Survey-weighted Unit-Level Small Area Estimation

Jan Pablo Burgard and Patricia Dörr

Abstract For evidence-based regional policy making, geographically differentiated estimates of socio-economic indicators are the basis. However, national surveys are often conducted under a complex sampling design due to diverse reasons. Often small sample sizes result within regions of interest leading to too inefficient classical design-based estimators for policy making. In this case, the methodology of small area estimation (SAE) is applicable. Classical SAE relies on the assumption of a multi-level regression model underlying the population data and presumes the sample design to be non-informative. These assumptions are hard to verify in practice. Under an informative sample design, estimated regression parameters are biased and the model-consistency of SAE gets lost. We correct for the sample informativeness in the parameter estimates, and construct design- and model-consistent estimates for regional indicators. Besides the estimation procedure we also propose a MSE estimator. In a simulation study, we illustrate the necessity of survey weights under the violation of typical SAE assumptions. Furthermore, we show that the proposed method is also applicable to generalized linear mixed model settings, allowing also for non-continuous dependent variables.

Key words: Small area estimation, generalized linear mixed models, survey-weighting

Jan Pablo Burgard
Research Institute for Official and Survey Statistics (RIFOSS), Trier University, Universitätsring 15, D-54295 Trier, e-mail: burgardj@uni-trier.de

Patricia Dörr
Research Institute for Official and Survey Statistics (RIFOSS), Trier University, Universitätsring 15, D-54295 Trier e-mail: doerr@uni-trier.de

1 Introduction

National Statistical Institutes often conduct surveys using a complex survey design, either due to costs or due to optimality considerations at the national level. This may lead to small sample sizes in certain geographic regions for which an estimate can be of interest, though. Small sample sizes lead to high variances of the classical design-based estimators and lead to the tradional set-up of the small area estimation (SAE) framework. In SAE borrowing strength across small domains and thus an increased efficiency is attained using a regression model. The inclusion of a random effects term - whose realization is area-specific - the regression model is called mixed. SAEs are usually composite of such random effects predictions and the realized sample's estimate. However, when the model - that is estimated on the realized sample - does not correspond to the population model for some reason, these procedure returns biased estimates. Complex survey designs or model misspecification may contribute to such non-correspondence between the sample and population model. In general, survey weights contain information about the sampling design and thus, their inclusion in the regression model component can reduce possible model bias.

We consider the case where the mixed model is estimated on the sampling unit, i.e. unit-level SAE in the sense of [2]. Usually, estimation is done using the sample log-likelihood, which requires integration over unit-likelihoods in a given area such that unit-specific weighting is not straight forward (cf, for example [13]). Existing proposals require units sharing one random effect to have the same weight, because the likelihood is expressed as a nested integral and elements within an integral cannot be weighted differently. This implies that the random effects structure reflects the sampling clustering and thus are nested [12], [13]. This is quite restrictive and furthermore, access to sampling stage specific inclusion probabilities is seldom for final data users. Therefore, a more general estimation procedure that allows for crossed and nested random effects and that requires only final survey weights is needed. The Expectation-Maximization (EM) methodology ([5]) that is applicable to mixed effects models in general ([7]) provides a framework that is applicable to these needs. However, as one could also think of dependent variables stemming from other exponential family distributions such as binary or count data, we employ a Monte-Carlo version (Monte-Carlo EM algorithm, MCEM) that replaces the E-step by a Monte-Carlo approximation ([10], [3]). The specific survey-weighting application is outlined in [4].

Section 2 introduces the algorithm and turns on problems of the MC-integration. Section 3 handles consistency considerations and Mean Squared Error (MSE) estimation. Afterwards, a simulation study demonstrates the possible gains of the survey-weighted SAE estimator. The final section discusses possible further research and concludes.

2 Proposed Estimators

2.1 Likelihood Set-up

We use the Monte-Carlo EM-algorithm adapted to survey-weighted Generalized Linear Mixed Models (GLMMs) as proposed in [4]. Here, we give a brief review of the set-up from which the SAE point estimator result. Consider as data generating process (DGP) a GLMM described through

$$\eta_i = \mathbf{x}_i^T \beta + \mathbf{z}_i^T \gamma \tag{1}$$

$$\mu_i = g(\eta_i) \tag{2}$$

$$Y_i \sim F(\mu_i, \varphi) \tag{3}$$

$$G \sim N(\mathbf{0}, \Sigma) \quad , \tag{4}$$

where γ is a realization of the mulitvariate normal random variable G (with normal density function $\phi(\cdot|\sigma)$), $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$ and $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_N)^T$ being matrices of explanatory variables in $\mathbb{R}^{N \times p}$ and $\mathbb{R}^{N \times q}$ respectively, β a superpopulation parameter vector of fixed effects and F a distribution from the exponential family. Consequently, g denotes the inverse link function between expectation μ_i and linear predictor η_i. We consider here the canonical link functions under which

$$\log f(y_i, \eta_i) = \frac{y_i \eta_i - b(\eta_i)}{a(\varphi)} + c(y_i, \varphi) \tag{5}$$

is the logarithmic density function. The covariance matrix Σ only depends on a parameter vector σ of length $q^\star$. Define the vector of model parameters by $\psi = (\beta^T, \sigma^T)^T$. Let $\mathscr{U}$, $|\mathscr{U}| = N$ be an index set and for $i \in \mathscr{U}$, let y_i be a realization of Y_i. Then, the population log-likelihood is

$$\mathscr{L}\mathscr{L}(\mathbf{y}, \psi, \gamma) = \sum_{i \in \mathscr{U}} \log f(y_i|\gamma, \beta) + \log \phi(\gamma|\sigma) \quad , \tag{6}$$

and if the random effect vector γ was known, a natural survey-weighted version of (6) would be the Horvitz-Thompson estimator for totals ([6]). However, this is not the case. But a γ simulated from the correct distribution may serve as a plug-in and then (6) can be estimated through a HT-estimator. Repeating the simulation and averaging yields then the (model) expectation of the HT-estimator (denoted by $\mathrm{E}(\widehat{\mathscr{L}\mathscr{L}_{\mathscr{S}}})$), that is maximized in the M-step. Computational and conceptual difficulties that must be taken into account with that procedure are discussed in [4]. As the EM-algorithm applied on the expectation of the estimand of (6) converges to a saddlepoint of the likelihood and the latter is the weighted sum of strictly convex unit-likelihood: The first summand is a generalized linear model component, whose maximum likelihood (ML) estimator is unique for the canonical link (cf. [16]) and the second summand in (6) is a normal log-density, whose ML is unique, too. The proposed procedure thus converges to the maximum likelihood (ML) estimator of the population likelihood if the survey-weights reflect the selection process into the sample. In contrast, an unweighted sample log-likelihood rather converges to the ML of the *sample's* data generating process, composed of the population DGP and the sample randomization. Thus, the survey weighting may protect against some violations of the SAE assumptions (cf. [14]): A non-ignorable sample design.

2.2 SAE Estimator

Having the estimates of the model components, $\hat{\psi}$, we can define our estimator, a weighted unit-level estimator (wUL) of the finite population mean of variable Y. For the pairwise disjoint subpopulations $\mathscr{U}_d$, $|\mathscr{U}_d| = N_d$ and $\mathscr{U} = \uplus_{d=1}^{D} \mathscr{U}_d$, we propose three alternative estimators for the finite population mean in domain d, $\bar{y}_d = N_d^{-1} \sum_{i \in \mathscr{U}_d} y_i$. The notation $\bar{y}_d$ is chosen in order to differentiate between the expectation under the superpopulation model, μ_d, and the finite population realization, which is the focus here. The first alternative is

$$\hat{\mu}_d^{wUL} = N_d^{-1} \sum_{i \in \mathscr{U}_d} \hat{\mu}_i, \tag{7}$$

which is similar to [14, eq 5.3.7] in the linear case and [14, eq 9.4.20] for binary data. This version does not include any finite population correction, which however, might be negligible in a small area setting where the sampling fractions are rather small. Another version that incorporates the sampled observations y_i is

$$\hat{\mu}_d^{wUL} = N_d^{-1} \left(\sum_{i \in \mathscr{U}_d} \hat{\mu}_i + \sum_{i \in \mathscr{S} \cap \mathscr{U}_d} (y_i - \hat{\mu}_i) \right). \tag{8}$$

In both equations, $\hat{\mu}_i$ is the prediction of individual i's variable Y_i expectation under the estimated vector $\hat{\psi}$ and the mode $\hat{\gamma}$ of the weighted sample likelihood under $\hat{\psi}$

$$\hat{\mu}_i = g(\mathbf{x}_i^T \hat{\beta} + \mathbf{z}_i^T \hat{\gamma}), \tag{9}$$

where the random effects only can be predicted for those areas that have at least one observation in the sample. In the LMM setting, (8) is a generalization of the classical unit-level estimator introduced in [2], which only incorporates a random intercept and does not include any additional survey information. That means, in [2], the survey weights are considered to be equal across areas and units.

Another option for SAE would be a model-assisted version

$$\hat{\mu}_d^{wUL} = N_d^{-1} \left(\sum_{i \in \mathscr{U}_d} \hat{\mu}_i + \sum_{i \in \mathscr{S} \cap \mathscr{U}_d} w_i (y_i - \hat{\mu}_i) \right). \tag{10}$$

Estimator (10) is similar to the Generalized Linear Regression Estimator proposed in [15] and [8] in the logistic regression setting. However, note that the version (10) is based on a GLMM in lieu of GLM and incorporates area-specific information through the random effect prediction $\hat{\gamma}$, which makes this model-assisted estimator especially applicable to SAE settings.

2.3 MSE Estimation

As our proposed estimators are smooth functions of ψ and γ, asymptotic MSE estimation fits into the framework of [11] who even deal with non-smooth SAE estimators for poverty analysis. Therefore, under the (common) regularity conditions given in [11] (that we also partially assume in the previous sections in order to establish design-consistency of the point estimators), an asymptotic MSE of (10) consists of the design variance of the model residuals in the domain under consideration. For the model-based estimators (7) and (8), however, MSE estimation is more difficult. We suggest a first-order Taylor approximation of the predictions $\hat{\mu}_i$ at the population parameters $(\beta^{pop}, \gamma^{\star})^T$ - $\gamma^{\star}$ being the population likelihood mode and then calculating the variance of the linearized predictions. This requires variance estimators for the fixed effects estimates and the random effect predictions. [9] gives a formula on how to approximate the Hessian of the observed data matrix in the EM-algorithm and its application for the fixed effects parameter is also proposed in [3]. A lower bound of the prediction error of $\hat{\gamma}$, on the other hand, is the inverse Hessian of the log-likelihood evaluated at the ML-estimates of $\hat{\beta}$. An approximation of the inverse Hessian is readily available from the specific MCEM-estimation algorithm discussed in [4]. However, note that this suggestion is only a lower bound for MSE estimation and hold only asymptotically.

3 Simulation Study

We present a small simulation study in order to demonstrate the necessity of survey-weighting when the non-informativeness assumption is violated. We therefore generate a fixed population $\mathscr{U}$, $|\mathscr{U}| = 3000$ under the following superpopulation model where the population is made up of 50 pairwise disjoint domains $\mathscr{U} = \dot{\cup}_{d=1}^{50} \mathscr{U}_d$, $|\mathscr{U}_d| = 60$, and $X_1 \sim N(6, 3^2)$ and $X_2 \sim \text{Exp}(3)$. The DGP is

$$\eta_i = 30 - 3x_{1,i} - 8x_{2,i} + \gamma_d, \quad i \in \mathscr{U}_d \tag{11}$$

$$\gamma_d \sim N(0, 2^2) \tag{12}$$

$$\mu_i = g(\eta_i), \qquad g \in \{\text{id}, \text{logit}^{-1}\} \tag{13}$$

$$Y_i \sim \begin{cases} N\left(\mu_i, (2.3)^2\right) \\ \text{Ber}(\mu_i) \end{cases} . \tag{14}$$

We draw $B = 1500$ (equally allocated) stratified samples $\mathscr{S}$ of size $n = 200$ of a finite population generated in this way. Consequently, $\mathscr{S} \cap \mathscr{U}_d = \mathscr{S}_d$ and $|\mathscr{S}_d| = n_d = 4$. This is a relatively easy set-up and if the sampling design is non-informative, the estimation conditions for the BHF ([2]) are optimal.

We contrast a πps design (which is under the correct model specification non-informative) where the inclusion probability π_i for a unit in domain d equals

$$\pi_i = \max\left\{\frac{x_{2,i}}{\sum_{j \in \mathscr{U}_d} x_{2,j}} \cdot n_d, 1\right\} \tag{15}$$

and an informative design where the inclusion probability of unit i in domain d is calculated in three steps:

$$e_i = y_i - \mu_i \tag{16}$$

$$q_i = \begin{cases} 0.1 \text{ if } e_i \text{ is below the } 0.25 \text{ quantile} \\ 0.2 \text{ if } e_i \text{ is between the } 0.25 \text{ and } 0.5 \text{ quantile} \\ 0.4 \text{ if } e_i \text{ is above the } 0.5 \text{ quantile} \end{cases} \tag{17}$$

$$\tilde{\pi}_i = \max\left\{\frac{q_i}{\sum_{j \in \mathscr{U}_d} q_j} \cdot n_d, 1\right\} \quad . \tag{18}$$

As observations with a bigger residual tend thus to be oversampled, the intercept estimator of the unweighted model estimation is expected to be overestimated which in return should yield biased predictions.

We compare the proposed estimator (8) (that has conceptually the closest similarity to the traditional BHF) to the Generalized Regression estimator (GREG), the BHF estimator and another survey-weighted SAE estimator, the You-Rao estimator (YR, cf. [17]). In the case of the binary outcome, we consider the logistic regression estimator (LGREG, cf. [8]), too, and the BHF is estimated like (8), but the underlying regression is a generalized linear mixed model estimated with the R-Package `lme4` ([1]). Due to construction, the GREG and the YR estimate a linear model for the binary outcome, too.

The quality criteria that we assess are the relativeempirical bias and the relative empirical mean squared error (MSE) of an estimator $\hat{\mu}$ over all domains:

$$\text{relBias} = \tfrac{1}{50}\sum_{d=1}^{50} \tfrac{1}{1500}\sum_{b=1}^{1500} \tfrac{\hat{\mu}_{d,b}-\mu_d}{\mu_d} \tag{19}$$

$$\text{relMSE} = \tfrac{1}{50}\sum_{d=1}^{50} \tfrac{1}{1500}\sum_{b=1}^{1500} \left(\tfrac{\hat{\mu}_{d,b}-\mu_d}{\mu_d}\right)^2 \tag{20}$$

Results are listed in table (1). The results under the non-informative design and the gaussian outcome variable are standard: Survey-weights are not needed for consistent estimation and inflate the estimators. Consequently, the BHF is the most efficient estimator with respect to relative mean squared error. However, we note that the loss of efficiency when applying survey weights is low and thus might be recommended as the model assumptions are usually not verifiable. Furthermore, we find that the proposed estimator wML can compete with YR.

Under a binary variable of interest, we find that results are similar to the linear mixed model case and both GREG and YR perform well though they employ a linear regression model. Nonetheless, the LGREG is the less biased estimator and has a lower relative MSE than the GREG.

Table 1 Simulation Results

Sampling design	Outcome Variable	Estimator	relBias	relMSE
Non-informative	Normal	GREG	0.00457	0.10959
		wML	0.00974	0.0248
		YR	0.0212	0.02504
		BHF	0.01264	0.01759
	Binary	GREG	0.00466	0.07273
		LGREG	0.00242	0.02196
		wML	0.00682	0.00456
		YR	0.01076	0.00752
		BHF	0.00647	0.00454
Informative	Normal	GREG	0.00312	0.04289
		wML	0.04555	0.02394
		YR	0.05136	0.02402
		BHF	0.12068	0.03189
	Binary	GREG	0.00594	0.04658
		LGREG	0.00742	0.02256
		wML	0.02744	0.00564
		YR	0.05692	0.01108
		BHF	0.06463	0.01033

Under the informative sampling design, results change remarkably, though. As expected, the unweighted traditional BHF has an unacceptable high relative bias, although the relative MSE may compete with the GREG. In the continuous dependent variable case, wML and YR return comparable results. But when the dependent variable is binary, the suggested method outperforms BHF due to the inclusion of survey weights and the YR due to the GLMM framework employed. Thus, wML gives any of the four presented cases a good balance between bias and variance.

4 Conclusion

In this paper, we propose the use of a GLMM estimation framework that allows the inclusion of unit-specific survey weights in the estimation process in order to protect unit-level based SAE estimators against sampling informativeness. A linearization and/ or residual based MSE estimation of the suggested estimators is discussed. Finally, a simulation study demonstrates the necessity of survey weights in the model estimation step in order to reduce the SAE estimators' bias, both in a continuous variable set-up as well as for a binary variable of interest. We find that the loss in efficiency when survey weights need not be included in the model estimation but are nonetheless, is marginal. In contrast there are important gains when the sampling is informative. To conclude, we would like to note that the informativeness of the survey design is hard to verify in real world application and may also depend on the analyzed variable. Thus, we highly recommend the inclusion of survey weights in SAE analysis.

References

1. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823 (2014)
2. Battese, G. E., Harter, R. M., Fuller, W. A.: An error-components model for prediction of county crop areas using survey and satellite data. J. Am. Stat. Assoc. **83**(401), 28–36 (1988)
3. Booth, J. G., Hobert, J. P.: Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. J. Royal Stat. Soc., Series B (Methodological), **61**(1), 265–285 (1999)
4. Burgard, J.P., Dörr, P.: Survey-weighted Generalized Linear Mixed Models. Research Papers in Economics 2018-01, University of Trier, Department of Economics (2018)
5. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc., Series B (Methodological), 1–38 (1977)
6. Horvitz, D. G., Thompson, D. J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc., **47**(260), 663–685 (1952)
7. Laird, N. M., Ware, J. H.: Random-effects models for longitudinal data. Biometrics, 963–974 (1982)
8. Lehtonen, R., Veijanen, A.: Logistic generalized regression estimators. Surv. Methodol., **24**, 51–56 (1998)
9. Louis, T. A.: Finding the observed information matrix when using the EM algorithm. J. Royal Stat. Soc., Series B (Methodological), 226–233 (1982)
10. McCulloch, C. E.: Maximum likelihood algorithms for generalized linear mixed models. J. Am. Stat. Assoc., **92**(437), 162–170 (1997)
11. Morales, D., del Mar Rueda, M., Esteban, D.: Model-Assisted Estimation of Small Area Poverty Measures: An Application within the Valencia Region in Spain. Soc. Indic. Res., 1–28 (2017)
12. Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., Rasbash, J.: Weighting for unequal selection probabilities in multilevel models. J. Royal Stat. Soc., Series B (Methodological), **60**(1), 34–40 (1998)
13. Rabe–Hesketh, S., Skrondal, A.: Multilevel modelling of complex survey data. J. Royal Stat. Soc., Series A (Statistics in Society), **169**(4), 805–827 (2006)
14. Rao, J. N. K.: Small Area Estimation. Wiley, New York (2003)
15. Rondon, L. M., Vanegas, L. H., Ferraz, C.: Finite population estimation under generalized linear model assistance. Comput. Stat. Data Anal., **56**(3), 680–697 (2012)
16. Wedderburn, R. W. M. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. Biometrika, **63**(1), 27–32 (1976)
17. You, Y., Rao, J. N. K.: A pseudoempirical best linear unbiased prediction approach to small area estimation using survey weights. Can. J. Stat., **30**(3), 431–439 (2002)