

Robust estimation for small domains in business surveys

Paul A. Smith¹, Chiara Bocci², Nikos Tzavidis¹, Sabine Krieg³, Marc J.E. Smeets³

¹ S3RI and Department of Social Statistics & Demography, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.

² Department of Statistics, Computer Science, Applications “G. Parenti”, University of Florence, Viale Morgagni 59, 50134 Firenze, Italy.

³ Statistics Netherlands, Process Development & Methodology, P.O. Box 4481, 6401CZ Heerlen, The Netherlands.

Abstract

Small area (or small domain) estimation is still rarely applied in business statistics, because of challenges arising from the skewness and variability of variables such as turnover. We examine a range of small area estimation methods as the basis for estimating the activity of industries within the retail sector in the Netherlands. We use tax register data and a sampling procedure which replicates the sampling for the retail sector of Statistics Netherlands’ Structural Business Survey as a basis for investigating the properties of small area estimators. In particular, we consider the use of the EBLUP under a random effects model and variations of the EBLUP derived under (a) a random effects model that includes a complex specification for the level 1 variance and (b) a random effects model that is fitted by using the survey weights. Although accounting for the survey weights in estimation is important, the impact of influential data points remains the main challenge in this case. The paper further explores the use of outlier robust estimators in business surveys, in particular a robust version of the EBLUP, M-regression based synthetic estimators, and M-quantile small area estimators. The latter family of small area estimators includes robust projective (without and with survey weights) and robust predictive versions. M-quantile methods have the lowest empirical mean squared error and are substantially better than direct estimators, though there is an open question about how to choose the tuning constant for bias adjustment in practice. The paper makes a further contribution by exploring a doubly robust approach comprising the use of survey weights in conjunction with outlier robust methods in small area estimation.

Key words

Small area estimation, M-quantile regression, bias adjustment, outliers

1. Introduction

Small area estimation encompasses a wide range of methods (Rao & Molina 2015) that have become a major component in the official statistician’s toolkit, with many applications to a wide range of variables and domains in many different types of surveys. (We will use the terms domain and area interchangeably.) Nevertheless, the application of small area estimation methods to variables derived from business surveys has proved to be more challenging, because of the differences in the types of sample designs used in business surveys (Rivière 2002): detailed stratification, non-negligible sampling fractions, large variations in estimation weights, etc, all generally connected with the skewed distribution of variables. It is not sufficient to obtain point estimates for small areas as official statistics, it is also necessary to have some estimate of their quality, generally as mean squared error (mse) to account for variability and bias. This then allows proper inference and appropriate use of the resulting estimates. In business surveys, the mse estimates are often very large because of the skewed distributions, and here we will explore approaches that reduce the mse’s.

Hidiroglou & Smith (2005) discuss some of the challenges in applying small area estimation in business surveys, and Burgard *et al.* (2014) demonstrate the effects of different sample designs as used in business surveys on the effectiveness of small area estimation methods. There are also few practical applications, and these are restricted to estimating proportions (Chandra *et al.* 2012) or numbers (Militino *et al.* 2015) of businesses with particular characteristics. Numbers and proportions work better with small area estimates in business surveys, as they are based on quantities whose underlying distributions are not (or at least much less) skewed. It is therefore easier (though sometimes still challenging) to construct a suitable model in these cases than when dealing with numeric variables such as turnover or employment, which may vary by several orders of magnitude. More recently Ferrante & Pacei (2017) have extended Fay-Herriot (area-level) models to deal with skewed data and Fabrizi, Ferrante & Trivisano (2018) have used Bayesian procedures applied to area-level models for the same purpose.

In this paper, however, we consider the use of unit-level rather than area-level models as the basis of small area estimation, for estimation of turnover in small domains from the retail sector of the Structural Business Survey (SBS) in the Netherlands. The retail sector is one of those where small area estimation is likely to be most practical within business surveys, since there are generally very many retail businesses (and outlets), so that there is enough data to fit and compare models. Contrast this with industrial activities such as shipbuilding, which in many countries include only a handful of businesses; they therefore suffer from small numbers of observations from which to build models as well as all the difficulties caused by the structural characteristics of business surveys.

The skewness of distributions of business survey variables means that samples are likely to include outliers. These outliers in turn have a large effect on the estimation of population quantities which require special treatment (Chambers 1986), and they have an even larger effect in small area (domain) estimation (SAE), where sample sizes are considerably smaller and model-dependent estimation is the norm. This problem remains when the small area estimator is an indirect one, e.g. an empirical best linear unbiased predictor (EBLUP), since the weights, which determine the balance between use of information from within and beyond the domain, show that data from the small area of interest will still be used. Also, the estimates of the model parameters underpinning the estimator will themselves be affected by the sample outliers. In order to address these challenges, Chambers *et al.* (2014) considered the ways in which robust survey estimation procedures could be adapted to small area estimation.

The first approaches to explicitly address the issue of outlier robustness for small area estimation use plug-in robust prediction. That is, they replace the parameter estimates in optimal, but outlier sensitive, predictors by outlier robust versions (Chambers *et al.* (2014) call this a *robust projective* approach). Chambers and Tzavidis (2006) proposed an approach based on fitting outlier robust M-quantile models to the survey data, and Sinha & Rao (2009) proposed outlier robust procedures for small area estimation with linear random effects models. These robust projective approaches aim to produce a low prediction variance, but may involve an unacceptable prediction bias because they assume that non-sampled values in the target population are drawn from a distribution with the same properties as the sample non-outliers – in particular that prediction errors cancel on average in small areas. To avoid this bias Chambers *et al.* (2014) consider a *robust predictive* approach where a bias correction is applied through the use of two influence functions – the first quite restrictive to robustly choose a working model uninfluenced by the outliers, and a second, less restrictive, one to produce a bias correction.

Some of these methods have been evaluated in empirical work by Chambers *et al.* (2014). In this application to the retail sector in the Netherlands we use tax data to provide a known population which can be used to evaluate the methods, and demonstrate how they fare in a real data situation.

This is consistent with a design-based evaluative framework as recommended by Tzavidis *et al.* (2018). Krieg *et al.* (2012) undertook a simulation study using tax data to approximate the population and auxiliary variables of the SBS, focussing on the situation where there are no strong auxiliary variables for individual units. They compared estimates from a direct estimator, the generalised regression estimator (GREG), with small area estimates from the EBLUP. The EBLUP was found to be biased because of the skewed distribution of turnover and the consequent prevalence and impact of outliers. Krieg *et al.* (2012) considered modelling transformations, albeit not data driven (see Rojas-Perilla *et al.*, 2019), of the turnover variable to deal with the skewness. This reduced the bias and the influence of the skewness on the resulting estimates, but increased their variability, such that the net effect was only a small improvement over the EBLUP.

In this paper we continue the investigation of the retail sector from the Dutch SBS, using tax data in different years as explanatory and outcome variables with samples based on the SBS sample design. The availability of information on the whole population of businesses from this administrative source allows us to evaluate the performance of different modelling methods against the true values. To compensate for the skewness of the turnover data, we consider models which are robust to unusual observations, comparing a variety of approaches which have been proposed in the literature. Of particular interest is the use of outlier robust methods, which reduce the impact of unusual observations, in conjunction with survey weighting, which ensures design-consistent estimators so that estimates for (at least) large domains make sense even if the selected model is inappropriate (Fabrizi *et al.* 2014). This joint approach may provide double robustness, but possibly at the cost of increased variance. The structure of the remainder of the paper is as follows. In Section 2 we describe the pseudo-SBS data which form the basis of the comparisons, and in Section 3 the range of estimation methods to be compared is presented. Section 4 describes the results of applying these methods in the retail sector of the SBS. Section 5 makes an assessment of the sensitivity of the results to different aspects of the data and methods, and we draw conclusions and discuss the general applicability of our findings in Section 6.

2. Tax-based population and pseudo-Structural Business Survey data

The data to be modelled are derived from Value Added Tax (in Dutch BTW - belasting toegevoegde waarde) data of enterprises classified in the retail sector. These administrative data are provided to Statistics Netherlands to support the construction of the business register and survey-taking. The tax-turnover data are strongly correlated ($\rho \approx 0.7$, see Table 1) with the turnover which is the actual target of the Structural Business Survey, an annual survey undertaken by Statistics Netherlands. By using the tax-turnover we have information for almost the whole population (which we treat as if it is the whole population), and therefore can assess the outputs of the methods against the true population value. For this study, any enterprises which do not have tax data available for both periods considered are removed. The data in this study are taken from 2006 and 2007, with the first year's data playing the role of the business register, and the second year's data playing the role of a survey response, which is therefore available for the whole of the considered population. The correlation between the 2006 and 2007 tax-turnover is larger than between the turnover and tax-turnover (Table 1), but otherwise the data structure is very similar to the SBS.

Table 1: Correlation between tax in 2006 and 2007 and turnover

	Tax 2007	turnover 2007
Tax 2006	0.996	0.732
Tax 2007		0.727

In a real application the lag between the survey and the tax data would be more likely to be two years, which may attenuate the correlations somewhat, though they are still likely to be strong.

The SBS has a stratified design, with strata defined by a combination of industrial classification according to NACE¹ (in twenty classes in the retail sector) and nine size classes. The largest businesses (in size classes 6-9, with employment 50 or greater) are completely enumerated. Since their information is in concept completely known, they are not included in the current study. The sample sizes in other strata were determined based on Neyman allocation with some additional constraints on subpopulations (including for the retailing sector); in this study we therefore use the design of the 2009 SBS excluding the completely enumerated strata. Within strata, samples are selected from the population of enterprises by simple random sampling without replacement. We have population size $N = 63981$, five size classes and 20 industries, and will denote the size class \times industry strata by h ; the total sample size is 5074, with sample sizes per industry varying from 21 to 769 (see Table S1 in the supplementary material). The design of the SBS used in this study is described in Krieg *et al.* (2012, Appendix A) – the smallest enterprises have the smallest inclusion probabilities and the largest weight, and there are also quite large differences in weights between industries.

Repeated samples were drawn from the population of tax-turnover data using the design of the SBS in the retail sector, and inclusion probabilities were retained with the data. Since the population is fixed, the simulation is over repeated sampling, and not over repeated realisations of a population from a model, and we are therefore assessing the design-based properties of these small area estimators as recommended by Tzavidis *et al.* (2018).

Four auxiliary variables are available with which to construct models to predict the 2007 tax-turnover:

- tax-turnover from 2006, $tax1$
- industrial classification, ind
- size class, based on the working persons in the business in bands 1, 2-4, 5-9, 10-19, 20-49, sc
- employment, that is the number of working persons, wp

and we additionally denote the tax-turnover to be estimated (for 2007) as tto , and the sampling (design) weight as d_i for business i . It is slightly curious to have two predictors derived from the same working persons information – both the size class and employment in its own right – and Statistics Netherlands' preferred models (Krieg *et al.* 2012) include *both* variables. We will consider whether this approach, which implies an element of nonlinearity in the effect of working persons, is a valuable strategy during our model fitting.

We have used tax data in different years as predictor and outcome variables, as this is the only possibility where the outcome is known in real data. It is typical in business surveys for there to be strong predictors for size-related outcome variables, derived from administrative data and often available operationally from the business register. For variables where the predictors are not so strong, small area estimation approaches will likely have a smaller impact on the rmse, and it seems likely that robust methods will only be effective if the outliers are the main causes of lack of fit. There are important outliers in this dataset, as we see in section 3.

For the purposes of evaluation we will consider the small areas to be the 20 industry classes within the retail sector, $ind = 1, \dots, 20$. In line with common practice in business surveys we will estimate totals for tto within the domains; (the equations for estimators in this paper are therefore for totals, in contrast to the more usual formulation for estimation of means). There are two potential challenges

¹ Nomenclature statistique des activités économiques dans la Communauté européenne, the standard European industrial classification.

in this situation. First, the sampling weights of observations within the industry classes vary, because selected businesses come from different original strata h . The sampling weights are usually assumed to be ignorable given the model in model-based estimation (which includes small area estimation), but the large variation in the sampling weights in business surveys makes this assumption problematic. The investigations below are assessing the effect of ignoring the weights as well as the effect of outliers among the responses, and we further consider whether using the weights is helpful in section 5.4. The second challenge is that when the survey is undertaken, some businesses may be found to have changed their size and/or industry classification; we do not consider this situation in this investigation. If we did wish to make estimates for domains defined by the new information, we would have less information on the domain sizes in the population, and therefore more variable estimators.

3. The small area methods

3.1 Direct estimation

Several different estimators are compared in this study. A natural comparator to use as a baseline for evaluation is the direct estimator which is often preferred by National Statistical Institutes (NSIs) because of its unbiasedness and objectivity (Rao 2011). The simplest of these is the Horvitz-Thompson (HT) estimator:

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

where the $d_i = 1/\pi_i$ are the inverse of the selection probabilities π_i . 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):

$$\hat{y}_{ind}^{GREG} = \hat{y}_{ind}^{HT} + \hat{\beta}^{GREG} [\mathbf{X}_{ind} - \hat{\mathbf{x}}_{ind}^{HT}] \quad (2)$$

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 (1) with y_i replaced by the appropriate auxiliary variables. There are many possible models, derived from suitable choices of the definition of \mathbf{X} and β , but in this case we use the actual approach from the Dutch SBS with auxiliary variables defined by $tax1 \times ind \times size$ class group, where the size class group is an aggregation of sc into 1-9 and 10-49 working persons. We note that it is possible to write such estimators in the same form as (1) as $\hat{y}_{ind}^{GREG} = \sum_{i \in s \cap ind} w_i y_i$ with modified weights w_i , and that domain estimation can then be used to obtain the required industry level estimates. Of course, this approach fails in the case that there are no sample units in the domain of interest, but the sample design ensures that this does not happen in our example.

3.2 Non-robust small area methods

The basis of small area estimation is to borrow strength across a wider pool of data than the domain in question (*indirect* estimators). In general this produces biased estimators, in exchange for a substantial reduction in the variance, with the trade-off between bias and variance chosen through the mean squared error. The classical approach when the microdata are available for modelling is to assume that the population data follow a random effects model (Battese, Harter & Fuller 1988, Rao & Molina 2015 section 4.3), and in this case the structure of the data with enterprises within industries suggests a two-level random effects model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \quad (3)$$

where β contains the fixed effects, the rows of \mathbf{X} contain the corresponding auxiliary data, the rows of \mathbf{Z} are dummy variables for industry, $\mathbf{u} \sim N(\mathbf{0}, \Sigma_u)$ is a vector of industry-specific random effects and $\mathbf{e} \sim N(\mathbf{0}, \Sigma_e)$ is a vector of individual-specific random effects. We have followed Krieg *et al.* (2012)

who in their study of the effect of transformations of the outcome variable considered a model using all the available variables

$$tto_{i,ind} = \beta_0 + \beta_1 tax1_{i,ind} + \beta_2 sc_{i,ind} + \beta_3 wp_{i,ind} + \beta_4 (tax1 \times wp)_{i,ind} + u_{ind} + e_{i,ind} . \quad (4)$$

(using the notations for specific variables introduced in section 2) and this is the basis of the main results. Note that this model contains both the working persons and the size classes derived from the same variable. We also explore the effect of using a smaller model using only the size class variable without the more detailed information on number of working persons:

$$tto_{i,ind} = \beta_0 + \beta_1 tax1_{i,ind} + \beta_2 sc_{i,ind} + u_{ind} + e_{i,ind} \quad (5)$$

This produces only small changes in the results (section 5.3), but the AIC for (4) is lower (see Table S2 in the supplementary material for model fit statistics), which suggests that there is a nonlinear effect of size class so that the incorporation of both *wp* and the *sc* variable derived from it is beneficial. We present detailed results from model (5) in the on-line supplementary material, but the results are not qualitatively different from model (4). Although the correlation between *tto* and *tax1* is high (Table 1) there are important effects of size, and models based only on tax do not fit as well as models (4) and (5). Similarly, fixed effects models also gave large variances, and apart from its use the GREG (section 3.1) we do not explore them further.

Note that (3) does not incorporate the sampling weights d_i , in contrast to (1) and (2). Including the variables (*ind*, *sc*) which explain the differences in the weights as predictors is one strategy to reduce the impact of the informative design (Pfeffermann 2011). We will assume that the sampling is approximately ignorable and the weights are not needed if these variables are included in the model (for some supporting evidence see supplementary material section S1 and Fig. S2), though inasmuch as this is not true it will have an impact on the repeated sampling results in section 4. However, one of the main features that distinguish business surveys is the importance of the weights to deal with differential sampling due to size differences. We therefore consider weighted versions of several methods to assess whether including the weights improves the properties of the small area estimates (see for example the end of this subsection).

Let $\hat{\beta}$ and \hat{u} denote estimates of the fixed and random effects in (3). The EBLUP of the industry *ind* total of the y_i ($= tto_j$ in our case) under (3) is then

$$\hat{y}_{ind}^{EBLUP} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} (\mathbf{x}_i^T \hat{\beta} + \mathbf{z}_i^T \hat{u}) \quad (6)$$

where \hat{u} denotes the vector of the estimated area-specific random effects and we use *s* and *r* to denote sample and non-sample units respectively.

It is important to check the diagnostics from fitting the assumed population model (3) to the sample data, specifically residual and normal probability plots, to ensure that the model assumptions are satisfied, before the model can safely be used for prediction. Diagnostics from one realised sample are shown in Fig. 1.

There are clearly some large and influential outliers in the residuals in Fig 1a. The industry-level errors are plausibly normal in Fig. 1c, particularly considering the small number of observations. However, it is clear that the distribution of the enterprise-level residuals in Fig. 1b departs substantially from normality, and this is a characteristic problem when applying such methods in business surveys, deriving from the skewed distributions of the data (Rivière 2002). By comparing Figs 1b and 1c it is apparent that random effects play a negligible role in this analysis, as they are so much smaller than the unit level residuals. The symmetry of Figs 1b&c also suggest that the outliers in this dataset will have an effect on the variance rather than the bias, and this is borne out in some of the analysis below.

Krieg *et al.* (2012) explored models fitted to transformations of tto ; in business surveys the variance often increases with the size of the business (which also looks true from inspection of Fig. 1a), so log

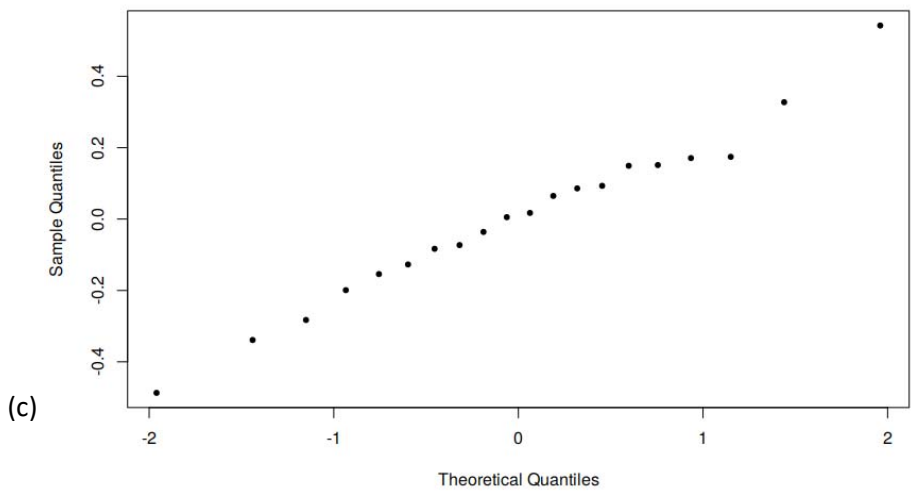
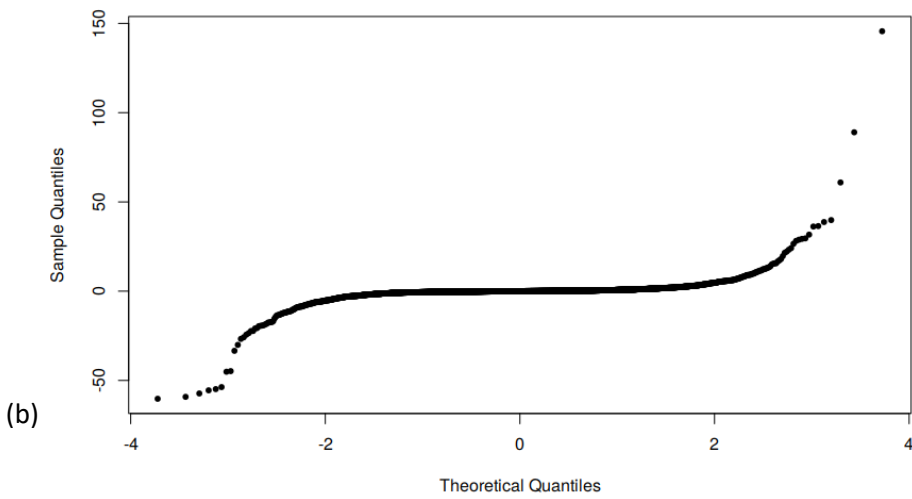
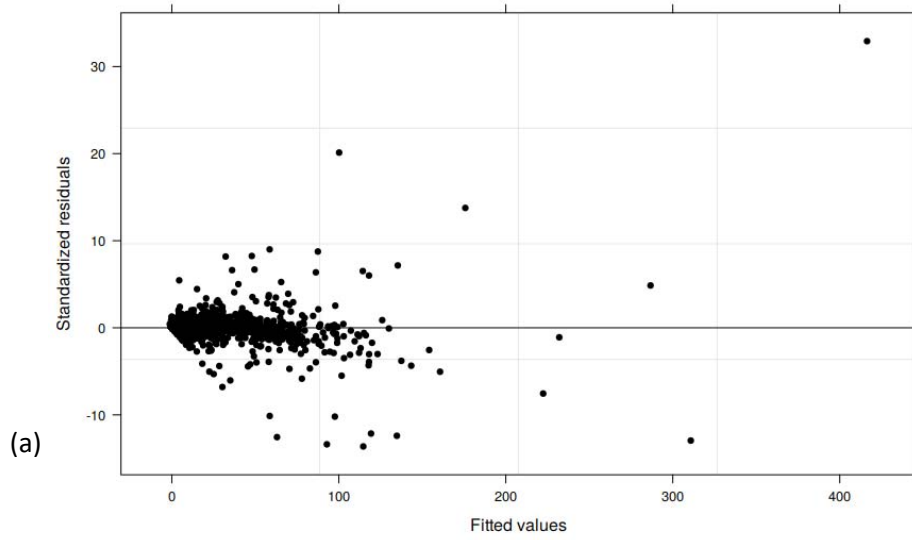


Fig. 1: Diagnostic plots from model (3) fitted to one sample from the SBS retail population. (a) Standardised residuals vs fitted values; (b) Normal probability plot for enterprise-level residuals; (c) Normal probability plot for industry-level residuals.

or root transformations are often appropriate to stabilise variances. Krieg *et al.* discover that using the square root transformation of tto in (3) results in negatively skewed residuals, and find that $tto^{1/3}$ produces approximately normal residuals. Nevertheless, following the approach for small area estimation for transformed variables developed by Chandra & Chambers (2011) based on $tto^{1/3}$ does not substantially reduce the mean squared error, which seems still to be affected by outliers after the transformation.

So one option is to extend the method to allow the enterprise-level variance to vary as a function of the size class by replacing \mathbf{e} in (3) by $\mathbf{e}_{sc} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{e,sc})$, or as a function of the number of working persons. In fact using wp^2 provides a better characterisation of the variance than wp (see Table S2 in the supplementary material for model fit statistics), so in this case replace \mathbf{e} in (3) by $\mathbf{e} = (\mathbf{wp}^2)^T \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$. These reduce the impact of outliers in larger enterprises by allowing for larger variances in larger size classes, but are not fully robust to outliers. In the results below, allowing the variance to change as a function of the size class gives smaller relative root mean square errors, so we restrict our presentation to this version.

Another way to improve the fit of the model is to account for the weighting in fitting the random effects model, which is a quite often used in modelling business survey data. This approach also bridges the two schools of thought in survey estimation, namely design-based and model-based estimation. One strategy to account for the weighting is to use the approach of You & Rao (2002), hereafter referred to as the pseudo-EBLUP,

$$\hat{y}_{ind}^{pEBLUP} = \hat{\gamma}_{ind,w} \sum_{i \in S_{ind}} w_{ind,i} y_{ind,i} + \left(\mathbf{x}_{ind} - \hat{\gamma}_{ind,w} \sum_{i \in S_{ind}} w_{ind,i} \mathbf{x}_{ind,i} \right) \hat{\boldsymbol{\beta}}_w \quad (7)$$

where $\hat{\gamma}_{ind,w} = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2 \delta_i)$ is the shrinkage factor which defines the weights for the direct and

regression synthetic parts of the estimator, with $\delta_i = \sum_{i \in S_{ind}} \left(\frac{w_{ind,i}}{\sum_{i \in S_{ind}} w_{ind,i}} \right)^2$ and σ_u^2 and σ_e^2 are the

(common) diagonal elements of $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$ respectively.

$$\hat{\boldsymbol{\beta}}_w = \left[\sum_{ind} \sum_{i \in S_{ind}} \frac{w_{ind,i}}{\sum_{i \in S_{ind}} w_{ind,i}} \mathbf{x}_{ind,i} (\mathbf{x}_{ind,i} - \gamma_{ind,w} \bar{\mathbf{x}}_{ind,i}) \right]^{-1} \left[\sum_{ind} \sum_{i \in S_{ind}} \frac{w_{ind,i}}{\sum_{i \in S_{ind}} w_{ind,i}} \mathbf{x}_{ind,i} (\mathbf{x}_{ind,i} - \gamma_{ind,w} \bar{\mathbf{x}}_{ind,i}) y_{ind,i} \right]$$

is the vector of regression coefficients fitted using the survey weights (for full details see You & Rao 2002, section 4.2).

3.3 Robust projective methods

Our analysis so far indicates that the impact of influential data points on the model fit is significant and thus it cannot be alleviated by just extending the structure of the random effects model. In our view the use of outlier robust fitting methods is essential when working with business survey data. In the last decade there has been a series of developments on outlier robust small area estimation, which we will summarise in the next sections.

The natural approach to outlier robust small area estimation is to extend the random effects model fit to control for the impact of outliers. Using work by Richardson & Welsh (1995) and Welsh & Richardson (1997), Sinha & Rao (2009) developed a robust EBLUP (REBLUP) estimator, by replacing the usual maximum likelihood equations with alternative versions which reduce the effect of outliers using a Huber function

$$\psi(a) = a \min(1, b_\psi / |a|) \quad (8)$$

(Huber 1973) where $b_\psi > 0$ is a tuning constant. For residuals with a normal distribution, the theoretical value $b_\psi = 1.345$ is generally considered to be optimal (derived from Holland & Welsh 1977), and we follow this strategy here, as it is the standard approach. Nevertheless, for skewed distributions such as those found in business surveys data other values may be more suitable (Dawber *et al.* in prep.). This function is applied to the model residuals, scaled by the inverse of their variances and leads to robust estimates of the parameters in (3) $\hat{\boldsymbol{\beta}}^{\psi,SR}$ and $\hat{\boldsymbol{u}}^{\psi,SR}$ where the ψ superscript denotes the dependence of the fitted parameters on the chosen influence function and SR denotes Sinha and Rao (2009). Using these parameters in the EBLUP derivation gives us the REBLUP estimates of the industry totals

$$\hat{y}_{ind}^{REBLUP} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} (\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{\psi,SR} + \mathbf{z}_i^T \hat{\boldsymbol{u}}^{\psi,SR}) \quad (9)$$

(for details of the derivation see Chambers *et al.* 2014).

Applying the REBLUP estimator with $b_\psi = 1.345$ to the SBS retail data has some interesting effects – the robustly estimated industry-level effect resulted in convergence problems because the between-industry variance component is close to the boundary of the parameter space. So the industry-level effect apparent in the EBLUP (and EBLUP with transformed data) is due entirely to the effect of the outliers. If we remove this term from the regression, we return to a standard linear model for which the REBLUP methodology is not appropriate and which leads back to a synthetic estimator. In this approach our assumed population model (3) and small area estimator (6) is clearly not appropriate based on the sample data. So if we want to use (3) as a basis for small area estimation in this example we must examine alternative robust estimation methods.

A second, simple robust projective approach is to use M-regression, which uses a Huber function (8) to reduce the influence of outlying residuals in model fitting. The resulting robust estimate of the regression parameters $\hat{\boldsymbol{\beta}}^{\psi,M}$, derived by an iterative fitting process (Draper & Smith 2014, pp569-572) can be used to calculate a simple robust synthetic estimator

$$\hat{y}_h^{RSYN} = \sum_{i \in h \cap s} y_i + \sum_{i \in h \cap r} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{\psi,M} \quad (10)$$

Note that M-regression contains only fixed effects; we use the same fixed effects as in model (4) in the paper (and (5) in the supplementary material); an industry fixed effect is not added, so for this model the weights are expected not to be ignorable. However, we also expect that as a result of reducing the impact of outliers, the importance of weighting will be reduced.

A further robust projective approach is the M-quantile regression-based method described by Chambers and Tzavidis (2006). M-quantile regression is a robust regression approach based on an influence function similar to (8), which allows us to fit a model which produces an optimal mixture of regressions for the mean and median, preserving some of the sensitivity of the mean while taking advantage of the robustness of the median (Breckling & Chambers 1988). We again use $b_\psi = 1.345$.

Chambers & Tzavidis's approach is based on a linear model for the M-quantile regression of \mathbf{y} on \mathbf{X} ,

$$m_q(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}_q^{\psi}$$

where $m_q(\mathbf{X})$ denotes the M-quantile of order q of the conditional distribution of y given \mathbf{X} . An estimate $\hat{\boldsymbol{\beta}}_q^{\psi}$ of $\boldsymbol{\beta}_q^{\psi}$ can be calculated for any value of q in the interval $(0,1)$, and for each unit in sample we define its unique M-quantile coefficient under this fitted model as the value q_i such that $y_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{q_i}^{\psi}$. Then for each small domain ind , we calculate the sample average of these coefficients and denote it by $\bar{q}_{ind} = \frac{1}{n_{ind}} \sum_{i \in ind} q_i$. The M-quantile estimate of the total of y in industry ind is then

$$\hat{y}_{ind}^{MQ} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\bar{q}_{ind}}^{\psi}. \quad (11)$$

Note that the M-quantile approach to small area estimation allows for different predicted values in different industries. This is in contrast to the robust synthetic estimator (10). We believe this approach to quantifying between industry variability may overcome the numerical issues we encountered when fitting the outlier robust version of the random effects model that resulted in a zero between-industry variance component. The robust projective approaches are called naïve estimators by Tzavidis *et al.* (2010), and we will refer to (11) as the naïve M-quantile estimator. This is because the working model, which is equivalent to an outlier rejection approach, is projected onto the non-sampled part of the population without accounting for the possibility that outliers also exist in this part too. Although this approach will reduce the variance of the estimates, we expect that it will also introduce bias in the small area estimates. The assumption that all the outliers are observed in the sample is a strong one. Methods that relax this assumption are presented in the next section.

As in the case of the EBLUP, we may be interested in using a weighted version of the M-quantile estimator. Arguing for design-consistency, Fabrizi *et al.* (2014) proposed weighted versions of M-quantile estimators, which we also include in our empirical assessments as we are interested in studying the use of survey weights in conjunction with outlier robust estimation, see section 5.4. The weighted version of the naïve M-quantile estimator is

$$\hat{y}_{ind}^{wMQ} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \mathbf{x}_i^T \boldsymbol{\beta}_{w, \bar{q}_{ind}}^{\psi} \quad (12)$$

where the w_i are the sampling weights and $\boldsymbol{\beta}_{w, \bar{q}_{ind}}^{\psi}$ is also estimated using the sampling weights, see equation (14) in Fabrizi *et al.* (2014).

3.4 Robust predictive methods

The naïve robust projective approaches of section 3.3 assume that all the non-sampled units follow the (robustly fitted) working model, whereas in practice there will in general be businesses like the outliers among the non-sampled units. Hence, using the M-quantile predictions for the out-of-sample units directly as in (11) leads to a biased estimator of the population total in each small domain. Using the ideas in Chambers (1986), Tzavidis *et al.* (2010) substitute a consistent estimator of the distribution function, using the approach of Chambers & Dunstan (1986), to derive a version of the M-quantile estimator adjusted for bias, and we label this MQCD:

$$\hat{y}_{ind}^{MQCD} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\bar{q}_{ind}}^{\psi} + \frac{N_{ind} - n_{ind}}{n_{ind}} \sum_{i \in ind \cap s} \left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\bar{q}_{ind}}^{\psi} \right) \quad (13)$$

In principle, this estimator works by adding a third term to correct for the potential bias at the cost of allowing the variance to increase. It is easy to note that this estimator is a model-based version of a GREG estimator. Fabrizi *et al.* (2014) also propose a weighted version of (13):

$$\hat{y}_{ind}^{wMQCD} = \sum_{i \in ind \cap s} w_i y_i + \left(\sum_{i \in ind \cap U} x_i^T - \sum_{i \in ind \cap s} w_i x_i^T \right) \hat{\beta}_{w, \bar{q}_{ind}}^{\psi} \quad (14)$$

with $\hat{\beta}_{w, \bar{q}_{ind}}^{\psi}$ again estimated using equation (14) from their paper. The disadvantage of (13) and (14) is that we cannot control the impact of the third term, which can lead to a large increase in the variance.

To obtain an estimator which accounts for this possibility, we must reduce the effect of large residuals in the third term of (13), but by less than in the second term, so that we allow a part of their effect into the bias adjustment. A further Huber function ϕ is therefore needed (defined as in (8)), with a different tuning constant from ψ such that $|\phi(\cdot)| \geq |\psi(\cdot)|$ (Tzavidis *et al.* 2010 p172), that is the tuning constant for ϕ , b_ϕ , takes larger values than b_ψ , and the estimation is less robust than for ψ . The (robust predictive) robustly bias-adjusted version of the M-quantile estimator (labelled MQWR because of its derivation from Welsh & Ronchetti (1998)) is:

$$\hat{y}_{ind}^{MQWR} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} x_i^T \hat{\beta}_{\bar{q}_{ind}}^{\psi} + \frac{N_{ind} - n_{ind}}{n_{ind}} \sum_{i \in ind \cap s} \omega_{ind}^{MQ} \phi \left\{ \frac{(y_i - x_i^T \hat{\beta}_{\bar{q}_{ind}}^{\psi})}{\omega_{ind}^{MQ}} \right\}, \quad (15)$$

where ω_{ind}^{MQ} is a robust estimator of the scale of the residuals $y_i - x_i^T \hat{\beta}_{\bar{q}_{ind}}^{\psi}$ in domain *ind*. The tuning constant for the Huber function in the third term of (15) codifies the trade-off between bias and variance. We use several values of this tuning constant ($b_\phi = 1, 2, 3$) to investigate its effect on the small domain estimates. Note that we are also testing values of $b_\phi < b_\psi$ in contradiction of the theory set out by Tzavidis *et al.* (2010). See also section 5.1 below. A similar robust predictive extension to the REBLUP estimator (9) is also available (Dongmo Jiongo *et al.* 2013), but since the REBLUP is not appropriate for our data we do not consider it.

4. Repeated sampling results

We start with the known retail sector business population derived from the tax data, and consider only size classes 1-5 since the others are completely enumerated and therefore fully known. We select 500 replicate samples using the sample design of the SBS. In some industries there remain some size classes which are completely enumerated, so some businesses are always included. In each of these samples we apply the set of estimators described in Table 2, using model (4); we explore the sensitivity to the model choice in section 5.3 and the supplementary material.

The residuals from the fixed part of model (4) are shown in Fig 2(a). This shows that there are outliers, but the three labelled ones dominate measures of Cook's distance (Fig. 2(b)). Observation 50010 is in one of the completely enumerated strata, so it is included in all the replicate samples. See section 5.2 for more on the effect of these observations.

For each sample we calculate estimates for the small domains formed by the 20 industries, with population sizes ranging from 110 to 8407 and sample sizes from 21 to 769. We then calculate the per cent relative bias in each estimator $\left(\frac{100}{500 y_{ind}} \sum_{k=1}^{500} \hat{y}_{ind,k} \right) - 100$, where \hat{y}_{ind} is any of the considered estimators and y_{ind} the true value for industry *ind* calculated from the population, with *k* indexing the repeated samples. The per cent empirical relative root mean squared error (rrmse) is similarly

calculated as $\frac{100}{y_{ind}} \left[\frac{1}{500} \sum_{k=1}^{500} (\hat{y}_{ind,k} - y_{ind})^2 \right]^{\frac{1}{2}}$. The results of the simulation for each industry ind are presented in Figs 3 and 4 and Tables S3 to S5; Table S3 gives summary information for the full range of methods.

The interpretation of the results from Figs 3 and 4 (and Tables S3 to S5 in the supplementary material) largely follows the expectations from the development of the theory in section 3, though with some interesting nuances. The HT estimator is approximately unbiased as anticipated, but in most industries with a very large relative rmse because of its high variance, which makes the estimates unusable. Introducing strong explanatory information through the GREG substantially improves the rrmse, and there is only a small increase in the relative bias resulting from the small sample sizes within industries. The EBLUP also reduces the rrmse, though not to an acceptable level, and is biased; in fact the GREG has better properties than the EBLUP in this example even though both are affected by outliers. Accounting for the variance structure in a more reasonable way in the EBLUP by allowing the variance to be different within different size classes ($var=f(sc)$) reduces the rrmse to less than half that of the HT estimator and improves somewhat on the GREG, though the estimator can be seen to have a small bias over repeated sampling.

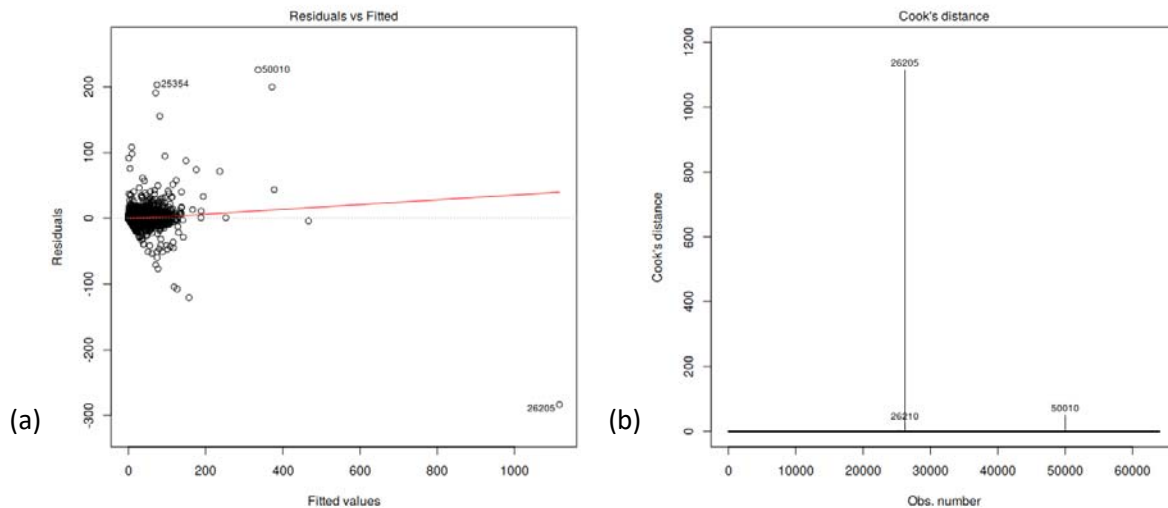


Fig. 2: (a) Plot of residuals against fitted values for the whole population from the fixed part of model (4). The three points with (by some way) the largest values of Cook's distance are labelled. (b) Cook's distances for each observation (compare with Fig. S3(b)).

Moving to the robust synthetic estimator using parameters fitted with M-regression has a substantial further impact on the rmse, partly from a reduction in the bias and partly from reduced variance. In particular it reduces the rrmse in industries with large rrmse values under other methods considered so far. This suggests that using methods which are robust to the outliers in the data may have a substantial impact on the quality of the estimates. Most of the remaining larger rrmse are dominated by the bias, however, so further exploration of bias correction seems potentially valuable.

The consistency of the pseudo-EBLUP brought about by using the survey weights helps to reduce the bias in some industries, but conversely increases it in others. The average effect is to increase the rmse over the robust synthetic estimator, so it seems that adding the sampling weights is not sufficient to deal completely with the differences in business sizes.

The naïve M-quantile approach reduces the average rrmse still further than with the robust synthetic estimator, though with a similar median rrmse. It works well in this dataset, even though it is not a consistent estimator in general, producing estimates with good properties in almost all industries.

There is a single larger rrmse for industry 52610, which is consistently challenging for all the approaches considered. This seems to be because 52610 includes one of the businesses with the largest Cook's distances (Fig. 2), and it is in a completely enumerated stratum in the SBS design, so this extreme outlier is included in *every* replicate sample. Introducing the bias correction to the M-quantile estimator to introduce consistency does not have the expected effect. It reduces the bias in

Table 2: Sequence of estimators considered for small area estimation; the estimators appear in the same order in Figs. 3 and 4 (where the estimators are identified by their superscripts) and the Tables in the supplementary material.

Estimator	Symbol	Equation	Method information			
			Fixed effects only or random effects model?	Regression type	Weighted?	Outlier robust?
Direct (HT)	\hat{y}_{ind}^{HT}	(1)	constant	ordinary	no	no
Direct (GREG)	\hat{y}_{ind}^{GREG}	(2)	fixed	ordinary	no	no
EBLUP	\hat{y}_{ind}^{EBLUP}	(6) with variables from (3) with $\mathbf{e} \sim N(\mathbf{0}, \Sigma_e)$	random	ordinary	no	no
EBLUP (var = f(sc))	\hat{y}_{ind}^{EBLUP}	(6) with variables from (3) with $\mathbf{e} \sim N(\mathbf{0}, \Sigma_{e,sc})$	random	ordinary	no	no
pseudo-EBLUP	\hat{y}_{ind}^{pEBLUP}	(7)	random	ordinary	yes	no
Robust synthetic	\hat{y}_h^{RSYN}	(10)	fixed	M-regression	no	projective
M-quantile naïve ($b_\psi = 1.345$)	\hat{y}_{ind}^{MQ}	(11)	fixed	M-quantile	no	projective
M-quantile bias-adjusted ($b_\psi = 1.345$)	\hat{y}_{ind}^{MQCD}	(13)	fixed	M-quantile	no	predictive
M-quantile robustly bias-adjusted ($b_\psi = 1.345, b_\phi = 1, 2, 3$)	\hat{y}_{ind}^{MQWR}	(15)	fixed	M-quantile	no	predictive
weighted M-quantile naïve ($b_\psi = 1.345$)	\hat{y}_{ind}^{WMQ}	(12)	fixed	M-quantile	yes	projective
weighted M-quantile bias-adjusted ($b_\psi = 1.345$)	\hat{y}_{ind}^{WMQCD}	(14)	fixed	M-quantile	yes	predictive

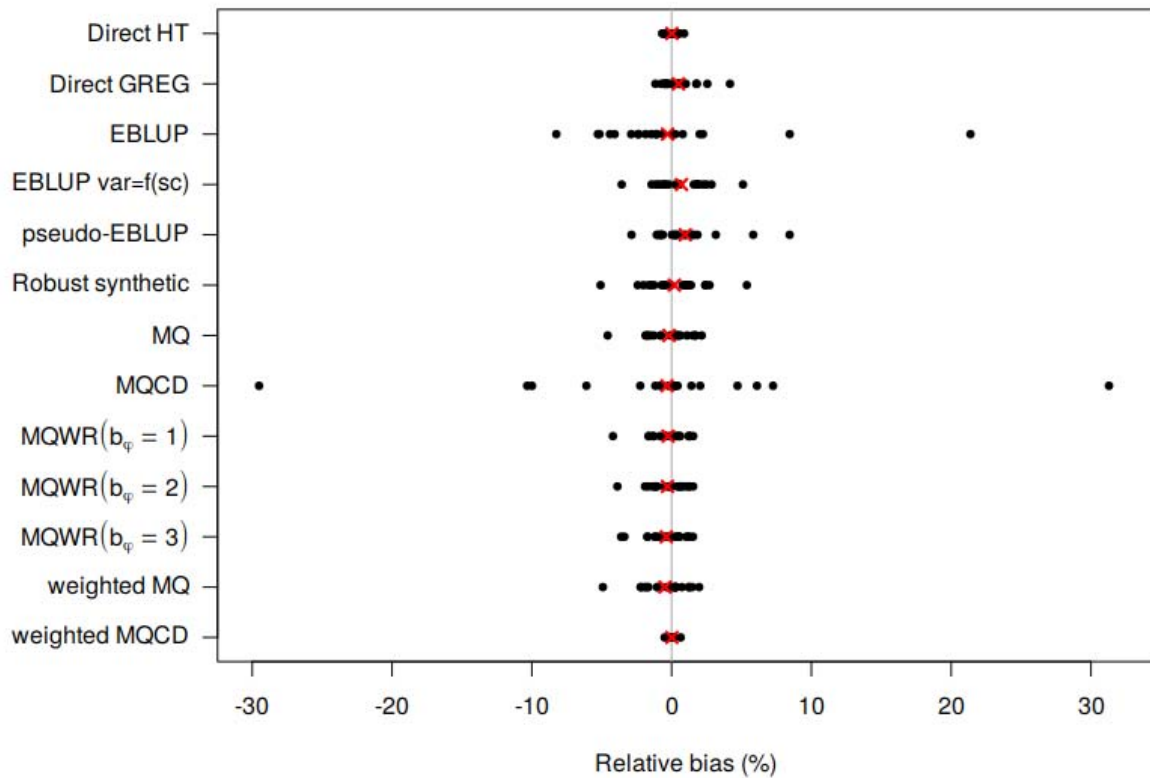


Fig. 3: Linear scatterplots of the relative bias of the different estimators by industry; the mean bias is shown by a cross. The labels are summarised in Table 2. The detailed results can be found in Table S4 in the supplementary materials.

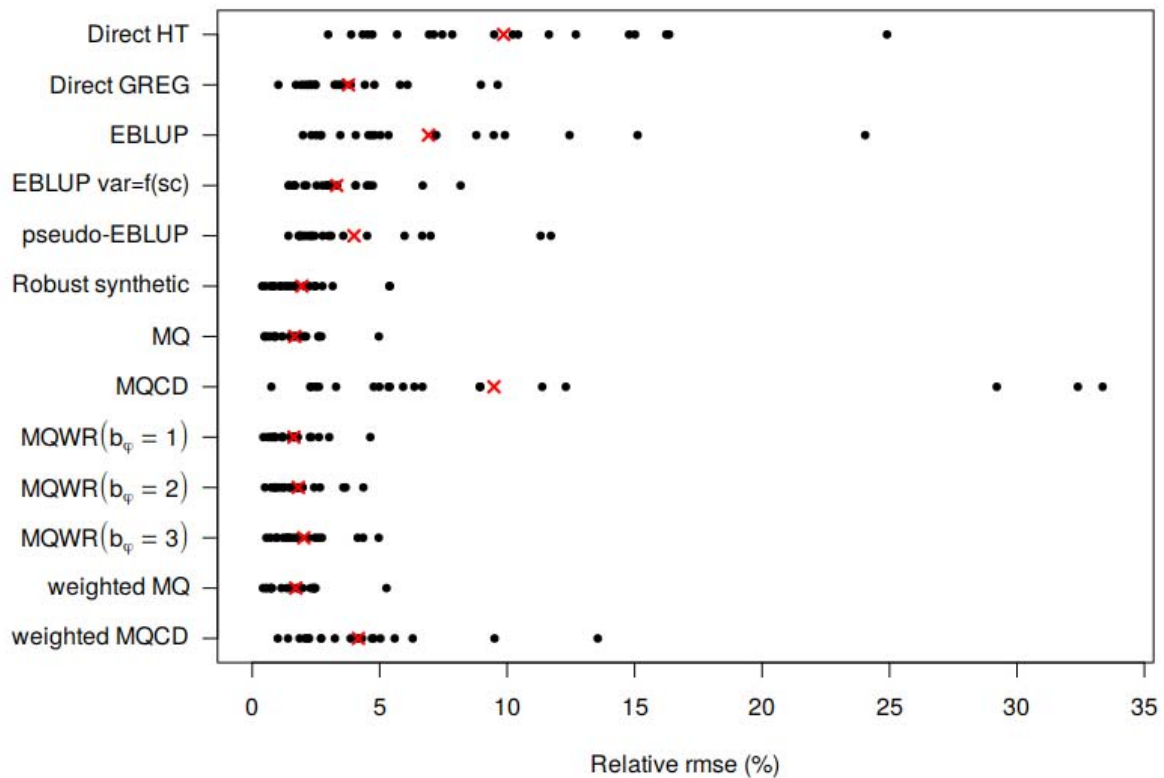


Fig. 4: Linear scatterplots of the relative rmse of the different estimators by industry; the mean rmse is shown by a cross. The labels are summarised in Table 2. The detailed results can be found in Table S5 in the supplementary materials.

many industries, but also substantially increases the bias in a smaller number of industry estimates, with the result that the average rmse over industries is also substantially increased, to an unacceptable level.

The weighted versions of the naïve and bias-adjusted M-quantile estimators should both be design-consistent, and indeed we find only small biases for both, though the average relative bias for the weighted naïve M-quantile is larger than for its unweighted version. The bias of the weighted bias-adjusted M-quantile estimator is particularly low, but the compensatory increase in the variance gives a large average rmse which means it is not satisfactory as a small area estimator in this case.

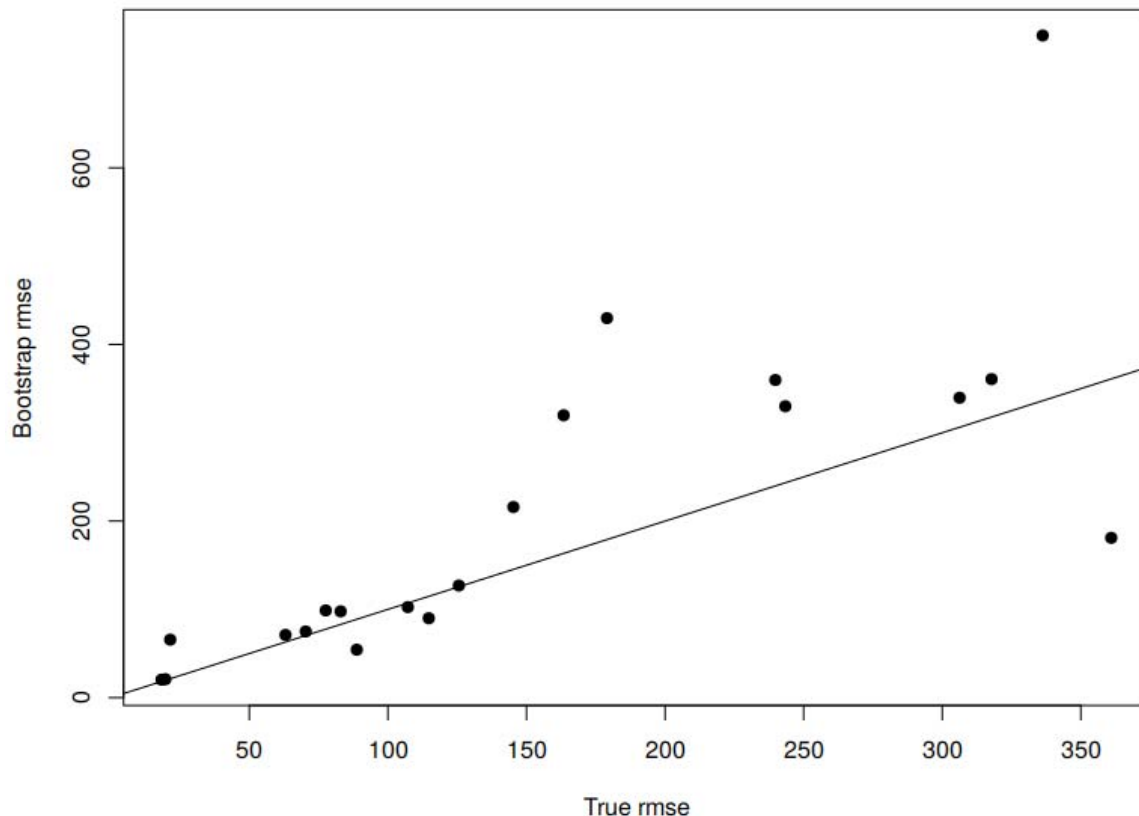


Fig. 5: The empirical true mse and mean of the bootstrap mse's for the M-quantile robustly bias-adjusted estimator, with the line of equality for reference. The detailed results can be found in Table S6 in the supplementary materials.

Finally the (unweighted, predictive) robustly bias-adjusted M-quantile estimators have similar properties to the (unweighted) naïve M-quantile estimator; the estimator with $b_\phi = 1$ is the best of the values considered, which is odd because $b_\phi < b_{\psi^*}$, contradicting expectations from theory; we return to this in section 5.1. The robustly bias-adjusted estimator with $b_\phi = 1$ generally causes reductions in the relative biases of the industries which have the largest relative biases with the naïve M-quantile estimator; industries with small relative biases may see an increase. But the offsetting changes in the variance from the bias correction term mean that some rmse's go up and others down; the net effect is a marginal improvement from the best of the alternative estimators. But it is debatable whether the additional complications from this machinery would be worthwhile in this case.

4.1 Estimating the mean squared errors of the robustly bias-adjusted M-quantile estimator

Mean squared error (mse) estimation is an important part of small area estimation. In this paper we consider the estimation of the mse of the predictive M-quantile robustly bias-adjusted estimator. It is possible to construct a linearisation-type estimator of the mean-squared error for some estimators which can be written as a weighted sum of the observed values (Tzavidis *et al.* 2010). But for the approaches considered here there is in general no linearisation estimator, and we must fall back on a bootstrap estimator of the rmse (also presented in Tzavidis *et al.* 2010). Since we have the simulations from the known population, we can evaluate the bootstrap rmse estimator based on single samples with the true rmse evaluated over the 500 selected samples. The results of this comparison are presented in Fig. 5 and Table S6. With the exception of few industries, the estimated mse tracks the empirical mse fairly well.

5. Sensitivity

We have examined a wide range of estimators in section 4 to assess which have the best repeated sampling properties for use with our example business survey dataset. This provides evidence of the most appropriate methods, and how these vary as the properties (particularly the presence of outliers) of the population information differ in different small areas. In addition to the guidance provided by such an investigation, the series of simulations also allows us to assess the sensitivity of the small area estimates to some further method and estimator properties, which we describe in this section.

5.1 Impact of b_ϕ in the robust predictive estimators

In a practical application of small area estimation in a business survey there would not be any objective way to set the tuning constant b_ϕ in the bias adjustment (for the robustly bias-adjusted REBLUP or M-quantile estimator). We therefore explored the sensitivity of the rmse to the range of values in more detail. We examine b_ϕ from 0.25 to 3 in steps of 0.25, using the same simulated samples as for the main results in section 4 and the relative rmses for some selected industries are shown in Table 3 (Table S13 in the supplementary material gives results for all industries, from which the median and mean are calculated).

Table 3: Relative rmse of the robustly bias-adjusted M-quantile estimator under repeated sampling over values of b_ϕ from 0.25 to 3 in steps of 0.25 for selected industries, and the median and mean rmse over all industries. (Results for all industries are given in the supplementary material Table S13.)

b_ϕ	52110	52321	52330	52413	52630	median	mean
0.25	0.84	1.81	2.70	2.30	1.47	1.55	1.63
0.50	0.81	1.75	2.78	2.14	1.48	1.55	1.61
0.75	0.80	1.75	2.90	2.17	1.49	1.55	1.62
1.00	0.80	1.78	3.02	2.33	1.49	1.56	1.64
1.25	0.81	1.83	3.16	2.57	1.48	1.53	1.67
1.50	0.83	1.88	3.32	2.88	1.47	1.53	1.71
1.75	0.84	1.93	3.49	3.22	1.46	1.59	1.76
2.00	0.86	1.98	3.66	3.57	1.45	1.60	1.81
2.25	0.88	2.03	3.83	3.93	1.44	1.62	1.86
2.50	0.91	2.09	4.00	4.28	1.43	1.64	1.92
2.75	0.93	2.14	4.17	4.63	1.42	1.66	1.97
3.00	0.96	2.19	4.35	4.96	1.42	1.68	2.02
max - min	0.16	0.45	1.65	2.82	0.07	0.15	0.41

We hope for a minimum of the rmse over the range of values we consider, and in our example there is a minimum in the mean of the relative rmse at $b_\phi = 0.5$ (Fig. 6a). However, there is a range of patterns in the individual industries (Tables 3 and S12 and Fig. S4). Half of the industries have the same pattern as Fig. 6(a) with a minimum at values of b_ϕ from 0.5 to 1.75. Seven are monotonic increasing over the range of b_ϕ values considered, and one is monotonic decreasing. Two industries show a *maximum* of the rmse (eg industry 52630, Fig. 4b), which is an unexpected pattern. Some uncertainty may be expected in the rmse estimates for the smallest values of b_ϕ , because the estimator is unstable when the Huber function affects many values. But the patterns in general seem to be quite stable and interpretable. It is not straightforward to pick a single value for use in all industries in this dataset, but perhaps $b_\phi = 0.75$ would be the best compromise giving reasonable bias adjustment in all industries (and indeed only values of $b_\phi < 1.25$ are better than the naïve M-quantile estimator). In most industries (eg 52110 in Table 3) the differences in the rmse over the range of values of b_ϕ is small. The largest difference, in industry 52413, is almost 3% in the rmse, which is sufficiently large that it may affect the inference on the estimates. Therefore in this example the choice of b_ϕ does seem to be important. We therefore suggest that the sensitivity of the estimates to different values of b_ϕ should be investigated when using the robustly bias-adjusted M-quantile estimator. Of course in other domains or datasets b_ϕ may have other impacts, and further experimentation with other business survey data and a range of estimators would be valuable.

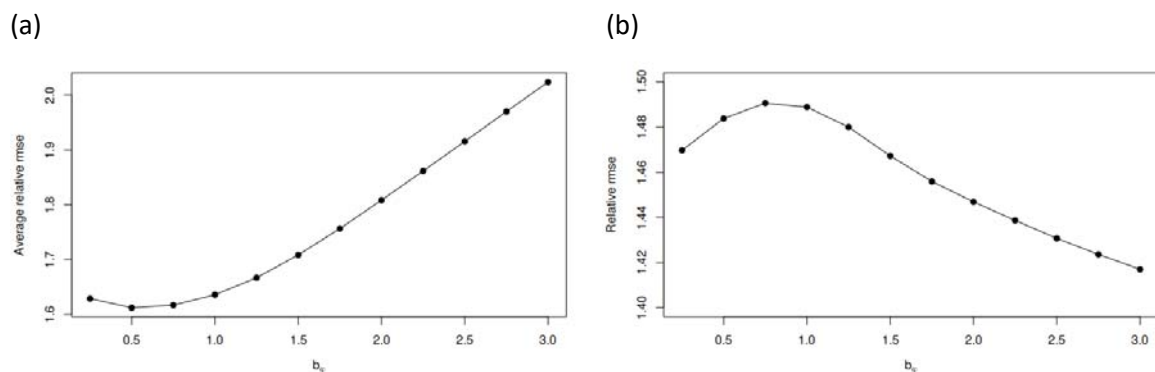


Fig. 6: The change in (a) average (over industries) relative rmse and (b) relative rmse for industry 52630, both as b_ϕ varies in the robustly bias-adjusted M-quantile estimator.

It is interesting that values of $b_\phi < b_{\psi^*}$, which are precluded by Tzavidis et al. (2010) – presumably because they imply *more* bias in the bias-variance trade-off in the adjustment and therefore do not in concept make a bias adjustment – seem to work well. In general, small values of b_ϕ give the lowest rmse. An alternative interpretation could be that bias is not the main influence in this optimisation, and that variance may be more important.

5.2 Sensitivity to extreme population outliers

The tax data that we have used as a population contain some very extreme outliers, and this is clearly a situation in which robust methods should be effective. But we are also interested in their performance with fewer, less extreme outliers. For this purpose we produced a version of the population where the points with the highest leverage in the working model formed from the fixed part of (4) were removed, the working model was refitted, and the influence diagnostics examined. If necessary, further high leverage points were removed. This led to a reduced population where 29 of the most extreme observations were removed. The reduced population has a less skewed distribution

of residuals, but the estimation of small area (industry) totals is still affected by the small sample sizes, so the same arguments for the use of small area estimators apply.

(Note that this is not proposed as a way to deal with outliers – as this process would have two significant challenges, in when to stop, and how to deal with the removed observations. But the reduced population does allow us to examine the properties of the considered estimators in a related dataset with fewer very extreme observations.)

The results of using the reduced population are presented in the supplementary material Tables S7-S9. Essentially they present the same picture as the results from the full population. The difference between the two best estimators, the naïve M-quantile estimator and the robustly bias-adjusted M-quantile estimator, is even slightly smaller. So our conclusions are not affected when the population of interest has fewer (and less extreme) outliers. Of course we are not able from this study to examine real data containing more (and more extreme) outliers, though it would be possible to introduce some artificial contaminating values and assess the sensitivity to these. We leave this for further work, although perhaps repeating the study with different populations would be a higher priority.

5.3 Sensitivity to model choice

Equations (4) and (5) present the two models which we have considered for this dataset, and a summary of the estimator properties using the two models (with the original, complete dataset) is shown in Table 4. The conclusions from the two models are substantively similar, though there are some small differences, which are biggest in the relative bias of the EBLUP and naïve M-quantile estimators. The full results from the simulation with model (5) are shown in Tables S10-S12 in the supplementary material.

Table 4: Comparison of RRMSE and bias properties for the same forms of estimator and the same data, but using model (4) or (5). The GREG is based on the SBS model variables, and the HT estimator has no predictors, so neither is linked to models (4) and (5). The methods with the best properties are highlighted in bold.

	mean rb		mean rrmse	
	(4)	(5)	(4)	(5)
Direct (HT)	0.02		9.86	
Direct (GREG)	0.49		3.78	
EBLUP	-0.29	-1.75	6.92	7.55
EBLUP (var = f(sc))	0.71	0.62	3.32	3.08
pseudo-EBLUP	0.98	0.97	3.99	3.98
Robust synthetic	0.20	0.26	1.95	1.95
M-quantile naïve ($b_{\psi} = 1.345$)	-0.19	0.01	1.67	1.60
M-quantile bias-adjusted ($b_{\psi} = 1.345$)	-0.33	-0.41	9.48	9.49
M-quantile robustly bias-adjusted ($b_{\psi} = 1.345, b_{\phi} = 1$)	-0.25	-0.11	1.64	1.58
M-quantile robustly bias-adjusted ($b_{\psi} = 1.345, b_{\phi} = 2$)	-0.30	-0.20	1.81	1.76
M-quantile robustly bias-adjusted ($b_{\psi} = 1.345, b_{\phi} = 3$)	-0.37	-0.30	2.02	1.98
weighted M-quantile naïve ($b_{\psi} = 1.345$)	-0.47	-0.41	1.71	1.67
weighted M-quantile bias-adjusted ($b_{\psi} = 1.345$)	0.02	0.02	4.16	4.17

5.4 Weights

The pseudo-EBLUP and weighted M-quantile methods use the sampling weights, which means that they give design-consistent estimators in small areas as the sample size approaches the population size. This provides some extra protection against model misspecification bias, but the use of the weights increases the variance. These effects are not obvious among the range of estimators that we consider – although the bias should be reduced, there are unweighted estimators that have small bias too. The variance does however seem to be larger, though the weighted naïve M-quantile estimator is competitive in terms of rmse in our example data. But the best-performing estimators in our example are unweighted. Applications of these estimators to further examples are needed to assess whether this is data specific or whether a more general result can be deduced.

We also considered a variation on the calculation of \bar{q} in the weighted versions of (11) and (13) to

use the survey weights, replacing \bar{q} by $\tilde{q}_{ind} = \frac{\sum_{i \in ind} w_i q_i}{\sum_{i \in ind} w_i}$; this seems to use the weights twice, and the

results (not shown) are not better than the original versions of the estimators, so we do not consider this approach further.

5.5 Pure prediction or out-of-sample prediction

Several of the estimators (see (6), (9), (10), (11), (15)) are of the form $\hat{y}_{ind} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \hat{y}_i$ for

various definitions of the predictor \hat{y}_i , and producing estimates in this way seems to be standard practice. There is an alternative approach however, in which even the sample values are predicted, that is $\hat{y}_{ind} = \sum_{i \in ind} \hat{y}_i$ which is sometimes used, particularly when the sample is a small proportion of

the population. We have used both approaches with the naïve M-quantile estimator (11) in unweighted and weighted versions. In both cases the average rmse is lower for the first formulation, and the effect is particularly noticeable in the presence of a large outlier (though when there is no outlier there are cases of pure prediction industry estimates with lower rmse's than the out-of-sample predictions). So (unsurprisingly) the pure prediction is rarely better than the real observation, and we recommend that for business surveys in particular the real data are always used directly when the estimator has this form.

6. Discussion

In this paper we have illustrated the effects of robust unit-level models as the basis for small-area estimation for business surveys with variables with skewed distributions; this contrasts with the adaptation of area-level models to business surveys in Ferrante & Pacei 2017 and Fabrizi *et al.* 2018. M-quantile methods which deal with outlying observations provide a strategy for small domain estimation which offers much reduced rmses over direct estimation when domains are small, and better properties than other robust small area methods in this example. The robustly bias-adjusted M-quantile estimator with $b_\phi = 0.5$ produces the best overall results, but the unweighted and weighted naïve M-quantile estimators have similar performance, and with these latter methods there are fewer complications from the need to choose an appropriate tuning parameter. The relative rmse's for industry-level small area estimates in Table 4 range from 0.44 to 4.63%, which makes estimates much more usable than the corresponding (direct) Horvitz-Thompson estimates which have relative rmse's from 2.98 to 24.90%, and even considerably better than the GREG which uses the strong auxiliary information in design-based estimation, where the industry rmse's range from 1.01 to 9.68%.

The robust synthetic estimator also has reasonable performance in this dataset, because of the strong correlation between the outcome variable and the auxiliary variable. The synthetic estimator is simpler to apply in practice than the other robust methods described above, so in practical situations may be a good compromise between accuracy and complexity, though it performs less well particularly in some industries in this application.

There is a question over the complexity of modelling that can be handled in the course of statistical production in National Statistical Institutes. As methods progress so does the technical capacity within NSIs, though often at a slower rate. The approaches investigated in this paper are involved, but software is available for model fitting, and it should be reasonable for these methods to be adopted in production. Further extensions, such as to the use of nonlinear estimators to deal directly with nonlinear relations in the data rather than using the transformations explored by Krieg *et al.* (2012) may become practical in the future.

After this research project was finished, Statistics Netherlands decided to change the design of the SBS, for reasons not connected with this research. As a consequence, there is less need for these types of small area estimation methods, and the approaches developed in this paper have not been implemented. Nevertheless the research is useful, as similar problems occur in other statistics and in other countries.

In the M-quantile methods of section (3.3) we used $b_\psi = 1.345$ as the tuning constant in (8), which is optimal for normally distributed errors under certain conditions (Holland & Welsch 1977, Dawber *et al.* in prep). However, we know that at least some of the difficulty with business surveys is that the errors are not normally distributed. Dawber *et al.* (in prep) propose numerical methods to deduce optimal values of the tuning constant for a range of distributions; these vary according to the value of q .

The robustly bias-adjusted M-quantile estimator is the best in our example, but shows only small gains over the naïve M-quantile estimator, and the result depends on the value chosen for the tuning constant b_ϕ in the bias adjustment. In practice there would be no simulation results to indicate the best value for the tuning constant, and in order to have an evidence base for this choice it would be helpful to have further studies with known business survey populations to examine how much this value varies. Until this is available we suggest assessing the sensitivity of estimates to the choice of this parameter. The examination of the average relative rmse with varying values of b_ϕ shows that the rmse's are sensitive to this parameter in some cases, and could have an impact on inferences in the most affected estimates. The choice of $b_\phi < b_\psi$ is best in our example, a situation not expected to be effective according to the way this theory is developed (Tzavidis *et al.* 2010).

There are some outstanding issues that would benefit from further research. We have used tax data in different years as predictor and outcome variables, but it would be interesting to see whether the results are reproduced with survey data, particularly where the survey response is a different variable from the tax data which act as a predictor. Where the available predictors are not so strong, small area estimation approaches will likely have a smaller impact on the rmse. Robust methods will be most effective if the outliers are the main causes of lack of fit. Further investigation of these robust approaches with different types of variables would therefore be valuable to gauge the limits of their usefulness.

We used small area estimation in isolation, to better understand its properties. In practice, however, we would like to enforce consistency with the main SBS outputs. Benchmarking to the direct estimators at higher levels of aggregation where the rmse is small could achieve this (Pfeffermann *et*

al. 2014), and might also have beneficial impacts on the bias and rmse of the small area estimators. Alternatively, if the aggregated small area estimates are more accurate than the direct estimate of the population total, then benchmarking may introduce too much of the random variation in the direct estimation. We leave this for further investigation.

Acknowledgements

We are grateful to the anonymous referees for some insightful comments on an earlier version of the paper. The opinions expressed are those of the authors, and do not necessarily reflect those of their organisations or of Statistics Netherlands.

7. References

- Battese, G.E., Harter, R.M. & 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** 28-36.
- Breckling, J. & Chambers, R. (1988) M-quantiles. *Biometrika* **75** 761–771.
- Burgard, J.P., Münnich, R. & Zimmermann, T. (2014) The impact of sampling designs on small area estimates for business data. *Journal of Official Statistics* **30** 749-771.
- Chambers, R.L. (1986) Outlier robust finite population estimation. *Journal of the American Statistical Association* **81** 1063-1069.
- Chambers, R., Chandra, H., Salvati, N. & Tzavidis, N. (2014) Outlier robust small area estimation. *Journal of the Royal Statistical Society, Series B* **76** 47-69.
- Chambers, R.L. & Dunstan, R. (1986) Estimating distribution functions from survey data. *Biometrika* **73** 597–604.
- Chambers, R. & Tzavidis, N. (2006) M-quantile models for small area estimation. *Biometrika* **93** 255–268.
- Chandra, H. & Chambers, R. (2011) Small area estimation under transformation to linearity. *Survey Methodology* **37** 39-51.
- Chandra, H., Chambers, R. & Salvati, N. (2012) Small area estimation of proportions in business surveys. *Journal of Statistical Computation and Simulation* **82** 783-795.
- Dawber, J., Salvati, N., Schmid, T. & Tzavidis, N. (in prep) Scale estimation and data-driven tuning constant selection for M-quantile regression.
- Dongmo Jiongo, V., Haziza, D. & Duchesne, P. (2013) Controlling the bias of robust small-area estimators. *Biometrika* **100** 843-858.
- Draper, N.R. & Smith, H. (2014) *Applied regression analysis*, 3rd edition. New York: John Wiley & Sons.
- Fabrizi, E., Ferrante, M.R. & Trivisano, C. (2018) Bayesian small area estimation for skewed business survey variables. *Applied Statistics* **67** 861-879.
- Fabrizi, E., Salvati, N., Pratesi, M. & Tzavidis, N. (2014) Outlier robust model-assisted small area estimation. *Biometrical Journal* **56** 157-175.
- Ferrante, M.R. & Pacei, S. (2017) Small domain estimation of business statistics by using multivariate skew normal models. *Journal of the Royal Statistical Society: Series A* **180** 1057-1088.
- Hidiroglou, M.A. & Smith, P. (2005) Developing small area estimates for business surveys at the ONS. *Statistics in Transition* **7** 527-539.
- Holland, P.W. & Welsch, R.E. (1977) Robust regression using iteratively reweighted least-squares. *Communications in Statistics – Theory and Methods* **6** 813-827.
- Huber, P.J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics* **1** 799-821.
- Krieg, S., Blaess, V. & Smeets, M. (2012) *Small area estimation of turnover of the Structural Business Survey*. Statistics Netherlands discussion paper 2012-03.
- Militino, A.F., Ugarte, M.D. & Goicoa, T. (2015) Deriving small area estimates from information technology business surveys. *Journal of the Royal Statistical Society, Series A* **178** 1051-1067.

- Pfeffermann, D. (2011) Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology* **37** 115-136.
- Pfeffermann, D., Sikov, A. & Tiller, R. (2014) Single- and two-stage cross-sectional and time series benchmarking procedures for small area estimation. *Test* **23** 631–666.
- Rao, J.N.K. (2011) Impact of frequentist and Bayesian methods on survey sampling practice: a selective appraisal. *Statistical Science* **26** 240-256.
- Rao, J.N.K. & Molina, I. (2015) *Small area estimation*. Hoboken: John Wiley & Sons.
- Richardson, A.M. & Welsh, A.H. (1995) Robust restricted maximum likelihood in mixed linear models. *Biometrics* **51** 1429-1439.
- Rivière, P. (2002) What makes business statistics special? *International Statistical Review* **70** 145-159.
- Rojas-Perilla, N., Pannier, S., Schmid, T. & Tzavidis, N. (2019) Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society, Series A* **183** 121-148. doi: 10.1111/rssa.12488.
- Särndal, C.-E., Swensson, B. & Wretman, J. (1992) *Model assisted survey sampling*. New York: Springer.
- Sinha, S.K. & Rao, J.N.K. (2009) Robust small area estimation. *Canadian Journal of Statistics* **37** 381-399.
- Tzavidis, N., Marchetti, S. & Chambers, R. (2010) Robust estimation of small-area means and quantiles. *Australian & New Zealand Journal of Statistics* **52** 167-186.
- Tzavidis, N., Zhang, L-C., Luna Hernandez, A., Schmid, T. & Rojas-Perilla, N. (2018) From start to finish: a framework for the production of small area official statistics (with discussion). *Journal of the Royal Statistical Society, Series A* **181** 927-979. doi: 10.1111/rssa.12364.
- Welsh, A.H., & Richardson, A.M. (1997). Approaches to the robust estimation of mixed models. *Handbook of Statistics* **15** 343-384.
- Welsh, A.H. & Ronchetti, E. (1998) Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society, Series B* **60** 413–428.
- You, Y. & Rao, J.N.K. (2002) A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics* **30** 431-439.