Coherent small area estimates for skewed business data

Thomas Zimmermann Ralf Münnich

Abstract

The demand for reliable business statistics at disaggregated levels such as NACE classes increased considerably in recent years. The interest in this topic has been reflected in the BLUE-ETS project, where work package 6 dealt with methods to enhance the quality of business statistics. One of the main areas of research in work package 6 was related to the modelling of business data to produce reliable small domain estimates.

Business data are frequently characterized by skewed distributions, with a few large enterprises which account for the majority of the total for the variable of interest, e.g. turnover. Moreover, the relationship between the variable of interest and the auxiliary variables is typically non-linear on the original scale. Non-linear transformations are commonly applied to the raw data, which remove the skewness and allow for the use of powerful linear models after transformation. This becomes even more complicated when being interested in subgroups In small area estimation, we are typically interested in estimating the mean or the total at the domain level. This may lead to a problem for estimators based on non-linear transformations, since the naïve back-transformation would be negatively biased in the case of the popular log-transformation. In order to overcome this issue, bias correction terms are generally included in estimators based on log-transformations.

Another challenge that NSIs face when using small area estimation techniques for business data, is that the sum of the small area estimates should equal the design-based estimate of the aggregate.. To achieve the coherence of small area and aggregate estimates benchmarking techniques might be applied. The question whether bias-correction and benchmarking should be combined for transformed small area estimators has been raised recently in the context of the BLUE-ETS project.

We consider unit-level and area-level estimators and compare estimators which use both bias-correction and benchmarking to estimators that concentrate on of these techniques. Our analysis is conducted by means of a large-scale design-based simulation study. This study is based on the fully synthetic TRItalia business data set, which was generated as part of the BLUE-ETS project. The research originated within the BLUE-ETS research project financed by the European Commission within FP7 (cf. http://www.blue-ets.eu).

1 Introduction

Business data are often characterized by skewed distributions with important outliers, thereby violating the assumptions present in classical small area models (cf. Fay and Herriot, 1979 or Battese et al., 1988). To address the issue of outliers robust estimators in the spirit of estimators as proposed by Sinha and Rao (2009) might be

employed. See Schmid (2012) for a comprehensive overview on robust small area estimation. Another approach which might recover some of the models assumptions is by the use of transformations. As most small area estimation problems are in some way related to estimating expected values of data, this leads to a problem if the transformations used are nonlinear. In this case, simple back-transformations lead to a bias due to Jensen's inequality as $\mathrm{E}(\log(\boldsymbol{z})) \leq \log(\mathrm{E}(\boldsymbol{z}))$ for any random variable $\boldsymbol{z}$. This bias may be reduced by using bias-correction techniques, some of them are compared in Chambers and Dorfman (2003). Employing bias-reduction techniques, however, may be accompanied by a loss in efficiency owing to a higher variation of these estimates. A thorough investigation of the behaviour of small area estimates using transformations for social statistics is given in Lehtonen et al. (2011). Shlomo and Priam (2013) discuss various benchmarking approaches for log-transformed estimators. In the following section we briefly introduce the small area estimators based on log-transformations used in our simulation study. Section 3 outlines our simulation study.

2 Model-based Estimators

Our exposition of model-based small area estimators based on log- transformations follows Zimmermann and Münnich (2013) closely. As a starting point we consider the unit-level mixed due to Battese et al. (1988)

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

where u_d denotes the domain-specific random effect, $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$, $\varepsilon_{dj} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$ refers to the individual error term and independence between u_d and ε_{dj} is assumed. Model (1) is commonly used in small domain estimation, but the assumption of a linear relationship between the variable of interest $\boldsymbol{y}$ and the covariates $\boldsymbol{x}$ might be to restrictive in the case of business surveys (cf. Chandra and Chambers, 2011). Frequently, the support of the dependent variable such as revenues or labour costs is strictly positive and a log-transformation can be applied on the variable of interest. In many cases this transformed dependent variable can be linked to a set of auxiliary variables by means of a linear model. Karlberg (2000) developed a bias corrected estimator based on a log-transformation of the dependent variable based. Their estimator is based on a lognormal model and allows for the efficient estimation of national statistics based on right-skewed variables. To allow for the simultaneous estimation for many domains, Berg and Chandra (2012) introduced a unit-level lognormal-mixed model

$$\log(y_{dj}) = \boldsymbol{x}_{dj}^T\boldsymbol{\beta} + u_d + \varepsilon_{dj}, \quad d = 1, \ldots, D,\ j = 1, \ldots, N_d \tag{2}$$

where $\boldsymbol{x}_{dj}$ includes an intercept and the other components of it are appropriately transformed. The assumptions on u_d and ε_{dj} are the same as in model (1).
An empirical best predictor under model (2) minimizing the MSE has been proposed by Berg and Chandra (2012). Their estimator is given by

$$\hat{\theta}_d^{EBLOG} = \frac{1}{N_d}\left(\sum_{j\in s_d} y_{dj} + \sum_{j\notin s_d} \hat{y}_{dj}^{EBLOG}\right) \text{ where} \tag{3}$$

$$\hat{y}_{dj}^{EBLOG} = \exp\left(\boldsymbol{x}_{dj}^T\hat{\boldsymbol{\beta}} + \hat{u}_d + 0.5\hat{\sigma}_\varepsilon^2(\hat{\gamma}_d/n_d + 1)\right). \tag{4}$$

Note that estimator (3) is domain-specific as the predictions for the non-sampled units depend on the estimated domain-specific random effect $\hat{u}_d = \hat{\gamma}_d \left(\bar{l}_d - \overline{\boldsymbol{x}}_d^T \hat{\boldsymbol{\beta}} \right)$ with $\bar{l}_d = \frac{1}{n_d} \sum_{j=1}^{n_d} \log(y_{dj})$. Furthermore, the empirical best predictor is also biased, owing to the nonlinear contribution of the parameter estimates $(\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2)^T$ to the predictions $\hat{y}_{dj}^{EBLOG}$ for the non-sampled units. To reduce this bias we consider applying bias corrections such as the ratio adjustment by sample total technique to the predictions (cf. Chambers and Dorfman (2003)). A derivation of (3) is given by Berg and Chandra (2012). In order to estimate the MSE for estimator (3) the double bootstrap procedure as outlined by Hall and Maiti (2006) may be considered.
An extension of Karlberg's estimator to model (2) was introduced by Berg and Chandra (2012). Their estimator is given by

$$\hat{\theta}_d^{ULSynth} = \frac{1}{N_d} \left(\sum_{j \in s_d} y_{dj} + \sum_{j \notin s_d} \hat{y}_{dj}^{ULSynth} \right) \tag{5}$$

where

$$\hat{y}_{dj}^{ULSynth} = \exp \left(\boldsymbol{x}_{dj}^T \hat{\boldsymbol{\beta}} + 0.5 \left(\hat{\sigma}_u^2 + \hat{\sigma}_e^2 - \boldsymbol{x}_{dj}^T \hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_{dj} - 0.25 \hat{\boldsymbol{V}}(\hat{\sigma}_u^2 + \hat{\sigma}_e^2) \right) \right). \tag{6}$$

This estimator is synthetic in the sense that it does not incorporate any domain-specific random effects. Moreover, it is particularly important since it is closely related to the log-transformed model-calibrated estimator developed by Chandra and Chambers (2011). Their estimator is of the form

$$\hat{\theta}_d^{TrMBD} = \frac{\sum_{j \in s_d} a_{dj} y_{dj}}{N_d}, \tag{7}$$

where a_{dj} refers to the model-based model-calibrated weight for unit j in domain d. This weight is obtained from minimizing the distance between itself and the inverse inclusion probability, obeying two calibration constraints:

$$\sum_{j \in s} a_j = N$$
$$\sum_{j \in s} a_j \hat{y}_j = \sum_{j \in U} \hat{y}_j.$$

The first of these constraints requires the sum of the design weights in the sample to agree with the total population size and the second constraints requires the weighted sum of the fitted values in the sample to agree with the total of the fitted values in the population. Note that both of these constraints are with respect to national values. Since the calculation of the model-based model-calibrated weights requires the computation and storage of large-scale matrices, we do not include estimator (7) in our simulation study.
In many cases, the researcher might not have access to unit-level information or important auxiliary information may be available at aggregate level only. In this case a transformed Fay-Herriot model might be used. An approximately bias-corrected estimator under the area-level lognormal mixed model was introduced by Maiti (2004) and is given by

$$\hat{\theta}_d^{ALLOG} = \exp \left(\overline{\boldsymbol{X}}_d^T \hat{\boldsymbol{\beta}} + \hat{u}_d + 0.5 \hat{\sigma}_u^2 (1 - \hat{\gamma}_d) \right). \tag{8}$$

To estimate the MSE of (8), Maiti (2004) proposes using the jackknife approach due to Jiang et al. (2002). This leads to an MSE estimator of the form

$$\widehat{\mathrm{MSE}}(\hat{\theta}_d^{ALLOG}) = \widehat{M}_{1d}^*(\boldsymbol{\delta}) + \widehat{M}_{2d}^*(\boldsymbol{\delta})$$

where $\widehat{M}_{1d}^*(\boldsymbol{\delta})$ and $\widehat{M}_{2d}^*(\boldsymbol{\delta})$ are defined as in Maiti (2004) and $\boldsymbol{\delta}$ is the vector of model parameters. Alternatively, the mean square error of (8) might be calculated using the second-order correct formulae given by Slud and Maiti (2006).

3 Simulation Study

The aim of our study, which is based on the TRItalia dataset described in Kolb et al. (2013), is to estimate the mean of the labour costs in each domain. We determined the domains as cross-classifications of NUTS 1 and the first digit of the industry classification, which leads to $D = 45$ domains. The population was stratified within each domain according to classified company size in terms of numbers of employees. We compare the performance of estimators based on transformations described in section 2 under different sampling designs by means of a design-based simulation study. The sampling designs considered comprise stratified random sampling using different allocation schemes and unequal probability sampling. The inclusion probabilities for unequal probability sampling were determined by using the number of employees as a size variable. The unequal probability samples were drawn within the strata as drawing the samples for the population as a whole was not feasible.

References

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

Berg, E. and **Chandra, H.** (**2012**): *Small area prediction for a unit level lognormal model.* Federal Committee on Statistical Methodology Research Conference.

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

Chambers, R. and **Dorfman, A. H.** (**2003**): *Transformed variables in survey sampling.* Joint Statistical Meetings - Section on Survey Research Methods.

Chandra, H. and **Chambers, R.** (**2011**): *Small area estimation under transformation to linearity.* Survey Methodology, 37, pp. 39–51.

Fay, R. E. and **Herriot, R. A.** (**1979**): *Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data.* Journal of the American Statistical Association, 74 (366), pp. 269–277.

Hall, P. and **Maiti, T.** (**2006**): *On Parametric Bootstrap Methods for Small Area Prediction.* Journal of the Royal Statistical Society, 68 (2), pp. 221–238.

Jiang, J., **Lahiri, P.** and **Wan, S.-M.** (**2002**): *A unified jackknife theory for empirical best prediction with M-estimation*. The Annals of Statistics, 30, pp. 1782–1810.

Karlberg, F. (**2000**): *Population total prediction under a lognormal superpopulation model*. Metron, LVIII, pp. 53–80.

Kolb, J.-P., **Münnich, R.**, **Volk, F.** and **Zimmermann, T.** (**2013**): *TRItalia dataset*. **Bernardini Papalia, R.**, **Bruch, C.**, **Enderle, T.**, **Falorsi, S.**, **Fasulo, A.**, **Fernandez-Vazquez, E.**, **Ferrante, M.**, **Kolb, J. P.**, **Münnich, R.**, **Pacei, S.**, **Priam, R.**, **Righi, P.**, **Schmid, T.**, **Shlomo, N.**, **Volk, F.** and **Zimmermann, T.** (editors) BLUE-ETS Deliverable D6.2: Best practice recommendations on variance estimation and small area estimation in business surveys, chapter 5, pp. 168 – 188.

Lehtonen, R., **Veijanen, A.**, **Myrskylä, M.** and **Valaste, M.** (**2011**): *Small Area Estimation of Indicators on Poverty and Social Exclusion*. Technical report, AMELI deliverable D2.2.

Maiti, T. (**2004**): *Applying Jackknife Method of Mean Squared Error Prediction Error Estimation in SAIPE*. Statistics in Transition, 6, pp. 685–695.

Schmid, T. (**2012**): Spatial Robust Small Area Estimation applied on Business Data. Ph.D. thesis, University of Trier.

Shlomo, N. and **Priam, R.** (**2013**): *Improving Estimation in Business Surveys*. **Bernardini Papalia, R.**, **Bruch, C.**, **Enderle, T.**, **Falorsi, S.**, **Fasulo, A.**, **Fernandez-Vazquez, E.**, **Ferrante, M.**, **Kolb, J. P.**, **Münnich, R.**, **Pacei, S.**, **Priam, R.**, **Righi, P.**, **Schmid, T.**, **Shlomo, N.**, **Volk, F.** and **Zimmermann, T.** (editors) BLUE-ETS Deliverable D6.2: Best practice recommendations on variance estimation and small area estimation in business surveys, chapter 4.2, pp. 52 – 70.

Sinha, S. K. and **Rao, J. N. K.** (**2009**): *Robust small area estimation*. Canadian Journal of Statistics, 37 (3), pp. 381–399, ISSN 1708-945X.

Slud, E. and **Maiti, T.** (**2006**): *Mean-squared error estimation in transformed Fay–Herriot models*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68 (2), pp. 239–257.

Zimmermann, T. and **Münnich, R.** (**2013**): *Impact of sampling designs on small area methods*. **Bernardini Papalia, R.**, **Bruch, C.**, **Enderle, T.**, **Falorsi, S.**, **Fasulo, A.**, **Fernandez-Vazquez, E.**, **Ferrante, M.**, **Kolb, J. P.**, **Münnich, R.**, **Pacei, S.**, **Priam, R.**, **Righi, P.**, **Schmid, T.**, **Shlomo, N.**, **Volk, F.** and **Zimmermann, T.** (editors) BLUE-ETS Deliverable D6.2: Best practice recommendations on variance estimation and small area estimation in business surveys, chapter 4.3, pp. 71 – 89.