Unit level small area estimation for business surveys: comparing transformation-based and robust models

Chiara Bocci* & Paul A. Smith†

Abstract

Small area estimation methods are generally based on models which have assumptions of normal errors, but many types of data do not have a normal distribution. Several approaches have been suggested to deal with skewed data, and here we investigate transformations (with and without bias correction) and compare them with previous work with robust models which are less affected by the tails of the distributions. We investigate the properties of these models with a real data set of Italian retail businesses which mimics a structural business survey. Transformation based approaches improve small area estimates, but are not as effective as the best robust approaches. The assessment of which robust approaches are best is qualitatively the same as in previous work, and corroborates the earlier findings with a different data set.

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.

Different strategies have been proposed for dealing with the special characteristics of business surveys in SAE applications. One approach is to use robust SAE based on M-estimators and M-quantile estimators. These deal with the challenge of outliers using robust models that are less affected by outlying observations, so that the assumption of normal errors is less problematic. A decision is still needed whether to model the unobserved parts of the data without outliers, or whether some allowance for outliers is needed in the predictions for the unobserved part of the population. Smith et al. (2021) examined a variety of approaches in a real dataset where the outcomes are known. These include some approaches which incorporate the sampling weights, and therefore also account for the informative sampling in business surveys. For an overview of small area estimation approaches for informative data see Parker et al. (2023).

Regression relationships among business survey data are often found to be multiplicative, both according to economic theory and from observation (Chandra and Chambers, 2011). They are therefore frequently analysed using a log transformation and a linear model assuming normally distributed errors. This is equivalent to the errors on the original scale being lognormally distributed. So a second approach for skewed data is to transform the data so that the regression errors are approximately normal, and then to apply the standard SAE models with the transformed data. This procedure has its own challenges,

*Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Italy. chiara.bocci@unifi.it

†S3RI & Dept. of Social Statistics and Demography, University of Southampton, UK. p.a.smith@soton.ac.uk

because predictions are made on the transformed scale, and naïve inverse transformation (back to the original scale) induces a bias. Some methods have been proposed for bias correction, including generating datatsets on the original scale using the transformed model.

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. In section 3 we present the variety of transformation-based approaches which have been proposed in SAE with a unified notation, consistent with Smith et al. (2021). In section 2 we summarise the setup of the dataset to which these methods will be applied. In section 4 we discuss the results of a design-based simulation from the data, and compare them with the robust methods from Smith et al. (2021).

2 Business survey data

Smith et al. (2021) used data from the retail sector in the Netherlands; these data were not available for the current research, so we have used a similar population of retail businesses in Italy derived from the AIDA dataset (Bureau van Dijk, 2015). We make some modifications to obtain a known population with complete information on turnover from 2020, which is needed in order to have 'the truth' against which to compare the estimates from the different small area procedures described below. 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 we use the information from two years previously as the register/frame information. The final dataset comprised 71568 retail businesses, divided into 36 industry groups defined by the NACE sectors, with auxiliary information from 2018 on (i) turnover, (ii) size class, and (iii) number of working persons.

We use the population size from 2020 and population variances calculated from the 2018 population data as the inputs to a Neyman allocation on strata defined by size classes and industry groups. Some variances are missing (usually because N = 1), so need to be imputed. There was one extreme standard deviation (for the small size stratum in NACE 4740), which was set to NA before any further processing, so that it too would be imputed. The allocated sample size was 3635.

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

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 the usual target for business surveys (and which is also consistent with Smith et al. (2021)). We compare outputs with the direct HT and GREG estimators, and with a naïve application of the empirical best linear unbiased predictor (EBLUP) approach (see Smith et al. (2021) for details of those).

There are two main strategies for transformation. One is to generate multiple simulated datasets data using errors 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 3.1. A second approach is to calculate the estimates with the transformed data and then to apply a suitable bias correction in the back transformation to obtain (approximately) unbiased estimates on the original scale; the properties of this approach are left for further investigation. Both approaches require a unit level model fitted to the transformed data $\mathbf{y}^*$

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{Z}\mathbf{u}^* + \mathbf{e}^*. \tag{1}$$

where a * designates the use of the transformation.

3.1 Empirical best prediction

The basic idea 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 Rojas-Perilla et al. (2020), and the stages are:

1. select a transformation, fitting the shift parameter to obtain $\hat{\lambda}$ if necessary, and obtain $y_i^* = T_{\hat{\lambda}}(y_i)$

2. use the transformed data in the unit level model (1) 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 $1, \ldots, L$

   - take random draws v_{ind} from $N\left(0, (1-\hat{\gamma}_{ind})\hat{\sigma}_u^2\right)$ for each value of ind and e_i from $N\left(0, \hat{\sigma}_e^2\right)$ for each value of i
   - obtain pseudopopulation l as $y_i^{*(l)} = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^* + \hat{u}_{ind}^* + v_{ind} + e_i^*$, choosing ind such that $i \in ind$
   - back-transform the pseudopopulation values to obtain $y_i^{(l)}$
   - 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, $\hat{y}_{ind}^{EBP} = \frac{1}{L}\sum \hat{y}_{ind}^{(l)}$.

Here we follow Rojas-Perilla et al. (2020) in considering four transformations (Table 1), although others are available. First is the standard log transformation. 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) 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), 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 fourth transformation, the dual power transformation (Yang, 2006) was developed to avoid the bounds of the Box-Cox transformation, but otherwise is rather similar in its behaviour. All four transformations are available with the EBP methodology in the R package `emdi` (Kreutzmann et al., 2019). In common with the log transformation, the Box-Cox and dual power transformations require strictly positive inputs, so if there are zero or negative values in the data a shift s, which should not be too small, must be applied. In the example data below we use the reasonably standard $y_i + 1$, but we note that the results may have a certain sensitivity to this value.

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

Table 1: Transformations considered 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.

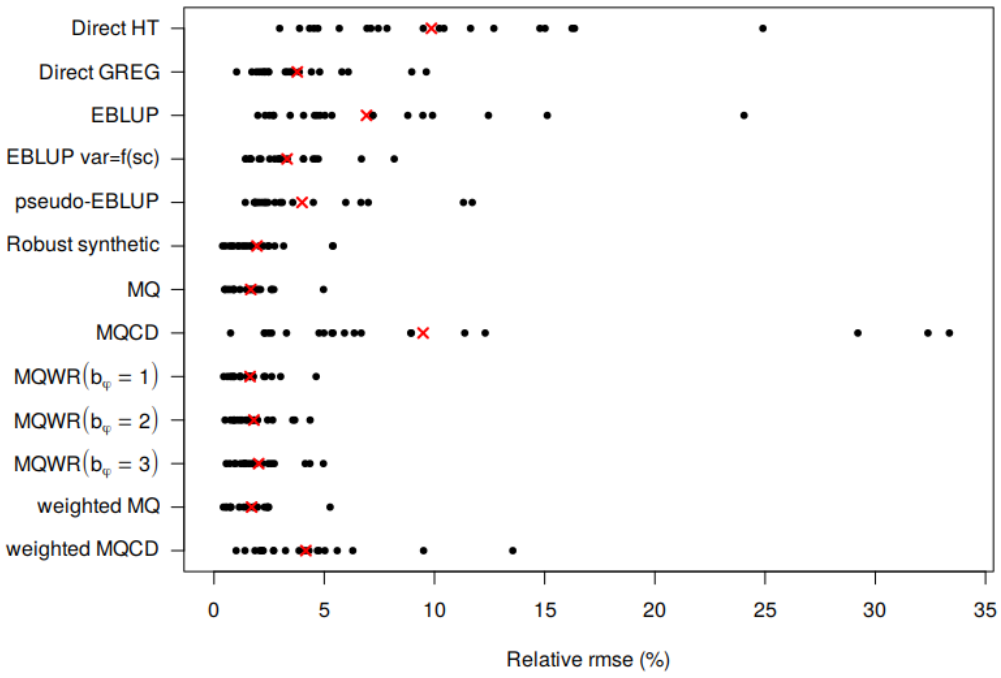
4 Application to AIDA data with known outcomes

Repeated sampling simulations with the data and design described in section 2 were undertaken, fitting a model with $\mathbf{X}\boldsymbol{\beta}^*$ in equation (1) given by

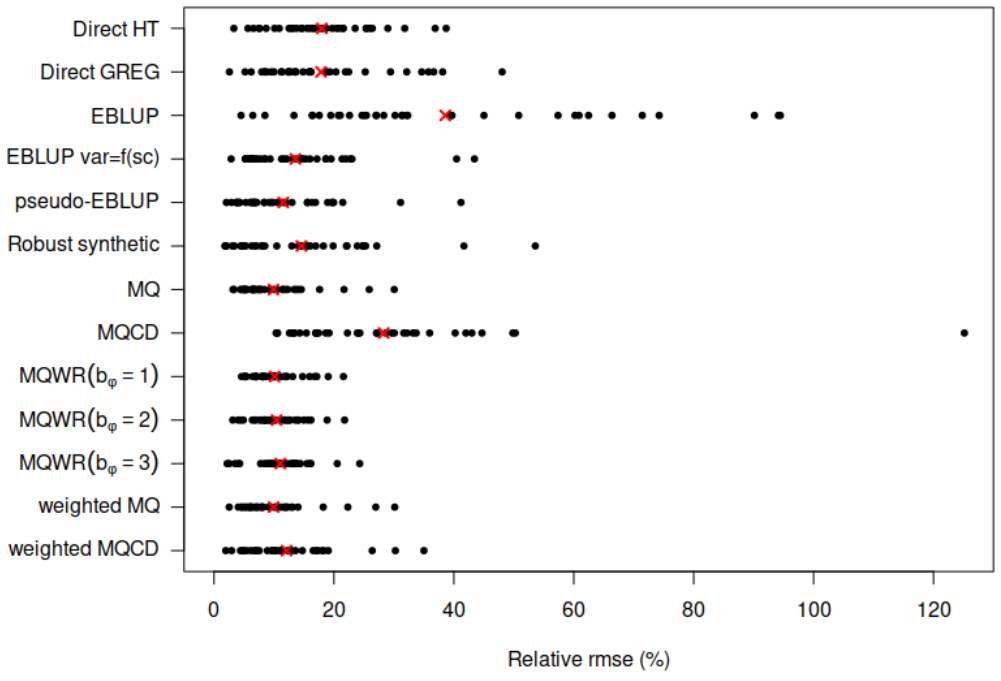
$$\mathbf{X}\boldsymbol{\beta}^* = \beta_0 + \beta_1 t_{i,ind}^{2018} + \beta_2 wp_{i,ind} + \beta_3 \left(t^{2018} \times wp\right)_{i,ind} \tag{2}$$

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

$$\mathbf{X}\boldsymbol{\beta}^* = \beta_0 + \beta_1 \log(t_{i,ind}^{2018}) + \beta_2 wp_{i,ind} + \beta_3 \left(\log(t^{2018}) \times wp\right)_{i,ind} \tag{3}$$



(a)



(b)

Figure 1: Mean squared errors under repeated sampling of a range of robust estimators (see Smith et al. (2021) for definitions) for (a) Dutch tax data and (b) Italian AIDA data.

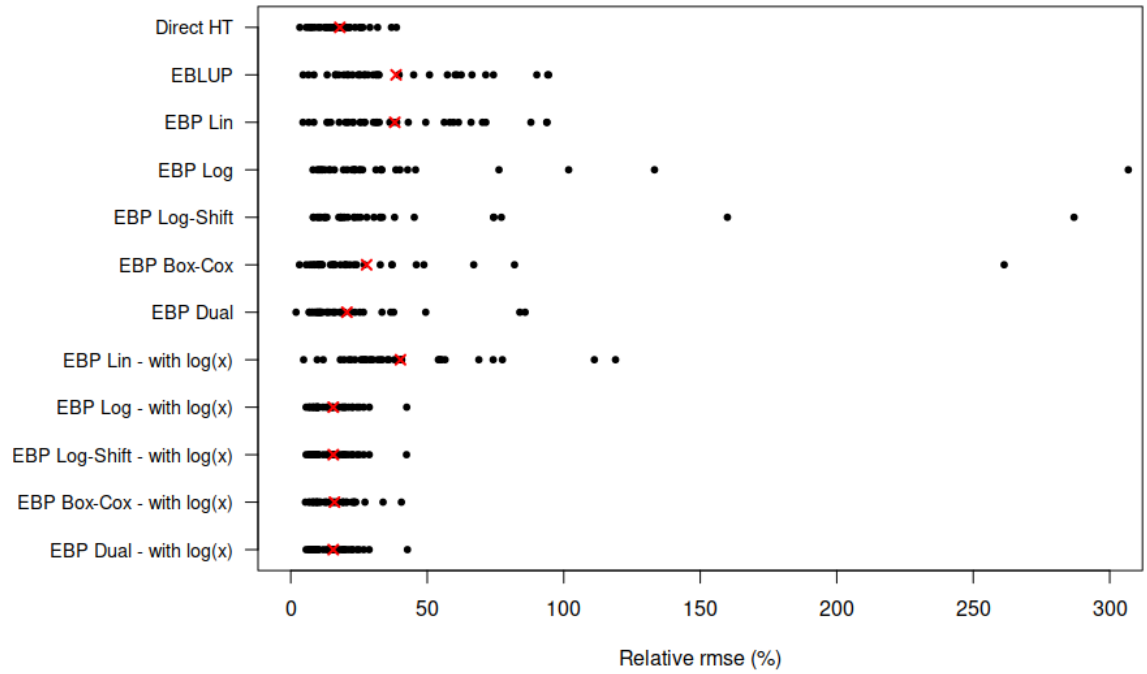


Figure 2: Mean squared errors under repeated sampling of the empirical best prediction (EBP) estimator with a range of transformations, using the Italian AIDA data.

when log(turnover) is used; wp represents the number of working persons. We use the relative bias $\left(\frac{100}{500 y_{ind}} \sum_{k=1}^{500} \hat{y}_{ind}^{k}\right) - 100$ and relative root mean square error $\frac{100}{y_{ind}} \left[\frac{1}{500} \sum_{k=1}^{500} \left(\hat{y}_{ind}^{k} - y_{ind}\right)^2\right]^{\frac{1}{2}}$ to evaluate the performance of the different transformations and estimators.

First we examine the mean squared errors of the variety of robust estimators explored by Smith et al. (2021) using the Netherlands tax dataset and the new Italian AIDA dataset. Summary plots are shown in Fig. 1. The GREG is not such an improvement over the HT estimator in the Italian data as it was in the Dutch example, and is only marginally better than HT. The naïve M-quantile (MQ) is good in both datasets, even though it is not a consistent estimator, and the consistent version MQCD has much larger MSEs. The Weighted M-quantile estimator is consistent and nearly as good as MQ; in fact the weighted estimators in general perform slightly better for the Italian data than they did for the DUtch data. For the Italian example the naïve M-quantile estimator just beats unweighted, bias-adjusted M-quantile (MQWR) with the best b_ϕ, though this best tuning parameter would be unknown in practice. So we conclude that the performance of the different estimators is essentially the same in both datasets, which is a useful replication of the findings of Smith et al. (2021). We note that the MSEs are consistently higher in the Italian dataset, probably reflecting a larger time difference between the auxiliary and the 'survey' information. The MQ and MQWR estimators are the best and have similar performance, and we therefore prefer MQ which delivers the good performance without the added complication of estimating a second tuning parameter.

Second, we produce a similar plot (Fig. 2) comparing the different versions of the EBP using the transformations from table 1, and using either x or $\log(x)$ as a predictor alongside the number of working persons and their interaction (equations (2) and (3)). The direct and EBLUP scatters are identical with those from Fig. 1(b), though the scale is different. The log and log-shift models which use turnover as a predictor have long tails of MSEs, and in a few instances these estimates can blow up to very large values (not plotted). The Box-Cox has only one very large MSE, and the dual shift estimator seems to be better. Nonetheless, the MSEs are substantially reduced in the models with log(turnover). The log-shift with a deterministic shift and the dual power transformation have the best average MSE performance, but all four transformations have similar MSEs.

We can compare the best performances among the EBP estimators with the best of the robust estimators; comparative statistics of the rrmse are shown in Table 2. It is clear that the best robust models considerably outperform the transformation-based estimators based on mse. This is interesting

	log (deterministic shift)	dual power	M-quantile naïve	weighted M-quantile naïve	bias-adjusted M-quantile ($b_\phi = 1$)
median rrmse	14.76	14.78	7.77	8.01	9.60
mean rrmse	15.56	15.55	9.91	9.94	10.11

Table 2: Comparison of man and median relative root mean squared error performance across industry-level domains in the AIDA dataset, with the best EBP estimators (both using log(turnover) in the predictor variables) and the best robust estimators.

because the robust models use turnover as a predictor, whereas the best transformation-based models always use log(turnover). The robust fitting procedures appear to deal effectively with the skewness in the predictors in the process of robustifying against the skewness of the residuals.

References

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.

Bureau van Dijk (2015). AIDA. Dati finanziari e software per l'analisi immediata delle aziende italiane. `https://www.bvdinfo.com/en-gb/-/media/brochure-library/aida.pdf`. Accessed: 25/7/2023.

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

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.

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`.

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

Parker, P. A., Janicki, R., and Holan, S. H. (2023). 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.

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.

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.

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.