Regional analysis of business surveys: methods and applications in the context of Small Area Statistics

Julia Manecke

Keywords: small area estimation; business surveys; stratum jumpers

1. INTRODUCTION

In recent years, the demand for results of business surveys broken down by region and content has increased significantly. The great interest is, among other things, due to the need for reliable data when making economic decisions or planning regional and structural policy measures. Often, however, the sampling design of a survey is only designed for a reliable design-based estimation at the state or federal level due to maximum permissible sample sizes. If additional estimated values are to be determined for regional or content-related subpopulations, insufficient sample sizes might lead to unacceptably high variances of the estimates. A subpopulation in which the sample size is not large enough for a direct design-based estimation of sufficient precision is called a *small area*. So-called *small area estimation* methods can be used to calculate precise estimates for the respective subpopulations. These mostly model-based approaches rest on the supportive inclusion of additional auxiliary information from other subpopulations using a statistical model.

However, the lack of timeliness of business registers usually used as a sampling frame for the design of business surveys or the time lag between the sampling design and the data collection itself lead to inconsistencies also referred to as *frame errors*. As a result, the variables contained in the business register might be obsolete. In addition, the register and the *target population* differ in terms of their composition and the number of businesses. Nevertheless, the register is a potential source of auxiliary variables for small area estimation methods. This however may create problems, as erroneous or obsolete auxiliary variables may cause small area estimation methods being even worse than classic design-based estimators.

Furthermore, the sampling design of business statistics usually includes a stratification by industry groups and size classes. Due to the lack of topicality of the *frame population* and the strong dynamics of business populations, however, *industry-specific* and *size-specific stratum jumpers* may result. These are businesses that would have been assigned to a different design stratum if the correct design information had been available at the design stage. Accordingly, the assumptions under which the original design weights were determined within the sampling design are no longer applicable.

Building on the challenges of business surveys elaborated above, the objective of this work is to analyse the potential to improve the estimations for small areas in business statistics. In this context, the inconsistencies between the available frame population and the target population referred to as frame errors should be considered in particular.

On the one hand, various small area estimation methods should be implemented and compared with one another with regard to their ability to improve the estimation quality despite outdated auxiliary information. On the other hand, as the assumptions under which the original design weights were determined within the sampling design no longer apply, various reweighting approaches should be developed and evaluated.

The comparison should be made taking into account different frame error scenarios. Therefore, the aim is to examine the extent to which small area estimation approaches using outdated auxiliary information and various reweighting approaches can achieve an improvement compared to the classic design-based *Horvitz-Thompson-Estimator* [1] in various realistic scenarios of stratum jumpers.

2. Methods

In the context of the present work, different approaches for the adjustment of design weights in the context of a reweighting are developed. These approaches can be divided into *case number-based* and *model-based reweighting methods*. In case number-based methods, the design weight is interpreted as the number of units in the population that have similar variable expressions to those of the unit surveyed. Taking into account the design strata, it is therefore tried to estimate the number of similar units as accurately as possible. Within the model-based reweighting methods, however, the relation between the design weights and the variables of interest is quantified using suitable models and used to stabilise the estimate ([2], [3] and [4]).

Besides the reweighting approaches, a literature-based selection of suitable small area estimation approaches is made. In addition to the design-based *Horvitz-Thompson-Estimator* and the *Hájek-Estimator* as well as the model-assisted *Generalised Regression Estimator*, further model-based approaches are to be investigated. These include the prominent area-level *Fay-Herriot-Estimator* [5], the unit-level *Battese-Harter-Fuller-Estimator* [6], the *You-Rao-Estimator* including sampling weights [7], a robust Unit-Level-Model and the *Measurement Error Model* for erroneous covariates [8].

The theoretically sound approaches will be applied in a simulation study and compared with each other. The simulation study is implemented following the annual retail survey of the federal state Hesse [9]. For this purpose, a synthetic data set of the retail companies contained in the Hessian business register 2014 is used. The aim of the simulation study is to estimate the total number of employees subject to social insurance contributions per subpopulation. The subpopulations are composed by combination of the retail industry groups used for stratification and the 26 NUTS 3 regions of the federal state Hesse.

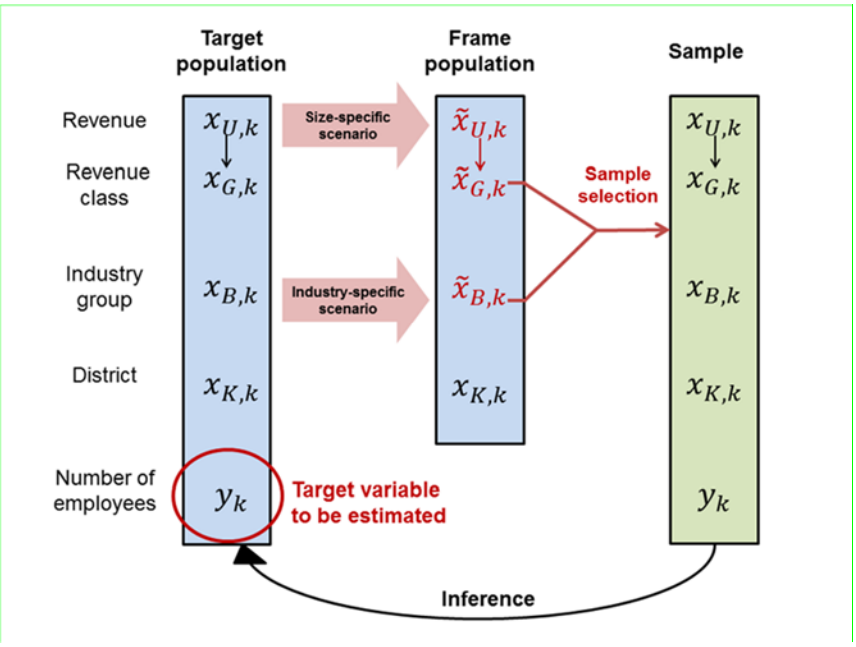

Figure 1. Definition of the frame population and sample selection in the simulation study

The setup of the simulation comprising 2,000 Monte Carlo replications is illustrated in Figure 1. The present data set is defined as the true target population, which is usually not known in reality. Based on this population, in each replication, a frame population that is assumed to be available and known in practice is generated. This frame population contains both size-specific and industry-specific stratum jumpers. These were generated according to a combination of seven different *scenarios of industry-specific stratum jumpers* and six *scenarios of size-specific stratum jumpers*. Finally, the sample is drawn from this frame population using the actual stratum-specific sampling fractions of the Hessian retail survey of the year 2014 in order to replicate the stratification, drawing and estimation process of this survey as realistically as possible. By means of the respective sample, the aim is to estimate the target variable of interest, i.e. the total number of employees, for each subpopulation of interest in the target population. This is carried out using the various small area methods as well as the reweighting approaches. In order to evaluate the potential of each method in the respective scenarios, the *relative bias* and the *relative root mean squared error* (RRMSE) is computed for each combination of estimation and weighting approach, size-specific scenario and industry-specific scenario.

3. RESULTS

When comparing different estimation methods in the simulation study, it becomes obvious that, especially in the case of strong size-specific stratum jump scenarios, the classic unbiased Horvitz-Thompson estimator is highly inefficient in comparison to the other methods investigated. For the model-based small area methods examined, the lack of timeliness of auxiliary variables and the skewed distribution of business data clearly pose a challenge. In particular, the investigated unit-level small area approaches that ignore the design weights of the sampled units show a clear bias in the form of overestimations.

However, the strong influence of outliers on the model parameter estimation can be reduced by logarithmising the target and the auxiliary variables. This adaptation makes the so-called Battese-Harter-Fuller-Estimator, the standard approach for auxiliary variables at individual level, conspicuously robust, even with considerably outdated auxiliary information in the context of strong stratum jump scenarios. By means of the logarithms, the Battese-Harter-Fuller-Estimator convinces throughout with both a relatively low bias and a low estimation variance.

The simulation study additionally shows that keeping the original design weights resulting from the sampling design leads to unbiased estimates even in strong stratum jump scenarios. By contrast, the estimation utilising the investigated case number-based and model-based reweighting approaches is not unbiased. With regard to the estimation variance, however, the estimation using the weights adjusted in a reweighting is able to achieve a significant improvement compared to the estimation based on the original design weights. While the quality of the estimation using the latter decreases in particular in the case of strong size-specific stratum jump scenarios, particularly model-based reweighting approaches appear to be relatively robust against different stratum jump scenarios.

In practice, however, size-specific stratum jumpers are often ignored in the course of a reweighting. This implies in concrete terms that a business that does not change the industry group but which would have to be allocated to a different size class is not defined as a stratum jumper. This is especially problematic if a supposedly small business, which is usually associated with a large design weight according to the sampling design, however has undergone a high increase in size. In this case, the large target variable is erroneously weighted by a large design weight, which might lead to considerable overestimations. Thus, different correction levels have been additionally compared. In the course of this, the simulation results emphasize the importance of not only respecting industry-specific stratum jumpers but also size-specific stratum jumpers when implementing a reweighting. The additional consideration of size-specific stratum jumpers leads to an increase of the bias, but, especially in the case of strong size-specific scenarios, also to a significant reduction of the estimation variance. This likewise leads to a significant reduction in the mean squared error of the estimation.

In official statistics, a high number of *spurious non-response units*, e.g. in the form of extinct businesses or businesses belonging to industries not of interest, has led to a significant underestimation by official business surveys and represents a further general difficulty in the evaluation of these surveys. Since these non-response units are part of the so-called *overcoverage* of the frame population and do not belong to the target population, they are not considered as real non-response and therefore do not contribute to the adjustment of the design weights of the other businesses surveyed. In this context, an alternative approach will be evaluated in which spurious non-response units are treated as real non-response units and compensated by increasing the weights of the remaining units surveyed. Thus, the simplified assumption is made that the number of businesses per stratum belonging to this overcoverage is in equilibrium with the number of businesses being part of the respective *undercoverage* of the frame population, i.e. newly established businesses. This redistribution of the design weights of spurious non-response units leads to a significant reduction of the previous underestimation. However, it also causes an increase in the estimation variance, so that no clear recommendation can be made for the alternative approach.

4. CONCLUSIONS

In conclusion, there is a clear conflict of objectives between the bias and the variance of the estimations both in the analysis of small area methods and in the comparison of reweighting approaches. The classic Horvitz-Thompson-Estimator while retaining the original design weights leads to unbiased estimates even under strong stratum jump structures. However, this unbiasedness is accompanied by an unacceptably high estimation variance. Although reweighting approaches and model-based small area methods lead to a bias of the estimation, in terms of the variance the estimation quality is in a large part significantly higher than the quality of the classic Horvitz-Thompson estimator. Therefore, despite existing inconsistencies and the use of inaccurate auxiliary variables, especially in small areas, a stabilized estimation can be achieved.

REFERENCES

[1] D. G. Horvitz and D. J. Thompson, A generalization of sampling without replacement from a finite universe, Journal of the American statistical Association, 1952, 47(260), 663-685.

[2] J.-F. Beaumont, A new approach to weighting and inference in sample surveys, Biometrika, 2008, 95(3), 539-553.

[3] J.-F. Beaumont and L.-P. Rivest, Dealing with outliers in survey data, Handbook of Statistics, 2009, 29, 247-279.

[4] J. Gershunskaya and M. Sverchkov, On weight smoothing in the current employment statistics survey, Bureau of Labour Statistics, JSM 2014 - Survey Research Methods Section, 2014.

[5] R. E. Fay and R. A. Herriot, Estimates of income for small places: an application of james-stein procedures to census data, Journal of the American Statistical Association, 1979, 74(366a), 269-277.

[6] G. E. Battese and R. M. Harter and W. A. Fuller, An error-components model for prediction of county crop areas using survey and satellite data, Journal of the American Statistical Association, 1988, 83(401), 28-36.

[7] Y. You and J. Rao, A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights, Canadian Journal of Statistics, 2002, 30(3), 431-439.

[8] L. M. Ybarra and S. L. Lohr, Small area estimation when auxiliary information is measured with error, Biometrika, 2008, 95(4), 919-931.

[9] Hessisches Statistisches Landesamt, Strukturdaten des Einzelhandels in Hessen im Jahr 2014, Ergebnisse der Jahreserhebung, Statistische Berichte, 2016.