Synthetic Generation of Business Statistics Micro Data

Kolb, Jan-Philipp*
E-mail: Kolb@uni-trier.de

Münnich, Ralf*
E-mail: Muennich@uni-trier.de

Zimmermann, Thomas*
E-mail: thzimmer@uni-trier.de

**University of Trier, Economic and Social Statistics*
Universitätsring 15
54290 Trier, Germany

ABSTRACT

The analysis of business statistics is of enormous importance for the evaluation of actions implemented in the new growth strategy Europe 2020. In order to appropriately measure growth and sustainability in this context, adequate indicators have to be constructed. Many of these indicators, however, are based on survey data. Hence, considering the statistical production process leads to applying proper estimation methods in order to obtain reliable figures. In case model-based methods are applied, the critique from Gelman *(2007) makes it necessary to also consider the sampling design and the corresponding weights. The evaluation of these methods, however, urges the need of incorporating a Monte-Carlo study in an appropriate realistic environment. The availability of micro data is a necessary condition to apply complex Monte Carlo simulations. However, business micro data are usually not available due to disclosure reasons. Thus, synthetic but realistic data have to be generated in order to allow the researcher to compare estimation strategies properly. The paper provides an overview of generation methods for synthetic business micro data. An example application to small area estimation methods for business data is given. The synthetically generated dataset is applied in a large close to reality Monte Carlo study within the BLUE-ETS project (http://www.BLUE-ETS.eu) in order to compare model and design based estimation methods under various sampling designs.*

Keywords: Synthetic data, microsimulation, business survey, small area estimation

Introduction

Modern policy support in business statistics faces the difficulty to gain information on small regions as well as on many NACE subclasses. This is also an object of the project on blue enterprise and trade statistics (BLUE-ETS) which is part of the MEETS initiative. One of the targets of the BLUE-ETS project is it to investigate the survey design in business statistics. Since business data have very specific characteristics, it is a very challenging task to improve the information collection in business statistics. Often small area estimation techniques are applied, when classical methods suffer from low reliability due to small sample sizes in at least some domains or areas. Münnich (2008) showed that focusing on asymptotic results in real applications may lead to inappropriate results. Therefore, it is useful to evaluate methods in a simulation framework with a close-to-reality environment. A familiar proceeding to control for the interplay between survey design and estimation techniques is the evaluation within a Monte Carlo simulation. ISTAT, BLUE-ETS project coordinator, kindly provided business statistics datasets to enable simulations within the BLUE-ETS project. The provided data sets, however, are very sensitive datasets which have to be protected. Thus, data

disclosure is an important topic of this work. Moreover, the information provided is not contained in a single dataset. None of the provided datasets contains the complete necessary information to carry out the simulation study. Thus techniques have to be found which allow the combination of all the information in one realistic and safe dataset. To generate a reasonable synthetic dataset, the peculiarities of business statistics in general and of the provided datasets in particular have to be studied. The information resulting from this examination is used for further steps in the proceeding. It should be made sure that complications which can occur in reality, are also likely to occur for the synthetic case.

Moreover, the aim of the work is to provide a safe dataset which is publicly availble to enable other researchers to reproduce the results and to make their own comparative research. This can be interpreted as an open source research philosophy (Burgard et al. 2011). A precondition for making a synthetic data set publicly available is that it is impossible to link synthetic entries with real life firms. However, a good solution has to be found which combines disclosure control and usability of data. Data utility is a very decisive keyword for simulation studies on synthetic data. Models build upon the synthetic data should resemble the models built upon real data. The usability of the synthetic data set will be tested in a simulation study with small area estimators.

The challenge working with business data

The world of business statistics is very rich in different kinds of enterprises which can show a huge variety of different characteristics. Heavily skewed distributions with outliers and highly different numbers of firms with respect to NACE classes lead to large problems when dealing with optimal sampling designs and estimation. For example the distributions of variables like turnovers or returns have many outliers and they are expected to be highly skewed. The revenue per employee can show extremely different values for firms of different branches. Moreover, heterogeneity also occurs for the European regions.Small enterprises are completely different from worldwide affiliated groups. This makes it complicated to develop an optimal sampling design. The European perspective makes the harmonized analysis of economy even more complicated, because different sampling schemes are applied which do not lead necessarily to the same quality of results. Therefore, it is important to get an impression of the impact of the different influence factors.

Another challenging topic is the fact that the spatial structure of business data plays an important role. It may happen that the economy of a whole region is shaped by one big enterprise. Many economic works cover the question how to encourage cluster activities. The understanding on this topic is decisive for the regional business development. Business demography is an important term in the context of this work. The topic gets even more complicated if a development over time should be evaluated, because entries and exits occur more often than it is the case for population data.

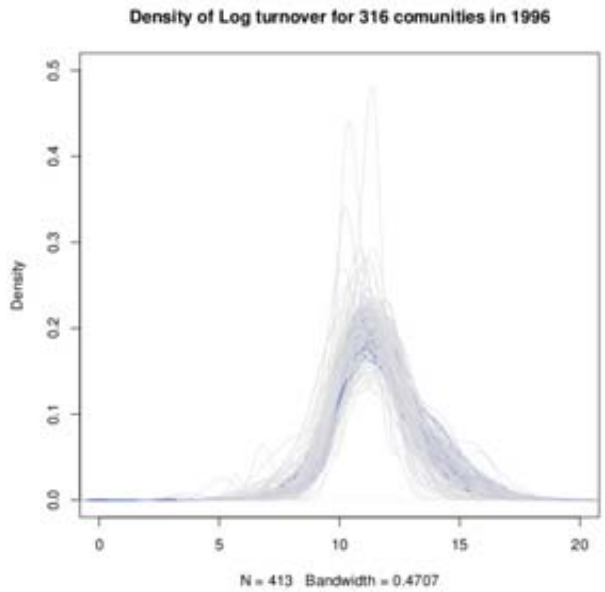
Requirements for a synthetic business dataset and methods for generating such a dataset

The most important requirement is that the synthetic data should be as close as possible to the real data. The idea to use synthetic micro data for simulation purposes is already known for while (see for example Rubin, 1993). In his work he also addresses confidentiality issues connected with the release of publicly available micro data. Rubin uses multiple imputation to fulfill the requirement of a synthetic data set which is as close as possible to reality. This approach has been first applied for population micro data. Drechsler et al. (2008) used a multiple imputation approach to create a synthetic business data set. His proceeding aims at providing micro data to researchers. In the present case the focus lies on the generation of evaluation datasets. The requirements for a universe of business entities which can be used for evaluation purposes are manifold. Logical constraints play an important

role. To list these constraints much more special knowledge is required as for the preparation of a list of logical constraints for a synthetic population of persons. Some patterns may occur for enterprises of one sector while they are completely excluded for other types of enterprises. The methodology underlying the generation procedure of the DACSEIS data set is an important starting point for the generation of the planned synthetic universe. This methodology is described by Münnich and Schürle (2003). For the generation of the present synthetic universe heterogeneities and the spatial structure are of particular importance. As for the generation of a synthetic dataset for persons it is also important to reproduce the most essential population characteristics for enterprises. The requirements for a synthetic population data set are described in Alfons et al. (2011a) and further developed in Burgard et al. (2011). A computing framework for the generation of populations was developed by Kraft (2009) and Alfons et al. (2011b). Methods to control for disclosure control can be found in Polettini (2003) as well as in Templ and Alfons (2010).

Peculiarities of the basic data set

As stated before the analysis of the starting point is an important first step for the generation of a synthetic data set. Hidiroglou and Laniel (2001) stated that the existence of an accurate and up-to-date business register is worthwhile. In the case at hand, the register used is the *Archivio Statistico Imprese Attive* (ASIA) data set which is available for the period between 1996 and 2008. A first analysis showed that the challenges described above are also relevant for this study. The median of the number of employees amounts 8, whereas the mean of the same number amounts 58 100. The comparison of these two statistics shows, that a heavily skewed distribution with many outliers exist, for this variable. In the figure below, it can be seen that huge discrepancies differences exist in log turnover for the different communities.



For the generation of a synthetic enterprise universe three different datasets are available. In the ASIA data set the ATECO classification is used, which classifies the economic activity. This classification is the national version of the European nomenclature NACE. Especially the cross classification of these economic classes with the regional identification is of great importance for the analysis of the Italian business. For the simulation the two types of subdivision to domains of interest are proposed, the enterprises are allocated to ATECO 5 areas in the first case, whereas they are distributed to domains per ATECO 4 and regions in the second case.

The planned simulation study

Three priorities have been declared in the Europe 2020 strategy. This is smart as well as sustainable and inclusive growth. At least for the evaluation of regional cohesion (the third priority), data on regional level is necessary. In the present simulation study which is planned as design based study, the average revenue per employee in the small domains is planned as the predictand. The estimators used in the simulation study will be described in brief in the following.

An often encountered problem in statistics is to obtain reliable estimates for the statistics of interest in cases of small sample sizes. The usage of traditional direct estimators may lead to problems in these situations because their variances tend to be rather high. This may cause difficulties in real world applications, where we try to estimate an unknown quantity with only one sample at hand. Small area estimation techniques may produce better results in such cases as they allow for *borrowing strength*, i.e. they increase the effective sample size by exploiting statistical relationships between the quantities of interest and available auxiliary information (cf. Rao 2003).

The generalized regression estimator (GREG) is a design-based estimator:

$$\hat{\mu}_{i,GREG} = \frac{1}{N_i}\left[\sum_{j\in U}\hat{y}_{ij} + \sum_{j\in s} w_{ij}(y_{ij}-\hat{y}_{ij})\right],$$

where $\hat{y}_{ij}$ denotes the fitted values of the underlying regression model. Thus, the GREG estimator of the population mean in area i is the sum of the fitted values of the population in area i corrected by sum of the weighted resdiuals of the sample in this area. GREG estimators can be constructed for various situations (see Lehtonen and Veijanen 2009), but we shall restrict our attention to a linear fixed effects model. Hence, in our case $\hat{y}_{ij} = \mathbf{x}_{ij}^T\hat{\beta}$, where $\hat{\beta}$ denotes the vector of regression parameters and $\mathbf{x}_{ij}$ as the vector of auxiliary information for person j in area i. Assuming access to auxiliary information for all elements in the universe, we may use a unit-level mixed model to estimate the small area means of interest. This approach was developed by Battese et al. (1988) (BHF) and their estimator is given by

$$\hat{\mu}_{i,BHF} = \bar{\mathbf{X}}_i^T\hat{\beta}_{GLS} + \hat{u}_i,$$

with $\bar{\mathbf{X}}_i = N_i^{-1}\sum_{j\in U}\mathbf{x}_{ij}$. Furthermore, $\hat{\beta}_{GLS}$ denotes the generalized least squares estimator of the regression parameters in the unit-level mixed model and $\hat{u}_i$ indicates the EBLUP estimator of the area-specific random effect in that model. In contrast to the GREG estimator, the BHF estimator is model-based, i.e. unbiased with respect to the underlying model, but it is not design consistent unless the sampling design is self-weighting. Thus, the BHF-estimator might encounter difficulties if a complex sampling design is used, which is often the case in business statistics. The fact that design consistency is usually considered a desirable property led to the derivation of model-based estimators, which incorporate weights into unit-level mixed models. An example is the estimator proposed by You and Rao (2002) (YR), who transformed the unit-level mixed model to a survey-weighted area-level model to obtain a design consistent estimator satisfying the benchmarking property (cf. Rao 2003). Their pseudo-EBLUP estimator is defined as

$$\hat{\mu}_{i,YR} = \bar{\mathbf{X}}_i^T\hat{\beta}_w + \hat{u}_{i,w},$$

where $\hat{\beta}_w$ denotes a design-weighted estimator of β and $\hat{u}_{i,w}$ refers to the EBLUP estimator of the area-specific random effect, both with respect to the survey-weighted area-level model, described in You and Rao (2002). The assumption of access to unit-level information for all elements in the population could be too restrictive in reality, which is why models that only require access to auxiliary variables on area-level are highly useful. A small area estimator under an area-level model was derived by Fay and Herriot (1979) (FH) and is defined as follows

$$\hat{\mu}_{i,FH} = \bar{\mathbf{X}}_i^T\hat{\beta}_{FH} + \hat{u}_{i,FH},$$

where $\hat{\beta}_{FH}$ and $\hat{u}_{i,FH}$ denote the estimators for the regression parameters and random effects under an area-level model. Besides requiring less detailed information about the auxiliary variables, the FH-estimator is also attractive in terms of the computational effort because the matrices to be inverted are of smaller dimension compared to unit-level estimators. Of course it is not guaranteed that one synthetic universe completely meets the real situation. Therefore, it is interesting to compare the results for different scenario variables. Scenarios are an important tool to get an impression which results are within the bounds of possibility and should always be included in simulation studies.

Conclusion

The target of this work was it to give a brief overview on the reasons why business statistics are so important for the European 2020 growth strategy. Further the aim was to present the challenges of simulations with business data, the requirements for a synthetic data set and the methods of generating a synthetic data set. The analysis of the datasets provided by ISTAT shows that the challenges presented at the beginning of the paper are also important for the case at hand. A simulation within the context of business statistics has to deal with extreme situations. The target of the approach in this paper is to generate a dataset which is freely available and which can be used by other researchers within an open research philosophy. In analogy to the AMELIA data set, it is the target to provide synthetic datasets on a website, which are completely safe from a disclosure control viewpoint. The results of the generation of a synthetic population and the results of the simulation study will be presented on the ISI conference in Dublin.

Acknowledgements

The research is done within the BLUE-ETS (http://www.BLUE-ETS.eu) research project which is financially supported within the seventh Framework Programme by the European Commission. The authors are grateful to ISTAT the BLUE-ETS project coordinator for the kind provision of business data and the very inspiring discussion on the data provided.

References

Alfons, A., **Filzmoser, P.**, **Hulliger, B.**, **Kolb, J.-P.**, **Kraft, S.**, **Münnich, R.** and **Templ, M.** (**2011**a): *Synthetic data generation of SILC data*. Deliverable 6.2, AMELI project.
URL http://ameli.surveystatistics.net

Alfons, A., **Kraft, S.**, **Templ, M.** and **Filzmoser, P.** (**2011**b): *Simulation of close-to-reality population data for household surveys with application to EU-SILC*. Statistical Methods & Applications, accepted for publication.

Battese, G. E., **Harter, R. M.** and **Fuller, W. E.** (**1988**): *An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data*. Journal of the American Statistical Association, 83, pp. 28–36.

Burgard, P., **Kolb, J.-P.** and **Münnich, R.** (**2011**): *Generation of Synthetic Universes for Micro-Simulations in Survey Statistics*. In submission.

Drechsler, J., **Bender, S.** and **Rässler, S.** (**2008**): *Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel*. Transactions on Data Privacy, 1 (3), pp. 105–130.

Fay, R. E. and **Herriot, R. A.** (**1979**): *Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data*. Journal of the American Statistical Association, 74, pp. 269–277.

Gelman, A. (**2007**): *Struggles with Survey Weighting and Regression Modeling*. Statistical Science, 22(2), pp. 153–164.

Hidiroglou, M. A. and **Laniel, N.** (**2001**): *Sampling and Estimation Issues for Annual and Sub-annual Canadian Business Surveys*. International statistical review, 69, pp. 487–504.

Kraft, S. (**2009**): Simulation of a population for the European Income and Living Conditions survey. Diploma thesis, Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria.

Lehtonen, R. and **Veijanen, A.** (**2009**): *Design-based methods of estimation for domains and small areas*. **Rao, C.** and **Pfeffermann, D.** (editors) Handbook of Statistics - Sample Surveys: Inference and Analysis, vol. 29b, pp. 219 – 249, Elsevier.

Münnich, R. (**2008**): *Varianzschätzung in komplexen Erhebungen*. Austrian Journal of Statistics, 37, pp. 319–334.

Münnich, R. and **Schürle, J.** (**2003**): *On the simulation of complex universes in the case of applying the German Microcensus*. DACSEIS research paper series No. 4, University of Tübingen.
URL http://w210.ub.uni-tuebingen.de/volltexte/2003/979/

Polettini, S. (**2003**): *Maximum entropy simulation for microdata protection*. Statistics and Computing, 13, pp. 307–320, ISSN 0960-3174, doi:10.1023/A:1025606604377.
URL http://portal.acm.org/citation.cfm?id=941652.941665

Rao, J. N. K. (**2003**): Small Area Estimation. Wiley Series in Survey Methodology. Hoboken, Wiley.

Rubin, D. (**1993**): *Discussion: Statistical disclosure limitation*. Journal of Official Statistics, 9 (2), pp. 461–468.

Templ, M. and **Alfons, A.** (**2010**): *Disclosure risk of synthetic population data with application in the case of EU-SILC*. **Domingo-Ferrer, J.** and **Magkos, E.** (editors) Privacy in Statistical Databases, *Lecture Notes in Computer Science*, vol. 6344, pp. 174–186, Heidelberg: Springer, ISBN 978-3-642-15837-7 Pick It!

You, Y. and **Rao, J. N. K.** (**2002**): *A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights*. Canadian Journal of Statistics, 30, pp. 431–439.