STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 541–562

SMALL AREA ESTIMATION FOR SKEWED DATA IN THE PRESENCE OF ZEROES

Forough Karlberg[1]

ABSTRACT

Skewed distributions with representative outliers pose a problem in many surveys. Various small area prediction approaches for skewed data based on transformation models have been proposed. However, in certain applications of those predictors, the fact that the survey data also contain a non-negligible number of zero-valued observations is sometimes dealt with rather crudely, for instance by arbitrarily adding a constant to each value (to allow zeroes to be considered as "positive observations, only smaller", instead of acknowledging their qualitatively different nature).

On the other hand, while a lognormal-logistic model has been proposed (to incorporate skewed distributions as well as zeroes), that model does not include any hierarchical aspects, and is therefore not explicitly adapted to small area prediction.

In this paper, we consolidate the two approaches by extending one of the already established log-transformation mixed small area prediction models to incorporate a logistic component. This allows for the simultaneous, systematic treatment of domain effects, outliers and zero-valued observations in a single framework. We benchmark the resulting model-based predictors (against relevant alternatives) in applications to simulated data as well as empirical data from the Australian Agricultural and Grazing Industries Survey.

Key words: small area estimation, representative outliers, zero-valued observations, lognormal-logistic mixture model.

1. Introduction

1.1. Estimation in the presence of skewed data

It is a well-known fact that survey data frequently are skewed (Huber 1981, Fuller 1991, Barnett and Lewis 1994). Examples include the income (Mincer 1970) and wealth (Huggett 1996) of private individuals as well as many of the variables observed in Business surveys (Chambers 1986, Thorburn 1993, Hidiroglou and

[1] Luxembourg Statistical Services. E-mail: Forough.Karlberg@LuxStat.eu.

Smith 2005, Zimmermann and Münnich 2013, Shlomo and Priam 2013). These extreme values are not erroneous; on the contrary, to take but one example, a large enterprise typically constitutes an important part of the local economy of a municipality – and to treat them as anomalies by merely eliminating them when they are encountered would be erroneous. Such extreme values are to be regarded as representative outliers in the terminology of Chambers (1986). Various methods have been developed to treat the issue of estimation in the presence of such outliers, e.g. by adjusting outlyingness, possibly in connection with determining a boundary (threshold) for the outliers (Searls 1966, Kokic 1998, Hubert and Van der Veeken 2007), as well as some methods with downweighting (Hidiroglou and Srinath 1981, Lee 1995, Sinha and Rao 2009). Historically, there are different approaches used for transforming the data (including important outliers) to linearity (Carroll and Ruppert 1988, Chen and Chen 1996, Chandra and Chambers 2011, Berg and Chandra 2012), with some applications concentrating on the finite population distribution of a survey variable (Royall 1982, Jiang and Lahiri 2006, Salvati *et al.*, 2012). Karlberg (2000a) conducts model-based estimation under a lognormal model and extends it to a lognormal-logistic (Karlberg, 2000b). This has the double advantage of moderating the impact of outliers that are in the sample and, in case no outliers are included, to adjust for their (assumed) presence in the population. However, there are also issues with lognormal models. First, the back-transformation introduces bias which must be corrected for; while technically challenging, this is manageable; bias-correction terms are provided by, e.g. Karlberg (2000b). More importantly, as with all model-based estimation, severe bias could result in case the presumed lognormal model does not hold.

By logical extension, small area estimation involving skewed variables is also a challenge, compounded by the fact that the samples for each domain are smaller, leading to an even higher sensitivity to outliers (Lehtonen *et al.*, 2003). Various methods, some of them including log-transformation of the data, have been proposed (Chambers and Dorfman 2003, Slud and Maiti (2006), Chandra and Chambers 2011, Berg and Chandra 2012, Zimmermann and Münnich 2013).

1.2. The added complexity of zero-valued observations

It is not infrequent to encounter skewed variables that, while considerably right-skewed, also contain a sizeable proportion of zero-valued observations (Lamberta 1992, Chen *et al.*, 2003). Obviously, estimation methods based on logarithmic transformation are no longer directly applicable to such variables. Sometimes, this is addressed by merely adding an arbitrary constant κ (κ=1 being common practice) to the variable (see Young and Young 1975), which then again becomes possible to logarithm. However, this manner of treating zero-valued observations is not unproblematic. First, from a technical point of view, it is hard to argue that the resulting logarithmed variable is normally distributed – it would rather be bimodal, with one mode at $\ln(\kappa)$, and definitely not continuous, with a large number of values assuming the exact same value $\ln(\kappa)$. Moreover, the choice of the constant κ is

arbitrary, with a different choice rendering different results. Finally, and most importantly, it could be argued that a variable assuming the value 0 is something more than a computational problem or a technical nuisance – sample units with zero-valued observations are in fact often qualitatively different from those with positive values. Taking wages as an example, a person with a wage figure of 0 is typically not "gainfully employed but with a salary of 0", but rather unemployed or otherwise out of the labour market. Similarly, a farm with a crop area of 0 does typically not belong to a crop farmer who just happens to not grow any crops, but rather to a farmer focusing on other activities, such as dairy, forestry or livestock.

1.3. Solutions investigated in this paper

The lognormal-logistic model discussed by Karlberg (2000b) seems to be a more appropriate way to address this issue. The estimator associated with that model first fits a logistic model (to deal with the zero-valued observations), and thereafter fits a lognormal model to the positive observations. However, the model in question is not directly designed to accommodate small area estimation. In this paper, we will therefore devote Section 2 to extending the model of Karlberg to incorporate hierarchical elements (or, put differently, extending the model of Berg and Chandra (2012) to incorporate a logistic element). This is achieved by straightforward, practical combinations of already existing tools (see Pfeffermann, 2013); this paper includes no major theoretical contributions. The empirical properties of the four resulting estimators are then examined in Section 3, for random lognormal-logistic data, as well as for data from the Australian Agricultural and Grazing Industries Survey (AAGIS). The findings are discussed in Section 4, which also brings up possible future lines of study.

2. Methods

2.1. The lognormal-logistic model

Under the lognormal-logistic model studied in this paper, we will, just like Karlberg (2000b), assume that Y_{ij} , the value of unit j for area i for the variable of interest(Y), is the product

$$Y_{ij} = \widetilde{Y}_{ij}\Delta_{ij}$$

of a "lognormal component" $\widetilde{Y}_{ij}$ and a binary (0 or 1) "logistic component" Δ_{ij} with independence between the two components.

2.1.1. The lognormal component

Letting X_{ij} denote a vector of auxiliary variables for unit j, we assume that

$$\ln\left(\widetilde{Y}_{ij}\right) = \tilde{Z}_{ij} = \mathbf{B}\mathbf{X}_{ij} + u_i + e_{ij}$$

where **B** is an unknown parameter, and, for the area-level effects, we have that they are i.i.d.

$$u_i \sim N(0, \sigma_u)$$

and for the residuals that they are i.i.d.

$$e_{ij} \sim N(0, \sigma_e)$$

with, furthermore, independence between any u_i and any e_{ij}.

2.1.2. The logistic component

Letting Ξ_{ij} denote a vector of auxiliary variables for unit j (possibly identical X_{ij}), we assume that the logistic component values are conditionally independently Bernoulli distributed:

$$\Delta_{ij} \sim \text{Bernoulli}\left(\frac{\exp\left(\boldsymbol{\beta}\boldsymbol{\Xi}_{ij} + \omega_i\right)}{1 + \exp\left(\boldsymbol{\beta}\boldsymbol{\Xi}_{ij} + \omega_i\right)}\right)$$

where **β** is an unknown parameter and the area-level effects are i.i.d.

$$\omega_i \sim N(0, \sigma_\omega).$$

2.1.3. Relationship with previous models

We see from the first column of Table 1 that estimators for unit-level lognormal models (without a logistic component) have been defined without area effects by Karlberg (2000a) and with area effect by Berg and Chandra (2012). From the two other columns (with stochastic Δ_{ij}), we see, however, that to date, only the simplest case (i.e. with no hierarchical components) has been treated; this corresponds to Karlberg (2000b).

In this paper, we will therefore proceed to investigate lognormal-logistic estimators of small area means corresponding to all four possible cases.

Table 1. Relationship between the model parameters and previously addressed models

	$\Delta_{ij} \equiv 1$ *(i.e. no logistic component)*	Δ_{ij} stochastic	
		$\sigma_\omega = 0$	$\sigma_\omega > 0$
$\sigma_u = 0$	Karlberg (2000a)	Karlberg (2000b)	–
$\sigma_u > 0$	Berg and Chandra (2012)	–	–

2.2. Fitting the model and estimation of small area means

2.2.1. Estimation of the model parameters and fitted area effects

In order to evaluate the various estimators, a simulation study has been conducted. Due to the availability of appropriate SAE packages in R, the study was set up through a couple of R scripts. For all four possible options, the estimation procedure proposed in this paper is as follows:

1. First, the logistic model parameters are estimated. Two cases are possible:
   a. If there is no logistic area effect (i.e. if $\sigma_\omega=0$), the logistic parameter $\boldsymbol{\beta}$ is estimated by means of logistic regression via the GLM function.
   b. If $\sigma_\omega > 0$, the parameters $\boldsymbol{\beta}$ and σ_ω are estimated (and the ω_i-values are fitted) using hierarchical logistic regression via the HGLM function (Rönnegård *et al*., 2010).
2. Based on the logistic regression outcome:
   a. Estimated probabilities are computed for each unit as

$$\hat{p}_{ij} = \frac{\exp\left(\hat{\boldsymbol{\beta}}\boldsymbol{\Xi}_{ij} + \hat{\omega}_i\right)}{1 + \exp\left(\hat{\boldsymbol{\beta}}\boldsymbol{\Xi}_{ij} + \hat{\omega}_i\right)},$$

   b. area frequencies with positive Y_{ij} values are estimated by

$$\hat{N}_{+i} = \sum_{j \in s_i} \Delta_{ij} + \sum_{j \in r_i} \hat{p}_{ij}\,, \text{ and}$$

   c. area auxiliary variable averages for the observations with positive Y_{ij} values are estimated by

$$\hat{\bar{\mathbf{X}}}_{+i} = \left(\sum_{j \in s_i} \Delta_{ij}\mathbf{X}_{ij} + \sum_{j \in r_i} \hat{p}_{ij}\mathbf{X}_{ij}\right)/\hat{N}_{+i}\,.$$

3. Thereafter, the lognormal model parameters are estimated.
   a. If there is no lognormal area effect (i.e. if $\sigma_u=0$), $\mathbf{B}$ and σ_e are fitted as in Karlberg (2000b).

b. If $\sigma_u > 0$, the parameters **B**, σ_u and σ_e are estimated (and the u_i-values are fitted) as in Battese, Harter and Fuller (1988) using the eblupBHF function (Molina and Marhuenda, 2013), i.e. the empirical best linear unbiased predictor (EBLUP; see Rao 2003, and Wang and Fuller 2003).

2.2.2. Prediction of unobserved values

If there is no lognormal area effect, then the lognormal component of each unobserved value is predicted, as in Karlberg (2000b) by the back-transformed predicted values of Z_{ij} multiplied by a bias correction factor:

$$\hat{\tilde{Y}}_{ij} = \exp\left(\hat{\tilde{Z}}_{ij}\right)\exp\left(\frac{\hat{\sigma}_e^2}{2}\left(1 - a_{ij}\right) + \frac{\hat{\sigma}_e^4}{4n_+}\right)$$

where n_+ is the number of positive observations in the sample (obtained as the sum of all observed values of Δ_{ij}),

$$a_{ij} = \mathbf{X}_i'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_j$$

and

$$\hat{\tilde{Z}}_{ij} = \hat{\mathbf{B}}\mathbf{X}_{ij} .$$

If the model incorporates lognormal area effects, then the lognormal components are instead predicted, as in Berg and Chandra (2012), by

$$\hat{\tilde{Y}}_{ij} = \exp\left(\hat{\tilde{Z}}_{ij}\right)\exp\left(\frac{\hat{\sigma}_e^2}{2}\left(\frac{\gamma_i}{n_{+i}} + 1\right)\right)$$

where the number of positive observations in area *i* is denoted by

$$n_{+i} = \sum_{j \in s_i} \Delta_{ij},$$

$$\gamma_i = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_{+i}),$$

and

$$\hat{\tilde{Z}}_{ij} = \hat{\mathbf{B}}\mathbf{X}_{ij} + \hat{u}_i .$$

Combining this with the logistic probability estimates, each unobserved value is predicted by

$$\hat{Y}_{ij} = \hat{\tilde{Y}}_{ij}\hat{p}_{ij} .$$

2.2.3. Estimation of small area means

Finally, based on the sum of the observed and predicted values, the small area means are simply estimated by:

$$\hat{\bar{Y}}_i = \frac{1}{n_i}\left(\sum_{j\in s_i} Y_{ij} + \sum_{j\in r_i} \hat{Y}_{ij}\right).$$

To distinguish between the four possible lognormal-logistic (LL) estimators, subscripts based on the hierarchical components are used, as indicated in Table 2.

Table 2. The four lognormal-logistic small area estimators obtained by combining the dispersion parameter models

Lognormal \ Logistic	$\sigma_\omega = 0$	$\sigma_\omega > 0$
$\sigma_u = 0$	$\hat{\bar{Y}}_i^{LL_{00}}$	$\hat{\bar{Y}}_i^{LL_{0\omega}}$
$\sigma_u > 0$	$\hat{\bar{Y}}_i^{LL_{u0}}$	$\hat{\bar{Y}}_i^{LL_{u\omega}}$

Letting $\hat{T}$ denote the population total estimator of Karlberg (2000b), we have that

$$\hat{T} = \sum_{i=1}^{a} N_i \hat{\bar{Y}}_i^{LL_{00}}$$

where a is the number of areas. As the exact same model is used, the variance estimator of Karlberg (2000b) is easily applicable to $\hat{\bar{Y}}_i^{LL_{00}}$.

3. Empirical evaluation of estimator properties

3.1. Estimators evaluated and benchmark estimators

The lognormal-logistic estimators of small area means have been evaluated against estimators based on the raw (unlogarithmed) Y_{ij} values. For real survey data, we used

(i) the direct estimator $\hat{\bar{Y}}_i^{DIR}$, as implemented in the SAE package (Molina and Marhuenda, 2013)

(ii) the synthetic unit-level regression estimator $\hat{\bar{Y}}_i^{REG}$ (thus without area effect), used for benchmarking purposes by Karlberg (2000b) and

(iii) the Battese, Harter, Fuller estimator (1988) $\hat{\bar{Y}}_i^{BHF}$ as implemented in the said SAE package.

For random data, we limited the set of benchmark estimators to (ii) and (iii), since there was no model misspecification for the lognormal-logistic estimators rendering the direct estimator superior in terms of unbiasedness. Since there are two sets of auxiliary information $\boldsymbol{\Xi}$ and **X** used by the lognormal-logistic estimators, we used the union of those matrices as auxiliary information for the benchmark estimators (ii) and (iii) using auxiliary information.

3.2. Stochastic data

3.2.1. Lognormal-logistic parameters

There are numerous ways to vary the ways in which stochastic data are generated. In this simulation study, we fixed most parameters, in essence only varying the small area sample size n_i and, directly or indirectly, the dispersion parameters of the two types of area-level effects (u_i and ω_i).

First, we limited the study to lognormal-logistic data, saving the investigation of possible model misspecification to the simulation study related to real survey data. In terms of size, we used only a=20 small areas, and fixed the ratio between small area (population) size and small area sample size to N_i/n_i=20, and also imposed the restriction that n_i be the same across all of the a areas. Considering the essence of auxiliary variables being sufficiently captured by one auxiliary variable for the purposes of this simulation study, we limited the $\boldsymbol{\Xi}$ and **X** matrices to contain (in addition to the requisite intercept dummies) a sole auxiliary variable each. We set these variables to be i.i.d. normal distributed, i.e. $\Xi_{1ij} \sim N(0,1)$ and $X_{1ij} \sim N(0,1)$ (thus having zero correlation between the two auxiliary variables; $\rho_{X\Xi} = 0$).

We invariably used the logistic regression parameter $\boldsymbol{\beta}$=(1,1); with the logistic intercept parameter β_0 thus equal to 1, the resulting number of non-zero Y_{ij} values is roughly equal to e/(1+e) ≈ ¾. We thus have roughly ¼ zero-valued observations in the population. We used the lognormal regression parameter **B**=(0,1) throughout.

3.2.2. Simulation study

With most parameters fixed, we tried out the Cartesian product of the following free parameters:

- We used two different area sample sizes n_i=20 and n_i=5.
- With the overall variance in the lognormal component fixed at

$$\sigma_{\bullet}^2 = \sigma_{\mathrm{u}}^2 + \sigma_{\mathrm{e}}^2 = 1,$$

we varied the area effect proportion

$$\mathrm{p}_{\sigma} = \sigma_{\mathrm{u}}^2 / \sigma_{\bullet}^2$$

in small increments from 0 to 0.2.

- We varied the logistic area effect standard deviation σ_ω in small increments from 0 to 1.5.

For each parameter combination, we generated K=100 random populations and drew a single stratified random sample from each of them. (However, if any sample with no positive observations at all for an entire area, i.e. where any n_{+i}=0, was encountered, the population was regenerated, and the sample was redrawn.) The three benchmark and four lognormal-logistic estimators were then used to estimate the small area averages, and for each area i and replicate k, the relative bias of the estimator EST was calculated as

$$\mathrm{RB}_{\mathrm{i(k)}}^{\mathrm{EST}} = \left(\widehat{\overline{\mathrm{Y}}}_{\mathrm{i(k)}}^{\mathrm{EST}} - \overline{\mathrm{Y}}_{\mathrm{i(k)}}\right)/\overline{\mathrm{Y}}_{\mathrm{i(k)}}$$

and the relative MSE of EST was obtained as

$$\mathrm{RMSE}_{\mathrm{i(k)}}^{\mathrm{EST}} = \left(\mathrm{RB}_{\mathrm{i(k)}}^{\mathrm{EST}}\right)^2.$$

Thereafter, in view of the fact that with the stochastic data, the small areas are interchangeable, the overall relative bias of the estimator EST is obtained by averaging $RB_{i(k)}$ across all areas as well as across all replicates as:

$$\mathrm{RB}_{\mathrm{EST}} = \frac{1}{\mathrm{aK}}\sum_{\mathrm{k=1}}^{\mathrm{K}}\sum_{\mathrm{i=1}}^{\mathrm{a}}\mathrm{RB}_{\mathrm{i(k)}}^{\mathrm{EST}}$$

and the overall relative root mean squared error is obtained as:

$$\mathrm{RRMSE}_{\mathrm{EST}} = \sqrt{\frac{1}{\mathrm{aK}}\sum_{\mathrm{k=1}}^{\mathrm{K}}\sum_{\mathrm{i=1}}^{\mathrm{a}}\mathrm{RMSE}_{\mathrm{i(k)}}^{\mathrm{EST}}}\ .$$

The relative efficiency of an estimator EST w.r.t. a benchmark estimator BNCH, can then be obtained as

$$\mathrm{RE}_{\mathrm{BNCH}}^{\mathrm{EST}} = \mathrm{RRMSE}_{\mathrm{BNCH}}^2/\mathrm{RRMSE}_{\mathrm{EST}}^2\ .$$

3.2.3. Results

In Figure 1, the observed relative efficiency at an area level sample size n_i=20 for each dispersion parameter combination is illustrated for each estimator/benchmark estimator (columns; orange labels / rows; green labels) pair. In essence, green colour coding indicates superiority w.r.t. the benchmark, and red-orange-yellow patterns indicate various degrees of inferiority. Given the multitude of comparisons that we perform below, we will, for compactness, use the index as a shorthand form to refer to an estimator in running text; for instance, we let LL_{00} denote the estimator

$$\widehat{\overline{\mathrm{Y}}}_{\mathrm{i}}^{\mathrm{LL_{00}}}$$

This largely corresponds to the row and column labels of the figures presenting the results (although the figures use "w" for ω, and have a leading "Y" for the estimators based on unlogarithmed values).

A reasonable conjecture is that there is monotonicity of the true relative efficiency w.r.t. to the dispersion parameters, meaning that if the number of replicate populations was larger, the colour regions would be contiguous. Match-ups where a colour mosaic is displayed are thus an indication of lack of precision in terms of RE estimation. The prevalence of such "mosaics" in Figure 1 thus means that we can only express ourselves in terms of general tendencies regarding the impact of dispersion parameters on the RE of an estimator w.r.t. another estimator. We would have to conduct a simulation study with somewhat more replicates to be able to more precisely define the boundaries at which one estimator becomes more efficient than the benchmark estimator.

However, already the general tendencies observed are quite informative. Starting out with the intra-class comparison among the lognormal-logistic estimators, we see, as expected, that if the logistic area dispersion parameter σ_ω increases (rightwards in each pane), the estimators incorporating σ_ω ($LL_{u\omega}$ and $LL_{0\omega}$) fare better than the corresponding estimators lacking those components (LL_{u0} and LL_{00}, respectively). The pairwise comparisons in question ($LL_{u\omega}$ vs. LL_{u0}; $LL_{0\omega}$ vs. LL_{00}) indicate that this superiority holds already for very small positive values of σ_ω, with the boundary somewhere around σ_ω=0.2. Similarly, an increase in the lognormal area effect proportion (upwards in each pane) renders the estimators incorporating a positive parameter σ_u ($LL_{u\omega}$ and LL_{u0}) more efficient than those that do not ($LL_{0\omega}$ and LL_{00}, respectively). The pairwise comparisons in question ($LL_{u\omega}$ vs. $LL_{0\omega}$; LL_{u0} vs. LL_{00}) indicate that this superiority occurs already at a very modest area effect proportion (the boundary seemingly falling somewhere around p_σ=0.025).

Turning our attention to comparisons with the design-unbiased (*DIR*) and model-based (*REG* and *BHF*) estimators based on raw, untransformed Y_{ij} values, it appears from Figure 1 that the lognormal-logistic estimator incorporating both variants of area-level effects, $LL_{u\omega}$, is more efficient than the estimators based on untransformed data, with the possible exception of situations where both σ_ω and σ_u are very small.

While Figure 1 presents the bottom line, i.e. the relative efficiency, it could also be interesting to explore the relative bias of the various estimators. The results (not shown here) indicate that, as expected, the relative bias of the direct estimator is invariably low regardless of the parameterisation – typically in the range of ±1%. At p_σ=0, as σ_ω increases from 0 to 1.5 the relative bias of the appropriate estimator $LL_{0\omega}$ increases only moderately (from 2% to 6%), whereas the bias of the estimator LL_{00}, which lacks a logistic area component, increases dramatically (from 2% to 30%). At p_σ=0.2 and σ_ω=0, the estimators lacking a lognormal area component have

a relative bias of 20%, compared to a modest relative bias of 5% for those that allow for a positive value of σ_u.

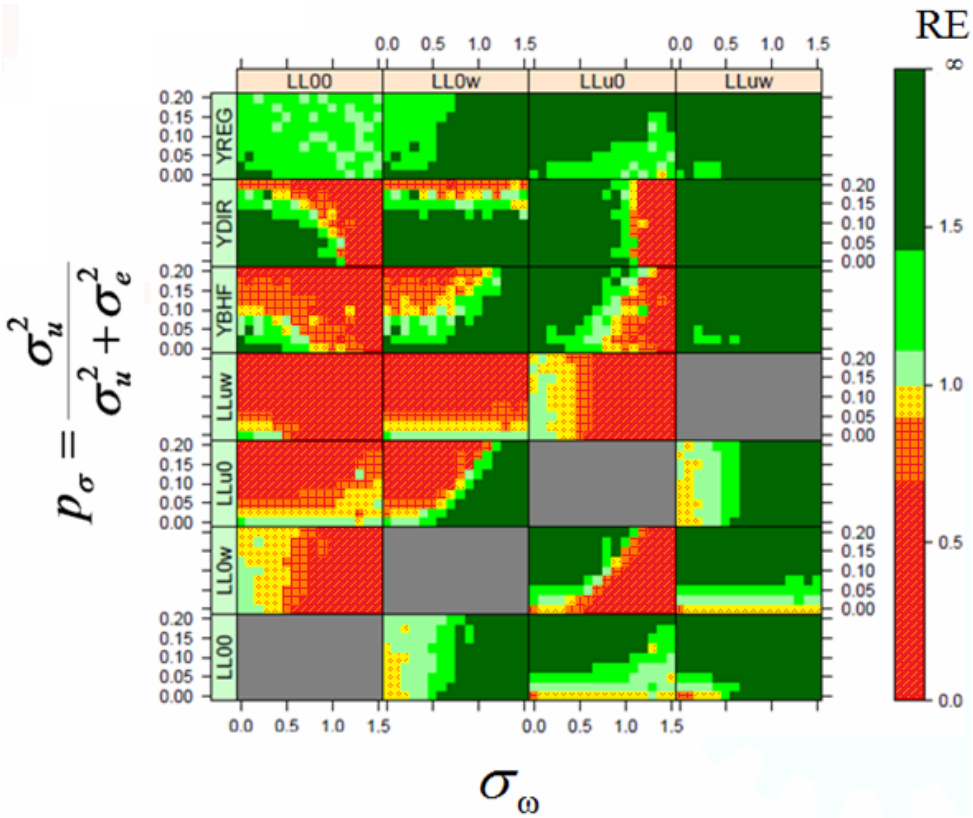

Figure 1. Relative Efficiency (*RE*) of each of the four evaluated estimators (columns) against seven benchmark estimators (rows) for various values of the lognormal-logistic parameters σ_ω and p_σ. Each rectangle corresponds to 100 stratified random samples; each of them drawn from a different lognormal-logistic data set. For each of the a=20 small areas (each with a size N_i=400), the sample size is n_i=20.

Figure 2 summarises the relative estimator efficiencies for random data with area level sample size of n_i=5 (with the sampling proportion remaining the same, the area population size N_i is 5·20=100 here, whereas it was 20·20=400 for the results summarised in Figure 1 above). To summarise the results for that very small sample size, we could say that the same general tendencies hold, but with the area-level dispersion parameter boundaries shifted upwards (to $\sigma_\omega \approx 0.3$ and $p_\sigma \approx 0.075$). However, Figure 2 is much more of a "mosaic" nature. This is due to the far more volatile nature of both numerator and denominator (in turn due to the high volatility of the small area estimators caused by the very low sample sizes for the small areas). A surprising finding is, however, that for very large values of the

logistic dispersion parameter ($\sigma_\omega \approx 1.5$) the direct estimator turns out to be superior to those based on lognormal-logistic models. This might be attributed to the very low number of non-zero observations used to estimate the lognormal distribution parameters and area effects.

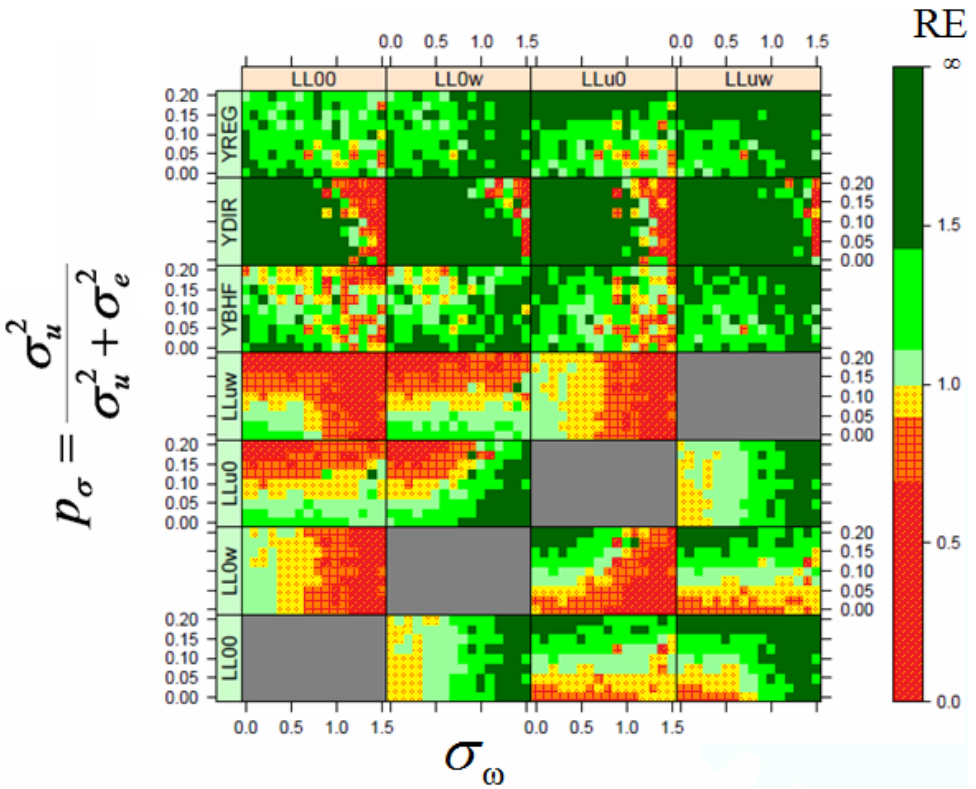

Figure 2. Relative Efficiency (*RE*) of each of the four evaluated estimators (columns) against seven benchmark estimators (rows) for various values of the lognormal-logistic parameters σ_ω and p_σ . Each rectangle corresponds to 100 stratified random samples; each of them drawn from a different lognormal-logistic data set. For each of the a=20 small areas (each with a size N_i=100), the sample size is n_i=5.

3.3. Survey data

3.3.1. The AAGIS data

Like, e.g. Chandra and Chambers (2005) and Chambers and Tzavidis (2006) and Molina (2009), we have applied our lognormal-logistic estimators data obtained from a sample of 1652 farms that participated in the Australian Agricultural and Grazing Industries Survey (AAGIS). This survey includes a number of variables with skewed distributions and a sizeable proportion of 0s,

lending itself well to lognormal-logistic modelling. Moreover, as the data are subdivided into 29 regions (areas), it is also useful for Small Area Estimation. Out of the 1652 observations, we have excluded one with a zero-valued observation for a possible auxiliary variable (to allow us to logarithm it if needed). Some basic characteristics of the variable Beef Cattle are provided in Appendix 1.

The only possible *Y* variable for our class of estimators is Beef Cattle, since the other variables with zero-valued observations have some areas for which there are no observations with positive values at all, rendering estimation with the current implementation of the BHF estimator in the SAE package impossible. (Obviously, this would have to be resolved before such lognormal-logistic estimators are implemented in production.) We have used Farm Area as the auxiliary variable for the logistic component as well as for the lognormal one.

In the simulation study, we have drawn stratified samples (treating the AAGIS data, albeit they are from a sample survey, as a population of size 1651). The only parameter varied has been n_i , for which we have used six different parameterisations, of two different types: (i) the same absolute number across areas (capped at a sample fraction of 50% per area) and (ii) a constant sample fraction per area (with a minimum absolute sample size of 1).

For each parameterisation, we have used 100 replicates. It should be underlined that in contrast to the evaluation of estimator performance for random data (where the areas could be considered interchangeable), the performance measures have been calculated area by area (across all replicates), and not across all small areas. The area-specific relative bias of area *i* is thus obtained as

$$\mathrm{RB}_{\mathrm{EST;i}} = \frac{1}{\mathrm{K}}\sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{RB}_{\mathrm{i(k)}}^{\mathrm{EST}}$$

and the other performance measures are obtained analogously.

3.3.2. Results

As could be seen from Figure 3, the bias is severe for LL_{00} and $LL_{u\omega}$ for certain small areas, with the relative bias sometimes extremely high. With *DIR* unbiased by design, this inevitably carries over into the direct estimator being superior in terms of relative efficiency for such areas, as illustrated by Figure 4. Taking area 1, the area with the smallest number of positive observations (N_{+1}=4) as an example, we have that the relative bias of LL_{00} is around 100, which, in spite of the high variance of *DIR*, carries over a relative efficiency of the direct estimator of approximately 10^4.

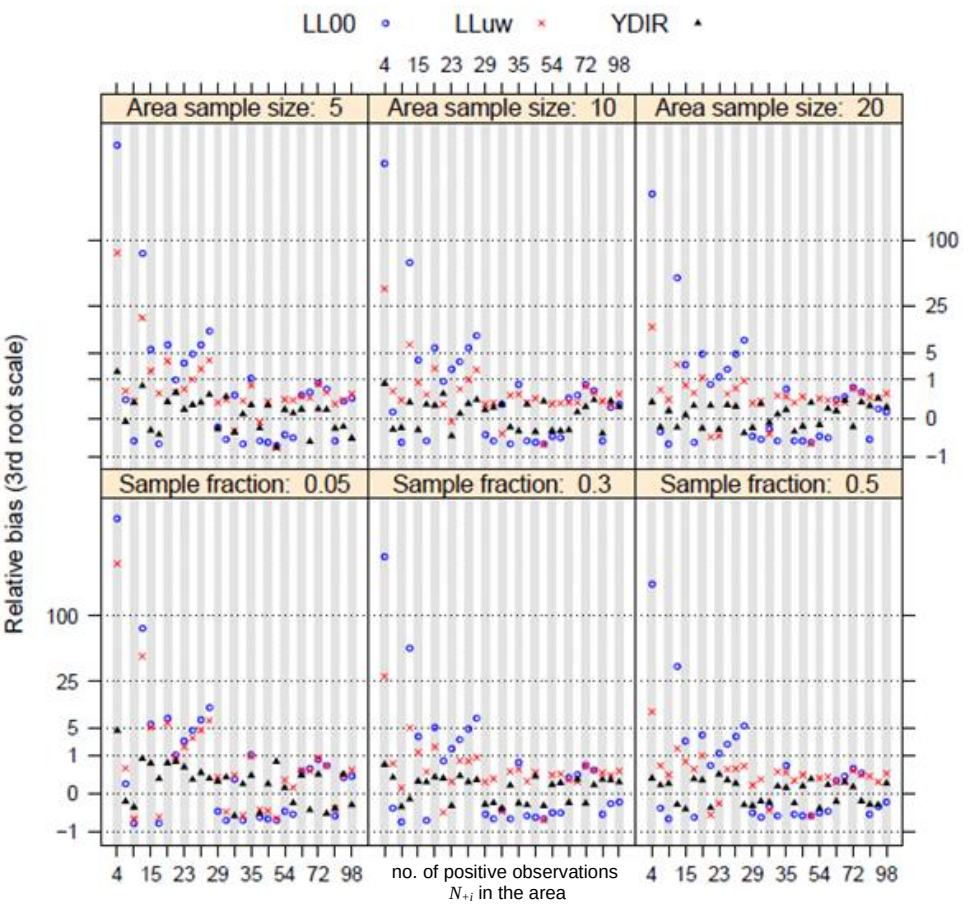

Figure 3. Average relative bias of the *DIR*, LL_{00} and $LL_{u\omega}$ estimators of Beef Cattle area means for various sample sizes. 100 replicates have been used for each sample size parameter.

Owing to these findings, we do not present findings regarding the other benchmark estimators or lognormal-logistic estimators here; if the lognormal-logistic estimators fail to outperform the direct estimators, their performance relative to each other and relative to other benchmark estimators becomes less interesting.

In Appendix 1, the drivers for these tendencies are investigated. In short, as is often the case for small area estimation (Chambers *et al.*, 2014), a model which works reasonably well at population level is found to be inappropriate at the area level.

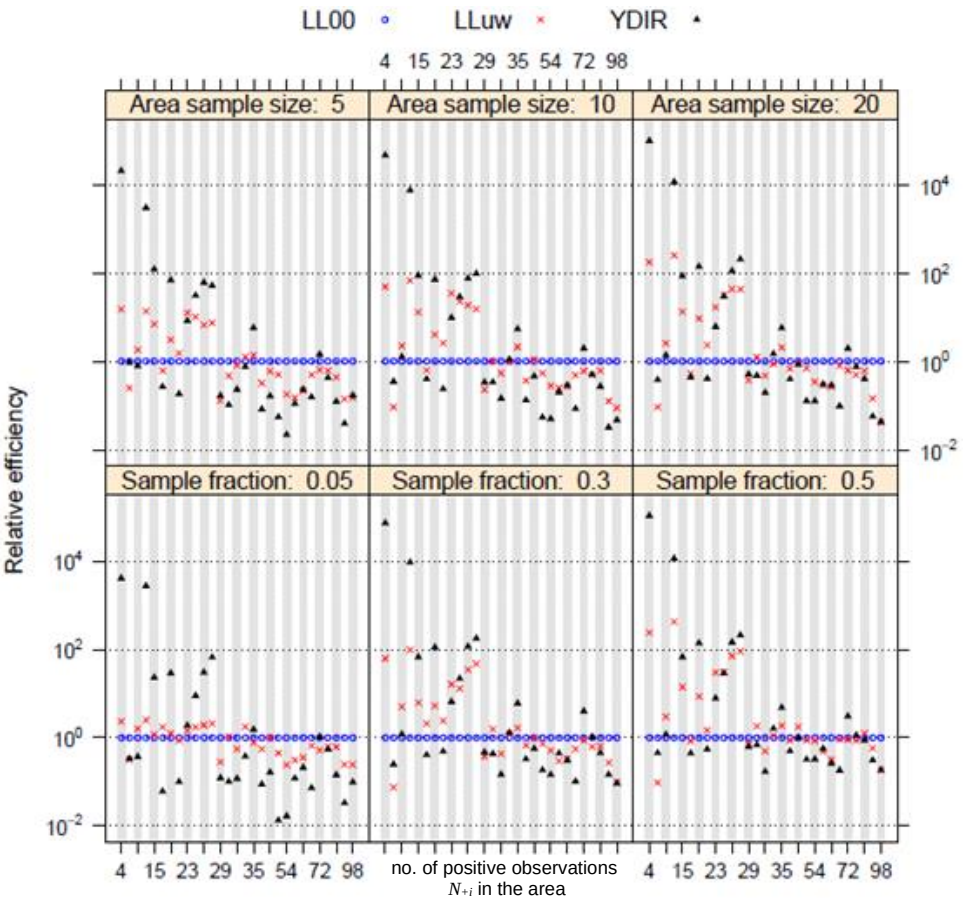

Figure 4. Relative efficiency of the *DIR* and $LL_{u\omega}$ estimators (w.r.t. LL_{00}) of Beef Cattle area means for various sample sizes. 100 replicates have been used for each sample size parameter

4. Conclusions

In Section 2 of this paper, we have arrived at four different lognormal-logistic estimators of small area means by combining the lognormal small area estimator of Berg and Chandra (2012) with the lognormal logistic model of Karlberg (2000b), and optionally incorporating hierarchical logistic regression.

We have conducted a simulation study to investigate the estimator properties under ideal circumstances, i.e. when the presumed lognormal-logistic model holds. As seen from Section 3.2, the estimators behave largely as predicted, i.e. when lognormal and/or logistic area-level effects are present, models incorporating such effects are superior, in terms of relative efficiency. Interestingly, this holds already

for rather small effects; the "penalty" for "unnecessarily" estimating a parameter when such a parameter is not present (thus introducing "white noise" into the estimation process) seems to be very modest. Using $LL_{u\omega}$ for lognormal-logistic data thus seems to be the best option (with the possible exception of situations with very low sample sizes (say $n_i \leq 5$) combined with large heterogeneity between the areas in terms of the proportion of positive observations (say $\sigma_\omega \geq 1.5$) when the direct estimator might be a safer option).

However, the model assumptions could be challenged. First, the assumption about independence between the lognormal and logistic components, made in Section 2.1, could be challenged; Pfefferman *et al.* (2008) convincingly argue for assuming a correlation between the two types of random effects; an extension of the model presented in this paper following the Bayesian approach proposed by Pfefferman *et al.* to relax the independence assumption. Even more critical is the fact that in real life data do not necessarily comply with a lognormal-logistic model, rendering the possible presence of correlation an issue of secondary importance. As could be seen from Section 3.3, the estimator's performance for the Beef Cattle variable of AAGIS is disastrous for certain small areas. This is studied in Appendix 1, where it is found that the small area estimation fails even if the model is fitted to the entire AAGIS data set, as going from national level to regional (area) level leads to severely biased estimates for some areas. Given this failure at small area population level, it is no surprise that the performance is bad when estimation is carried out for random samples. The situation is somewhat improved when area-level random effects are introduced – but an intolerable bias level remains for many areas.

It would be interesting to evaluate whether this is an artefact of the AAGIS data, i.e. if there are other real data sets where the lognormal-logistic estimators fare better, and what the properties of such data sets are (e.g. larger "small areas", or more highly correlated variables) – or if this poor performance is all but unavoidable. It could be argued that the performance issues are not so much related to the data as to the model, and there are a number of possible improvements of the lognormal-logistic models, such as somehow integrating it into the robust weighted mixed model of Chandra and Chambers (2011), which might be worth exploring.

Minor possible improvements also include a more formal treatment of the bias correction factor (currently simply carried over from Berg and Chandra; 2012), and the development of a proper model-based variance estimator (currently only readily available for LL_{00}), possibly even with an uncertainty measure for this variance (see Royall and Cumberland 1978 and Fellner 1986). Practical extensions to allow for some $n_{+i}=0$, and extensions to also allow negative values of Y_{ij} are also worth considering.

REFERENCES

BARNETT, V., LEWIS, T., (1994). Outliers in Statistical Data, 3rd ed. John Wiley & Sons.

BATTESE, G.E., HARTER, R.M., 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, Vol. 83, pp. 28–36.

BERG, E., CHANDRA, H., (2012). Small area prediction for a unit level lognormal model, Federal Committee on Statistical Methodology Research Conference.

CARROLL, R., RUPPERT, D., (1988). Transformation and Weighting in Regression, Chapman and Hall.

CHAMBERS, R. L., (1986). Outlier robust finite population estimation, Journal of the American Statistical Association, Vol. 81, pp. 1063–1069.

CHAMBERS, R. L., CHANDRA, H., SALVATI, N., TZAVIDIS. N., (2014). Outlier robust small area estimation, Journal of the Royal Statistical Society Series B: Statistical Methodology, Vol. 76, pp. 47–69.

CHAMBERS, R. L., DORFMAN, A. H., (2003). Transformed variables in survey sampling, Joint Statistical Meetings, Section on Survey Research Methods.

CHAMBERS, R. L., TZAVIDIS, N., (2006). M-quantile models for small area estimation., Biometrika, Vol. 93, pp. 255–268.

CHANDRA, H., CHAMBERS, R. L., (2005). Comparing EBLUP and C-EBLUP for Small Area Estimation, Statistics in Transition, Vol. 7, pp. 637–648.

CHANDRA, H., CHAMBERS, R. L., (2011). Small area estimation under transformation to linearity, Survey Methodology, Vol. 37, pp. 39–51.

CHEN, G., CHEN, J., (1996). A Transformation Method for Finite Population Sampling Calibrated with Empirical Likelihood, Survey Methodology, Vol. 22, pp. 139–146.

CHEN, J., CHEN, S.-Y., RAO, J. N. K., (2003). Empirical Likelihood Confidence Intervals for the Mean of a Population Containing Many Zero Values, The Canadian Journal of Statistics, Vol. 31, pp. 53–68.

FELLNER, W. H., (1986). Robust estimation of variance components, Technometrics, Vol. 28, pp. 51–60.

FULLER, W. A., (1991), Simple estimators for the mean of Skewed populations, Statistica Sinica, Vol. 1, pp. 137–158.

HIDIROGLOU, M. A., SMITH, P. A., (2005). Developing Small Area Estimates for Business Surveys at the ONS, Statistics in Transition, Vol. 7, pp. 527-539.

HIDIROGLOU, M. A., SRINATH, K. P., (1981). Some estimators of a population total from simple random samples containing large units, Journal of the American Statistical Association Vol. 76, pp. 690-695.

HUBER, P. J., (1981). Robust Statistics, John Wiley.

HUBERT, M., VAN DER VEEKEN, S., (2007). Outlier detection for skewed data, Journal of Chemometrics Vol. 22, pp. 235–246.

HUGGETT, M., (1996). Wealth distribution in life-cycle economies, Journal of Monetary Economics, Vol. 38, pp. 469–494.

JIANG, J., LAHIRI, P., (2006). Estimation of Finite Population Domain Means: A Model-Assisted Empirical Best Prediction Approach, Journal of the American Statistical Association, Vol. 101, pp. 301–311.

KARLBERG, F., (2000a). Population Total Prediction Under a Lognormal Superpopulation Model, Metron, Vol. LVIII, pp. 53–80.

KARLBERG, F., (2000b). Survey Estimation for Highly Skewed Populations in the Presence of Zeroes, Journal of Official Statistics, Vol. 16, pp. 229–241.

KOKIC, P. N., (1998). On Winsorisation in Business Surveys, SSC Annual Meeting, Proceedings of the Survey Methods Section.

LAMBERTA, D., (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, Technometrics, 34, pp. 1–14.

LEE, H. L., (1995). Outliers in Business Surveys, In Business Surveys Methods, edited by Cox, Binder, Chinnappa, Christianson, Colledege and Kott, Chapter 26. John Wiley.

LEHTONEN, R., SÄRNDAL C. E., VEIJANEN, A., (2003). The effect of model choice in estimation for domains, including small domains, Survey Methodology, Vol. 29, pp. 33–44.

MINCER, J., (1970). The Distribution of Labor Incomes: A Survey With Special Reference to the Human Capital Approach, Journal of Economic Literature 8, pp. 1–26.

MOLINA, I., (2009). Uncertainty under a multivariate nested-error regression model with logarithmic transformation, Journal of Multivariate Analysis, Vol. 100, pp. 963–980.

MOLINA, I., Marhuenda, Y., (2013). Package ‘sae’, http://cran.r project.org/web/packages/sae/sae.pdf

PFEFFERMANN, D., (2013). New Important Developments in Small Area Estimation, Statistical Science 28, pp. 40–68.

PFEFFERMANN, D., Terryn, B. Moura, F. A. S., (2008). Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries, Survey Methodology, Vol. 34, pp. 235–249.

RAO, J. N. K., (2003), Small Area Estimation, Wiley.

ROYALL, R. M., (1982), Finite populations (Sampling from), Entry in the Encyclopedia of Statistical Sciences.

ROYALL, R. M., CUMBERLAND, W. G., (1978). Variance estimation in finite population Sampling, Journal of the American Statistical Association Vol. 71, pp. 351–358.

RÖNNEGÅRD, L., SHEN, X. ALAM, M., (2010). hglm: A Package for Fitting Hierarchical Generalized Linear Models, The R Journal Vol. 2, pp. 20-28, http://journal.r-project.org/archive/2010-2/RJournal_2010-2_Roennegaard~et~al.

SALVATI, N., Chandra, H., Chambers, R. L., (2012). Model Based Direct Estimation of Small Area Distributions, Australian & New Zealand Journal of Statistics 54, pp. 103–123.

SEARLS, D. T., (1966). An estimator which reduces large true observations, Journal of American Statistical Association, Vol. 61, pp. 1200–1204.

SINHA, S. K., RAO, J. N. K., (2009). Robust small area estimation, Canadian Journal of Statistics, Vol. 37, pp. 381–399.

SHLOMO, N., PRIAM, R., (2013). Improving Estimation in Business Surveys. Chapter 4.2, 52–70 in BLUE-ETS Deliverable D6.2: Best practice recommendations on variance estimation and small area estimation in business surveys, edited by 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.

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

THORBURN, D., (1993). The treatment of outliers in economic statistics, Proceedings of the International Conference on Establishment Surveys, Buffalo, New York.

WANG, J., FULLER W. A., (2003). The Mean Squared Error of Small Area Predictors Constructed with Estimated Area Variances." Journal of the American Statistical Association, Vol. 98, pp. 716–723.

YOUNG, K. H., YOUNG, L. Y., (1975). Estimation of Regressions Involving Logarithmic Transformation of Zero Values in the Dependent Variable, The American Statistician , Vol. 29, pp. 118–120.

ZIMMERMANN, T., Münnich, R., (2013). Coherent small area estimates for skewed business data, Proceedings of the 2013 European Establishment Statistics Workshop.

APPENDIX

Appendix 1. Methodological details

A.1. Regression line fit at population level

In an attempt to identify the root cause of the poor performance of the lognormal-logistic estimators, we started out by fitting the model associated with LL_{00} to the entire population, i.e. the 1651 AAGIS observations. As could be seen from Figure A1, the model fits the data reasonably well; this is corroborated by the performance of $\hat{T}$ for the very same variables observed by Karlberg (2000b).

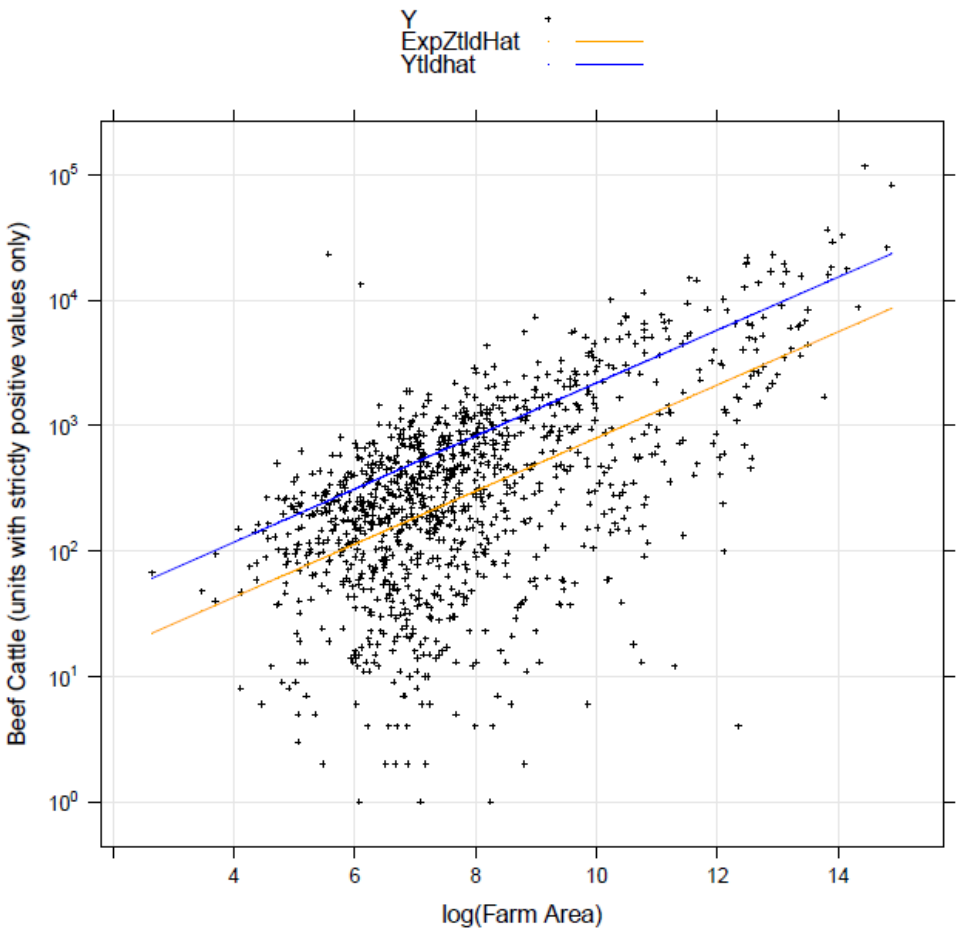

Figure A1. Regression line (red) fit to the logarithmed positive values of Beef Cattle for all 1651 observations (black) of AAGIS. The application of the bias correction factor is illustrated by the blue line.

A.2. Estimator performance at area level

Figure A2 demonstrates the effect of proceeding to the area level. There, we see that sometimes (taking the 4/43 area with 4 positive and 43 zero observations at the bottom left as an example) the entire area is composed of observations far from the regression line. Further investigations (not explicitly presented here) demonstrate that even if the large heterogeneity between areas in terms of zero valued observations (ranging from 0% to 91%, as could be seen from Table A1) is disregarded, the model completely fails to capture the structure of the positive values in a number of areas.

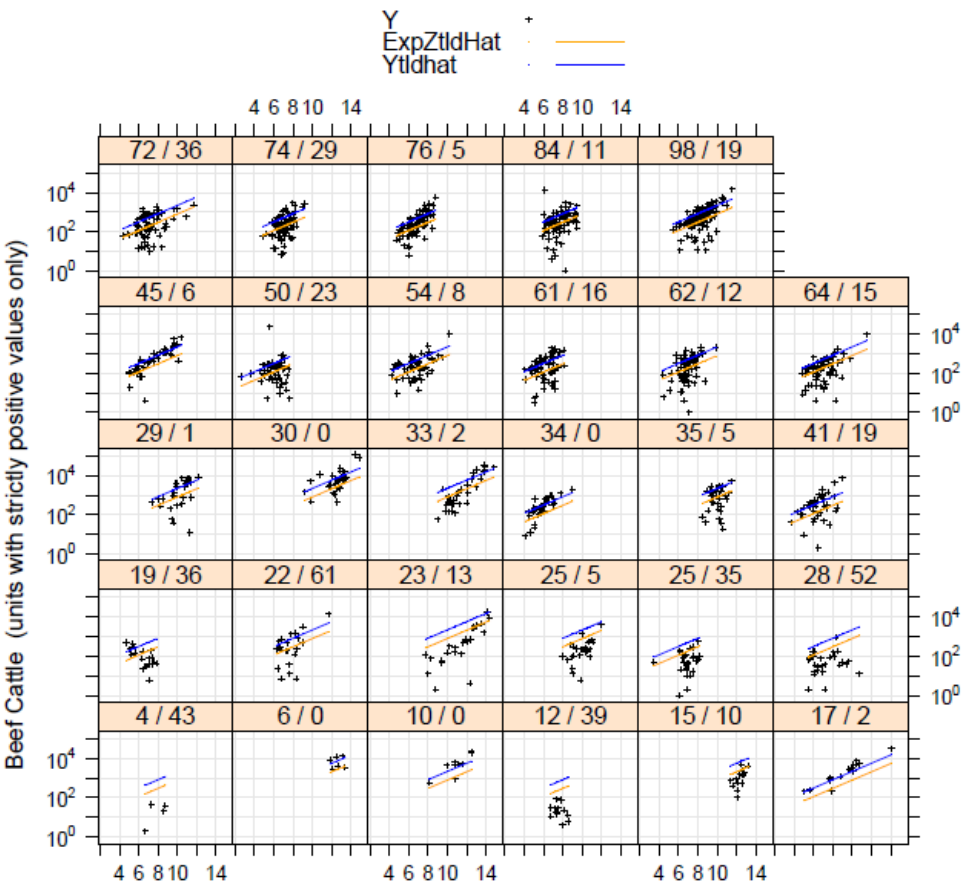

Figure A2. Regression line (red) fit to the logarithmed positive values of Beef Cattle for all 1651 observations of AAGIS, illustrated together with the observations (black) area by area. The number of positive/zero observation per area is indicated in the red strip above each area.

Obviously, if there is a severe bias even in an ideal situation, even with the model fit to the entire population, this is what could be expected to hold on average

for samples drawn from that population as well. This is precisely what we observe in Figure 3 for certain of the areas in the simulation study.

As the incorporation of area effects allows the fitting of a model that is closer to the values observed for each area, the bias of LL_{00} is, as could be seen from Figure 3, somewhat less severe across most areas, in particular the smaller ones. However, the performance is still unacceptable for that estimator as well.

Table A1. Some characteristics of the AAGIS variable Beef Cattle

Area i	No. of farms N_i	$(N_i - N_{+i})/N_i$	$\sum_{j=1}^{N_i} Y_j/N_{+i}$
1	47	91%	26.5
2	6	0%	7523.5
3	10	0%	8945.7
4	51	76%	28.8
5	25	40%	1554.7
6	19	11%	4285.6
7	55	65%	136.6
8	83	73%	1148.9
9	36	36%	1985.5
10	30	17%	430.1
11	60	58%	100.2
12	80	65%	97.8
13	30	3%	2774.7
14	30	0%	12903.0
15	35	6%	5878.8
16	34	0%	404.5
17	40	13%	1129.4
18	60	32%	670.5
19	51	12%	1139.6
20	73	32%	643.6
21	62	13%	530.9
22	77	21%	387.0
23	74	16%	390.7
24	79	19%	434.8
25	108	33%	435.2
26	103	28%	415.5
27	81	6%	526.6
28	95	12%	632.5
29	117	16%	980.6
All areas	**1651**	**30%**	**1308.5**