Impact of Sampling on Small Area Estimation in Business Surveys

Jan Pablo Burgard, Thomas Zimmermann and Ralf T. Münnich
University of Trier, Economic and Social Statistics Department,
Universitätsring 15, 54296 Trier, Germany
Corresponding author: Ralf T. Münnich, e-mail: muennich@uni-trier.de

Abstract

Modern Business statistics often faces the difficulty that an increasing demand of information on sub-levels defined by regions or cross-classifications of variables such as industry classes and business size can be observed, eg for measures of competitiveness by policy makers. In order to enable data producers to provide estimates on those sub-levels, sophisticated stratifications are implemented in the sampling design. These detailed stratifications may produce two difficulties. First, many strata contain only very few elements and, hence, make it difficult to derive optimal sample sizes. Second, statistical model building may suffer from the survey weights derived under these constraints. Additionally, optimization of sampling designs may have a strong impact on the accuracy of different estimation strategies. The aim of the paper is to evaluate different sampling designs in the context of estimation on sub-levels by regions and cross-classifications and their impact on these domain estimates. As estimators of interest, the Horvitz-Thompson- and generalized regression estimator as design-based methods as well as the Battese-Harter-Fuller-, the You-Rao-, and the augmented estimators are considered. The analysis is performed by means of a Monte Carlo study based on Italian business data.

Keywords: optimal sampling design, model-based estimation, design-based estimation

1 Introduction

Starting with papers by Fay and Herriot (1979) and Battese et al. (1988) modern small area estimation techniques gained popularity in data production in several fields of statistics. However, in business statistics the small area estimation techniques were not applicable for a long time because their assumptions are strongly violated by the underlying population of interest. The specific hitch of business statistics is that the distribution of key variables, such as return or turnover are highly skewed with many outliers that need to be taken into account in the estimation process. Classical large sample approximations may be questioned as the assumption of a normally distributed estimate often does not hold even for considerably large sample sizes. Specially, when distributions with fat-tails or extreme outliers are present, the distribution of design-based point and variance estimates, resemble more a multimodal distribution. In small area estimation the assumption of normally distributed model errors and random effects is often violated, in particular when the covariates at hand are not explaining enough of the variation of the dependent variable. Further, for model-based small area estimation the sampling design may impose critical difficulties. This is heavily criticized by Gelman (2007) with the words *weighting is a mess*.

In recent years there are some developments of model-based small area estimation techniques which try to cope with complex survey designs and weighting. These seem attractive for the application in business statistics. Therefore, it is of interest to study the performance of both classical design-based and model-based small area estimator and the model-based small area estimators coping for complex survey designs.

Chapter 2 is devoted to explaining survey designs and estimation techniques for small domain estimation in business statistics. Chapter 3 comprises the set-up for our simulation study which is used to compare different small domain estimators under a variety of survey designs. Chapter 4 gives an outlook on present research of the paper.

2 Design and estimation in business statistics

The sampling designs used in business statistics typically use stratified random sampling at some stage, where the strata are often constructed as cross-classifications of industry classes, enterprise size and a geograpical information. In the following we briefly describe some of the stratified allocations used in our study. A more detailed account on these allocation procedures is given in Bernardini Papalia et al. (2013, Ch. 4.3). A very basic procedure is the equal allocation which allocates the sample size to all the strata, i.e.

$$(1) \quad n_h^{\text{Equal}} = \frac{n}{L}.$$

Note that (1) may lead to highly different sampling fractions in the strata, if the stratum sizes are highly variable. If a constant sampling fraction within the strata is desired and a variation of the sample sizes is accepted, the proportional allocation may be used. It is given by

$$(2) \quad n_h^{\text{Proportional}} = n\,\frac{N_h}{N}.$$

While the equal allocation may be suitable for domain estimation due to a guaranteed sample size in the strata, the proportional allocation may be more efficient for the estimation at national level as it allocates more sample size to the larger strata. Thus, a convex combination of the allocations (1) and (2) as proposed by Costa et al. (2004) may give a reasonable balance between these goals. If the sole purpose of a survey is to produce efficient estimates at the national level and the stratum specific variances are known, the optimal allocation due to Neyman and Tschuprow may be preferred as it minimizes the variance under stratified random sampling. The optimal allocation assuming large N_h so that the finite population correction may be ignored, follows as

$$(3) \quad n_h^{\text{Optimal}} = n\,\frac{N_h\sigma_h}{\sum_{k=1}^{L} N_k\sigma_k},$$

where σ_h refers to the standard deviation in stratum h of our variable of interest. The optimality of allocation (3) for national design-based estimation may come at the expense of reliable estimation on the domain-level as very small sample sizes may

result. Further, the optimal allocation may result in highly different design weights, which may be an issue for model-based small domain estimation strategies. These disadvantages of the optimal allocation are dealt with by introducing box-constraints for the stratum-specific sample sizes into the optimization procedure as proposed by Gabler et al. (2012). They consider the following optimization problem:

$$
\begin{aligned}
&\min_{n_h} \quad ||\boldsymbol{RRMSE}_{<\cdot>}(\hat{\mu})||_2 = \sqrt[2]{\sum_{h=1}^{L} \mathrm{RRMSE}(\hat{\mu}_{<h>})} \\
&\text{s.t.} \quad m_h \leq n_h \leq M_h, \quad h = 1, \ldots, L \\
&\qquad \sum_{h=1}^{L} n_h \leq n,
\end{aligned} \tag{4}
$$

where m_h and M_h denote the lower and upper bound of the sample size in stratum h. In the solution of problem (4), the set of stratum indices is split into three parts. For the first group of strata, the unconstrained optimal sample size n_h^{Optimal} would be smaller than the lower bound m_h. As this violates the condition $m_h \leq n_h$, the resulting sample size is set to the lower bound. For the second group of strata $n_h^{\text{Optimal}} > M_h$, and to fulfil the constraint $n_h \leq M_h$, n_h is set to M_h. In the remaining strata, the optimal allocation (3) is applied using the remaining sample size. In addition to these allocations, also sampling strategies may be considered, where the allocation is determined as to be optimal for a given design-based estimation procedure. Examples for this approach in the context of stratified sampling are the allocations due to Longford (2006) and Choudhry et al. (2012).

Small domain estimators are used to produce estimates of a domain-specific quantity, such as the domain mean or the domain total for a variable of interest. In the following we focus on estimating the domain mean, which is defined as $\mu_d = \frac{1}{N_d} y_{dj}, \quad d = 1, \ldots, D$, where y_{dj} denotes the response of unit j in domain d and N_d refers to the population size in domain d. The different small domain estimators may be derived from a design-based perspective or a model-based perspective. A widely used design-based estimator in survey sampling is the weighted sample mean, which is given by

$$
\hat{\mu}_{d,Direct} = \frac{\sum_{j=1}^{n_d} w_{dj} y_{dj}}{\sum_{j=1}^{n_d} w_{dj}}, \tag{5}
$$

where w_{dj} denotes the design weight of unit j in domain d. Estimator (5) has good design-based properties, but it may not be efficient for the estimation of small domains as it does not use any auxiliary information. The family of generalized regression estimators (GREG) allows to incoporate auxiliary information through an assisting model to reduce the variance compared to the direct estimator (5). In a small domain setting, the general form of the GREG is given by

$$
\hat{\mu}_{d,GREG} = \frac{1}{N_d} \left[\sum_{j=1}^{N_d} \hat{y}_{dj} + \sum_{j=1}^{n_d} w_{dj} \left(y_{dj} - \hat{y}_{dj} \right) \right], \tag{6}
$$

where the $\hat{y}_{dj}$ refer to the predicted values under the assumed assisting model. It can be seen that the sum of the predicted values for all units in a particular domain is corrected by the weighted residuals from all sampled units in that domain. Note that formula (6) accounts for various specifications of the assisting model (cf. Lehtonen and Veijanen, 2009).

In the case of model-based estimators we assume that the following unit-level mixed model holds

$$y_{dj} = \boldsymbol{x}_{dj}^T\boldsymbol{\beta} + u_d + \varepsilon_{dj}, \quad d = 1, \ldots, D,\ j = 1, \ldots, N_d, \tag{7}$$

where $\boldsymbol{x}_{dj}$ refers to the p-dimensional vector of covariates for unit j in domain d, $\boldsymbol{\beta}$ denotes the p-dimensional vector of estimated fixed effects, $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$, $\varepsilon_{dj} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$ and the domain specific random effects u_d are independent from the sampling errors ε_{dj}. An empirical best linear unbiased predictor (EBLUP) under model (7) has been derived by Battese et al. (1988). Their predictor is given by

$$\begin{aligned} \hat{\mu}_d^{ULEBLUP} &= \hat{\gamma}_d \left[\overline{y}_d + (\overline{\boldsymbol{X}}_d - \overline{\boldsymbol{x}}_d)^T\hat{\boldsymbol{\beta}}\right] + (1 - \hat{\gamma}_d)\overline{\boldsymbol{X}}_d^T\hat{\boldsymbol{\beta}} \\ &= \overline{\boldsymbol{X}}_d^T\hat{\boldsymbol{\beta}} + \hat{u}_d. \end{aligned} \tag{8}$$

It can be seen that (8) is a convex combination of the unweighted survey regression estimator $\left[\overline{y}_d + (\overline{\boldsymbol{X}}_d - \overline{\boldsymbol{x}}_d)^T\hat{\boldsymbol{\beta}}\right]$ and the regression synthetic component $\overline{\boldsymbol{X}}_d^T\hat{\boldsymbol{\beta}}$. As the local sample size n_d increases, $\hat{\gamma}_d$ tends to 1 and more weight is attached to the survey regression estimator. For small n_d and small estimated variances of the random effect, $\hat{\sigma}_u^2$, the shrinkage coefficient $\hat{\gamma}_d$ is close to 0, and the estimator tends to its synthetic component. The estimator is not design-consistent unless the sampling design is simple random sampling within the domains. This assumption, however, is violated in most business surveys as discussed above.

An estimator based on model (7) which is nonetheless design-consistent is the pseudo EBLUP due to You and Rao (2002). They transform the unit-level mixed model (7) to a survey-weighted area-level model, where the weights are normalized within each area. The pseudo EBLUP under the survey-weighted area-level model follows as:

$$\begin{aligned} &\hat{\mu}_{d,YR} = \bar{\boldsymbol{X}}_d^T\hat{\boldsymbol{\beta}}_{YR} + \hat{u}_{d,YR} \text{ with} \\ &\hat{u}_{d,YR} = \hat{\gamma}_{dw}\left(\bar{y}_{dw} - \bar{\boldsymbol{x}}_{dw}^T\hat{\boldsymbol{\beta}}_{YR}\right), \end{aligned} \tag{9}$$

where $\hat{\boldsymbol{\beta}}_{YR}$, $\hat{\gamma}_{dw}$, $\bar{y}_{dw}$ and $\bar{\boldsymbol{x}}_{dw}$ are estimated under the survey-weighted area-level model. In addition to being design-consistent, estimator (9) also satisfies the benchmarking property automatically (cf. You and Rao, 2002).

An implicit assumption when deriving estimators (8) and (9) is that the sampling design is non-informative, i.e. that the model which holds for the population holds for the sample as well (cf. Pfeffermann and Sverchkov, 2009). In business surveys, this assumption may be violated as the sampling weights might be related to the variable of interest after conditioning on the covariates. In some situations the bias due to an informative sampling design might be overcome by including the design weights as an additional covariate in the statistical model. Verret et al. (2010) consider using augmented models for the estimators (8) and (9).

3 Outlook on the simulation study

Our design-based simulation study is conducted using the fully synthetic TRItalia dataset described in Bernardini Papalia et al. (2013, Ch. 5). TRItalia focuses on small and medium enterprises from the Italian population of businesses in 2003. Our variable of interest is the mean of labour costs in each domain. The domains are defined as cross-classifications of the first digit of the industry code and the twenty Italian provinces (NUTS 2), yielding $D = 180$ domains. Our sampling design is stratified random sampling, where the strata are defined as cross-classifications of the domain and the four size classes in terms of the number of employees, giving us $L = 720$ strata. As allocations we consider the equal allocation, the proportional allocation, the optimal allocation and box-constraint optimal allocations with Gelman factors (cf. Burgard et al., 2012) constrained to 10 and 50. Due to the fact that about 90% of all enterprises belong to the group with one to five employees, the stratum sizes are extremely variable. In order to facilitate proper variance estimation for design-based methods, we require at least two units to be drawn from each stratum giving rise to large sampling fractions in very small strata as indicated by Figure 1.

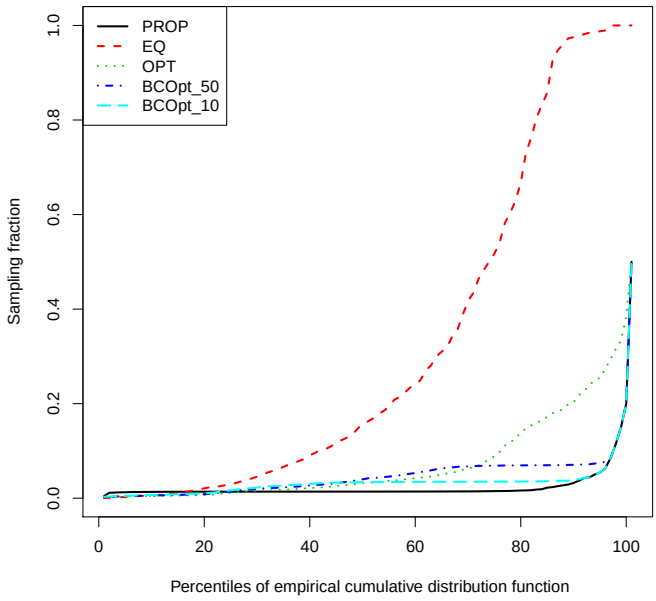

Figure 1: Sampling fractions

Acknowledgements

The research was conducted within the EU-FP7 project BLUE-ETS (`http://www.blue-ets.eu`). The authors express their gratitude to ISTAT for kindly providing the original data which were used to build TRItalia.

References

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

R. Bernardini Papalia, C. Bruch, T. Enderle, S. Falorsi, A. Fasulo, E. Fernandez-Vazquez, M. Ferrante, J. P. Kolb, R. Münnich, S. Pacei, R. Priam, P. Righi, T. Schmid, N. Shlomo, F. Volk, and T. Zimmermann. Best practice recommendations on variance estimation and small area estimation in business surveys. Technical report, BLUE-ETS, deliverable D6.2, 2013.

J. P. Burgard, R. Münnich, and T. Zimmermann. Small area modelling under complex survey designs for business data. In *Proceedings of the Fourth International Conference of Establishment Surveys, June 11 - 14, 2012, Montréal, Canada*, 2012.

G. H. Choudhry, J. N. K. Rao, and M. A. Hidiroglou. On sample allocation for efficient domain estimation. *Survey Methodology*, 38(1):23–29, 2012.

A. Costa, A. Satorra, and E. Ventura. Improving both domain and total area estimation by composition. *Statistics and Operations Research Transactions*, 28(1): 69–86, 2004.

R. E. Fay and R. A. Herriot. Estimation of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74 (366):269–277, 1979.

S. Gabler, M. Ganninger, and R. Münnich. Optimal allocation of the sample size to strata under box constraints. *Metrika*, 75(2):151–161, February 2012.

A. Gelman. Struggles with survey weighting and regression modeling. *Statistical Science*, 22:153–164, 2007.

R. Lehtonen and A. Veijanen. Design-based methods of estimation for domains and small areas. In D. Pfeffermann and C. R. Rao, editors, *Handbook of Statistics*, volume 29B, chapter 31, pages 219 –249. Elsevier, New York, 2009.

N. T. Longford. Sample size calculation for small area estimation. *Survey Methodology*, 32(1):87–96, 2006.

D. Pfeffermann and M. Sverchkov. Inference under informative sampling. In D. Pfeffermann and C. R. Rao, editors, *Handbook of Statistics*, volume 29B, chapter 39, pages 455–487. Elsevier, New York, 2009.

F. Verret, M. A. Hidiroglou, and J. N. K. Rao. Small area estimation under informative sampling. In *Proceedings of the Survey Methods Section SSC Annual Meeting, May 2010*, 2010.

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