Small area estimation for business surveys: a comparison of transformation-based unit level models

Chiara Bocci[1] and Paul A. Smith[2]

[1] Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Viale Morgagni 59, 50134 Firenze, Italy. Email: chiara.bocci@unifi.it

[2] S3RI & Department of Social Statistics and Demography, University of Southampton, Southampton, SO17 1BJ, UK. Email: p.a.smith@soton.ac.uk

Abstract

Small area estimation methods are generally based on models which have assumptions of normal errors, but many types of data do not follow a normal distribution. Several approaches have been suggested to deal with skewed data, including transformations (with and without bias correction), robust models which are less affected by the tails of the distributions and building models directly with skewed error distributions. We investigate the properties of models for transformed data with a real data set which mimics a structural business survey. This contributes to the understanding of which tools are best for small area estimation with skewed data. We also investigate the sensitivity of results to different shift parameters (commonly used to make methods practical when data contain zeroes) and transformation parameters. The empirical best predictor (EBP) approach is found to be a flexible way to fit transformation-based models without the need for development of bias adjustments in back transformation. We prefer the EBP log-shift and EBP dual power which have good performance in our example (noting that the variables affecting the weighting are included in the model) because of their adaptability to new datasets. The bias-corrected empirical best (EBbc) estimator has similar performance in our example, but is tailored to the log transformation.

1 Introduction

Small area estimation (SAE) is a well established methodology with many adaptations to particular situations and topic areas. The application of small area estimation to business surveys has however lagged behind its use in other topic areas for two main reasons. First, the main approach to SAE is through the use of multilevel models which assume that the errors are normally distributed, whereas variables in business surveys are characterised by skewed distributions (Cox and Chinnappa, 1995; Rivière, 2002) which often give rise to skewed distributions of residuals in fitted models. Therefore the model assumptions are violated, and the skewed distributions generate outliers with respect to the model which affect the fits. Secondly, sampling in business surveys is informative, because the largest units are completely enumerated, and larger units have a higher probability to be included than smaller units. However, most model-based approaches assume that sampling is noninformative.

Business surveys, however, also have some characteristics which are helpful in SAE, at least from the perspective of a national statistical office (Rivière, 2002). There is a business register which contains some auxiliary variables at the unit level which can be used for modelling, and these auxiliary variables are known for all the units in the population. Here we focus on unit level models which can make use of this detailed information. There are relatively few applications of unit level small area estimation in business surveys, and a short overview is given in Smith et al. (2021).

Several strategies are available for unit level small area modelling with business survey data, with papers describing a range of modelling approaches within each strategy. These are:

1. Transformation-based approaches. Transformations can often reduce the effect of extreme values, but present an additional set of challenges because modelling with transformed data and then back-transforming generally results in biased estimates. These kinds of approaches have been investigated in practice by Krieg et al. (2012), but there has also been some subsequent development, particularly by Rojas-Perilla et al. (2020).

2. Robust modelling approaches. The effectiveness of different robust models was investigated by Smith et al. (2021) with a dataset based on tax data from the Netherlands. They found that the naïve M-quantile estimator was the best practical approach (even though it is not a consistent estimator because it does not use the survey weights). The weighted M-quantile estimator which does use the weights and is consistent is nearly as good. And the best estimator overall is the bias-adjusted M-quantile method (using the approach of Welsh and Ronchetti (1998)) if the optimum value of the second tuning constant is known. However, it is normally unknown in practice, and the difference from the naïve M-quantile method is small, so that the additional effort required for this kind of model seems not worthwhile, at least in the given example.

3. Dealing directly with skewed distributions of residuals in multilevel models. Models with residuals having generalised Beta distributions, skew normal distributions and mixtures of normal distributions have all been proposed. We expect to review and evaluate these approaches in the future.

4. Using tree-based models (Krennmair and Schmid, 2022; Krennmair et al., 2026) rather than directly fitting a parametric model. It is known that trees can capture complex data structures, particularly if there is a sufficiently large number of predictor variables, though the number of predictors is often quite small in business surveys. We leave an investigation of these approaches for further research

5. If the extreme values from different businesses tend to cancel out in aggregation (that is, high values of the outcome of interest are balanced by low values within the same domains of interest), then changing to an area level model may also be an effective strategy for robustifying small area estimation. However, since our main focus is on unit level models, and since we suspect that this kind of cancellation does not happen often within the small domains of interest, we do not consider this approach further.

In this paper we examine in detail the transformation-based approaches, and use a real data example, where the outcomes are known, to evaluate the performance of the different approaches within this broad strategy. Our intention is to produce a counterpart for the assessment of robust approaches in Smith et al. (2021). We investigate two classes of methods. One is the empirical best predictor (EBP) of Molina and Rao (2010), and extensions to this allow different transformations to be plugged in; in this paper we consider the log, log-shift, Box-Cox and dual power transformations. The second class of models is fitted to the transformed data, and here attention has been restricted to the log transformation; a bias correction is needed in back transformation from these models, and we investigate different forms of the bias correction. Both classes of estimators can also be extended to use survey weights to deal with informative sampling, and we additionally consider these weighted estimators.

In business surveys, 0 is a common response for some variables, and we therefore need to add a shift to the responses to obtain admissible values from the transformations. We therefore investigate the sensitivity of the different approaches to the shift parameter in the transformation. In some cases the shift parameter can be fitted, in others it has to be assumed.

This range of estimators and transformations leads us to compare 23 strategies. In section 2 we present all these transformation-based approaches, with a unified notation consistent with Smith et al. (2021). In section 3 we describe the dataset to which these methods will be applied, and the approach to repeated sampling simulation which we use to evaluate the estimators. In section 4 we give the results of the design-based simulation from the data, and examine the properties of the estimators. In section 5 we discuss the results.

2 Transformation-based approaches to small area estimation for skewed data

A number of approaches based on transformations have been proposed in the literature, accompanied by procedures designed to maintain the unbiasedness of the estimators under back-transformation. Lyu et al. (2020, section 1.3) give a brief overview of the methods developed for small area estimation based on transformations. We follow their approach, but provide more detail of the different estimators, presented with a standard notation based on estimating a population total, which is one of the most important target parameters for business surveys (and which is also consistent with Smith et al. (2021)).

We begin by setting out direct estimators in section 2.1, and then introduce the range of transformation-based estimators in section 2.2, starting with empirical best prediction (EBP) (section 2.2.1), and then moving to models fitted directly to the transformed data (section 2.2.2). The form of the bias correction

for these models is investigated in section 2.2.3. Then weighted versions of both the EBP and the models for the transformed data are introduced in section 2.3.

2.1 Estimators on the original scale

We follow the strategy of Smith et al. (2021) in choosing design-based direct estimators as a baseline for the evaluation of the different approaches. In the simulation application below we use the industrial classes as the domains of interest, so we use *ind* to represent them in the equations; the equations are, however, general. Design-based and direct estimators are often preferred by National Statistical Institutes (NSIs) because of their unbiasedness and objectivity (Rao, 2011). However, they often also have large variances for estimates in small areas, so comparing them with SAE methods will generally demonstrate the mean squared error gains to be made by using small area methods. The simplest direct estimator is the Horvitz–Thompson (HT) estimator:

$$\hat{y}_{ind}^{HT} = \sum_{i \in ind \cap s} d_i y_i \tag{1}$$

where $d_i = 1/\pi_i$ is the inverse of the selection probability, y the variable of interest, s represents the sample and *ind* labels the industrial classification stratum. Note that throughout we provide estimators for the population *totals* in small domains, in line with typical targets of inference in business surveys (and consistent with Smith et al. (2021)), which differs from most of the small area literature which uses domain means as the target.

Auxiliary information is typically very valuable in estimation in business surveys, and although the HT estimator is used, it is unusual; therefore we also consider the generalised regression (GREG) estimator (Särndal et al., 1992, chapter 6):

$$\hat{y}_{ind}^{GREG} = \hat{y}_{ind}^{HT} + \boldsymbol{\beta}^{GREG} \left(\mathbf{X}_{ind} - \hat{\mathbf{x}}_{ind}^{HT} \right) \tag{2}$$

where $\mathbf{X}_{ind}$ is a vector of known totals of auxiliary variables and $\hat{\mathbf{x}}_{ind}^{HT}$ is the vector of HT estimates of these variables from the sample calculated using Equation (1) with y_i replaced by the appropriate auxiliary variables. It is possible to write such estimators in the same form as Equation (1) as $\hat{y}_{ind}^{GREG} = \sum_{i \in ind \cap s} w_i y_i$ with modified weights w_i, and then domain estimation can be used to obtain the required industry level estimates. Direct estimation approaches fail in the case that there are no sample units in the domain of interest.

We also consider a standard application of the unit level model (Battese et al., 1988)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{3}$$

with untransformed y, where $u \sim N(0, \sigma_u^2)$ and $e \sim N(0, \sigma_e^2)$. Let $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ denote estimates of the fixed and random effects in equation (3). Then the EBLUPs of the totals for industry classifications are given by

$$\hat{y}_{ind}^{EBLUP} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \left(\mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}} + \mathbf{z}_i^{\mathrm{T}} \hat{\mathbf{u}} \right) \tag{4}$$

where s and r denote the sample and non-sample units respectively. When y has a skewed distribution, the assumptions of normally distributed errors for this model are violated, and we expect that estimates will be biased. Naïve application of this model will also provide a point of reference in the assessment of models that do account for the skewed distributions of the data.

2.2 Transformation-based estimators

There are different strategies available for transformation. One is to generate multiple simulated datasets using errors generated from the distributions on the transformed scale, then back-transform and calculate the estimates with the simulated data on the original scale, thereby avoiding any bias through back-transforming estimates calculated on the transformed data; this approach is covered in section 2.2.1. A second approach is to calculate the estimates with the transformed data, an approach which has been developed principally with the log transformation; however, simply taking the exponential of the fitted values is known to give biased estimates (Finney, 1941) because this gives a geometric mean which is always smaller than the target arithmetic mean behaviour (conditional on explanatory variables). A suitable bias correction can be made to obtain (approximately) unbiased estimates on the original scale,

and this approach is addressed in section 2.2.2. Both approaches require a unit level model fitted to the transformed data $\mathbf{y}^*$

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{Z}\mathbf{u}^* + \mathbf{e}^*; \tag{5}$$

throughout we use a * in equations to designate a transformed response or a model parameter fitted to a transformed response.

The bias correction can take several forms (Flewelling and Pienaar, 1981; Zeng and Tang, 2011). Flewelling and Pienaar (1981) consider six different bias corrections, and deduce that for sample sizes > 30 the maximum relative difference between estimates made with these corrections is $exp(\frac{3}{2s^2})$ in most situations, where s^2 is the sum of the area-level and unit-level variances from (5). Zeng and Tang (2011) give two types of first-order correction, $1 + \frac{s^2}{2}$ and $\exp\left(\frac{s^2}{2}\right)$, and a second order correction from Finney (1941), but find negligible differences between them.

The bias correction is derived from the fitted model, so the sample size applying to the correction will be the sample size across all parts of the data which contribute to the model. Therefore the equivalence of the different bias adjustments is not called into question through small sample sizes even in the case of small area estimation. In order to check this conclusion empirically we check this with the Karlberg-type estimation in section 2.2.3.

The corrections require information on the values of the auxiliary data in the population. In business surveys these data are often available from the business register from which the survey is sampled. But there may be situations in which only aggregate information about the population is available, and Würz et al. (2022) propose an approach for this situation based on kernel density estimation to derive the appropriate distribution for use in the correction.

2.2.1 Empirical best prediction

The basic idea of empirical best prediction (EBP) comes from Molina and Rao (2010), where the expected value of the conditional distribution of the unobserved data given the sample data is approximated efficiently by a numerical procedure. This approach is extended for transformations by Guadarrama et al. (2016) and Rojas-Perilla et al. (2020), and the stages are:

1. select a transformation, fitting the shift or shape parameter $\hat{\lambda}$ if necessary, and obtain $y_i^* = T_{\hat{\lambda}}(y_i)$
2. use the transformed data in the unit level model (5) to estimate $\hat{\boldsymbol{\beta}}^*$, and the variance components $\hat{\sigma}_u^{*2}$ and $\hat{\sigma}_e^{*2}$; calculate $\hat{\gamma}_{ind}^* = \hat{\sigma}_u^{*2} / \left(\hat{\sigma}_u^{*2} + n_{ind}^{-1}\hat{\sigma}_e^{*2}\right)$
3. for l in $1, \ldots, L$
   (a) take random draws v_{ind}^* from $N\left(0, (1 - \hat{\gamma}_{ind}^*)\hat{\sigma}_u^{*2}\right)$ for each value of ind included in the sample, or from $N\left(0, \hat{\sigma}_u^{*2}\right)$ for any out-of-sample values of ind
   (b) take random draws of e_i^* from $N\left(0, \hat{\sigma}_e^{*2}\right)$
   (c) obtain pseudopopulation l as $y_i^{*(l)} = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^* + \hat{u}_{ind}^* + v_{ind}^* + e_i^*$ when ind is included in the sample, choosing ind such that $i \in ind$, and as $y_i^{*(l)} = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^* + v_{ind}^* + e_i^*$ when i belongs to an out-of-sample value of ind
   (d) back-transform the pseudopopulation values to obtain $y_i^{(l)}$
   (e) calculate the estimate of interest for each ind with pseudopopulation l, $\hat{y}_{ind}^{(l)}$
4. take the average of the statistic of interest for each ind over the l replicates, $\hat{y}_{ind}^{EBP} = \frac{1}{L}\sum \hat{y}_{ind}^{(l)}$.

Steps 3(b-d) can either be undertaken on the whole population (the "census EBP"), or with the out-of-sample units only, with the observed data being used to complete each pseudopopulation l. The implementation of the EBP in `emdi` and `povmap` is a census EBP (Skarke et al., 2021; Edochie et al., 2023). However, we demonstrated that using the real values where they are available is better with robust estimators (Smith et al., 2021, section 5.5), and expect that that will continue to hold with the EBP. We therefore modify the implementation to predict only for out-of-sample units in the population.

Here we follow Rojas-Perilla et al. (2020) in considering four transformations (Table 1), although others are available. First is the standard log transformation (EBP log). Since zero values are present in the dataset, we need to shift the data by an amount s, deterministically chosen so that $y_i + s > 0$; it is important that $min(y_i + s)$ is not too small to avoid creating high leverage points on the transformed

scale. Second is the log-shift transformation (Yang, 1995) (EBP log-shift) which is basically the same except that the shift parameter, labelled λ in line with the data-driven transformation notation above, is fitted from the data (for details see Rojas-Perilla et al. (2020)).

The third transformation is the Box-Cox transformation (Box and Cox, 1964) (EBP Box-Cox), where the shape of the transformation is driven by the data. The deterministic shift is again needed to ensure that the data are positive. Under the Box-Cox transformation, y_i^* is bounded below by $1/\lambda$ if $\lambda > 0$ and above by $-1/\lambda$ if $\lambda < 0$. The inverse transformation can then be impossible if a value beyond the bound is obtained from the linear predictor. The fourth transformation, the dual power transformation (Yang, 2006) (EBP dual power) was developed to avoid the bounds of the Box-Cox transformation, but otherwise is rather similar in its behaviour; it too needs a deterministic shift to ensure strictly positive inputs. All four transformations are available with the EBP methodology in the R package `emdi` (Kreutzmann et al., 2019). The `povmap` extension to `emdi` additionally includes rank-order and arcsin transformations (Edochie et al., 2023), but we do not consider these in this research.

The log, Box-Cox and dual power transformations require a deterministic shift s when the data contain zero or negative values; s should not be too small, because small values generate outliers in $\log(s)$. With the example data below we use the reasonably standard $s = 1$, the default value in `povmap` and `emdi`, but we note that the results may be sensitive to this value, and explore this in more detail by comparing estimates across a grid of values $s = \{0.0001, 0.5, 1, 2, 5\}$ in section 4.2.2.

transformation	T_λ
log†	$\log(y_i + s)$
log-shift	$\log(y_i + \lambda)$
Box-Cox†	$\begin{cases} \frac{(y_i+s)^\lambda - 1}{\lambda} & \text{if} \lambda \neq 0 \\ log\,(y_i + s) & \text{if} \lambda = 0 \end{cases}$
dual power†	$\begin{cases} \frac{(y_i+s)^\lambda - (y_i+s)^{-\lambda}}{2\lambda} & \text{if} \lambda > 0 \\ log\,(y_i + s) & \text{if} \lambda = 0 \end{cases}$

Table 1: Transformations considered for use with the EBP for business survey data and their corresponding functions. Those labelled † require a deterministic shift parameter s in the case of zero and/or negative values in order to ensure that the functions are defined on the range of the data. Parameters λ are fitted from the data.

The implementations in `emdi` and `povmap` involve fitting $\hat{\lambda}$ together with $\hat{\boldsymbol{\beta}}^*$ so that there is some interdependence between the transformation parameter and the model fitting (rather than the two step approach outlined in the algorithm above which suggests that the estimation of $\hat{\lambda}$ is done independently first). We return to this in section 2.3.1 below.

2.2.2 Models fitted to the transformed data

A more traditional approach is to produce estimates (directly) from the model fitted to the transformed data, which is expected to conform better with the distributional assumptions on the errors. All these estimators, fitted on the transformed data, need to be transformed back to the original scale, and this may result in complicated functions of the mean and variance on the transformed scale (Sugasawa and Kubokawa, 2019). Here we consider only the log transformation; other transformations could be used, with suitable functions for the back-transformation. Li and Lahiri (2007) initiated this with a proposal for the Box-Cox transformation (noting that the back-transformation may be impossible for some values, since the transformation does not cover the whole real line), and Sugasawa and Kubokawa (2019) have extended this for the dual power and sin-arcsin transformations, which both span the real line, although the small area estimators do not have closed form solutions. Li et al. (2019) propose a different transformation $y^*(\lambda) = \text{sign}(y)|y|^\lambda$ and develop small area estimators which do have a closed form. We do not consider these transformation estimators here.

We focus on the log transformation $y^* = \log(y + s)$, for which small area estimators have been developed by several authors (and in fact we use $s = 0$ for simplicity of notation; substituting a different value is straightforward). The complex function for back-transformation for the log is handled with a bias correction. Chandra and Chambers (2011) already note that a first order correction is not sufficient, and propose a second order correction. The developed estimators for the log transformed data are analogues

for three strategies for the linear model – the synthetic estimator, the model-based direct estimator and the EBLUP (see Berg et al. (2016), section 15.3 for a description of these estimators and how the analogues are developed).

The first of these estimators on the log-transformed data is a synthetic estimator, which Chandra and Chambers (2011) call a Karlberg-type estimator because it extends the work of Karlberg (2000), so we label it K (rather than the somewhat cumbersome *SYN-EP* used by Chandra and Chambers (2016)). It is given by

$$\hat{y}_{ind}^{K} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \left(\hat{c}_i^K\right)^{-1} exp\left(\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^* + \frac{\hat{\sigma}_u^{*2} + \hat{\sigma}_e^{*2}}{2}\right). \tag{6}$$

with

$$\hat{c}_i^K = exp\left[\frac{1}{2}\left(\mathbf{x}_i^{\mathrm{T}}\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}^*)\mathbf{x}_i + \frac{1}{4}\hat{V}(\hat{\sigma}_u^{*2} + \hat{\sigma}_e^{*2})\right)\right] \tag{7}$$

(corrected from Chandra and Chambers (2016, below equation (2)), but in line with Berg et al. (2016, equation (15.10))) where the form of the conditional mean of y_i given $\mathbf{x}_i$ and $\hat{\boldsymbol{\beta}}^*$ is based on the moment generating function of a normal distribution, and $\hat{c}_i^K$ is a second-order correction for back transformation bias. $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}^*)$ and $\hat{V}(\hat{\sigma}_u^{*2} + \hat{\sigma}_e^{*2})$ are estimates of the variances of $\hat{\boldsymbol{\beta}}^*$ and $\hat{\sigma}_u^{*2} + \hat{\sigma}_e^{*2}$ respectively (Berg et al., 2016). Note that $\hat{\boldsymbol{\beta}}^*$ is the same as that obtained in section 2.2.1 with the log transformed data, since the transformation is the same, and therefore the modelling process is the same.

The second estimator is a version of the model based direct (MBD) estimator, originally developed for the linear model by Chandra and Chambers (2009), and adapted to the log-transformed data by Chandra and Chambers (2011). MBD estimators derive weights (different from the sampling weights, which are not considered as sampling is assumed to be ignorable) which are used with the observed responses $y_i, i \in s$ to generate optimal predictors of population totals. The MBD estimator with the log transformed data uses model calibration to define the set of weights. This process needs predictions of the population values under a model, and we use the bias-corrected predictions from the Karlberg-type synthetic estimator in equation (6), $\hat{y}_i^K = \left(\hat{c}_i^K\right)^{-1} exp\left(\mathbf{x}_i^{\mathrm{T}}\hat{\beta}^* + \frac{\hat{\sigma}_u^{*2} + \hat{\sigma}_e^{*2}}{2}\right)$; since these already have a second-order bias correction, the MBD is also second order corrected. We therefore have the MBD as

$$\hat{y}_{ind}^{MBD} = \sum_{i \in ind \cap s} w_i^{MBD} y_i. \tag{8}$$

Note that this is a *direct* estimator, so a sample of ≥ 1 unit is needed in each small domain to generate an estimate. The weights can be derived for a nonlinear model through model calibration (Wu and Sitter, 2001), a calibration process on the predicted values from the model. The weights are given by

$$\mathbf{w}^{MBD} = \mathbf{1}_s + \hat{\mathbf{H}}_s^{\mathrm{T}}\left(\hat{\mathbf{M}}_U^{\mathrm{T}}\mathbf{1}_U - \hat{\mathbf{M}}_s^{\mathrm{T}}\mathbf{1}_s\right) + \left(\mathbf{I}_s - \hat{\mathbf{H}}_s^{\mathrm{T}}\hat{\mathbf{M}}_s^{\mathrm{T}}\right)\hat{\mathbf{V}}_{ss}^{-1}\hat{\mathbf{V}}_{sr}\mathbf{1}_r \tag{9}$$

with $\hat{\mathbf{H}}_s = \left(\hat{\mathbf{M}}_s^{\mathrm{T}}\hat{\mathbf{V}}_{ss}^{-1}\hat{\mathbf{M}}_s\right)^{-1}\hat{\mathbf{M}}_s^{\mathrm{T}}\hat{\mathbf{V}}_{ss}^{-1}$, $\mathbf{M_U} = (\hat{\mathbf{M}}_s^{\mathrm{T}}, \hat{\mathbf{M}}_r^{\mathrm{T}})^{\mathrm{T}} = \left[(\mathbf{1}_s^{\mathrm{T}}, \mathbf{1}_r^{\mathrm{T}})^{\mathrm{T}}, \left((\hat{\mathbf{y}}_s^K)^{\mathrm{T}}, (\hat{\mathbf{y}}_r^K)^{\mathrm{T}}\right)^{\mathrm{T}}\right]$, a two-column matrix with the first column with all the elements 1 and the second column with elements $\hat{y}_i^K$, of which the first s rows (submatrix $\hat{\mathbf{M}}_s$) represent the sample units and the remaining r rows (submatrix $\hat{\mathbf{M}}_r$) the non-sample units. $\hat{\mathbf{V}}_{ss}$ and $\hat{\mathbf{V}}_{sr}$ are components of $\hat{\mathbf{V}}(\mathbf{y}_U) = \begin{pmatrix} \hat{\mathbf{V}}_{ss} & \hat{\mathbf{V}}_{sr} \\ \hat{\mathbf{V}}_{rs} & \hat{\mathbf{V}}_{rr} \end{pmatrix}$. For full details see Chandra and Chambers (2011), Berg et al. (2016) or Chandra and Chambers (2016).

The third estimator is the empirical best (EB) predictor, which is developed by Berg and Chandra (2014) as the minimum mean squared error predictor under the lognormal model (that is, the model for the log-transformed data):

$$\hat{y}_{ind}^{EB} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \exp\left\{\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^* + \hat{\gamma}_{ind}^*\left(\bar{y}_{ind}^* - \bar{\mathbf{x}}_{ind}^{\mathrm{T}}\hat{\boldsymbol{\beta}}^*\right) + \frac{1}{2}\hat{\sigma}_e^{*2}\left(\frac{\hat{\gamma}_{ind}^*}{n_{ind}} + 1\right)\right\}. \tag{10}$$

where $\bar{y}_{ind}^* = n_{ind}^{-1}\sum_{i \in ind \cap s}\log(y_i)$ is the sample mean of $\log(y)$ in ind, $\bar{\mathbf{x}}_{ind} = n_{ind}^{-1}\sum_{i \in ind \cap s}\mathbf{x}_i$ is the corresponding sample mean of $\mathbf{x}$ and $\hat{\gamma}_{ind}^* = \hat{\sigma}_u^{*2}/\left(\hat{\sigma}_u^{*2} + n_{ind}^{-1}\hat{\sigma}_e^{*2}\right)$ as in section 2.2.1.

(10) contains only a first-order correction, and Berg and Chandra (2014) demonstrate that the estimator is biased. Therefore, they develop a multiplicative bias correction (the estimator is labelled EBbc) derived from a second-order Taylor expansion:

$$\hat{y}_{ind}^{EBbc} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \left(\hat{c}_i^{EB}\right)^{-1}\hat{y}_i^{EB} \tag{11}$$

with $\hat{y}_i^{EB} = \exp\left\{\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^* + \hat{\gamma}_{ind}^*\left(\bar{y}_{ind}^* - \bar{\mathbf{x}}_{ind}^{\mathrm{T}}\hat{\boldsymbol{\beta}}^*\right) + \frac{1}{2}\hat{\sigma}_e^{*2}\left(\frac{\hat{\gamma}_{ind}^*}{n_{ind}} + 1\right)\right\}$ and

$$\hat{c}_i^{EB} = \exp\left[\frac{1}{2}\left(\hat{\mathbf{a}}_i + \hat{c}_{1,ind}\hat{V}(\hat{\sigma}_e^{*2}) + \hat{c}_{2,ind}\hat{V}(\hat{\sigma}_u^{*2}) + 2\hat{c}_{3,ind}\hat{Cov}(\hat{\sigma}_e^{*2}, \hat{\sigma}_u^{*2})\right)\right]. \tag{12}$$

Here

$$\hat{\mathbf{a}}_i = \left(\mathbf{x}_i^{\mathrm{T}} - \hat{\gamma}_{ind}^*\bar{\mathbf{x}}_{ind}\right)^{\mathrm{T}}\hat{V}\left(\hat{\boldsymbol{\beta}}^*\right)\left(\mathbf{x}_i^{\mathrm{T}} - \hat{\gamma}_{ind}^*\bar{\mathbf{x}}_{ind}\right)$$

and, writing $\hat{d}_{ind} = \bar{y}_{ind}^* - \bar{\mathbf{x}}_{ind}\hat{\boldsymbol{\beta}}^*$,

$$\hat{c}_{1,ind} = \left\{\frac{1}{2} + \frac{\hat{\gamma}_{ind}^{*2}}{n_{ind}}\left(\frac{1}{2} - \frac{\hat{d}_{ind}}{\hat{\sigma}_u^{*2}}\right)\right\} - \left\{\frac{\hat{\gamma}_{ind}^{*3}}{n_{ind}^2\hat{\sigma}_u^2}\left(1 - \frac{2\hat{d}_{ind}}{\hat{\sigma}_u^{*2}}\right)\right\}$$

$$\begin{aligned}
\hat{c}_{2,ind} = &\left\{\frac{\left(1-\hat{\gamma}_{ind}^*\right)^2}{2} + \frac{\hat{\gamma}_{ind}^*\left(1-\hat{\gamma}_{ind}^*\right)\hat{d}_{ind}}{\hat{\sigma}_u^{*2}}\right\} \\
&- \left\{\frac{\hat{\gamma}_{ind}^*\left(1-\hat{\gamma}_{ind}^*\right)}{\hat{\sigma}_u^{*2}}\left[\left(1-\hat{\gamma}_{ind}^*\right) + \frac{2\hat{\gamma}_{ind}^*\hat{d}_{ind}}{\hat{\sigma}_u^{*2}}\right]\right\} \\
\hat{c}_{3,ind} = &\left\{\frac{1}{2} + \frac{\hat{\gamma}_{ind}^{*2}}{n_{ind}}\left(\frac{1}{2} - \frac{\hat{d}_{ind}}{\hat{\sigma}_u^{*2}}\right)\right\}\left\{\frac{\left(1-\hat{\gamma}_{ind}^*\right)^2}{2} + \frac{\hat{\gamma}_{ind}^*\left(1-\hat{\gamma}_{ind}^*\right)\hat{d}_{ind}}{\hat{\sigma}_u^{*2}}\right\} \\
&+ \left\{\frac{\hat{\gamma}_{ind}^{*2}\left(1-\hat{\gamma}_{ind}^*\right)}{n_{ind}\hat{\sigma}_u^{*2}} - \frac{\hat{\gamma}_{ind}^{*2}\left(1-2\hat{\gamma}_{ind}^*\right)\hat{d}_{ind}}{n_{ind}\hat{\sigma}_u^{*4}}\right\},
\end{aligned}$$

(Berg and Chandra, 2014; Chandra and Chambers, 2016).

Molina (2009) considers a simpler version of equation (10) based on estimating the exponentials of mixed effects, which do not account for the individual errors in estimating the values of unobserved units in the population. Molina and Martín (2018) examine this estimator in a small simulation study in which it performs poorly compared to (10), so we do not consider it further here.

2.2.3 Forms of the bias correction

Different forms of the first-order bias correction are available (Flewelling and Pienaar, 1981; Zeng and Tang, 2011), and we implement an alternative in the Karlberg estimator (6) based on Zeng and Tang (2011) as

$$\tilde{y}_{ind}^K = \sum_{i\in ind\cap s} y_i + \sum_{i\in ind\cap r}\left(\tilde{c}_i^K\right)^{-1}\exp\left(\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^*\right)\left(1 + \frac{\hat{\sigma}_u^{*2} + \hat{\sigma}_e^{*2}}{2}\right). \tag{13}$$

We follow the approach in Chandra and Chambers (2011) to derive a second order bias correction for use with this adjustment, which turns out to be

$$\tilde{c}_i^K = \exp\left[\frac{1}{2}\left(\mathbf{x}_i^{\mathrm{T}}\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}^*)\mathbf{x}_i\right)\right], \tag{14}$$

that is, without the second term in the second order correction (7).

Returning to the standard first order bias correction in the Karlberg estimator (6), we note that Chandra and Chambers (2011) derive a different form of the second-order correction, its relation to (7) analogous to the two forms of the first-order correction noted in section 2.2. As an example of the sensitivity of the results to the form of the bias correction, we also use this correction in the Karlberg estimator, which replaces $\hat{c}_i^K$ in equation (7) by

$$\hat{c}_i^{K\text{-}alt} = 1 + \frac{1}{2}\left(\mathbf{x}_i^{\mathrm{T}}\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}^*)\mathbf{x}_i + \frac{1}{4}\hat{V}(\hat{\sigma}_u^{*2} + \hat{\sigma}_e^{*2})\right). \tag{15}$$

2.3 Weighted estimators

The stratified designs commonly employed in business surveys often give rise to situations where the response is not independent of the selection probability, a situation known as non-ignorable sampling. Under non-ignorability the model for the sampled units (only) does not apply to the rest of the population, and we need further adaptation. The ignorability actually depends on which variables are included in the model, and including weights when the sampling is ignorable is not desirable as it may inflate the variance. One strategy therefore is to test for ignorability before deciding whether to employ a weighted model. We present the options for weighted approaches below.

2.3.1 Pseudo-Empirical Best Prediction

A weighted EBP estimator is proposed by Guadarrama et al. (2018), who call it the pseudo-empirical best predictor (PEBP); the basic procedure as described in section 2.2.1 is retained, but

- the estimates are conditioned on the weighted means. Skarke et al. (2021) say that this means that the procedure does not correct for nonignorable sampling in out-of-sample areas, but we did not find any evidence for this and consider that it does make this correction; and
- the parameters are derived from a weighted unit level model. Weighted estimators may be fitted directly using maximum likelihood (Pfeffermann and Sverchkov, 2007), or using the method of moments of You and Rao (2002). The latter approach is implemented in the `emdi` package (Kreutzmann et al., 2019; Skarke et al., 2021) in `R` (R Core Team, 2021); these are labelled PEBP MM below, with the appropriate transformation appended. `povmap` (Edochie et al., 2023), which builds on `emdi`, additionally allows the weights to be used in the model fitting procedure. This is a maximum likelihood procedure, so this is likely to be analogous to the approach of Pfeffermann and Sverchkov (2007). We do not have sufficient information about the implementation to assess whether it is exactly the same approach. These estimators are labelled PEBP ML below, with the appropriate transformation appended.

The specific changes in the steps in section 2.2.1 are that $\hat{\gamma}^{*}_{ind}$ is replaced by $\hat{\gamma}^{*w}_{ind} = \hat{\sigma}^{*2}_u / \left(\hat{\sigma}^{*2}_u + \delta^2_{ind}\hat{\sigma}^{*2}_e\right)$ where $\delta^2_{ind} = \sum_{i\in ind\cap s} w_i^2 / \left(\sum_{i\in ind\cap s} w_i\right)^2$; and there is an extra step between steps 2 and 3 where weighted versions of the fixed effects parameters are estimated as

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{*w} = &\left(\sum_{ind}\sum_{i\in ind\cap s} w_{ind,i}\mathbf{x}_{ind,i}(\mathbf{x}_{ind,i} - \hat{\gamma}^{*w}_{ind}\bar{\mathbf{x}}^{w}_{ind})^{\mathrm{T}}\right)^{-1} \\ &\times \left(\sum_{ind}\sum_{i\in ind\cap s} w_{ind,i}(\mathbf{x}_{ind,i} - \hat{\gamma}^{*w}_{ind}\bar{\mathbf{x}}^{w}_{ind})y^{*}_{ind,i}\right)\end{aligned} \tag{16}$$

where $\bar{\mathbf{x}}^{w}_{ind} = \sum_{i\in ind\cap s} w_{ind,i}\mathbf{x}_{ind,i} / \left(\sum_{i\in ind\cap s} w_{ind,i}\right)$. $\hat{\gamma}^{*w}_{ind}$ and $\bar{\mathbf{x}}^{w}_{ind}$ are used in step 3. We denote the resulting estimator by $\hat{y}^{PEBP}_{ind}$. Guadarrama et al. (2018) develop the PEBP methodology only for untransformed data and the log transformation, and these are implemented in `emdi` (Skarke et al., 2021). In `povmap`, Edochie et al. (2023)'s maximum likelihood procedure additionally incorporates the PEBP for the log-shift, Box-Cox and dual power transformations. The maximum likelihood procedure also allows the λ parameter in these transformations (Table 1) to be estimated taking account of the survey weights, and we denote this estimate by $\hat{\lambda}^{w}$. The implementation in `povmap` allows the unweighted or weighted transformation parameter in conjunction with the PEBP.

As described in section 2.2.1, the transformation parameter λ is fitted with the optimisation of the $\hat{\boldsymbol{\beta}}^{*w}$, so to the extent that these parameters are correlated, the fitting of λ will be influenced. We therefore denote the fitted (unweighted) transformation parameter from the (weighted) PEBP as $\tilde{\lambda}$, and expect that, even though the underlying data is the same, the influence of the fitting of the other parameters means that $\tilde{\lambda} \neq \hat{\lambda}$.

2.3.2 Weighted models fitted to the transformed data

Zimmermann and Münnich (2018) consider three approaches for informative sampling for use with log transformation: one is based on including the design (eg stratification) variables as predictors in the models, and a second strategy involves augmenting the fitted model with a variable which is a function of the selection probabilities (Verret et al., 2015). Both of these approaches require that the model

variables (either the design variables or the augmenting variable) are available for all the non-sampled units in the population; this will often be true because of the availability of a business register, and the privacy constraints which mean that it is often the organisation designing the survey which undertakes the estimation. Nevertheless Verret et al. (2015)'s procedure requires the selection probabilities for non-sample units, which are not typically stored, so we do not consider this approach further here.

The third procedure needs only the weights for sampled observations, and follows the approach of You and Rao (2002) adapted to the log-transformed data by Zimmermann and Münnich (2018), independently developing almost the same model as Guadarrama et al. (2018) used in the PEBP in section 2.3.1. First estimate the weighted model parameters $\boldsymbol{\beta}^{*SWEE}$ using a survey weighted estimation equation (and the estimator is therefore labelled SWEE):

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{*SWEE} = &\left(\sum_{ind}\sum_{i\in ind\cap s} w_{ind,i}\mathbf{x}_{ind,i}(\mathbf{x}_{ind,i} - \hat{\gamma}^{*w}_{ind}\bar{\mathbf{x}}^{w}_{ind})^{\mathrm{T}}\right)^{-1} \\ &\times\left(\sum_{ind}\sum_{i\in ind\cap s} w_{ind,i}(\mathbf{x}_{ind,i} - \hat{\gamma}^{*w}_{ind}\bar{\mathbf{x}}^{w}_{ind})y^{*}_{ind,i}\right)\end{aligned} \tag{17}$$

where the last component of the last bracket on the right hand side has been corrected from Zimmermann and Münnich (2018, equation (9)). Then estimate the random effect accounting for the survey weights using:

$$\tilde{u}^{*}_{ind} = \hat{\gamma}^{*w}_{ind}\left(\bar{y}^{*w}_{ind} - (\bar{\mathbf{x}}^{w}_{ind})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{*SWEE}\right) \tag{18}$$

where $\bar{y}^{*w}_{ind} = \sum_{i\in s_{ind}} w_{ind,i}y^{*}_{ind,i}/\left(\sum_{i\in s_{ind}} w_{ind,i}\right)$ and $\hat{\gamma}^{*w}_{ind}$ is as before. Finally, the weighted small area estimator using predictions for the out-of-sample units is

$$\hat{y}^{SWEE}_{ind} = \sum_{i\in ind\cap s} y_i + \sum_{i\in ind\cap r}\exp\left[\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{*SWEE} + \tilde{u}^{*}_{ind} + \frac{1}{2}\hat{\sigma}^{*2}_{e}\left(\hat{\gamma}^{*w}_{ind}\delta^{2}_{ind} + 1\right)\right]. \tag{19}$$

3 AIDA dataset and SBS design

3.1 A population of retail businesses

Smith et al. (2021) used data from the retail sector in the Netherlands; these data were not available for further research, so we have used a similar population of retail businesses in Italy derived from the AIDA database (Bureau van Dijk, 2015). AIDA stands for "Analisi Informatizzata delle Aziende Italiane" (which can be translated as "Italian company information and business intelligence"). AIDA is a database of information on Italian companies, now created and distributed by Moody's, containing the balance sheets, registry and product data of all active Italian companies, excluding banks, insurance companies and public bodies (https://www.moodys.com/web/en/us/capabilities/company-reference-data/orbis/aida-orbis-for-italy.html). It therefore covers many, but not all, of the businesses in Italy. Such companies have a higher number of employees on average than sole proprietors or partnerships, so AIDA probably does not completely reflect the true distribution of all Italian businesses, although we have not attempted a formal comparison with ASIA, the Italian Business Register, which is not publicly available. Although it is a subgroup, it contains very many micro businesses, and is composed of a wide range of company types. Therefore we believe that it provides a realistic business population with which to evaluate small area estimation methods.

The AIDA database is not publicly available, but is available by subscription. It contains ten years of data at any given time, and is regularly updated. In using it, we have therefore needed to take a snapshot at a particular moment, which does not necessarily reflect the live information in the database. To continue the investigations from Smith et al. (2021) we looked for a similar dataset, so have extracted the information for retail businesses in Italy in operation in 2018-2020 (92123 observations in total). A wide range of variables is available in the dataset, but we focus our attention on 'revenue' (which we take to approximate the statistical definition of turnover).

We make some modifications to obtain a known population with complete information, which is needed in order to have 'the truth' against which to compare the estimates from the different small area procedures described in section 2. We are interested in the values of turnover and number of employees in 2020, so we first exclude any businesses with missing values on any of these variables (16324 observations deleted). Businesses are classified according to NACE Rev. 2 (Eurostat, 2008), and retail businesses fall

within the 2-digit NACE code 47 "Retail trade". Upon further investigation we discover that petrol retailers (NACE 4730) have unusual characteristics, specifically (because of their structure) no or almost no employees and very high turnover, so we exclude this industry from our dataset, reasoning that their activity is not typical and would generally be completely known in any business survey (2220 are removed). We also remove the businesses with more than 50 employees in all the remaining industries, since these would be completely enumerated in a survey and therefore fully known, in line with the approach we previously took with the Dutch dataset (1305 observations deleted).

It is necessary to do a light-touch editing of the population, and specifically we remove businesses with a NACE code at the 2- or 3-digit level only (there are only 485 such businesses), and improbable cases – businesses which are extremely large in 2018 and very small in 2020 (there are only 221 such businesses).

We also need some auxiliary information on which to base a sample design and sample selection, and in keeping with practices for Structural Business Surveys in several countries (including UK, Netherlands) we use the information from two years previously (2018) as the basis of the sample design (see below). The sample is drawn from strata defined by NACE $\times$ size class in 2020. The revenue from 2018 is then used as an explanatory variable in the model. There are records in our dataset, however, which do not have a matching record in 2018 (14754 businesses). To avoid missing values in the auxiliary information (due to nonresponse or non-existence, which are not distinguished), we impute 0 for all businesses that were non-existent in the database before 2019 (13955 observations, the great majority of them missing units) and for those existing but with revenues equal to zero in the previous years (118 observations). For the remaining units with missing values (681 businesses) the revenue value is imputed with the average of revenue in 2019 and 2017.

The net result is a dataset of 71568 businesses, with no missing values for the outcome variables from 2020, or in the auxiliary variables from 2018. At the end of this process, we treat the resulting list of businesses as if it were the complete population, from which we will subsample to simulate the survey process. The characteristics of the population are summarised in Table S1 in the supplementary material.

The resulting dataset includes the following variables, with which to construct models to predict the 2020 turnover:

- turnover in 2018, t^{2018}
- working persons in 2020 (that is the number of actual employees, not full-time equivalents), wp
- size class in 2020, based on the number of employees in the business in bands 1, 2–4, 5–9, 10–19, 20–49, sc
- industrial classification, ind

The choice of 2018 and 2020 for our analysis dataset means that the response variable is affected by the period of COVID disruption and lockdowns, when the predictive performance of the models might be expected to be more uncertain than usual (Smith and Lorenc, 2021). A summary of correlations between the numeric variables is given in Table 2.

	t^{2018}	$\log(t^{2020}+1)$	$\log(t^{2018}+1)$	wp
t^{2020}	0.871	0.657	0.547	0.622
t^{2018}		0.567	0.627	0.576
$\log(t^{2020}+1)$			0.790	0.722
$\log(t^{2018}+1)$				0.626

Table 2: Correlation between turnover in 2020 and 2018, its log-transformed versions and employment in 2020. The symbols are defined in the text.

3.2 A SBS-like design

Once the population is defined, we subset the 2018 information on 'revenue' within strata defined by industrial classification (NACE 4-digit code) and number of employees, using fixed values 1, 2–4, 5–9, 10–19, 20–49 of the latter in line with the previously used Dutch SBS design (Krieg et al. 2012, Appendix A; Smith et al. 2021). This gives 36 industries and 5 size strata. We then calculate estimates of the population variance of revenue from 2018 in each stratum. Several of these estimates are missing (usually

because the population size in stratum h, $N_h < 2$), and one estimate (for the smallest size stratum in NACE 4740) is extreme because it is affected by outliers, so we also set this value to missing. We impute the missing variances using a simple linear model with fixed effects for industries and size classes. Because we are estimating variances from the population we may do better in design than in a real-life situation where they are estimated from a previous sample. We do not concern ourselves overmuch with optimising the design, however, since allocation in stratified samples is known to have a flat optimum (Brewer and Gregoire, 2009, p28).

The population variances (estimated using 2018 revenue) and the population sizes within strata defined in the same way, but using 2020 classification and employment information, are then used as inputs to a Neyman allocation (Neyman, 1934; Cochran, 1977) with a sample size of 5,000, again similar to the previous study. We impose a minimum sample size of 5 in each stratum (or N_h if $N_h < 5$). This generates a design with sampling fractions realistically similar to those actually used in SBS. The resulting allocation is used in all our repeated sampling; a summary of the sample design is given in table S2 in the on-line supplementary material.

4 Application to example data with known outcomes

The simulation is design-based, with each replicate a probability selection from the AIDA dataset described in section 3.1 using the stratified design from section 3.2 and Table S2. We use 1000 replicates. The design ensures that some sample observations are obtained in every industry *ind*, so the pseudopopulation generation in the EBP uses only the in-sample approach in step 3(a) in section 2.2.1. We briefly assess the out of sample approach (the "census EBP") in section S5.2.

With each replicate sample we fit a model (the same as used by Smith et al. (2021)) to turnover in 2020, t^{2020}, which we treat as the survey response, with the i^{th} component of $\mathbf{X}\boldsymbol{\beta}^*$ in equation (5) given by

$$\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* = \beta_0 + \beta_1 t_{i,ind}^{2018} + \boldsymbol{\beta}_2 \mathbf{sc}_{i,ind} + \beta_3 wp_{i,ind} + \beta_4 \left(t^{2018} \times wp\right)_{i,ind} \tag{20}$$

when 2018 turnover (t^{2018}) is used directly as a predictor or

$$\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* = \beta_0 + \beta_1 \log(t_{i,ind}^{2018} + 1) + \boldsymbol{\beta}_2 \mathbf{sc}_{i,ind} + \beta_3 wp_{i,ind} + \beta_4 \left(\log(t^{2018} + 1) \times wp\right)_{i,ind} \tag{21}$$

when $\log(t^{2018} + 1)$ is used. In the AIDA dataset, revenue ("turnover") is given in €, but in (20) and (21) we divide by 100,000, as this makes the estimation procedure much more stable. In particular this means that when we use eg $\log(t^{2018} + 1)$, the +1 represents +€100,000 (see also section S5.4 in the supplementary material). **sc** represents the dummy variables of size classes of the businesses, which are the same as the strata set up in section 3.2, with classes 1-5 defined by 1, 2–4, 5–9, 10–19, 20–49 working persons *wp* respectively. Although it seems odd to include both *wp* and **sc** based on the same underlying variable, using both variables improves the quality of the small area estimates. It seems likely that this is because **sc** is a design variable, so adding it means that the design variables are fully included in the model (and this applies equally to Smith et al. (2021) where this point was not made).

We make estimates for domains which are industries. Although these are designed domains (part of the sample design), they still suffer from small sample sizes (sample sizes range from 12 to 905, with 44% with a sample of 30 or fewer), so there is a strong case for using small area estimation. Our primary aim is to illustrate the performance of the various estimators on realistic data and estimation scenarios. We use the relative bias $rb(\hat{y}_{ind}^{\circ}) = \frac{100}{y_{ind}}\left[\frac{1}{1000}\sum_{k=1}^{1000}\left(\hat{y}_{ind,k}^{\circ} - y_{ind}\right)\right]$ and relative root mean square error $rrmse(\hat{y}_{ind}^{\circ}) = \frac{100}{y_{ind}}\left[\frac{1}{1000}\sum_{k=1}^{1000}\left(\hat{y}_{ind,k}^{\circ} - y_{ind}\right)^2\right]^{\frac{1}{2}}$, where $\hat{y}_{ind}^{\circ}$ represents any of the estimators of the total for industry *ind*, to evaluate the performance of the different transformations and estimators. The estimators are compared by computing their empirical efficiency relative to the HT estimator, as

$$\frac{mse(\hat{y}_{ind}^{\circ})}{mse(\hat{y}_{ind}^{HT})}100 = \frac{\sum_{k=1}^{1000}\left(\hat{y}_{ind,k}^{\circ} - y_{ind}\right)^2}{\sum_{k=1}^{1000}\left(\hat{y}_{ind,k}^{HT} - y_{ind}\right)^2}100.$$

4.1 Comparison of the estimators

The estimators used in the repeated sampling simulations are summarised in Table 3. For all the transformations except the log-shift a deterministic shift parameter is required, initially taken as $s = 1$ (= €100,000), though we explore the sensitivity to this in section 4.2.2.

Estimator	transformation	acronym	symbol	description	weighted
HT	none	HT	$\hat{y}_{ind}^{HT}$	(1)	yes
GREG	none*	GREG	$\hat{y}_{ind}^{GREG}$	(2)	yes
EBLUP	none*	EBLUP	$\hat{y}_{ind}^{EBLUP}$	(4)	no
EBP	none*	EBP linear	$\hat{y}_{ind}^{EBP}$	section 2.2.1	no
EBP	log	EBP log		& table 1	no
EBP	log-shift	EBP log-shift			no
EBP	Box-Cox	EBP Box-Cox			no
EBP	dual power	EBP dual power			no
Karlberg-type synthetic	log	K	$\hat{y}_{ind}^{K}$	(6)	no
Model based direct	log	MBD	$\hat{y}_{ind}^{MBD}$	(8)	no
EB	log	EB	$\hat{y}_{ind}^{EB}$	(10)	no
EB bias corrected	log	EBbc	$\hat{y}_{ind}^{EBbc}$	(11)	no
PEBP MM	none*	PEBP MM linear	$\hat{y}_{ind}^{PEBP}$	section 2.3.1	yes
PEBP MM	log	PEBP MM log			yes
PEBP ML	none*	PEBP ML linear			yes
PEBP ML	log	PEBP ML log			yes
PEBP ML	log-shift	PEBP ML log-shift			model only
PEBP ML-$\hat{\lambda}^w$	log-shift	PEBP ML-$\hat{\lambda}^w$ log-shift			$\hat{\lambda}^w$ & model
PEBP ML	Box-Cox	PEBP ML Box-Cox			model only
PEBP ML-$\hat{\lambda}^w$	Box-Cox	PEBP ML-$\hat{\lambda}^w$ Box-Cox			$\hat{\lambda}^w$ & model
PEBP ML	dual power	PEBP ML dual power			model only
PEBP ML-$\hat{\lambda}^w$	dual power	PEBP ML-$\hat{\lambda}^w$ dual power			$\hat{\lambda}^w$ & model
SWEE	log	SWEE	$\hat{y}_{ind}^{SWEE}$	(19)	yes

Table 3: Overview of the estimators used in the repeated sampling simulations. For abbreviations see text. No transformation ("none") is described as a linear transformation in the figures. With PEBP, MM indicates the method of moments estimator of You and Rao (2002) as implemented in `emdi` and ML indicates maximum likelihood estimation as implemented in `povmap`. * indicates a model using t^{2018} as the predictor; otherwise $\log(t^{2018}+1)$ is used.

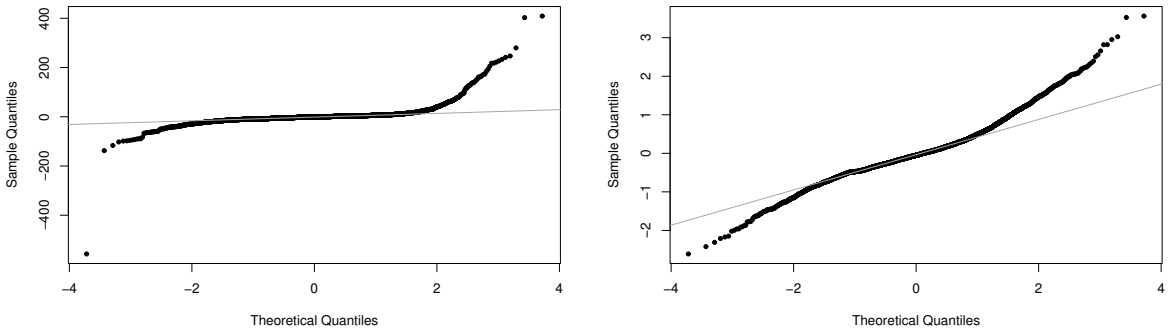
Some pairs of estimators are based on the same underlying model, but fitted in a different way. The EB with log transformed data and the EBP with log transformation both use the same model, but the EBP draws from the posterior distribution of the parameters given the data, and the EB relies on a model fitted to the transformed data and a back transformation. The PEBP and SWEE estimators are similarly closely related versions of the same model (as demonstrated in sections 2.3.1 and 2.3.2). We find that the results with these pairs differ only in very minor ways.

The diagnostic plots for model (20) (Fig. 1(a) and S4(a) (in the supplementary material)) show that the residuals do not have a Normal distribution, particularly deviating in the tails. Across many of the estimators, there are some domains where the use of t^{2018} as a predictor (that is, model (20)) results in extreme outliers in some samples, leading to very large mean squared error estimates. We present a summary of the relative bias and relative root mean squared error (rrmse) results for these in table S3 (in the supplementary material). By contrast, model (21), using $\log(t^{2018}+1)$ as a predictor, has stable behaviour for repeated samples in all the domains, and is therefore strongly preferred. We focus mainly on this model for the remaining results, though reverting occasionally to model (20). Summary information on the relative bias and rrmse of the different estimators when using model (21) are presented in Tables 4 and 5. Note that the mean and median are sometimes quite different (eg for the MBD), reflecting extreme relative bias or rrmse estimates in some industries.

The relative bias and relative rmse (rrmse) from the simulations are presented in Figs. 2 and 3 respectively. We investigated many minor variants of the PEBP (Table 3), and these are discussed in more detail in section 4.3.2. The bias is a major component of the rrmse, so that methods with better bias performance are also likely to have the lowest rrmse's. The estimators in Figs. 2 and 3 fall into seven groups based on their performance. This relates closely to their definitions in section 2, with variants of the same approach behaving more similarly than different underlying approaches. The plots show that the HT estimator is unbiased but rather variable, as expected, and the average rrmse is larger than would be desirable for publication of estimates. The GREG is already a substantial improvement, remaining essentially unbiased and with smaller rrmse than the HT estimator, even though it is potentially affected

Estimator		relative bias						
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Mean abs(rb)
HT		-0.9113	-0.2332	**-0.1147**	**0.0049**	0.2443	1.3500	0.3850
GREG*		-5.6826	-1.3926	**-0.2970**	**-0.3501**	-0.0195	7.4605	1.5378
EBLUP*		-95.9123	-51.6444	-26.9159	-29.5224	-14.1188	84.6590	35.8382
EBP	linear*	-95.3930	-51.5722	-26.9163	-29.5045	-14.2220	84.6420	35.8220
	log	-17.3265	-3.9598	3.4005	2.2144	9.2117	21.7481	7.8448
	log-shift	-17.3259	-3.9572	3.3930	2.2132	9.2078	21.7432	7.8438
	Box-Cox	-18.5211	-3.0907	3.5429	2.6517	9.2999	20.7777	8.2642
	dual power	-17.3291	-3.9592	3.4021	2.2128	9.2090	21.7315	7.8429
K		-37.4524	-10.3724	4.4645	7.6107	20.9458	81.6673	18.9938
MBD		-11.1641	-7.3472	-0.6744	15.8435	8.5725	243.0725	22.4253
EB		-17.3243	-3.9630	3.3832	2.2120	9.2284	21.6684	7.8404
EBbc		-17.3357	-3.9678	3.3354	2.1711	9.2186	21.4127	7.8200
PEBP MM	linear*	-11.3908	-1.9310	**0.0711**	2.8821	6.8037	27.0054	6.2846
	log	-22.6906	-0.0814	7.3507	6.7280	12.5928	42.0591	10.9370
PEBP ML	linear*	-10.3718	-1.2437	0.9249	4.1416	8.0965	39.8806	6.5609
	log	-24.4002	-2.8648	3.3759	3.3766	10.9408	37.6324	9.2163
	log-shift	-24.3988	-2.8665	3.3833	3.3760	10.9472	37.6291	9.2162
	Box-Cox	-24.1796	-0.5940	5.2995	5.4318	13.1991	35.4517	10.2556
	dual power	-24.4028	-2.8636	3.3817	3.3769	10.9439	37.6199	9.2171
PEBP ML-$\hat{\lambda}^w$	log-shift	-23.2886	-4.8675	1.1904	1.6720	7.0143	41.5010	8.8870
	Box-Cox	-22.3000	-5.3214	1.2807	3.2196	8.0862	48.1940	9.6037
	dual power	-22.2239	-4.6653	2.2394	4.0959	9.0204	49.4219	10.0154
SWEE		-22.6885	-0.0694	7.3284	6.7247	12.6050	41.9644	10.9310

Table 4: Summary information on the relative bias (from Tables S4 and S5) for the models listed in Table 3. * indicates a model using t^{2018} as the predictor; otherwise $\log(t^{2018}+1)$ is used. The best-performing models are indicated in bold.



(a) (b)

Figure 1: QQ plots of the enterprise-level (level 1) residuals from fitted models of the form of (3) with (a) the untransformed response and original (untransformed) predictors, (b) log-transformed response ($\log(t^{2020}+1)$) and using $\log(t^{2018}+1)$ in the predictors. All the plots are derived from the same sample, which is typical.

Estimator		relative rmse					
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
HT		2.7191	10.2471	14.9788	15.9882	20.0971	39.2614
GREG*		1.6916	5.8467	10.6431	12.3899	15.2088	40.4484
EBLUP*		4.3965	20.1616	29.8129	38.7345	56.9779	97.2907
EBP	linear*	4.3933	20.2698	30.1418	38.9980	57.0768	98.7193
	log	2.6073	6.8213	11.1612	**11.0608**	14.1698	25.4700
	log-shift	2.6085	6.8203	11.1543	**11.0602**	14.1727	25.4688
	Box-Cox	2.8242	7.8534	**10.9051**	11.5834	14.4741	22.7288
	dual power	2.6072	6.8240	11.1585	**11.0607**	14.1728	25.4540
K		1.8149	8.3213	16.3445	19.3313	24.6323	81.8067
MBD		3.7768	10.7936	14.3537	30.1459	19.0184	250.8476
EB		2.6004	6.8039	11.1509	**11.0345**	14.1422	25.3358
EBbc		2.5979	6.8096	11.1286	**11.0263**	14.1208	25.1259
PEBP MM	linear*	1.6848	5.9647	**8.6978**	**10.3777**	12.5745	35.4960
	log	2.6812	8.0585	12.3098	14.0707	16.9809	44.4042
PEBP ML	linear*	1.6308	5.5431	**8.3767**	**10.3252**	12.1076	42.2580
	log	2.3820	9.0740	12.1454	13.0585	15.5052	40.0942
	log-shift	2.3816	9.0782	12.1404	13.0586	15.5100	40.0933
	Box-Cox	3.6529	8.4648	13.1001	14.4017	18.4613	38.1119
	dual power	2.3831	9.0775	12.1480	13.0600	15.5109	40.0828
PEBP ML-$\hat{\lambda}^w$	log-shift	3.5108	8.2663	10.2779	12.3523	14.6613	43.8365
	Box-Cox	3.5369	7.8787	10.8167	12.8102	15.2398	50.1551
	dual power	4.0371	7.6026	11.4195	13.1071	15.3426	51.3172
SWEE		2.6785	8.0517	12.2524	14.0390	16.9688	44.2692

Table 5: Summary information on relative rmse (from Tables S6 and S7) for the models listed in Table 3. * indicates a model using t^{2018} as the predictor; otherwise $\log(t^{2018}+1)$ is used. The best-performing models are indicated in bold.

Estimator		empirical efficiency					
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
HT		100	100	100	100	100	100
GREG*		14.7	38.1	49.6	61.0	75.3	243.0
EBLUP*		13.6	193.7	366.4	2045.0	1061.0	30021.5
EBP	linear*	14.6	200.2	368.8	2051.5	1073.9	30032.1
	log	11.8	24.3	42.0	100.2	91.0	661.3
	log-shift	11.8	24.4	41.9	100.2	91.0	661.3
	Box-Cox	11.8	25.8	44.3	147.9	117.9	1263.9
	dual power	11.8	24.4	42.0	100.2	91.0	661.5
K		2.0	25.2	105.2	500.3	384.1	3544.7
MBD		15.5	57.6	89.0	920.3	167.1	18114.1
EB		11.8	24.3	42.1	100.0	90.7	661.2
EBbc		11.8	24.4	42.2	99.9	90.2	662.0
PEBP MM	linear*	4.6	28.8	38.7	55.2	50.5	385.4
	log	9.5	41.0	69.6	135.8	169.8	1126.9
PEBP ML	linear*	5.7	24.7	37.5	49.6	52.4	194.2
	log	10.1	32.1	71.5	125.3	111.5	1300.8
	log-shift	10.1	32.1	71.4	125.3	111.6	1300.7
	Box-Cox	8.5	37.6	76.7	181.7	153.2	1867.1
	dual power	10.1	32.2	71.5	125.3	111.6	1301.1
PEBP ML-$\hat{\lambda}^w$	log-shift	13.5	28.2	54.0	114.3	106.9	1185.6
	Box-Cox	11.7	29.6	60.4	120.1	128.2	1088.1
	dual power	10.5	31.5	64.4	122.1	146.3	1080.9
SWEE		9.4	40.8	69.1	135.5	168.8	1126.7

Table 6: Summary information on the empirical efficiency (from Tables S8 and S9) for the models listed in Table 3. * indicates a model using t^{2018} as the predictor; otherwise $\log(t^{2018}+1)$ is used.

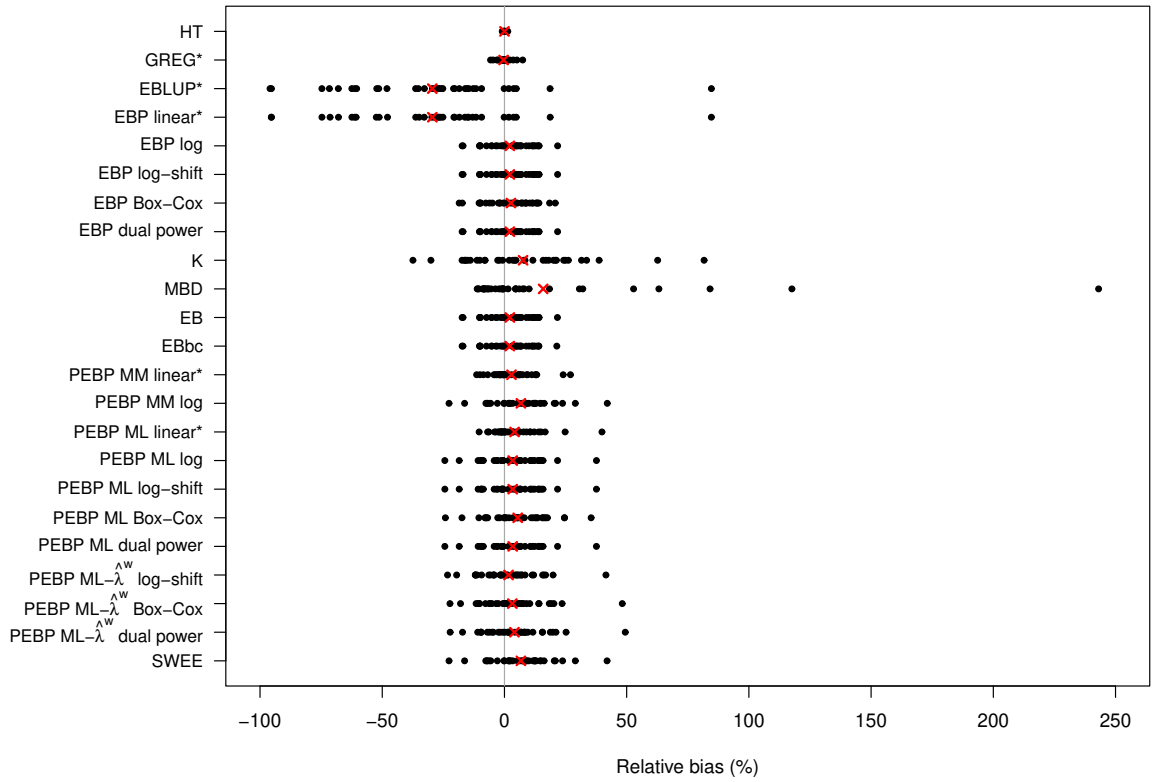


Figure 2: Linear scatterplots of the relative bias of the different estimators by industry; the mean bias across all industries is shown by a cross. The labels are summarised in Table 3. The detailed results can be found in Tables S4 and S5 in the supplementary material. * indicates a model using t^{2018} as a predictor; otherwise $\log(t^{2018}+1)$ is used.

by the outliers in the data.

Naïve application of the EBLUP estimator (3), widely used for small area estimation, clearly produces estimates with large biases and hence large rrmse's, and this situation is repeated (with very minor differences) using the EBP with untransformed data (EBP linear). The inclusion of the weights in the PEBP ameliorates the effect of the skewed data when using the untransformed predictor (PEBP linear). In fact this model has the lowest mean and median rrmse across the industry estimates of any of the tested approaches, including the PEBP with the transformed data, and this is largely because the bias in some estimates is reduced. The most extreme rrmse is however larger for this approach than for some other models. We discuss the impact of weights further in section 4.3.2.

Of the unweighted estimators, the Karlberg-type and model-based direct estimators both suffer from large biases in some industries which give them poor rrmse performance overall. The EBP has good performance, comparable with the PEBP, but the industries where the estimator is worst are not so extreme as in the PEBP. The EBP shows similar performance with all the transformations considered, largely because the fitted transformations are close to the log transformation for this dataset (see Table 7). The EB and EBbc estimators have similar performance to the EBP (again because the log transformation seems to be appropriate for these data), and the second-order bias correction of the EBbc gives a minor improvement in the mean and median rrmse.

Table 6 shows the effect of the small area estimation. Our example includes a range of area sizes. The small area estimates may not be an improvement where the sample size is adequate for direct estimation, and this is reflected in high maximum empirical efficiencies and consequent high values for the mean empirical efficiency. The median and (in the best performing approaches) third quartile of the empirical efficiencies are, however, < 100. Therefore, for most of the small areas the small area estimation approaches result in reduced rmses as expected. The relationship between the empirical efficiency and sample size is shown for the EBP dual power and PEBP linear in supplementary material Figs. S1 and S2.

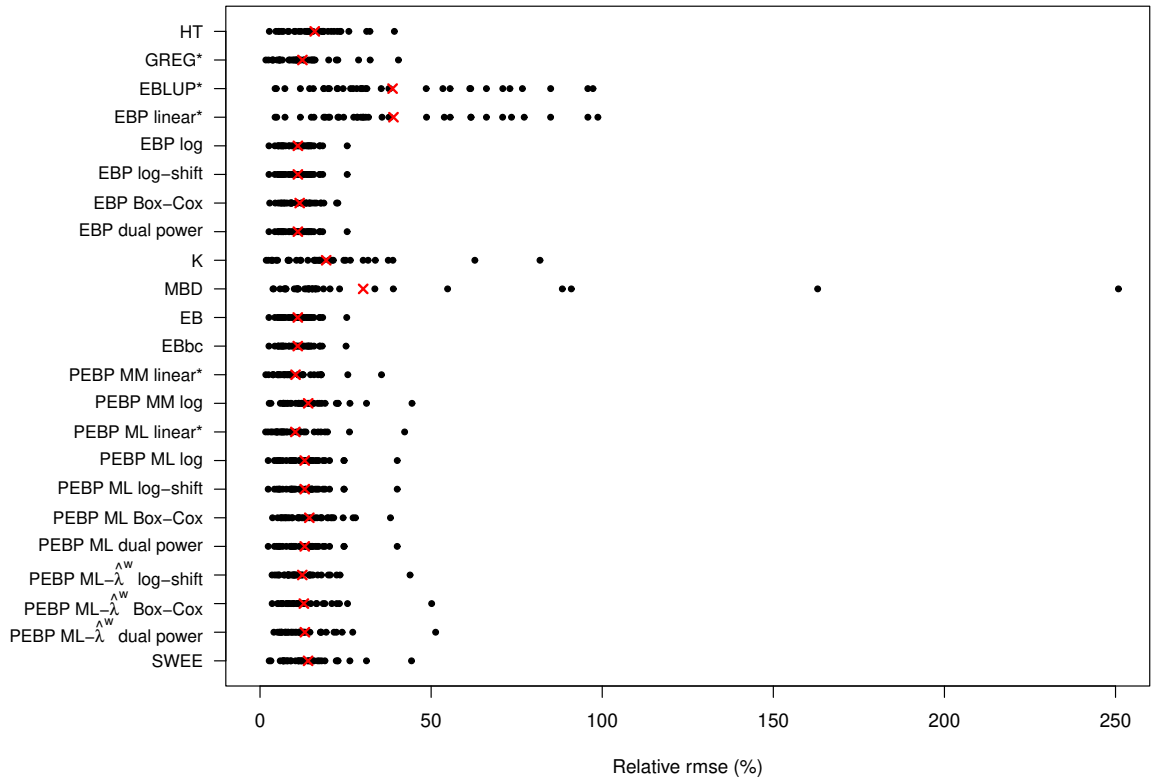


Figure 3: Linear scatterplots of the relative rmse of the different estimators by industry; the mean rrmse is shown by a cross. The labels are summarised in Table 3. The detailed results can be found in Tables S6 and S7 in the supplementary material. * indicates a model using t^{2018} as a predictor; otherwise $\log(t^{2018}+1)$ is used.

transformation	$\hat{\lambda}$	$\hat{\lambda}^w$
log-shift	1.000055	1.992512
Box-Cox	−0.062347	0.178979
dual power	0.000060	0.305633

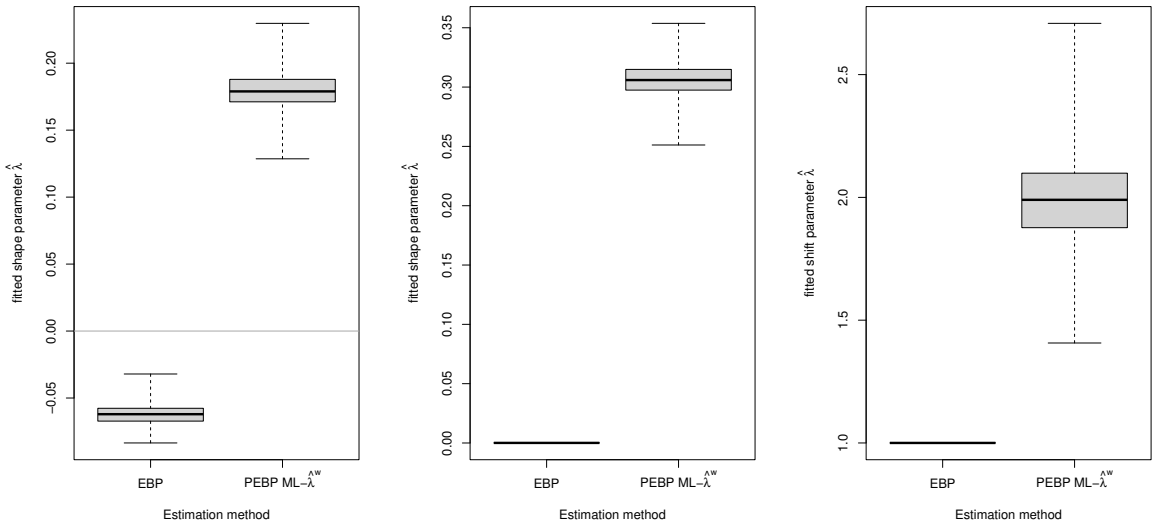
Table 7: Mean over repeated sampling simulations of the fitted value of the unweighted transformation parameter $\hat{\lambda}$ and weighted transformation parameter $\hat{\lambda}^w$. For the Box-Cox and dual power transformations the shift is deterministically set at 1; see section 4.2.2 for an assessment of the sensitivity to this choice. Recall (section 4) that a shift of +1 actually means +€100,000.

4.2 Transformation and the distributions of data and residuals

One of the assumptions behind the multilevel models used within the unit model in small area estimation is that the residuals (at both unit and area levels) follow normal distributions. The basic idea behind transformations is to produce data which more nearly have a normal distribution of residuals from the unit level model. Note that the transformation is aimed at the *residuals* and not at the original data. It is not always necessary for the empirical distribution of the residuals to be specifically normal, as the models are reasonably robust to departures, but approximate symmetry of the empirical distribution is important (as suggested by Rojas-Perilla et al. (2020)).

4.2.1 Fitted transformations and their effects

First we consider the effect of the fitted transformation parameters. Fig. 4(a) & (b) show the fitted values of $\hat{\lambda}$ and $\hat{\lambda}^w$ for the Box-Cox and dual power transformations. The fitted $\hat{\lambda}$ are close to 0 and therefore indicate that the distribution is most suitably transformed with something very close to the log. Although in section 2.3.1 we expected that the influence of the fitting of the other parameters would



(a) (b) (c)

Figure 4: Boxplots of the fitted values of the unweighted transformation parameter $\hat{\lambda}$ and weighted transformation parameter $\hat{\lambda}^w$ in (a) the Box-Cox, (b) the dual power and (c) the log-shift transformation over the repeated samples.

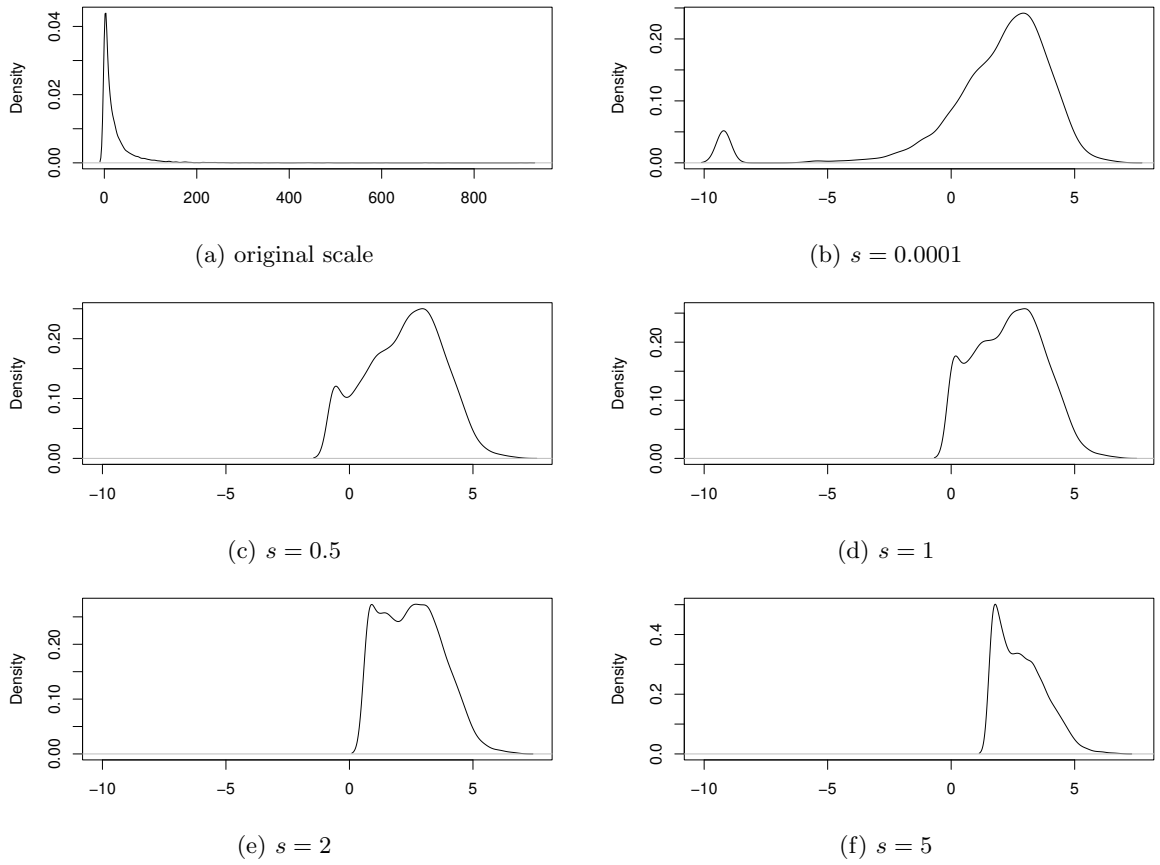
mean that $\tilde{\lambda}$ would differ from $\hat{\lambda}$, in practice we discover that this is not so, and the fitted values $\tilde{\lambda} = \hat{\lambda}$, so they are not shown separately. $\hat{\lambda}^w$ are however different from zero in both transformations, indicating that the weighting has an influence on the shape of the distribution of the response variable leading to a transformation different from the standard logarithmic one.

Since the target of the transformation is to make the residuals approximately normal, or at least symmetric, we show example QQ plots in Fig. 1, focusing on the element level (the area level QQ plots are shown in supplementary material Fig. S4). These demonstrate that before transformation the residuals are skewed and, particularly in the tails, very far from normally distributed. The models with log and other transformations still depart from normality in the tails, but less than before transformation. Their distributions are still asymmetric, however – the excess kurtosis in Fig. 1(a) is 2.514 and in Fig. 1(b) is 2.503, so the skewness before and after transformation are practically indistinguishable. (The log, Box-Cox and dual power transformation QQ plots are practically indistinguishable because the Box-Cox ($\hat{\lambda} = -0.062$) and dual power ($\hat{\lambda} = 0.00006$) transformations closely approximate the log transformation, which is attained when $\lambda = 0$. The QQ plots for Box-Cox and dual power are shown in supplementary material Fig. S3.)

4.2.2 Sensitivity to the shift parameter in transformations

The transformations in section 2.2.1 require that the input data are strictly positive, so the data need to be shifted by adding a positive constant if there are zeroes and/or negative values. The shift is fitted (with or without weights) in the log-shift model (Edochie et al., 2023), but otherwise it needs to be provided as an input. With the AIDA business survey dataset, the unweighted estimation in the log-shift model produces $\hat{\lambda} \approx 1$ (= €100,000) consistently across selected samples, and the weighted estimation produces $\hat{\lambda}^w \approx 2$ (Table 7) with some variation between samples; the distribution over the simulations is summarised in Fig. 4(c).

But for other transformations where the shift is provided as an input it is interesting to look at the effects of the shifts. We consider a grid of values $\{0.0001, 0.5, 1, 2, 5\}$. The minimal constant 0.0001, naïvely aimed at having the smallest impact on the original data, creates a spike in the distribution of the transformed data (Fig. 5(b)). All the values give a simple translation of the original data before transformation, but the impact of this on the distribution of the transformed data is surprising – Fig. 5(b)-(f) shows the distribution going from left- to right-skewed over our small set of values. We might say that the transformed distribution is approximately symmetric for $s = 0.5, 1$ and 2. The effect is relatively pronounced in our example population, because there are about 8% zero values. An alternative approach might be to use a zero-inflated model as the basis for small area estimation (Krieg et al., 2016); we leave



(a) original scale

(b) $s = 0.0001$

(c) $s = 0.5$

(d) $s = 1$

(e) $s = 2$

(f) $s = 5$

Figure 5: Examples from one simulation of the sample distribution (smoothed with a kernel density estimate) of the AIDA turnover data (a) on the original scale; and log-transformed with deterministic shift (b) $s = 0.0001$, (c) $s = 0.5$, (d) $s = 1$, (e) $s = 2$ and (f) $s = 5$.

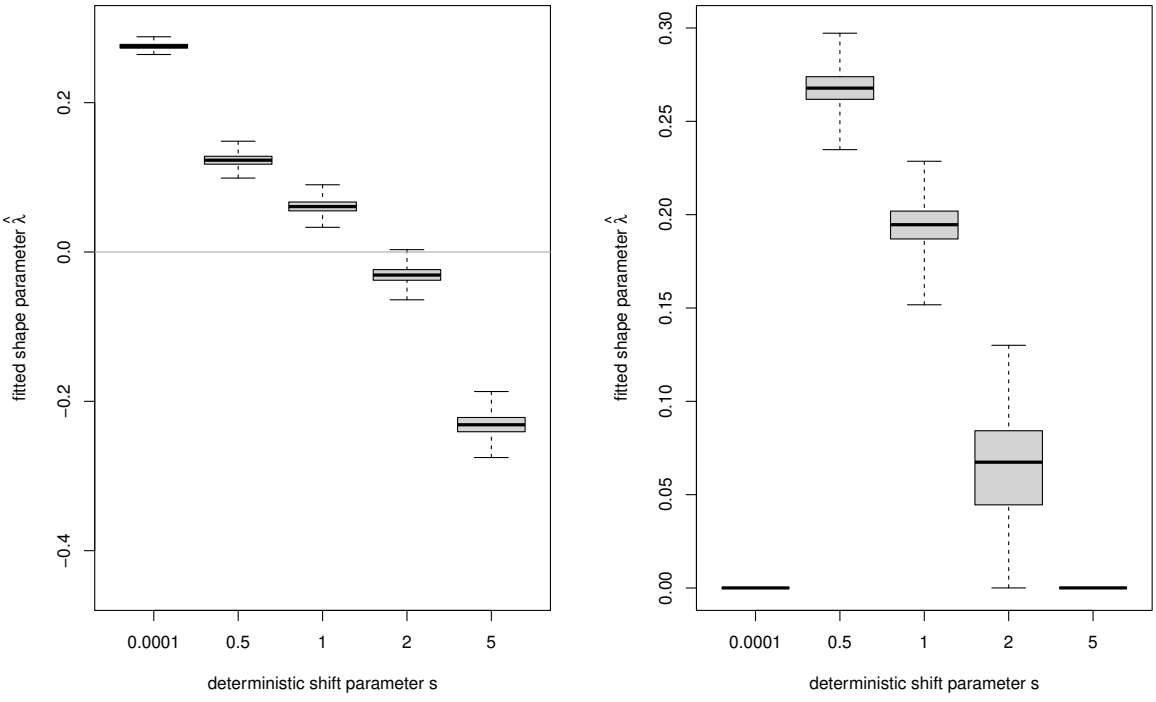
this for further research.

The consequence of treating these zero values is that fitted shape parameters in the Box-Cox and dual power transformations are quite different with the different shift parameters. Fig. 6 shows the distributions of these fitted parameters over the simulations. Because the shape parameters are fitted with the models, they differ for different choices of predictors (in this case with t^{2018} or $\log(t^{2018} + 1)$), even though the transformation works on the response variable which is the same in each case.

The Box-Cox shape parameter shows a monotonic decrease as the shift parameter increases, similarly for the models with t^{2018} and $\log(t^{2018} + 1)$. So the transformation changes from one which adjusts the data less than logarithmically to one that makes a greater adjustment, which reduces the impact of the larger values (and increases the impact of smaller values) more. The pattern for the dual power transformation is less easy to interpret. The model with t^{2018} has a similar decrease in the shape parameter for $s = 0.5, 1, 2, 5$, but both $s = 5$ and $s = 0.0001$ have a shape parameter $\hat{\lambda} \approx 0$ consistently across the repeated samples, indicating that the log transformation is appropriate. With $\log(t^{2018} + 1)$ the log transformation is consistently fitted for all the shifts except for $s = 0.5$.

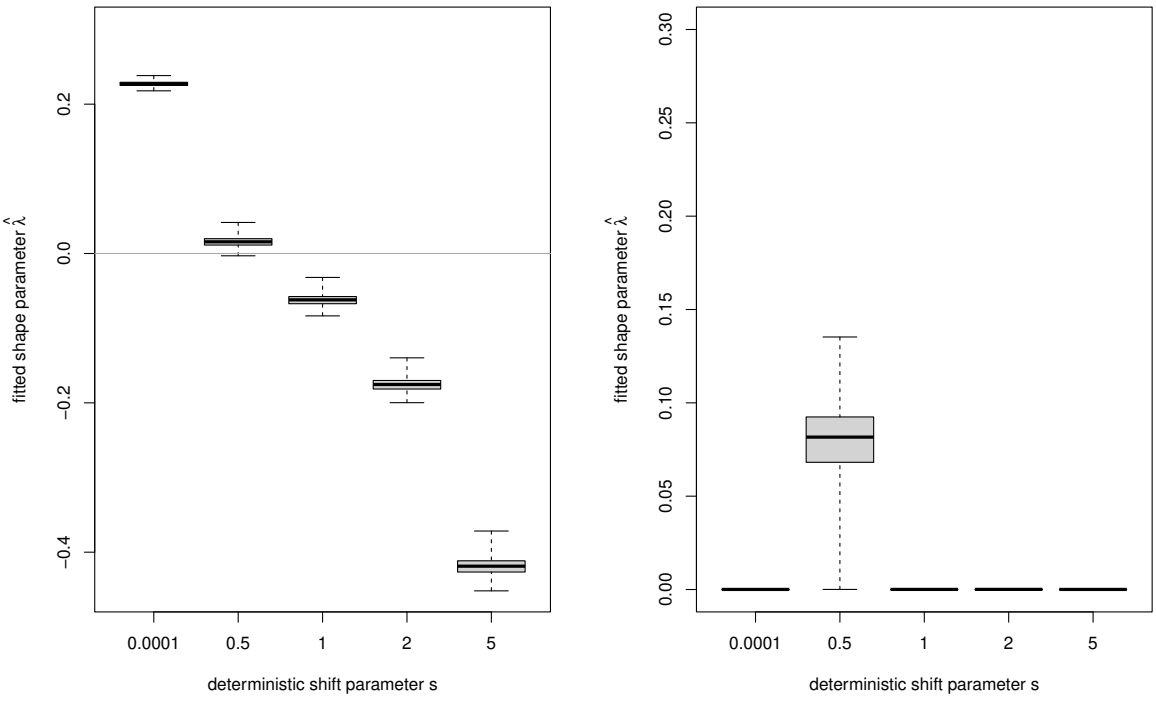
All these effects are on the distribution of the data, but it is the distribution of the residuals which affects the effectiveness of the small area models. We therefore present example QQ plots from the models fitted with the transformed data with different shifts in Fig. 7. The use of a minimal shift parameter clearly induces a large departure from normality, but the other shifts produce better results, though still with some departures in the tails, which are more symmetric but show slightly larger departures to the right. There is a small suggestion that the residuals are closer to normal as s increases, but there is not enough information to turn this into a conclusion at this stage. The standard $s = 1$ looks reasonable as a rule of thumb.

The effect of the different values of the shift parameter on the bias of the estimates from the EBP is shown in Fig. 8, and the resulting root mean squared errors (which are not very different as they are dominated by the bias component) in Fig. 9. There seems to be a risk to the models with a shift



(a)

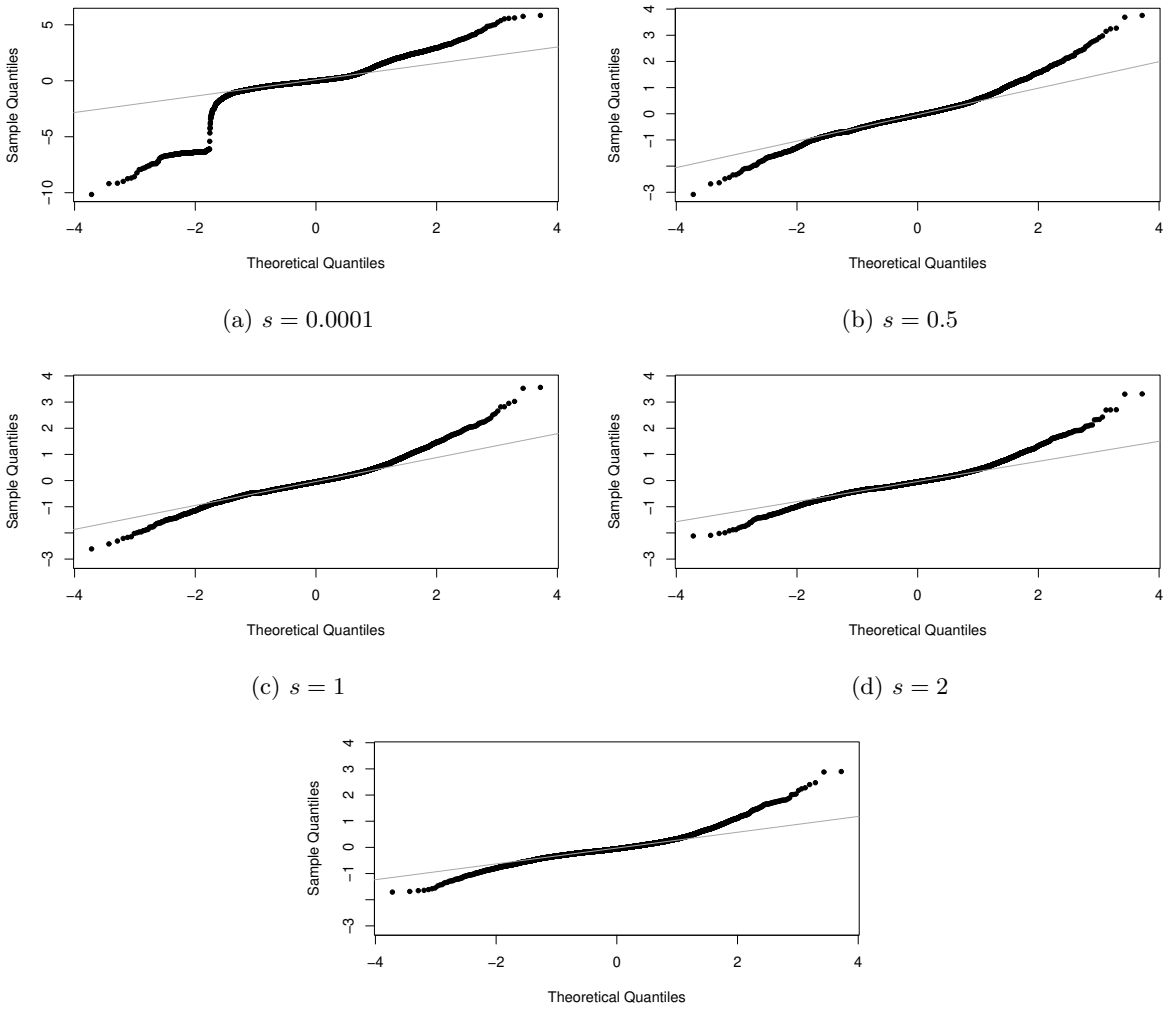
(b)



(c)

(d)

Figure 6: Boxplots of the fitted shape parameter $\hat{\lambda}$ in the EBP (see Table 1) with the Box-Cox (a and c) and dual power transformations (b and d), over repeated sampling simulations when using the deterministic shift parameter $s \in \{0.0001, 0.5, 1, 2, 5\}$. (a) and (b) are fitted on the models with t^{2018} and (c) and (d) are for the model with $\log(t^{2018} + 1)$.



(a) $s = 0.0001$

(b) $s = 0.5$

(c) $s = 1$

(d) $s = 2$

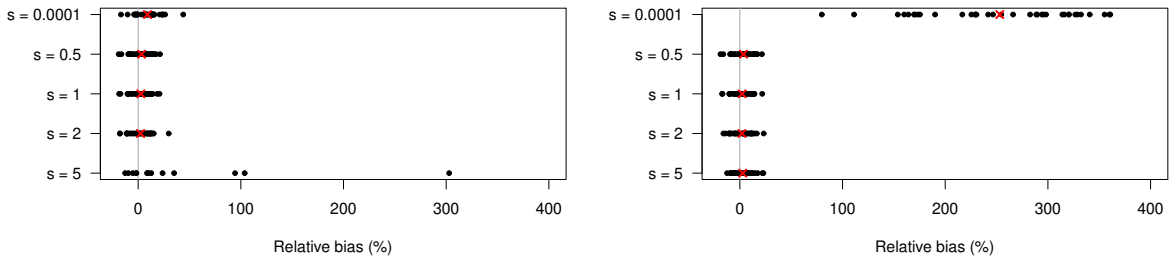
(e) $s = 5$

Figure 7: QQ plots of residuals from fitted models of the form of (3) with log-transformed data using the deterministic shift parameter $s \in \{0.0001, 0.5, 1, 2, 5\}$. All the plots are derived from the same sample (and the same as that used in Fig. 1, which is typical). Note that the y axis scale is different for the $s = 0.0001$ plot.

parameter that is too large or too small, where some of the estimates may have large biases and therefore large rrmse's, whereas there seems little impact on the quality of estimates from values nearer to 1 in this dataset. It is not clear how far this can be generalised to other data, but the differences are instructive.

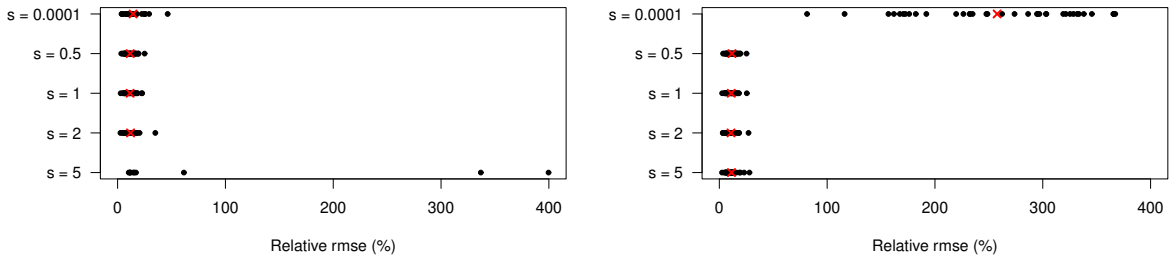
4.3 Sensitivity

Smith et al. (2021) investigated the sensitivity of models to extreme outliers by using a population with the most extreme outliers with respect to the model removed; this did not change their assessment of the relative performance of the different approaches, although there were some differences of detail in the estimates. We therefore do not repeat that investigation with the current set of models. There are, however, some further areas of sensitivity (in addition to the one already addressed in section 4.2.2) where we do make an assessment – a comparison of the EBP with direct modelling in subsection 4.3.1, an examination of different variants of the PEBP in subsection 4.3.2, a comparison of bias correction approaches in subsection 4.3.3, and a comparison of out-of-sample predictions with the census EBP in section S5.2 in the supplementary material.



(a) (b)

Figure 8: Linear scatterplots of the relative bias of the EBP with the Box-Cox (a) and dual power transformations (b) using the deterministic shift parameter $s \in \{0.0001, 0.5, 1, 2, 5\}$. The mean relative bias across industries is represented by a red cross. Note that many points and the mean relative bias are beyond the range of the plot for $s = 5$ for the Box-Cox transformation.



(a) (b)

Figure 9: Linear scatterplots of the rrmse of the EBP with the Box-Cox (a) and dual power transformations (b) using the deterministic shift parameter $s \in \{0.0001, 0.5, 1, 2, 5\}$. The mean rrmse across industries is represented by a red cross. Note that many points and the mean rrmse are beyond the range of the plot for $s = 5$ for the Box-Cox transformation.

4.3.1 EBP vs direct modelling

Where the same underlying models are used in (i) the EBP and (ii) directly modelling the transformed data with an appropriate bias correction in the back transformation, the results are qualitatively indistinguishable. It therefore seems that the back transformation is working very well in our dataset and introducing at most minimal error. It also suggests that EBP and PEBP provide a simple and practical approach to fitting the more complex models with a reduced need for theoretical derivation. We do, however, note that the census EBP produces outcomes with larger differences to the direct modelling of the transformed data, because of the process of replacing the sample observations with estimates (see section S5.2).

4.3.2 Weights

The PEBP with untransformed data is the best performing model (Tables 4-6, Fig. 3, section 4.1), which suggests that the weights are more important in compensating for the skewness of the response variable than the transformations. We reason that this is because the weights are calibrated to known population totals on the original scale, so that they are less effective at adjustment within the PEBP when used with the transformed data.

In the PEBP we also have the choice of whether to use the weights in the calculation of either the shift parameter of the logshift transformation or the shape parameters of the Box-Cox and dual power transformations. Intuitively it feels appropriate that the weights should be applied consistently in all the parameters of the model including the shift/shape parameters. The number of potential combinations is

quite large, and Fig. S7 in the supplementary material includes some additional to Table 3. The results (see Tables 4 and 5, with more details in Tables S5 and S7 in the supplementary materials) are not easy to interpret, but we deduce that

- there is some evidence (in Table 5) that the maximum likelihood method (Edochie et al., 2023, `povmap`) produces estimates with lower mean rrmses across industries than the method of moments approach implemented in `emdi` (Kreutzmann et al., 2019; Skarke et al., 2021), though this is not completely consistent (the PEBP with untransformed response has larger outlying rrmse's under the maximum likelihood approach).

- using the weighted $\hat{\lambda}^w$ with the untransformed predictor t^{2018} produces good results with the Box-Cox and dual power transformations, but not with the log-shift which is affected by poor results in several industries (results not shown, but available from the authors on request).

- using the weighted $\hat{\lambda}^w$ is generally better (in mean rrmse) than using the unweighted $\hat{\lambda}$ when the log predictor and transformations of the response variable are used (Table 5, though there is one exception). But the biggest rrmse is larger when using $\hat{\lambda}^w$.

- the log-shift with fitted weighted $\hat{\lambda}^w$ produces the best results across the combinations with log-transformed predictor and transformation of the response (Table 5). It suggests a larger shift than the default of 1, with the average of the fitted shift parameter $\hat{\lambda}^w = 1.9925$ (Table 7).

4.3.3 Sensitivity to different bias correction approaches

Different forms of the bias adjustment for back-transforming the log transformed data are possible (Flewelling and Pienaar, 1981; Zeng and Tang, 2011), and we implemented the Karlberg-type estimator, equation (6), with two different forms of the first-order bias correction, both with a corresponding second-order correction (see section 2.2.3). We find that there are small differences in the RMSE of the estimates, with the larger RMSEs increased and the smaller ones reduced using the alternative estimator (13) (see supplementary material Fig. S6).

We also investigated the different forms of the second-order bias correction, equations (7) and (15). The estimates and RMSEs are practically indistinguishable with these two different corrections (results not shown).

We interpret that the form of the first-order bias adjustment is not important in the small area models that we consider, and that the choice between the forms of the second-order adjustment does not matter.

5 Discussion

With the turnover variable from the AIDA dataset, our results show that the log transformation is effective, and using data-driven transformations based on the Box-Cox or dual power transformation returns parameters which indicate a transformation very similar to the log transformation. This conclusion is affected by the need to shift the data in order to accommodate a significant proportion of zero values in the dataset, but follows with a shift of +1 which is standard, and is also the value suggested by an unweighted fit of the parameter in the log-shift transformation. The weighted parameter is slightly larger (≈ 2), but using this shift has only a small impact on the quality of the industry ("small area") estimates.

A summary of the bias and rrmse properties of the best transformation methods is given in Table 8. The best of the approaches which we consider in this paper is not in fact transformation-based, but is the PEBP with untransformed auxiliary variables, which benefits from weights calculated on the same scale as the auxiliary variables to compensate for the skewed distributions of the response. In the industries where this works least well, the rrmse's are a little larger than for some transformation based approaches. There is therefore also a case for the best of these – the unweighted EBP, which has essentially the same outcomes for any of the transformations considered. One option would be to prefer the log transformation on the grounds of simplicity. However, in the absence of further similar studies, we prefer the log-shift to assess the required shift and the dual power transformation which is adaptive, which should provide some robustness to different distributions in case other datasets require it. The bias corrected empirical best estimator, EBbc, has the same performance in our example where the log transformation is appropriate, but it is not adaptive, so should be used only when the form of the transformation has been assessed.

Among the transformation-based methods, we note that the EBP and PEBP are very flexible and straightforward model-fitting procedures. They can be used to produce essentially the same estimates

estimator	relative bias		relative rmse	
	median	mean	median	mean
EBP log shift	3.39	2.21	11.15	11.06
EBP dual power	3.40	2.21	11.16	11.06
EBbc	3.34	2.17	11.13	11.03
PEBP ML linear	0.92	4.14	8.38	10.33

Table 8: Relative bias and rrmse properties of the best transformation-based estimators from Figs. 2 and 3.

as have been developed through careful derivation, but without requiring the theoretical development (see Li and Lahiri (2007) and Sugasawa and Kubokawa (2019) for examples of theoretical approaches to the Box-Cox and dual power transformations. But such development is not needed when the EBP accommodates the transformation directly).

It is interesting to consider the effect of informativeness and the weights. Models (20) and (21) include all the design variables which we used in our stratified design, so we can expect that this handles the informativeness, and therefore it is not so surprising that the EBP (without weights) is the best of the transformation-based approaches. It appears that additionally including the weights with the PEBP increases the variance, as might be expected. In a situation where the design variables are not included in the model, we would expect the PEBP to perform relatively much better. The same situation occurred in Smith et al. (2021), but there the weighted estimators performed almost as well as the unweighted ones.

It is difficult to say how far these results generalise beyond the AIDA data. Bocci and Smith (2023) find that a group of robust models have consistent perfoirmance across two different business survey datatsets, so we think there is a good chance that the conclusions derived here are generalisable. But we do not at this stage have the evidence to support that. Similarly the conclusion that the PEBP with untransformed data is better than using transformations is likely to hold across datasets, but more investigation is needed to confirm this. For a review of small area estimation approaches under informative sampling see Parker et al. (2023a,b).

Transformation is a well-known statistical procedure, and the log transformation is perhaps the most widely used. Nevertheless, we find that the interaction of the shift parameter and the log transformation can have unexpected effects on the transformed distributions. The effects of the shift parameters on the shape of the transformed distributions are illustrated. The properties of the small area estimators do not seem to be greatly affected by shift values around 1, which is commonly used, but this parameter can also be fitted. We do however note substantial effects for minimal ($s = 0.0001$) and larger ($s = 5$) values, which suggest that it is important to choose the shift parameter carefully, and not to rely on default behaviour. This is why we recommend using the log-shift model, to fit this parameter. A possible strategy when a transformation different to the log is appropriate would be to do some preliminary analysis to examine the distribution and Q-Q plot of the residuals in the sample, with the the aim of choosing a shift value to give at least a symmetrical distribution of the errors, and something as close to normally distributed as possible. We may consider modelling the 0 values directly with a zero-inflated small area approach (Krieg et al., 2016), but since this is a distinct strategy for small area estimation with business data we leave such a comparison for further investigation.

Acknowledgements

We are grateful to two anonymous referees for detailed comments which substantially improved the paper.

Data availability

The AIDA database is not publicly available, but is available by subscription. It is regularly updated and has a rolling window, so a current version will not correspond exactly with the data used here. Anyone interested in the specific dataset from this study should contact the authors.

Funding

This work was supported by the Italian Ministry of University and Research (MUR), Department of Excellence project 2023-2027 ReDS “Rethinking Data Science” – Department of Statistics, Computer

Science, Applications – University of Florence.

Conflict of Interest Statement

The authors declare no conflict of interest.

References

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

Berg, E. and Chandra, H. (2014). Small area prediction for a unit-level lognormal model. *Computational Statistics & Data Analysis*, 78:159–175.

Berg, E., Chandra, H., and Chambers, R. (2016). Small area estimation for lognormal data. In Pratesi, M., editor, *Analysis of Poverty Data by Small Area Estimation*, pages 279–298. Wiley, Chichester.

Bocci, C. and Smith, P. A. (2023). Unit level small area estimation for business surveys: comparing transformation-based and robust models. In *Proceedings of the 64th ISI World Statistics Congress, Ottawa, Canada*, pages 1–6. ISI, The Hague. https://www.isi-next.org/media/abstracts/ottawa-2023_3903096beb8f35de497178bb04bd605b.pdf.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, 26(2):211–243.

Brewer, K. and Gregoire, T. G. (2009). Introduction to survey sampling. In *Handbook of Statistics*, volume 29, pages 9–37. Elsevier.

Bureau van Dijk (2015). Aida dati finanziari e software per l'analisi immediate delle aziende italiane. https://www.bvdinfo.com/en-gb/-/media/brochure-library/aida.pdf. Accessed 25 July 2023.

Chandra, H. and Chambers, R. (2009). Multipurpose weighting for small area estimation. *Journal of Official Statistics*, 25(3):379–395.

Chandra, H. and Chambers, R. (2011). Small area estimation under transformation to linearity. *Survey Methodology*, 37(1):39–51.

Chandra, H. and Chambers, R. (2016). Small area estimation for semicontinuous data. *Biometrical Journal*, 58(2):303–319.

Cochran, W. G. (1977). *Sampling techniques*. John Wiley& Sons, New York.

Cox, B. G. and Chinnappa, B. N. (1995). Unique features of business surveys. In Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S., editors, *Business survey methods*, pages 1–17. Wiley, New York.

Edochie, I., Newhouse, D., Schmid, T., and Würz, N. (2023). `povmap`: Extension to the `emdi` package for small area estimation. https://mirrors.aliyun.com/CRAN/web/packages/povmap/vignettes/povmap.pdf. accessed 18 April 2024.

Eurostat (2008). *NACE Rev. 2*. Office for Official Publications of the European Communities, Luxembourg.

Finney, D. (1941). On the distribution of a variate whose logarithm is normally distributed. *Supplement to the Journal of the Royal Statistical Society*, 7(2):155–161.

Flewelling, J. W. and Pienaar, L. (1981). Multiplicative regression with lognormal errors. *Forest Science*, 27(2):281–289.

Guadarrama, M., Molina, I., and Rao, J. (2016). A comparison of small area estimation methods for poverty mapping. *Statistics in Transition new series*, 17(1):41–66.

Guadarrama, M., Molina, I., and Rao, J. N. K. (2018). Small area estimation of general parameters under complex sampling designs. *Computational Statistics & Data Analysis*, 121:20–40.

Karlberg, F. (2000). Population total prediction under a lognormal superpopulation model. *Metron*, 58(3/4):53–80.

Krennmair, P. and Schmid, T. (2022). Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(5):1865–1894.

Krennmair, P., Würz, N., Schmid, T., and Tzavidis, N. (2026). Random forests and mixed effects random forests for small area estimation of general parameters: a poverty mapping case study in Mozambique. *Annals of Applied Statistics*, in press:1–23.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N. (2019). The R package `emdi` for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, 91(7):1–33. https://www.jstatsoft.org/index.php/jss/article/view/v091i07.

Krieg, S., Blaess, V., and Smeets, M. (2012). Small area estimation of turnover of the Structural Business Survey. Statistics Netherlands discussion paper 2012-03.

Krieg, S., Boonstra, H. J., and Smeets, M. (2016). Small-area estimation with zero-inflated data–a simulation study. *Journal of Official Statistics*, 32(4):963–986.

Li, H., Liu, Y., and Zhang, R. (2019). Small area estimation under transformed nested-error regression models. *Statistical Papers*, 60(4):1397–1418.

Li, Y. and Lahiri, P. (2007). Robust model-based and model-assisted predictors of the finite population total. *Journal of the American Statistical Association*, 102(478):664–673.

Lyu, X., Berg, E. J., and Hofmann, H. (2020). Empirical Bayes small area prediction under a zero-inflated lognormal model with correlated random area effects. *Biometrical Journal*, 62(8):1859–1878.

Molina, I. (2009). Uncertainty under a multivariate nested-error regression model with logarithmic transformation. *Journal of Multivariate Analysis*, 100(5):963–980.

Molina, I. and Martín, N. (2018). Empirical best prediction under a nested error model with log transformation. *The Annals of Statistics*, 46(5):1961–1993.

Molina, I. and Rao, J. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38(3):369–385.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–606.

Parker, P. A., Janicki, R., and Holan, S. H. (2023a). A comprehensive overview of unit-level modeling of survey data for small area estimation under informative sampling. *Journal of Survey Statistics and Methodology*, 11:829–857.

Parker, P. A., Janicki, R., and Holan, S. H. (2023b). Comparison of unit-level small area estimation modeling approaches for survey data under informative sampling. *Journal of Survey Statistics and Methodology*, 11:858–872.

Pfeffermann, D. and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102(480):1427–1439.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, J. N. K. (2011). Impact of frequentist and Bayesian methods on survey sampling practice: A selective appraisal. *Statistical Science*, 26(2):240–256.

Rivière, P. (2002). What makes business statistics special? *International Statistical Review*, 70(1):145–159.

Rojas-Perilla, N., Pannier, S., Schmid, T., and Tzavidis, N. (2020). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society, Series A*, 183(1):121–148.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag, New York.

Skarke, F., Kreutzmann, A.-K., and Würz, N. (2021). Extensions to the `ebp` function in the R package `emdi`. https://mirror.las.iastate.edu/CRAN/web/packages/emdi/vignettes/vignette_ebp2.pdf. accessed 31 October 2023.

Smith, P. A., Bocci, C., Tzavidis, N., Krieg, S., and Smeets, M. J. E. (2021). Robust estimation for small domains in business surveys. *Journal of the Royal Statistical Society: Series C*, 70(2):312–334.

Smith, P. A. and Lorenc, B. (2021). Robust official business statistics methodology during covid-19-related and other economic downturns. *Statistical Journal of the IAOS*, 37(4):1079–1084.

Sugasawa, S. and Kubokawa, T. (2019). Adaptively transformed mixed-model prediction of general finite-population parameters. *Scandinavian Journal of Statistics*, 46(4):1025–1046.

Verret, F., Rao, J. N. K., and Hidiroglou, M. A. (2015). Model-based small area estimation under informative sampling. *Survey Methodology*, 41(2):333–348.

Welsh, A. H. and Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society, Series B*, 60(2):413–428.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193.

Würz, N., Schmid, T., and Tzavidis, N. (2022). Estimating regional income indicators under transformations and access to limited population auxiliary information. *Journal of the Royal Statistical Society, Series A*, 185(4):1679–1706.

Yang, L. (1995). *Transformation-density estimation*. PhD thesis, University of North Carolina, Chapel Hill.

Yang, Z. (2006). A modified family of power transformations. *Economics Letters*, 92(1):14–19.

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

Zeng, W. S. and Tang, S. Z. (2011). Bias correction in logarithmic regression and comparison with weighted regression for nonlinear models. *Nature Precedings*, pages 1–11. https://doi.org/10.1038/npre.2011.6708.1.

Zimmermann, T. and Münnich, R. T. (2018). Small area estimation with a lognormal mixed model under informative sampling. *Journal of Official Statistics*, 34(2):523–542.