal states and 6 employment size classes.

The scenario of a database cross match can mathematical be modeled as a multi-criteria assignment problem that can be transferred to a linear assignment problem with a target function that has to be optimized. The main task consists now in a suitable choice of the coefficients. Since a data intruder has information for several waves, it is self-evident to reproduce this complex structure in the coefficients. Lenz, 2008 proposes four methods for the determination of the coefficients: Conventional distance based approach, correlation based approach, collinearity approach and Chi Square approach. Since each approach exploits only partially the available information in the data, it seems to be advisable to combine the different approaches. The easiest way to do this is the hybrid matching. This means that all approaches are conducted and the resulting parameters are combined to a convex combination (i. e. a kind of weighting). We use the equal weight of 1 for the parameters of all approaches. For technical details, see Lenz, 2008.

Even the successful assignment of a unit may be a fruitless disclosure attempt, precisely when the interesting individual value (or the interesting individual information) deviates too much from the actual original value. We assume that a disclosure has a benefit for the intruder if an anonymized value *x* deviates at most 10%. In this case we call *x* a *useful value*.

We introduce some notations for the following disclosure risk analysis:

Let *A* be the external database (in our case the original data), *B* the anonymized data, *b* the number of blocks, $(n_A)_i$, $i=1,..,b$ the number of units of A in block *i*, $(n_B)_i$, $i=1,..,b$ the number of units of B in block *i* that are assigned to an unit of *A*.
We call $h_i = (n_B)_i / (n_A)_i$, $i=1,..,b$ the *hit rate* for block *i*. Let *t* be the number of variables that are anonymized, u_i the number of useful values in block *i*. Then we calculate the disclosure risk dr_i for block *i* as $dr_i = (h_i * u_i) / ((n_B)_i * t)$.

Model									
	11	12	13	14	21	22	23	24	25
# of blocks with $dr_i > 10$	197	197	197	196	29	34	24	30	57
# of correct matched enterprises in these blocks	8263	8575	8459	8862	133	175	114	146	405
# of blocks with $dr_i > 20$	197	197	197	196	7	8	9	6	9
# of correct matched enterprises in these blocks	8263	8575	8459	8862	16	25	27	16	57

Table 2. Blocks and enterprises with high disclosure risk

Table 2 shows the number of blocks having a disclosure risk above 10 respective above 20 percent and the number of correct matched enterprises contained within these blocks at risk.

It is evident that the disclosure risk of the anonymized data generated by approach 1 is not acceptable. We have 197 blocks in total. That means that – besides one exception for model 14 – all blocks have a disclosure risk above 20 percent. The number of correct matched enterprises is about 65 percent of all enterprises (13.227). The proportions of blocks and enterprises at risk for approach 2 are substantially lower. Even for the regarding the analytic potential most promising model 25, there are only about 3% of the enterprises that hold a disclosure risk exceeding 20 percent.

7 Conclusion and Future Work

So far, none of the tested models is satisfying concerning both the analytical potential and the disclosure risk of the anonymized data. While the approach 1 models have a high analytical potential and a high disclosure risk, it is the other way around for the approach 2 models: low disclosure risk but also huge deviations from the original data.

Most promising is the robust method for approach 2. Modifications of the assignment of the units between the two samples may lead to an increase of the utility of the data. Furthermore, the correlation structure between the waves needs to be better reproduced. Until now, a huge drawback of the robust method is the computing time. Future research has to develop an implementation to carry out the robust estimation of the parameters of an LMM in an acceptable time.

References

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-8.

Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. (2011). *Applied Longitudinal Analysis (2nd ed.).* Hoboken.

Fritsch, M., Görzig, B., Hennchen, O. & Stephan, A. (2004). *Cost Structure Surveys for Germany*. *Schmollers Jahrbuch* **124**, 557-566.

Hafner, H.-P. (2008) *Die Qualität der Angriffsdatenbank für die Matchingexperimente mit den Daten des KSE-Panels 1999 – 2002*. Mimeo. IAB. Nürnberg.

Hafner, H.-P., Lenz, R. (2011). *Some aspects concerning analytical validity and disclosure risk of CART generated synthetic data. UNECE/Eurostat Work Session on Statistical Data Confidentiality*. Tarragona.

Koller, M. (2014). *robustlmm: Robust Estimating Equations and Examples.* *https://cran.r-project.org/web/packages/robustlmm/robustlmm.pdf*.

Koller, M. (2015). *robustlmm: Robust Linear Mixed Effect Models.* R package version 1.7-4.

Lenz, R. (2008). *Risk Assessment Methodology for Longitudinal Business Micro Data, Journal of the German Statistical Society (Wirtschafts- und Sozialstatistisches Archiv), vol. 2,* 241-257.

Maronna, R.A., Martin, R.D. & Yohai, J.V. (2006). *Robust Statistics.* Chichester.

Müller, S., Scealy, J.L. & Welsh, A.H. (2013). *Model Selection in Linear Mixed Models. Statistical Science* **28**(2), 135-167.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* Vienna.

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

Verbeke, G., Molenberghs, G. (2009). *Linear Mixed Models for Longitudinal Data.* New York.