Statistical Methods and Applications manuscript No.
(will be inserted by the editor)

Small area estimation based on M-quantile models in presence of outliers in auxiliary variables

Received: date / Accepted: date

Abstract When using small area estimation models, the presence of outlying observations in the response and/or in the auxiliary variables can severely affect the estimates of the model parameters, which can in turn affect the small area estimates produced using these models. In this paper we propose an M-quantile estimator of the small area mean that is robust to the presence of outliers in the response variable and in the continuous auxiliary variables. To estimate the variability of this estimator we propose a non-parametric bootstrap estimator. The performance of the proposed estimator is evaluated by means of model- and design-based simulations and by an application to real data. In these comparisons we also include the extension of the Robust EBLUP able to down-weight the outliers in the auxiliary variables. The results show that in the presence of outliers in the auxiliary variables the proposed estimator outperforms its traditional version that takes into account the presence of outliers only in the response variable.

Keywords Robust estimation · Robust EBLUP · Bootstrap estimation · Trisquared redescending function · Unit level models

1 Introduction

In the past years there has been growing interest among policy makers and public administrators at both national and local level concerning the need for accurate, timely and reliable data as a prerequisite for good planning, proper management and effective governance (ESS, 2014).

Survey sampling has been recognized as the effective way to obtain reliable estimates at national level or at a given level of aggregation, i.e. sub-national areas. Usually the areas (or domains) not designed in the sample have sample sizes not large enough to allow for reliable direct estimation - i.e. estimates based only on area observations. In these cases one can resort to small area estimation (SAE) techniques.

Various efficient mixed-model-based SAE methods has been developed (Rao and Molina, 2015; Jiang and Lahiri, 2006). For example, the empirical best linear unbiased predictor (EBLUP) based on a linear mixed model is often recommended when the target of inference is the small area average of a continuously distributed variable (Battese et al, 1988; Prasad and Rao, 1990).

Since the presence of outliers is a common feature in real data applications, the issue of outlier robust model-based estimation has recently attracted some interest in the small area estimation literature (Chambers and Tzavidis, 2006; Sinha and Rao, 2009; Tzavidis et al, 2010; Dongmo Jiongo et al, 2013; Chambers et al, 2014). Chambers (1986) differentiates between two typologies of outliers, representative and non-representative. A non-representative outlier is a gross error in the sample data, caused by deficiencies in survey processing (e.g. miscoding). A representative outlier is, instead, a sample outlier that is potentially drawn from a group of population outliers and hence cannot be unit-weighted in estimation.

Chambers and Tzavidis (2006) and Sinha and Rao (2009) proposed robust techniques that can be used to down-weight outliers in the response variable when fitting the underlying model. Chambers and Tzavidis (2006) addressed the issue of outlier robustness in SAE using an approach based on fitting M-quantile models (Breckling and Chambers, 1988) to the survey data. Sinha and Rao (2009) addressed this issue from the perspective of linear mixed models. A comparison of these two alternative approaches can be found in Giusti et al (2014). Chambers et al (2014) defined such methods as robust projective, since they project the behaviour of the robust working model of the sample onto the non-sampled part of the population. Tzavidis et al (2010) and Chambers et al (2014) proposed methods that allow for contributions from representative sample outliers and they are defined as robust predictive methods since they attempt to predict the contribution of the population outliers to target parameters. Gershunskaya (2010) proposed a slight modification of a classical linear mixed model assuming that the underlying distribution is a scale mixture of two normal distributions, where outliers are assumed to have a larger variance than the regular observations. This model explicitly describes the behaviour of the outlying observations relative to the other units; thus, it automatically produces estimates (e.g., using MLE) that account for outliers.

None of the above authors have studied the presence of outlying values in the auxiliary variables, except for a final remark in Sinha and Rao (2009). Outlying values in the auxiliary variables may occur in many real data applications, as representative or non-representative outliers. For example, representative outlying values can be present in income data which can be used to predict consumption expenditure. Non representative outliers can be present in the data when, for example, measurement errors are systematically present but it is not possible to identify them (e.g. satellite data, big data). The presence of outliers in the auxiliary variables may influence the fitting process and the prediction in SAE. It can lead to biased estimates because they can be high leverage points in the factor space (Cook and Weisberg, 1980).

Sinha and Rao (2009) proposed to use a weight function based on the Mahalanobis distance to down-weight any outlier in the continuous auxiliary variables, x's, when estimating the target parameters (see also Maronna et al, 2006; Hubert et al, 2008; Filzmosera et al, 2008; Carroll and Pederson, 1993). However, the authors did not investigate the performance of the small area predictor including this weight function in the auxiliary variables neither by simulation experiments nor by a real data application.

In this article we propose a M-quantile-based estimator for small area means that is robust to the presence of outliers both in the response and auxiliary variables. In particular, the proposed estimator is based on a robust method that down-weights outliers in auxiliary variables by using the trisquared redescending weight function (Carroll and Pederson, 1993). Moreover, we also evaluate the performance of the Robust EBLUP defined following the proposal in Sinha and Rao (2009) by using the trisquared redescending weight function. The main aim is to check if these two new estimators outperform their traditional version when outliers are present in the auxiliary variables, and not to make a comparison between them. The study is limited to the case of continuous x's values.

The rest of the paper is organized as follows. In Section 2 we review the linear mixed model and the EBLUP of the small area mean under this model. In this Section we also show the robust version of the EBLUP proposed by Sinha and Rao (2009) and their suggestion to down-weight outliers in the auxiliary variables. In Section 3 we describe the M-quantile approach to small area estimation. In Section 4 we present the proposed M-quantile-based estimator. In Section 5 the traditional and outlier-robust small area methodologies are contrasted using model- and design-based simulations that employ a wide range of contamination mechanisms for generating population data with outliers in the response variable and in the covariates. In Section 6 the traditional and outlier-robust small area estimators are applied to real data on the production of olives in 53 agrarian regions of Tuscany, Italy. Finally, in Section 7 we summarize our main findings and provide suggestions for further work.

2 Small area estimation using linear mixed model estimators

In a typical small area estimation problem we consider a population U of size N divided into D non-overlapping subsets U_i (domains of study or areas) of size N_i, $i = 1, ..., D$. The unit level models commonly assume that a vector $\boldsymbol{x}_{ij}$ of p auxiliary variables is known without error for each unit j belonging to area i, while the values of the variable of interest y_{ij} are available in each area only for a sample of population units, $s_i \subset U_i$ of size $n_i \geq 0, i = 1, \dots, D$. The set $r_i \subset U_i$ contains the $N_i - n_i$ indices of the non-sampled units in area i. Thus, the y_{ij} ($j \in s_i$, $i = 1, \dots, D$) values are observed only for the sample units belonging to the sets s_i, while they are unknown for non-sample units in the sets r_i. The goal is to use these data to estimate the mean m_i for the variable y in each area. However, when the number of sample units in area i is limited, estimating the area means of y using a direct estimator would usually lead to unreliable estimates and small area methodologies should be used.

Roughly speaking, model-based small area estimation methods "borrow strength" from all the available observations trough an explicit model to obtain more precise estimates for small areas. The most popular approach to model-based small area estimation is based on random effects models, also known as linear mixed models, that include random area effects to account for between area variations (Battese et al, 1988). These models are based on the hypothesis that the following nested error regression model holds at the unit level:

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}, \tag{1}$$

where, $i = 1, \ldots, D$, $j \in U_i$, $\boldsymbol{\beta}$ is a vector of fixed effects, v_i is an area-level random effect and e_{ij} is a unit-level random error. It is also assumed that the sample data obey the same model (with n_i replacing N_i) thus implying that the sampling design is ignorable given the auxiliary variables included in the model. For the error terms it is assumed that $v_i \overset{i.i.d.}{\sim} N(0, \sigma_v^2)$ and $e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_e^2)$. Under this model the EBLUP of the mean is

$$\hat{m}_i^{EBLUP} = N_i^{-1} \left\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} (\boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{v}_i) \right\}, \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{v}_i$ are maximum likelihood or restricted maximum likelihood (ML or REML) estimates for the unknown model parameters $\boldsymbol{\beta}$ and v_i (further details in Rao and Molina, 2015). The mean squared error (MSE) of (2) can be estimated with the estimator proposed by Prasad and Rao (1990), based on a Taylor approximation.

It has been well documented (Fellner, 1986; Huggins, 1993; Richardson and Welsh, 1995) that the estimators of $\boldsymbol{\beta}$ and of the variance components in (1) are sensitive to outliers in the distribution of the random area effect and of the random errors (and then in the response variable), which in turn can affect the small area estimates. For this reason Sinha and Rao (2009) proposed a small area estimator of the mean using an outlier robust version of (2). Denoting by subscript ψ the robust estimates of the fixed and the random effects, the robust version of EBLUP, called Robust EBLUP (REBLUP), is

$$\hat{m}_i^{REBLUP} = N_i^{-1} \left\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} (\boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}_\psi + \hat{v}_{i,\psi}) \right\}. \tag{3}$$

In this expression the fixed effects and variance components are obtained using the Robust ML Proposal II in Richardson and Welsh (1995). The robust random effects are estimated using the equation by Fellner (1986). For estimating the MSE of (3) Sinha and Rao (2009) proposed a parametric bootstrap based on the paper of Hall and Maiti (2006). Estimator (3) is based on the assumption that non-sample data follows the assumed working model. Therefore, this estimator aims to estimating robustly the expected value of the non-sample sum (or mean) of the study variable under this working model on the basis of the outlier-contaminated sample data. This assumption typically leads to biased estimators. Chambers et al (2014) proposed a robust-bias corrected version of (3).

Sinha and Rao (2009) in the final discussion also drew the attention to the presence of outliers in the auxiliary variables. In particular, they briefly suggested to modify the robust ML estimating equation of $\boldsymbol{\beta}_\psi$ in (3) by applying to the $\boldsymbol{x}$ some weight function in order to down-weight the outliers in the auxiliary variables. When the $\boldsymbol{x}$ are continuous Sinha and Rao (2009) suggested to choose a weight function that is based on the Mahalanobis distance $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$, where $\boldsymbol{\mu}$ and $\boldsymbol{V}$ denote some robust estimates of the location and scale of $\boldsymbol{x}$. However, they did not present any additional detail or result for these proposed estimators.

3 The M-quantile approach to small area estimation

Chambers and Tzavidis (2006) proposed an alternative approach to small area estimation that relaxes the parametric assumptions of mixed effects models underlying the EBLUP. This method is based on M-quantile regression and it guarantees robustness with respect to outlier observations in the target variable. In this part of the Section we define the linear M-quantile regression model.

The M-quantile of order $q \in (0,1)$ for the conditional density of y given the set of covariates $\boldsymbol{x}$, $f(y|\boldsymbol{x})$, is defined as the solution $Q_y(q|\boldsymbol{x},\psi)$ of an estimating equation $\int \psi_q\{y - Q_y(q|\boldsymbol{x},\psi)\} f(y|\boldsymbol{x})\,dy = 0$, where ψ_q denotes an asymmetric influence function, which is the derivative of an asymmetric loss function ρ_q. In particular, a linear M-quantile regression model for y_{ij} given $\boldsymbol{x}_{ij}$ is one where we assume that

$$Q_y(q|\boldsymbol{x}_{ij},\psi) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q). \tag{4}$$

That is, we allow a different set of p regression parameters for each value of $q \in (0,1)$. The estimator of $\boldsymbol{\beta}_\psi(q)$ can be obtained by solving $\sum_{i=1}^D \sum_{j\in s_i} \psi_q(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q))\boldsymbol{x}_{ij} = \boldsymbol{0}$ with respect to $\boldsymbol{\beta}_\psi(q)$, assuming that $\psi_q(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q)) = 2\psi\{s^{-1}(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q))\}\{qI(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q) > 0) + (1-q)I(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q) \le 0)\}$, where s is a suitable robust estimate of scale, e.g. the MAD estimate $s = median\left|y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q)\right|/0.6745$. A popular choice for the influence function is the Huber, $\psi(u) = uI(|u| \le c) + c\,\mathrm{sgn}(u)I(|u| > c)$ (Chambers and Tzavidis, 2006). However, other influence functions are also possible. Provided that the tuning constant c is strictly greater than zero, estimates of $\boldsymbol{\beta}_\psi(q)$ are obtained using iterative weighted least squares (IWLS).

Chambers and Tzavidis (2006) extended the use of M-quantile regression models to small area estimation. They characterized the conditional variability across the population of interest by the M-quantile coefficients of the population units. For unit j in area i this coefficient is the value q_{ij} such that $Q_y(q_{ij}|\boldsymbol{x}_{ij}, \psi) = y_{ij}$. The M-quantile coefficients are determined at the population level. Consequently, if a hierarchical structure does explain part of the variability in the population data, then we expect units within clusters defined by this hierarchy to have similar M-quantile coefficients.

When the conditional M-quantiles are assumed to follow the linear model (4), with $\boldsymbol{\beta}_\psi(q)$ a sufficiently smooth function of q, the M-quantile (MQ) estimator of the mean in area i can be defined as

$$\hat{m}_i^{MQ} = N_i^{-1}\Big\{ \sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) \Big\}, \tag{5}$$

where $\hat{\theta}_i$ is an estimate of the average value of the M-quantile coefficients of the units in area i. See Chambers and Tzavidis (2006) for further details on the estimation of the M-quantile coefficients at unit level and for the computation of the small area M-quantile coefficients. Chambers and Tzavidis (2006) observed that the M-quantile mean estimator (5) can be biased, especially in the presence of heteroskedastic and/or asymmetric errors. This observation motivated the work in Tzavidis et al (2010) and Chambers et al (2014).

Analytic estimators of the MSE of (5) are described in Chambers and Tzavidis (2006), Chambers et al (2011) and Chambers et al (2014). The estimator proposed by Chambers et al (2011) is a first order approximation to the MSE of (5) and has been proved to be bias robust against model misspecification; however, its main criticism is that it can be unstable especially with small area-specific sample sizes. Chambers et al (2014) proposed an alternative analytic estimator conditional on the realized values of the area effects, which is defined by the solution of a set of robust estimating equations. Under this approach, the MSE of (5) is shown to be the sum of a prediction variance, a squared bias term and a correction term accounting for the sampling variability of parameter estimates. Tzavidis et al (2010) proposed a non-parametric bootstrap approach to estimate the MSE of (5), based on the approach of Lombardía et al (2003). This bootstrap estimator can be applied both for small area means and for non-linear quantities such as small area quantiles and poverty indicators estimated using M-quantile estimators (Marchetti et al, 2012). Although computationally intensive, the non-parametric bootstrap MSE estimator has smaller bias and is more stable than the analytic MSE estimator proposed by Chambers et al (2011).

4 M-quantile estimators with outliers in the auxiliary variables

When auxiliary variables are affected by the presence of outlying observations, traditional M-quantile estimators of the small area mean may be not appropriate, since they can lead to biased estimates. In this Section we propose a M-quantile estimator of the small area mean that is robust to the presence of outliers in the auxiliary variables.

Under two alternative scenarios we suppose that a proportion ζ of the $\boldsymbol{x}_{ij}$ in the population (scenario 1) or in the sample (scenario 2) are contaminated, so we observe $\boldsymbol{x}^*_{ij} = \{(1-\zeta)\boldsymbol{x}_{ij}, \zeta(\boldsymbol{x}_{ij} + \eta_{ij})\}$ where $\eta_{ij} \sim N(\boldsymbol{\gamma}, \sigma^2_\eta)$. That is, a proportion ζ of the observed covariates $\boldsymbol{x}^*_{ij}$ is affected by the presence of outlying observations that cause a shift both in location and variability with respect to the population or sampled $\boldsymbol{x}_{ij}$. We propose to deal with the presence of these outlying observations using a robust method that down-weights on the basis of extreme leverage values (Carroll and Pederson, 1993). The qth M-quantile $Q_y(q|\boldsymbol{x}^*, \psi)$ of the conditional distribution of y given $\boldsymbol{x}^*$ in the population satisfies:

$$Q_y(q|\boldsymbol{x}^*_{ij}, \psi) = \boldsymbol{x}^{*T}_{ij}\boldsymbol{\beta}_{\psi,w}(q) \tag{6}$$

where ψ denotes the influence function associated with the M-quantile, and w is the weighted function used to down-weight the extreme values of the auxiliary variables.

We choose the *Trisquared redescending* weight function (Carroll and Pederson, 1993) to down-weight the values of the auxiliary variables - excluding the intercept - that is defined as follows

$$w(t) = \{1 - (t/k)^2\}^3 I(|t| \leq k),$$

where k is a tuning constant. A weight function $w(t)$ depends on the influence function $\psi(t)$ since $w(t) = \psi(t)/t$. A ψ-function is called redescending if $\psi(x) = 0$ for all $x \geq x_r$ for $x_r < \infty$, and x_r is often called rejection point. The advantages on the use of redescending functions in M-estimation is explained in Maronna et al (2006).

To down-weight outlying values of the auxiliary variables we use a robust Mahalanobis distance, as suggested in Sinha and Rao (2009). In particular, we define for each observation the value $z_{ij} = (\boldsymbol{x}^*_{ij} - \boldsymbol{\mu})^T \boldsymbol{V}^{-1}(\boldsymbol{x}^*_{ij} - \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is a $p-1$ vector of robust estimates of the centres of the $p-1$ auxiliary variables (intercept is excluded) and $\boldsymbol{V}$ is a robust estimate of the $(p-1) \times (p-1)$ covariance matrix of the auxiliary variables. Defining $u^2_{ij} = z_{ij}/(p-1)$, the weight used to down-weight the extreme values is

$$w(u_{ij}) = w(\boldsymbol{x}^*_{ij}) = \{1 - (u_{ij}/k)^2\}^3 I(|u_{ij}| \leq k), \tag{7}$$

where we underline that u_{ij} is a function of $\boldsymbol{x}^*_{ij}$. Given that u_{ij} is a distance from the centre, choosing $k = 8$ gives weight 0.75 or greater to those points with distance less than or equal to 2.5, weight 0.5 or greater if the distance is less than or equal to 3.6, and weight 0.25 or greater if the distance is less than 5.0 (Carroll and Pederson, 1993). The value of the tuning constant can be chosen using a cross-validation technique as the one presented in Section 5.

We have also compared the weights obtained from the Trisquared function with the widely used Bisquare function and also with other weighting functions such as the Hampel, the Generalized Gaussian Weight (GGW), the Linear Quadratic Quadratic (LQQ) and the Welsh. All these funtions are in the class of redescending functions (Maronna et al, 2006; Koller and Stahel, 2011; Hampel et al, 1986). The results of this comparison are discussed in Section 5.

The estimator of $\boldsymbol{\beta}_{\psi,w}(q)$ can be obtained by solving

$$\sum_{i=1}^{D}\sum_{j\in s_i}\psi_q(y_{ij}-\boldsymbol{x}_{ij}^{*T}\boldsymbol{\beta}_{\psi,w}(q))w(\boldsymbol{x}_{ij}^{*})\boldsymbol{x}_{ij}^{*T}=\mathbf{0}$$

with respect to $\boldsymbol{\beta}_{\psi,w}(q)$. For fixed q the estimator of $\boldsymbol{\beta}_{\psi,w}(q)$ is as follows

$$\hat{\boldsymbol{\beta}}_{\psi,w}(q)=\left\{\boldsymbol{X}^{*T}\boldsymbol{W}^{*}(q,\boldsymbol{x}^{*})\boldsymbol{X}^{*}\right\}^{-1}\boldsymbol{X}^{*T}\boldsymbol{W}^{*}(q,\boldsymbol{x}^{*})\boldsymbol{y}, \tag{8}$$

where $\boldsymbol{y}$ is the n-vector of response values, $\boldsymbol{X}^{*}$ is the corresponding $n\times p$ matrix of the observed auxiliary variables and $\boldsymbol{W}^{*}(q,\boldsymbol{x}^{*})=\boldsymbol{W}_1(\boldsymbol{X}^{*})\boldsymbol{W}_2(q)$ is a diagonal matrix of order n which contains the final set of weights produced by the IWLS algorithm used to compute $\hat{\boldsymbol{\beta}}_{\psi,w}(q)$. Here, $\boldsymbol{W}_1(\boldsymbol{X}^{*})$ is an $n\times n$ diagonal matrix with elements $w(\boldsymbol{x}_{ij}^{*})$ and $\boldsymbol{W}_2(q)$ is a diagonal matrix of order n with elements $\psi_q(y_{ij}-\boldsymbol{x}_{ij}^{*T}\boldsymbol{\beta}_{\psi,w}(q))/(y_{ij}-\boldsymbol{x}_{ij}^{*T}\boldsymbol{\beta}_{\psi,w}(q))$. The equation (8) takes into account both the down-weight of the influence points for the response and auxiliary variables. Moreover, the IWLS algorithm guarantees the convergence to a unique solution (Kokic et al, 1997; Chambers and Tzavidis, 2006).

When (6) holds, the proposed M-quantile estimator (MQ/RobX) of m_i that accounts for the presence of outliers in both the response and the auxiliary variables is as follows

$$\hat{m}_i^{MQ/RobX}=N_i^{-1}\left[\sum_{j\in s_i}y_{ij}+\frac{N_i}{\sum_{j\in r_i}w(\boldsymbol{x}_{ij}^{*})}\sum_{j\in r_i}w(\boldsymbol{x}_{ij}^{*})\boldsymbol{x}_{ij}^{*T}\hat{\boldsymbol{\beta}}_{\psi,w}(\hat{\theta}_i)\right]. \tag{9}$$

Equation (9) follows from the assumption that the mean of the down-weighted auxiliary variables is equal to the mean of the unobservable uncontaminated auxiliary variables:

$$\left(\sum_{j\in r_i}w(\boldsymbol{x}_{ij}^{*})\right)^{-1}\sum_{j\in r_i}w(\boldsymbol{x}_{ij}^{*})\boldsymbol{x}_{ij}^{*}=N_i^{-1}\sum_{j\in r_i}\boldsymbol{x}_{ij}\,. \tag{10}$$

Using the equality (10) in (5) it is straightforward to obtain the estimator in (9).

Note that estimators $\hat{m}_i^{MQ}$ and $\hat{m}_i^{MQ/RobX}$ differ in the method used to estimate the regression coefficients ($\boldsymbol{\beta}_{\psi}$ in (5) and $\boldsymbol{\beta}_{\psi,w}$ in (9)), and in the use of the weighted x in equation (9).

We also defined and tested an alternative version of estimator in (9) by including the bias adjustment proposed in Tzavidis et al (2010). However, this adjustment though reducing the bias, caused a great increase in the variability of the estimator resulting in a worse MSE when compared to (9). This behaviour has been already noticed in the literature (Chambers et al, 2014; Giusti et al, 2014).

The estimator proposed in (9) may provide systematic bias if outliers in auxiliary variables are concentrated only in few areas. If this is the case, we propose to compute the weights $w(\boldsymbol{x}^*_{ij})$ conditional to the areas (or group of areas) - i.e. use only area i units in (7). So for those areas where there are no outliers the weights should be approximately equal to one and for those areas where there are outliers in the auxiliary variables the weights will down-weight the outlying values.

Applying the weight function (7) to the REBLUP estimator and following the proposal in Section 8 of Sinha and Rao (2009) we obtain a REBLUP that is robust also for outliers in the auxiliary variables, hereafter REBLUP/RobX.

We remark that estimators MQ/RobX and REBLUP/RobX depend on unit level auxiliary variables for non-sampled units unlike the ordinary MQ and REBLUP estimators where only area level totals are needed for small area estimation of the mean.

The contamination model can be considered similar to models that assume one ore more auxiliary variables subject to measurement error. Ybarra and Lohr (2008) proposed an area-level small area estimator based on the Fay-Herriot model that accounts for sampling variability in the auxiliary information, and derived its properties, in particular showing that it is approximately unbiased. Ghosh et al (2006) proposed a nested error linear regression population model with an area-level covariate, x, subject to a structural measurement error, but the proposed Bayes predictor does not involve sample information on the covariate. Torabi et al (2009) proposed a full efficient Bayes predictor and then the EB predictor. In the last two papers the authors assume that the true value of covariate (say x_i) remains the same for all the units in a small area, and it is measured with error as $x_{ij} = x_i + \eta_{ij}$ for the j-th sample unit in small area i. This is the main difference with our proposal, since we assume the presence of outliers in auxiliary variables only for some units of the population. The use of the models proposed by Ghosh et al (2006) and Torabi et al (2009) in presence of outliers could be an aim of future research.

4.1 Mean-squared error estimation

In this Section we propose a non-parametric bootstrap technique to estimate the mean squared error of $\hat{m}_i^{MQ/RobX}$. This technique follows the work in Marchetti et al (2012).

We start from the following M-quantile model:

$$y_{ij} = \boldsymbol{x}^{*T}_{ij} \boldsymbol{\beta}_{\psi,w}(\theta_i) + e_{ij} \,, \tag{11}$$

where θ_i is the unknown M-quantile of area i, $\boldsymbol{\beta}_{\psi,w}$ is the unknown vector of regression parameters for unknown M-quantile θ_i and e_{ij} is a vector of random errors.

Estimate the parameters of (11) according to the method described in Section 4, then compute the residuals $\hat{e}_{ij} = y_{ij} - \boldsymbol{x}^{*T}_{ij} \hat{\boldsymbol{\beta}}_{\psi,w}(\hat{\theta}_i)$. Generate B

bootstrap populations as follows:

$$y_{ij}^{\star b} = w(\boldsymbol{x}_{ij}^{*})\boldsymbol{x}_{ij}^{*T}\hat{\boldsymbol{\beta}}_{\psi,w}(\hat{\theta}_i) + \hat{e}_{ij}^{\star b} \quad j \in s_i, i = 1,\dots,D \tag{12}$$

$$y_{ik}^{\star b} = w(\boldsymbol{x}_{ik}^{*})\boldsymbol{x}_{ik}^{*T}\hat{\boldsymbol{\beta}}_{\psi,w}(\hat{\theta}_i) + \hat{e}_{ik}^{\star b} \quad k \in r_i, i = 1,\dots,D \tag{13}$$

where $\hat{e}_{ij}^{\star b}$ and $\hat{e}_{ik}^{\star b}$ are obtained sampling with replacement from a smoothed distribution of the re-centred model residuals $\hat{e}_{ij}$. The tuning constant k of the weight function in (12) and (13) is computed minimizing $\sum_{i=1}^{D}\sum_{j\in s_i}(y_{ij} - \hat{y}_{ij}^{\star b})^2$ with respect to k. Since $\hat{\boldsymbol{\beta}}_{\psi,w}$ is held fixed, this minimization is fast.

From each of the B bootstrap populations generate L samples $(y^{\star bl}, \boldsymbol{x}^{*bl})$ with the same sample sizes of the original sample as follow

$$y_{ij}^{\star bl} = \boldsymbol{x}_{ij}^{*T}\hat{\boldsymbol{\beta}}_{\psi,w}(\hat{\theta}_i) + \hat{e}_{ij}^{\star bl},$$

where $\hat{e}_{ij}^{\star bl}$ is drown conditionally to the areas from the re-centred residuals $\hat{e}_{ij}$ (Chambers et al, 2016).

The estimator of the mean in area i for the bootstrap population b using the bootstrap sample l $(l = 1,\dots,L)$, $\hat{m}_i^{MQ/RobX,bl}$, is obtained applying equation (9) on $(y^{\star bl}, \boldsymbol{x}^{*bl})$. For the bootstrap population b the true mean in area i is $m_i^b = N_i^{-1}\sum_{j\in U_i} y_{ij}^{\star b}$.

The bootstrap estimators of the bias and variance of the estimated target small area parameter, $\hat{m}_i^{MQ/RobX}$, are defined respectively by

$$\widehat{Bias}\left(\hat{m}_i^{MQ/RobX}\right) = B^{-1}L^{-1}\sum_{b=1}^{B}\sum_{l=1}^{L}\left(\hat{m}_i^{MQ/RobX,bl} - m_i^b\right)$$

$$\widehat{Var}\left(\hat{m}_i^{MQ/RobX}\right) = B^{-1}L^{-1}\sum_{b=1}^{B}\sum_{l=1}^{L}\left(\hat{m}_i^{MQ/RobX,bl} - \bar{\hat{m}}_i^{MQ/RobX,bl}\right)^2,$$

where $\bar{\hat{m}}_i^{MQ/RobX,bl} = L^{-1}\sum_{l=1}^{L}\hat{m}_i^{MQ/RobX,bl}$.

Finally, the bootstrap MSE estimator of the estimated small area target parameter is then defined as follows:

$$mse\left(\hat{m}_i^{MQ/RobX}\right) = \widehat{Var}\left(\hat{m}_i^{MQ/RobX}\right) + \widehat{Bias}\left(\hat{m}_i^{MQ/RobX}\right)^2. \tag{14}$$

A less computationally demanding MSE estimator could be developed following the ideas that are set out in Chambers et al (2011) and Chambers et al (2014). The first proposal is a bias robust MSE estimator that is based on the pseudolinearization approach of Chambers et al (2011). The second method is based on first-order approximations to the variances of solutions of estimating equations (Chambers et al, 2014). However, this extension is not an objective of this paper and its research worth pursuing in the future.

5 Empirical evaluation of the performance of the proposed estimators

In this Section we use model-based Monte-Carlo simulations and a design-based simulation to provide empirical evidence of the performance of the proposed small area robust estimators of the mean. Moreover, we also evaluate the performance of the proposed bootstrap MSE estimator of the MQ/RobX mean estimator.

5.1 Model-based simulations

The behaviour of the mean estimators by means of model-based simulations is assessed under four different scenarios: A) non-representative outliers in the auxiliary variable and absence of outliers in the target variable; B) non-representative outliers in the auxiliary variable and presence of outliers in the target variable; C) representative outliers in the auxiliary variable and absence of outliers in the target variable; D) representative outliers in the auxiliary variable and presence of outliers in the target variable.

In what follows subscript i identifies small areas, $i = 1, \dots, D$ and subscript j identifies units in a given area. Population data for $D = 30$ areas are generated under a random intercept model,

$$y_{ij} = 1 + 2x_{ij} + v_i + e_{ij}.$$

In each of the four scenarios the area population size is set to $N_i = 100$ units $\forall\ i$ and a simple random sample of size $n_i = 5$ and 10 is selected from each area. The use of two alternative sample sizes enables us to assess the effect of the domain sample sizes both on the bias and the stability of the small area mean estimator.

In scenarios A and B the population values of the auxiliary variable are generated from the Normal distribution with mean 5 and variance 1. Then, the sampled values of the auxiliary variable x_{ij} are contaminated with a proportion of $\zeta = 0.01, 0.03, 0.05$ values drawn from a Normal distribution with mean 15 and variance 1. In scenarios C and D the population values of the auxiliary variable are generated from a contaminated distribution with a proportion of $(1-\zeta)$ values drawn from a Normal distribution with mean 5 and variance 1, and the remaining ζ values drawn from a Normal distribution with mean 15 and variance 1 (table 1).

Table 1 Auxiliary variable generation scheme in the four simulation scenarios

	x_{ij} (population)	x^*_{ij} (sample)
Scenario A and B	$N(5,1)$	$(1-\zeta)N(5,1) + \zeta N(15,1)$
Scenario C and D	$(1-\zeta)N(5,1) + \zeta N(15,1)$	–

In scenarios B and D the area-specific random effects v_i are generated from a contaminated distribution. The notation used in table 2 for scenarios B and D means that a given proportion (0.85) of the v_i's are generated from the underlying "true" distribution $N(0,3)$ and the remaining proportion (0.15) of the random effects are generated from the contaminated distribution $N(0,30)$. The same technique is used to generate unit random effects.

Table 2 Area and unit random errors generation schemes in the four simulation scenarios.

	v_i	e_{ij}
Scenario A and C	$N(0,3)$	$N(0,6)$
Scenario B and D	$0.85N(0,3) + 0.15N(0,30)$	$0.9N(0,6) + 0.1N(0,150)$

Combining alternative sample size values ($n_i = 5, 10$) with three different contamination proportions ($\zeta = 0.01, 0.03, 0.05$) for each of the four scenarios (A, B, C and D) results in a total of 24 alternative simulation configurations.

For the estimator MQ/RobX we choose ψ as the Huber function (with tuning constant equal to 1.345). The tuning constant in equation (7) is set to $k = 10, 12, 14$ respectively to the contaminated proportions of the auxiliary variable $\zeta = 0.05, 0.03, 0.01$. As an alternative, to choose the tuning constant value in practical applications we propose a cross validation criterion, similar in spirit to the one proposed by Rudemo (1982), Stone (1984) and Bowman (1984) (see also Stone (1974)). The idea is to use the leave-one-out cross validation to minimize a squared loss function with respect to the tuning constant k. Let $\hat{m}_i^{RobX}$ be the mean estimates in area i and let $\hat{m}_{i,-j}^{RobX}$ be the mean estimate in area i estimated without observation j. Note that leaving one observation out affects the estimates in all the areas given that some parameters (i.e. the $\boldsymbol{\beta}_{\psi,w}$) are estimated using all the sample values. The optimal parameter k minimizes the following cross validation function:

$$\sum_{i=1}^{D} \sum_{j \in s_i} (\hat{m}_i^{RobX} - \hat{m}_{i,-j}^{RobX})^2 .$$

Given that this criterion is computationally intensive, it is not feasible to compute the k parameter for each Monte Carlo run. This is why we set three fixed values for the three corresponding levels of contamination for the auxiliary variable.

We run in total 1000 Monte-Carlo simulations. Denoting by m_i the true mean in small area i and by $\hat{m}_i$ a corresponding estimate, the performance of the estimators is evaluated using the Relative Bias (RB) and the efficiency with respect to the empirical RMSE of the EBLUP (EFF). Their definition is as follows

$$RB(\hat{m}_i) = H^{-1} \sum_{h=1}^{H} \Big(\frac{\hat{m}_{i,h} - m_{i,h}}{m_{i,h}} \Big), \tag{15}$$

$$EFF(\hat{m}_i) = \frac{\sqrt{\sum_{h=1}^{H} (\hat{m}_{i,h} - m_{i,h})^2 / H^{-1}}}{\sqrt{\sum_{h=1}^{H} (\hat{m}_{i,h}^{EBLUP} - m_{i,h})^2 / H^{-1}}}, \tag{16}$$

where for area i and Monte Carlo iteration $h = 1, \dots, H = 1000$, $m_{i,h}$ and $\hat{m}_{i,h}$ are respectively the population mean and an estimated mean of the y variable (MQ, MQ/RobX, REBLUP, REBLUP/RobX, EBLUP).

In tables 3, 4, 5 and 6 we show the results of the simulations in terms of percentage RB and percentage EFF for scenarios A, B, C and D respectively. Our aim is to check if the MQ/RobX performs better than the MQ, and if the REBLUP/RobX performs better than the REBLUP. The EBLUP estimator is used as benchmark. Furthermore, the use of MQ, REBLUP and EBLUP estimators allows us to check if the contamination of the auxiliary variable influences those estimators that do not take into account the presence of outliers in the auxiliary variables.

Table 3 Scenario A. Averages over areas and simulations of the RB(%) and the EFF (%) for EBLUP, MQ, REBLUP, MQ/RobX, REBLUP/RobX.

	$\zeta = 0.01$		$\zeta = 0.03$		$\zeta = 0.05$	
	RB%	EFF%	RB%	EFF%	RB%	EFF%
			n_i=5			
EBLUP	-0.66	100.0	-1.21	100.0	-1.41	100.0
MQ	-0.40	94.8	-1.24	98.3	-1.45	102.6
REBLUP	-0.27	91.9	-1.05	100.6	-1.37	104.8
MQ/RobX	-0.21	91.3	-0.77	85.4	-1.25	87.2
REBLUP/RobX	-0.09	90.2	-0.25	84.1	-0.42	87.1
			n_i=10			
EBLUP	-0.88	100.0	-1.32	100.0	-1.41	100.0
MQ	-0.59	101.9	-1.31	106.7	-1.37	108.5
REBLUP	-0.41	88.8	-1.18	101.6	-1.38	104.6
MQ/RobX	-0.31	98.5	-0.83	94.0	-1.16	95.4
REBLUP/RobX	-0.09	85.1	-0.19	80.0	-0.32	83.2

We can see that estimators robust to outliers in the auxiliary variables always show a gain in efficiency with respect to the corresponding non robust (against outliers in auxiliary variables) estimators.

In more details, the MQ/RobX estimator shows a gain in efficiency (smaller EFF) with respect to the MQ while the relative bias (RB) is similar. The same applies to the REBLUP/RobX and the REBLUP. In scenarios A and C - where there are no outliers in the response variable - the MQ is sometime worse than the EBLUP while the MQ/RobX is always more efficient that the EBLUP, but the gain in efficiency with respect to the EBLUP is small. Also the REBLUP in some cases is outperformed by the EBLUP, while the REBLUP/RobX is

Table 4 Scenario B. Averages over areas and simulations of the RB(%) and the EFF (%) for the EBLUP, MQ, REBLUP, MQ/RobX, REBLUP/RobX.

	$\zeta = 0.01$		$\zeta = 0.03$		$\zeta = 0.05$	
	RB%	EFF%	RB%	EFF%	RB%	EFF%
			n_i=5			
EBLUP	-0.80	100.0	-1.40	100.0	-1.57	100.0
MQ	-0.47	64.5	-1.32	70.1	-1.54	72.7
REBLUP	-0.38	58.7	-1.24	66.8	-1.51	69.3
MQ/RobX	-0.22	62.7	-0.85	63.6	-1.26	66.6
REBLUP/RobX	-0.12	57.6	-0.32	58.9	-0.47	63.2
			n_i=10			
EBLUP	-0.75	100.0	-1.18	100.0	-1.25	100.0
MQ	-0.55	68.8	-1.16	73.2	-1.25	74.4
REBLUP	-0.40	55.6	-1.10	64.4	-1.33	66.8
MQ/RobX	-0.26	67.2	-0.88	69.1	-1.19	70.7
REBLUP/RobX	-0.03	53.5	-0.10	54.5	-0.16	58.2

Table 5 Scenario C. Averages over areas and simulations of the RB(%) and the EFF (%) for the EBLUP, MQ, REBLUP, MQ/RobX, REBLUP/RobX.

	$\zeta = 0.01$		$\zeta = 0.03$		$\zeta = 0.05$	
	RB%	EFF%	RB%	EFF%	RB%	EFF%
			n_i=5			
EBLUP	0.28	100.0	0.19	100.0	0.05	100.0
MQ	0.64	98.8	0.76	102.7	0.33	105.1
REBLUP	0.85	97.1	0.86	103.4	0.30	104.9
MQ/RobX	-0.11	90.9	-0.54	86.4	-1.27	89.6
REBLUP/RobX	0.09	89.6	0.00	86.1	-0.35	92.9
			n_i=10			
EBLUP	0.21	100.0	0.06	100.0	0.30	100.0
MQ	0.75	107.0	0.54	106.8	0.59	107.1
REBLUP	0.92	95.9	0.60	104.1	0.52	104.8
MQ/RobX	0.02	99.9	-0.46	95.9	-0.89	96.2
REBLUP/RobX	0.21	85.9	0.14	82.8	-0.12	83.7

always better if we look at the efficiency. Looking at scenarios B and D the results are different. The MQ is always more efficient than the EBLUP, as expected. The same applies to the REBLUP. Moreover, the simulations show an additional gain in efficiency of the MQ/RobX and the REBLUP/RobX with respect to the MQ and the REBLUP respectively. As expected, the smaller EFF is showed by the REBLUP/RobX, given that the simulation configurations are based on a linear mixed model with symmetric error terms. Different results are obtained under the design-based simulation where the generation process of the data is unknown.

From the model-based simulation results we observe that as the contamination proportion ζ grows, the variability of the estimators also becomes higher, as expected (these results are not showed here). A higher small area sample size n_i and the absence of outliers in the target variable have instead the effect of decreasing the variability of the estimators, all the other factors being constant.

Table 6 Scenario D. Averages over areas and simulations of the RB(%) and the EFF (%) for the EBLUP, MQ, REBLUP, MQ/RobX, REBLUP/RobX.

	$\zeta = 0.01$		$\zeta = 0.03$		$\zeta = 0.05$	
	RB%	EFF%	RB%	EFF%	RB%	EFF%
			n_i=5			
EBLUP	0.34	100.0	0.14	100.0	0.22	100.0
MQ	0.60	66.1	0.58	72.8	0.35	73.9
REBLUP	0.75	61.9	0.65	68.8	0.35	70.5
MQ/RobX	-0.09	62.2	-0.60	65.2	-1.11	68.1
REBLUP/RobX	0.11	58.0	0.05	60.9	-0.28	67.9
			n_i=10			
EBLUP	0.12	100.0	0.18	100.0	0.25	100.0
MQ	0.52	69.1	0.58	73.1	0.48	74.0
REBLUP	0.69	57.1	0.68	64.5	0.42	66.1
MQ/RobX	-0.14	66.4	-0.47	68.4	-0.98	71.5
REBLUP/RobX	0.11	52.8	0.29	55.1	-0.02	57.6

5.2 Design-based simulation

The design-based simulation is based on an artificial population of enterprises generated using real tax-turnover data coming from the Structural Business Survey carried out by the Central Bureau of Statistics, Netherlands. The target variable is the *tax-turnover* in the sector of retail trade. The auxiliary variable is the *turnover*, which is highly correlated with the target. The target of estimation is the average tax-turnover in 20 industry domains.

The original artificial population has about 64000 units. Since the R function we use to estimate the REBLUP and REBLUP/RobX performs slowly, we decide to select randomly a sub set from the original artificial population. The population sub set has been selected with a stratified random sample where the strata are the 20 industry domains, obtaining a domain size that varies between 100 and 840 units for a total of about 6600 units that represent our target population. The correlation between target and auxiliary variable is about 0.96 and there is presence of heteroschedasticity.

In our artificial population the target variables Y –i.e. the tax-turnover– varies between -0.017 to 261.8 with an average of 4.5 (standard deviation 10) and a median of 1.7. The auxiliary variable X –i.e. the turnover– varies between -0.014 to 121.6 with an average of 4.3 (standard deviation 9) and a median equal to 1.7.

To evaluate the effect of outliers in the auxiliary variable we randomly select 5% of units in the population and we add to them a constant value equal to 25. The contaminated units in the population are held fixed in the simulation. The correlation between target and auxiliary variables decreases to 0.82 after the contamination. The contaminated X variable has the same range of the uncontaminated one, a median of 1.8 and an average of 5.6 (standard deviation 10.5).

From the new artificial population we select 500 samples with a sample size that is the 5% of the population size, so it varies between 5 and 42 units with an average sample size of 16. On each sample we estimate the average tax-turnover in the 20 industry domains using the EBLUP, REBLUP, MQ, REBLUP/RobX and MQ/RobX.

The simulation results are summarized in table 7 where we report the percentage relative bias and the percentage efficiency computed accordingly to (15) and (16) respectively. We can see that the less biased estimator is the MQ/RobX (-3%) followed by the EBLUP (4%) and the MQ (8%). The REBLUP and REBLUP/RobX in this simulation are highly biased (32% and 26% respectively). These results are probably due to the bias of the estimates of the random area effects, because, with these data, the normality assumption of the linear mixed model is violated. Looking at the efficiency with respect to the EBLUP, we can see that the MQ/RobX is more efficient than the MQ (84.6% vs 89%), the REBLUP/RobX is more efficient than the REBLUP (85.9% vs 97%) and the MQ/RobX performs the best. All the estimators are more efficient than the EBLUP.

The results of the design-based simulation justify the use of the MQ/RobX estimator as an alternative to the REBLUP/RobX.

Table 7 Design-based simulation. Averages over areas and simulations of the RB (%) and the EFF (%) for EBLUP, MQ, REBLUP, MQ/RobX, REBLUP, REBLUP/RobX.

	RB%	EFF%
EBLUP	4.03	100.0
MQ	8.39	89.0
MQ/RobX	-3.02	84.6
REBLUP	32.2	97.0
REBLUP/RobX	26.9	85.9

We conclude that taking into account for the outliers in the auxiliary variables can significantly increase the efficiency of the proposed estimators, in particular when there are outliers both on the target and on the auxiliary variables.

5.3 Sensitivity to the choice of the weight function

The robust function we use to down-weight the auxiliary variable seems to work properly on both the MQ/RobX and on the REBLUP/RobX estimators. We carry out a comparison among the different weight functions, computing the down-scale weights on a variable generated similarly to the auxiliary of the model-based simulations. We tune the weight functions constants so that each influence function has an asymptotic efficiency of the regression estimator equal to 95%. What emerges from this comparison is that the resulting weights are very similar to each other, with thin differences. The correlation coefficients between the Trisquared weights and the other redescending functions range between 0.999 (Bisquared) to 0.979 (Hampel). The mean absolute difference ranges from 0.003 (Bisquared) to 0.044 (Hampel). Since the Hampel weights seem to be the most different from the Trisquared ones, we carry out model-based simulations on scenarios A, B, C and D using the Hampel function to down-weight the x. We observe a very similar behaviour of MQ/RobX and REBLUP/RobX estimators with respect to themselves with Trisquared weights (results are not reported here). However, using the Hampel there is a little bit less efficiency in both MQ/RobX and REBLUP/RobX. Since the GGW and LQQ functions are more similar to the Hampel while the Bisquared and Welsh functions are more similar to the Trisquared (being, however, all very similar to each other), we suggest to use the latter redescending weight functions as an alternative to the Trisquared.

5.4 Evaluation of the performance of the mean squared error estimator

As concerns the estimation of the MSE, we evaluate the bootstrap estimator (14) proposed in Section 4.1. We use one bootstrap population ($B = 1$) from which we draw 500 bootstrap samples. Since the evaluation of the bootstrap MSE estimator is taking place within a Monte-Carlo framework, the generation of a new Monte-Carlo population and of a new bootstrap population at each iteration imitates the generation of many bootstrap populations. For more comments about the effect of the number of bootstrap populations B and on the choice of the value of B in real applications refer to Marchetti et al (2012). As specified in Section 4.1, we draw residuals from a smoothed version of the error distribution unconditionally to the areas (for further details on this technique see Marchetti et al (2012)). In order to smooth the model residuals we use the Epanechnikov kernel density, $k(t) = (3/4)(1 - t^2)I(|t| < 1)$, where the smoothing parameter is chosen so that it minimizes the cross-validation criterion suggested by Bowman et al (1998). To compute the smoothing parameter we use the `np` package (Hayfield and Racine, 2008) in the R environment (R Development Core Team, 2010). In particular, we use the function `npudensbw` to compute the smoothing parameter and the function `npudist` to derive the smoothed distribution function of the residuals.

To check the performance of the proposed bootstrap estimator (14) we compute the relative bias (RB) and the relative root mean squared error (RRMSE),

$$RB\big(mse(\hat{m}_i)\big) = H^{-1}\sum_{h=1}^{H}\left(\frac{mse(\hat{m}_{i,h}) - MSE(\hat{m}_{i,h})}{MSE(\hat{m}_{i,h})}\right), \tag{17}$$

$$RRMSE\big(mse(\hat{m}_i)\big) = \frac{\Big\{\sum_{h=1}^{H}(mse(\hat{m}_{i,h}) - MSE(\hat{m}_{i,h}))^2/H^{-1}\Big\}^{1/2}}{\Big\{MSE(\hat{m}_{i,h})\Big\}^{1/2}}, \tag{18}$$

where $mse(\hat{m}_{i,h})$ is the bootstrap MSE estimator of the MQ/RobX estimator in area i and simulation h and $MSE(\hat{m}_{i,h})$ is the empirical MSE of the

MQ/RobX estimator observed over the Monte Carlo simulations (considered as the "true" MSE of the estimator).

The results of the simulation are summarized in table 8 in terms of percentage RB and percentage RRMSE. Since we did not always observe a reduction of the RRMSE increasing the sample size from 5 to 10 (scenario C and D, $\zeta = 5\%$), we also run a simulation with sample size equal 20.

The values of the percentage RRMSE varies between 21% to 43%, with the vast majority of them above the 35%. The increase of the sample size reduces the RRMSE, particularly when $n_i = 20$. We have also tested the effect of further increments of the sample size, obtaining a further reduction of the RRMSE, as expected (results are not reported here).

The percentage RB is between -10% and 10% with the exception of scenario D for sample size equal 20 and $\zeta = 5\%$ (RB=14.7%). In general, when the sample size increases also the RB increases leading to conservative confidence intervals. So we evaluated the limit of this growth with $n_i = 30$ (results are not reported here), and we observed results similar to $n_i = 20$.

The performance of $mse(\hat{m}_i)$ is also evaluated by computing for each small area the average coverage rate of nominal 95% confidence intervals. The MSE estimation provides reasonable average coverage performance, though there is clear evidence that the intervals based on the bootstrap MSE estimator exhibit under coverage (around 90%) with small sample size, $n_i = 5, 10$. When the n_i increases the bootstrap MSE estimator improves its performance with a slight over coverage. The results (not reported here) are available from the authors upon request. However, the use of estimated MSEs to construct normal theory confidence intervals, though widespread, has been criticized by Hall and Maiti (2006). The results we obtained are consistent with their discussion.

The proposed bootstrap technique seems to work properly to evaluate the uncertainty in the MQ/RobX estimates. The bias is small and the second order properties are adequate.

6 Application: yield of olives in Tuscany

In this Section we apply the traditional and newly proposed estimators to real data, to check if the presence of outliers in the auxiliary variables leads to sub-optimal model-based small area estimators, as shown in model-based simulations. We use agricultural data to estimate the yield of olives in Tuscany agrarian regions (small areas). The Tuscan production of olive oil made by local olives jointly with the production of grape and wine is an example of the rich and excellent tradition of agricultural production in Italy and all over the world.

In the application we use the official data on agricultural production collected by the Italian National Statistical Insitute (Istat). We access the microdata of the sample of farms used by the Survey on the Structure and Production of Italian farms (Struttura e Produzione delle aziende Agricole, Istat 2003), jointly with the individual Italian Farms Census data (Censimento

Table 8 Averages over areas and simulations of the RB% and RRMSE% of the bootstrap MSE estimator of the MQ/RobX estimator

	$\zeta = 0.01$		$\zeta = 0.03$		$\zeta = 0.05$	
	RB%	RRMSE%	RB%	RRMSE%	RB%	RRMSE%
			Scenario A			
$n_i = 5$	-10.68	27.48	-10.12	28.20	-10.15	29.53
$n_i = 10$	-3.14	23.96	-0.84	26.41	0.61	29.68
$n_i = 20$	0.36	20.86	1.81	21.90	3.52	24.09
			Scenario B			
$n_i = 5$	-8.20	35.35	-8.26	35.22	-9.56	36.28
$n_i = 10$	-6.91	32.28	-5.50	34.00	-4.61	35.50
$n_i = 20$	0.52	28.95	2.43	31.24	3.93	32.97
			Scenario C			
$n_i = 5$	-5.27	27.43	-0.97	31.07	-0.28	33.89
$n_i = 10$	0.86	24.94	6.04	30.32	10.44	36.77
$n_i = 20$	5.47	21.70	9.66	25.67	14.44	31.40
			Scenario D			
$n_i = 5$	-4.22	36.59	-4.10	37.98	-4.39	39.08
$n_i = 10$	-0.63	34.36	2.93	36.34	10.27	43.26
$n_i = 20$	5.95	29.61	11.06	33.51	14.68	38.66

dell'Agricoltura, Istat 2000). The data considered here are those of the Tuscany region, which is divided into 53 agrarian regions. The sample size within the 53 agrarian regions varies between 2 and 185 farms (mean 45.34, median 34), while the population size varies between 475 and 8661 farms (mean 2592, median 2064).

We analyse the available data to identify the working model to carry out the model-based small area estimates. Among the set of variables in common between survey and census data, only the extension of the agricultural surface area devoted to the production of olives (OAS for short) is chosen as predictive to model the yield of olives (OY for short).

To detect the presence of outliers we use the multivariate outliers detection approach based on Projection Pursuit (PP). The PP approach consists in looking for low dimensional linear projections that are susceptible to reveal outlying observations. More details are available in Ruiz-Gazen et al (2010). The results of the PP outliers detection are shown in figure 1 together with boxplot of OAS and OY to graphically detect outliers in a univariate perspective.

Looking at figure 1 the presence of outliers is evident in both the variables. Also using a classical multivariate Mahalanobis distance detection approach (results not shown here) there is evidence of outliers in target and auxiliary variables.

The preliminary analysis suggests that the REBLUP/RobX and the MQ/RobX should be more efficient (in terms of MSE) than the REBLUP, the MQ and the EBLUP. We applied all the five estimators to our agricultural data. The results are in tables 9 and 10.

The MSEs of all the compared estimators are estimated using the bootstrap, accordingly to the techniques proposed in Section 4.1, and adapted where necessary. Substantially, we estimate the model and generate a boot-

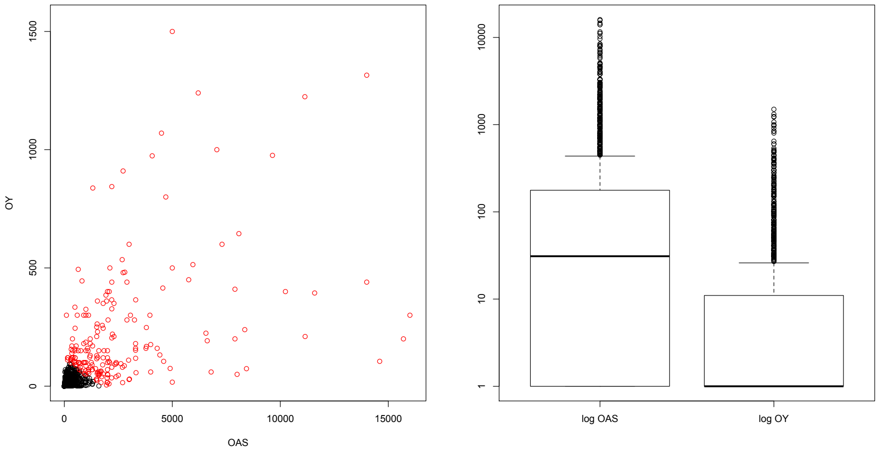

Fig. 1 Left: Multivariate outliers detection using PP approach, outliers in red. Right: log scale box-plot of OAS and OY

Table 9 Distribution over 53 agrarian regions of the estimated average yield of olives and its *mse* and correlation between small area and direct estimates (ρ)

	Point estimates			Root *mse*			ρ
	Q. 0.25	Median	Q. 0.75	Q. 0.25	Median	Q. 0.75	
EBLUP	2.77	4.91	9.80	11.46	14.62	17.46	0.77
EBLUPlog	2.47	5.80	10.82	2.66	5.65	9.11	0.86
REBLUP	2.76	5.81	11.84	2.33	2.84	3.67	0.84
MQ	1.38	5.44	10.35	0.57	1.04	1.63	0.87
REBLUP/RobX	2.12	4.93	8.37	1.43	1.64	2.06	0.77
MQ/RobX	1.43	4.84	10.00	0.50	0.84	1.22	0.86

strap population for each estimator. Then we generate 100 samples on which we estimate the target, obtaining the MSE estimates.

Table 9 shows the distribution of the average yield of oil and its *mse* in the 53 agrarian regions of Tuscany. Point estimates have a similar distribution, similar also to that of the direct estimates, as we can see from the correlation coefficients ρ between direct and small area estimates. The exceptions are the EBLUP and the REBLUP/RobX, which have a correlation less than 0.8. Point estimates have been benchmarked using the ratio adjustment (Rao and Molina, 2015). It consists of multiplying each point estimate by a common adjustment factor given by $\sum_{i=1}^{D} w_i \hat{m}_i^{DIR} / \sum_{i=1}^{D} w_i \hat{m}_i^{SAE}$, where $\hat{m}_i^{DIR}$ is the direct estimate for area i, $\hat{m}_i^{SAE}$ is a small area estimate for area i and $w_i = N_i/N$.

For what concerns the uncertainty, we can see that the EBLUP estimator shows high root *mse*s (*rmse*) with respect to the others estimators. This result was expected, given that there is presence of very big outliers, both in target and auxiliary variables. The MQ and REBLUP estimators account for the outliers in the target variable, so they show *rmse*s remarkably lower than the EBLUP. Among them the MQ shows the best performance, showing *rmse*s that halve the *rmse*s of REBLUP. An alternative to avoid the effect of the outliers on the target is to use a log transformation. We have tried such transformation on the EBLUP (EBLUPlog in table 9), and we have obtained better results than the EBLUP on the row scale. However, the EBLUPlog is still not comparable in term of *rmse* with the REBLUP. On the contrary, by using the log transformation on the REBLUP (results are not reported here) the estimates have a very high variability.

The MQ/RobX and the REBLUP/RobX estimators account for outliers both in the target and in the auxiliary variable. As expected, the REBLUP/RobX shows *rmse*s lower than the REBLUP and the same happens for the MQ/RobX with respect to the MQ. Furthermore, the MQ/RobX shows the best performance among all the estimators. Point estimates together with their *rmse* for all the agrarian regions are reported in the appendix.

The model parameters used to predict the small area means are reported in table 10. The OAS coefficient increases for REBLUP/RobX and MQ/RobX with respect to their non-robust-on-auxiliary-variable versions. The group of observations outlying in the OAS variable "pull down" the regression line, while this effect is mitigated estimating the coefficient with the proposed method. Consequently, the intercept decreases for REBLUP/RobX and MQ/RobX with respect to REBLUP and MQ respectively.

Table 10 Application: model parameters. The regression coefficients for MQ and MQ/RobX are estimated at the median.

	Intercept	OAS	σ_u	σ_e
EBLUP	7.448	0.051	18.927	73.449
EBLUPlog	1.024	0.001	0.635	1.335
REBLUP	1.242	0.037	2.247	6.198
MQ	0.423	0.037	-	-
REBLUP/RobX	0.394	0.056	0.721	1.813
MQ/RobX	0.181	0.054	-	-

In conclusion, we use survey and census agricultural data to estimate the average production of olives in 53 agrarian regions (small areas) in Tuscany. Exploratory analyses show the presence of outliers in both the target and the auxiliary variables, giving us the opportunity to test different small area estimators on a real data application. The results show that the EBLUP, which doesn't take into account the outliers, show the worst performance in terms of MSE. The MQ and the REBLUP, which take into account the outliers in the target variable, outperform the EBLUP, but are worse than the MQ/RobX and the REBLUP/RobX, which take into account the outliers both in target and auxiliary variables. As a note, by using the log transformation the EBLUP improves in precision; however, it still does not reach the efficiency of the other estimators.

7 Conclusions

The presence of outliers in the auxiliary variables may lead to biased small area estimates. In this paper we propose an M-quantile estimator of the small area mean robust to the presence of outliers both in target and in continuous auxiliary variables. We also define a similar estimator under the linear mixed models approach to SAE, following the suggestion in Sinha and Rao (2009).

The results of the model- and design-based simulation studies show that in the presence of outliers in the auxiliary variable the new estimators (MQ/RobX and REBLUP/RobX) outperform their traditional version in terms of relative bias and efficiency. Particularly, they both outperform the EBLUP. The REBLUP/RobX shows the best performance in model-based simulations, where unit and area level errors are generated symmetrically. Notwithstanding this, the performance of the MQ/RobX and the REBLUP/RobX is similar in terms of bias and MSE. On the contrary, the MQ/RobX shows the best performance in the design-based simulation. These results suggest that the presence of outliers in the auxiliary variables can significantly impact upon the small area estimates, recommending the use of appropriate small area methodologies. These conclusions are also supported by the results of the application, based on agricultural survey and census data collected in Tuscany (Italy).

In the model-based simulation studies we also evaluate the performance of the bootstrap estimator proposed to estimate the variability of the MQ/RobX estimator. The results obtained suggest that the bootstrap estimator has adequate second order properties and it is able to correctly reproduce the variability of the small area mean MQ/RobX estimator.

Future developments will deal with the choice of the tuning constant for the ψ influence function. An 'optimal' c-value could be potentially achieved by using a cross-validation criterion (Bianchi et al, 2015).

As in many cases official survey data are discrete, a further topic of interest will be the extension of the proposed methodology to discrete auxiliary variables.

Acknowledgements The authors are grateful to the Associate Editor, the referee and the Editor for their helpful comments which have led to substantial improvements in the paper. The work of Salvati has been developed under the support of the project PRIN-SURWEY http://www.sp.unipg.it/surwey/ (Grant 2012F42NS8, Italy).

Appendix

Table 11 Estimates (and *rmse*) of the average yield of olive oil in the agrarian regions of Tuscany presented in Section 6. DIR are direct estimates.

DIR.	EBLUP	EBLUP log	REBLUP	REBLUP RobX	MQ	MQ RobX
0.65	2.37 (12.08)	1.53 (17.12)	1.58 (2.67)	1.22 (2.02)	0.48 (0.65)	0.7 (0.5)
1.67	1.86 (9.75)	2.01 (5.17)	2.06 (2.18)	2.52 (1.35)	2.56 (0.15)	2.43 (0.8)
0.43	2.74 (11.93)	1.5 (2.09)	1.34 (1.83)	1.51 (1.39)	0.6 (0.31)	0.75 (0.29)
0.81	3.52 (8.3)	2.62 (16.19)	2.76 (2.89)	2.34 (1.04)	1.79 (0.58)	2.44 (0.23)
0	4.92 (7.22)	1.93 (7.99)	2.83 (3.21)	1.31 (1.05)	0.08 (0.59)	0 (1.22)
6.43	4.91 (9.46)	5.53 (6.87)	6.26 (5.59)	5.53 (1.34)	1.28 (1.93)	1.69 (1.02)
0.89	4.17 (7.31)	2.09 (4.35)	2.49 (2.27)	1.94 (0.85)	0.17 (0.78)	0.49 (0.51)
4.91	5.03 (34.34)	8.02 (1.64)	6.06 (9.24)	7.24 (6.37)	6.16 (2.31)	6.28 (1.17)
0	4.93 (4.59)	2.56 (3.34)	2.22 (1.31)	1.86 (0.98)	0 (0.4)	0 (0.36)
1.12	0.96 (16.33)	0.77 (2.45)	1.55 (2.29)	2.12 (1.85)	0.55 (0.17)	0.72 (0.77)
3.01	2.61 (9.64)	3.8 (2.8)	4.35 (2.84)	3.53 (1.64)	2.62 (0.22)	2.64 (0.19)
3.8	0 (17.07)	0.86 (8.93)	1.6 (4.63)	1.87 (1.48)	0.83 (0.38)	0.85 (1.14)
3.5	3.06 (17.6)	1.09 (5.65)	3.42 (2.54)	3.22 (2.26)	1.38 (0.68)	1.32 (1.72)
6.95	3.6 (10.01)	3.53 (4.39)	5.68 (2.5)	3.71 (1.62)	3.54 (0.27)	2.59 (0.98)
0	3.04 (12.28)	0.66 (3.98)	1.89 (2.15)	0.96 (1.18)	0.05 (0.71)	0 (0.84)
41.74	71.79 (24.13)	29.2 (2.39)	18.91 (6.77)	27.51 (2.44)	38.18 (4.5)	47.51 (3.73)
8.6	0 (19.13)	10.6 (10.79)	10.26 (2.05)	7.09 (2.23)	12.81 (0.53)	8.37 (1.73)
5.14	9.8 (12.38)	8.32 (2.7)	11.84 (3.93)	6.67 (1.22)	13.69 (0.91)	7.63 (2.22)
48.11	9.91 (12.27)	17.87 (21.74)	14.75 (2.1)	6.6 (2.06)	21.28 (1.27)	14.46 (3.37)
21.62	13.77 (24.12)	23.53 (1.82)	18 (2.78)	11.4 (2.02)	18.73 (1.88)	16.98 (0.98)
27.95	18.56 (17.46)	22.89 (2.3)	18.59 (2.98)	19.68 (2.96)	38.85 (0.97)	41 (3.24)
19.44	5.17 (15.05)	13 (8.43)	13.02 (3.4)	6.58 (1.47)	18.92 (1.87)	10.91 (0.75)
4.25	2.77 (17.75)	7.81 (10.73)	4.37 (1.97)	2.64 (4.29)	4.34 (1.38)	4.36 (0.33)
20.85	9.82 (84.52)	10.82 (2.21)	11.85 (3.67)	12.14 (1.77)	8.53 (1.53)	9.65 (1.17)
28.36	30.49 (16.17)	25.5 (8.75)	25.35 (3.51)	35.71 (2)	27.47 (5.29)	22.74 (3.28)
4	4.04 (6.42)	4.27 (13.9)	2.38 (4.11)	1.29 (1.51)	1.32 (1.45)	1.72 (0.82)
4.7	3.06 (14.62)	4.19 (2.66)	6.41 (4.26)	5.44 (1.45)	4.95 (0.45)	4.42 (0.5)
4.94	8.62 (12.24)	5.93 (2.03)	9.02 (2.65)	7.01 (1.8)	5.97 (0.62)	6.11 (0.87)
0	2.05 (11.46)	0.59 (20.04)	1.38 (3.26)	1.24 (1.54)	0.01 (0.81)	0.01 (1.64)
13.2	9.05 (9.73)	8.19 (4.34)	9.07 (4.73)	7.57 (1.56)	10.35 (8.57)	9.71 (3.6)
16.5	15.11 (13.89)	22.85 (8.45)	21.92 (3.72)	28.09 (1.56)	30.35 (4.78)	21.92 (1.28)
2.91	1.15 (12.59)	0.96 (2.55)	1.4 (2.33)	1.35 (1.76)	0.48 (0.49)	0.89 (0.2)
6.75	2.45 (12.97)	2.47 (1.94)	4.72 (4.38)	2.92 (2.2)	2.23 (0.32)	1.43 (1.3)
0	3.32 (14.79)	0.97 (4.32)	1.3 (2.47)	1.02 (2.1)	0 (0.47)	0 (0.23)
4.52	9.71 (16.17)	6.99 (17.26)	4.85 (2.62)	3.81 (1.7)	6.06 (0.65)	5.8 (0.53)
0.92	2.84 (6.59)	2.22 (4.4)	1.76 (1.79)	1.59 (0.91)	1.45 (0.54)	1.39 (0.33)
3.64	6.61 (17.02)	6.08 (9.27)	7.25 (2.45)	7.22 (1.47)	5.44 (0.99)	4.84 (0.41)
5.99	8.09 (14.64)	9.48 (18.59)	10.95 (3.1)	11.24 (1.81)	7.41 (1.35)	7.01 (0.68)
2.5	4.18 (8.48)	6.53 (1.81)	4.59 (2.9)	1.96 (1.29)	3.12 (1.78)	2.62 (0.5)
4.61	3.72 (15.35)	5.8 (2.38)	5.81 (5.27)	4.57 (1.45)	6.64 (0.85)	5.98 (1.18)
46.97	48.17 (18.18)	28.18 (2.79)	31.8 (4.32)	45.41 (4.22)	36.62 (3.12)	57.42 (4.77)
4.93	1.9 (17.89)	3.6 (9.11)	5.37 (3.55)	3.83 (1.73)	4.21 (0.48)	3.8 (1.25)
7.77	5.13 (15.71)	6.6 (2.58)	5.75 (3.2)	6.11 (1.84)	7.48 (0.5)	10 (1.11)
3.07	5.32 (12.44)	4.63 (19.75)	6.08 (2.28)	5.01 (1.63)	4.8 (0.99)	4.37 (0.46)
5.29	0.74 (15.1)	5.16 (18.01)	4.01 (2.48)	3.16 (1.83)	5.74 (0.61)	5.47 (0.84)
11.97	8.71 (14.23)	9.89 (8.44)	9.67 (2.48)	8.37 (1.43)	9.6 (0.84)	8.85 (0.59)
14.05	18.61 (15.35)	22.71 (7.95)	17.91 (2.95)	20.51 (2.09)	20.69 (2.84)	22.12 (0.8)
10.31	11.15 (20.12)	13.47 (8.41)	11.42 (2.29)	12.8 (4.57)	9.47 (1.88)	13.24 (0.91)
8.46	14.93 (19.34)	12.5 (17.66)	12.88 (2.74)	10 (1.27)	13.47 (1.5)	12.24 (0.43)
10.78	19.28 (18.34)	14.59 (6.64)	14.5 (3.52)	17.36 (1.55)	14.32 (1.41)	18.15 (0.75)
15.72	18.39 (12.76)	10.84 (6.3)	13.35 (1.87)	16.96 (2.32)	10.32 (0.88)	15.82 (0.38)
3.36	4.67 (6.79)	4.78 (6.45)	4.44 (2.48)	2.45 (1.18)	2.39 (1.61)	2.39 (0.52)
5.49	0 (20.38)	3.69 (3.92)	7.46 (7.55)	4.93 (3.57)	6.75 (1.92)	6.47 (1.02)

References

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

Bianchi A, Fabrizi E, Salvati N, Tzavidis N (2015) M-quantile regression: diagnostics and parametric representation of the model. Working paper pp http://www.sp.unipg.it/surwey/dowload/publications/24–mq–diagn.html

Bowman A (1984) An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71:353–360

Bowman A, Hall P, Prvan T (1998) Bandwidth selection for the smoothing of distribution functions. Biometrika 85:799–808

Breckling J, Chambers R (1988) M-quantiles. Biometrika 75(4):761–771

Carroll R, Pederson S (1993) On robustness in the logistic regression model. Journal of the Royal Statistical Society Series B (55):693–706

Chambers R (1986) Outlier robust finite population estimation. J Am Statist Ass 81:1063–1069

Chambers R, Tzavidis N (2006) M-quantile models for small area estimation. Biometrika 93(2):255–68

Chambers R, Chandra H, Tzavidis N (2011) On bias-robust mean squared error estimation for pseudo-linear small area estimators. Survey Methodology 37(2):153–170

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, Salvati N, Tzavidis N (2016) Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the uk. Journal of the Royal Statistical Society - Series A (179):453–479

Cook R, Weisberg S (1980) Characterization of an empirical influence function for detecting influential cases in regression. Technometrics 22:495–508

Dongmo Jiongo V, Haziza D, Duchesne P (2013) Controlling the bias of robust small area estimators. Biometrika (100):843–858

ESS (2014) The european statistical system vision 2020. Tech. rep., Eurostat, URL `http://ec.europa.eu/eurostat/documents/10186/756730/ESS-Vision-2020.pdf/8d97506b-b802-439e-9ea4-303e905f4255`

Fellner W (1986) Robust estimation of variance components. Technometrics 28:51–60

Filzmosera P, Maronna R, Werner M (2008) Outlier identification in high dimensions. Computational Statistics and Data Analysis 52:1694–1711

Gershunskaya J (2010) Robust small area estimation using a mixture model. In: Proceedings of the Joint Statistical Meeting 2010, American Statistical Association

Ghosh M, Sinha K, Kim D (2006) Empirical and hierarchical bayesian estimation in finite population sampling under structural measurement error model. Scandinavian Journal of Statistics 33(3):560–568

Giusti C, Tzavidis N, Pratesi M, Salvati N (2014) Resistance to outliers of m-quantile and robust random effects small area models. Communication in Statistics – Simulation and Computation 43(3):549–568

Hall P, Maiti T (2006) On parametric bootstrap methods for small area prediction. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(2):221–238, DOI 10.1111/j.1467-9868.2006.00541.x, URL http://dx.doi.org/10.1111/j.1467-9868.2006.00541.x

Hampel F, Ronchetti E, Rousseuw P, Stahel W (1986) Robust Statistics: The Approach Based on Influence Functions. Wiley, New York

Hayfield T, Racine JS (2008) Nonparametric econometrics: The np package. Journal of Statistical Software 27(5), URL http://www.jstatsoft.org/v27/i05/

Hubert M, Rousseeuw P, Van Aelst S (2008) High-breakdown robust multivariate methods. Statistical Science 23(1):92–119

Huggins R (1993) A robust approach to the analysis of repeated measures. Biometrics 49:255–268

Jiang J, Lahiri P (2006) Mixed model prediction and small area estimation. Test 15:1–96

Kokic P, Chambers R, Breckling J, Beare S (1997) A measure of production performance. Journal of Business and Economic Statistics 15(4):445–51

Koller PJ, Stahel WA (2011) Sharpening wald-type inference in robust regression for small samples. Computational Statistics and Data Analysis 55(8):2504–2515

Lombardía M, González-Manteiga W, Prada-Sánchez J (2003) Bootstrapping the chambers-dunstan estimate of finite population distribution function. Journal of Statistical Planning and Inference 116:367–388

Marchetti S, Tzavidis N, Pratesi M (2012) Non-parametric bootstrap mean squared error estimation for m-quantile estimators of small area averages, quantiles and poverty indicators. Computational Statistics and Data Analysis (56):2889–2902

Maronna AR, Martin R, Yohai V (2006) Robust Statistics Theory and Methods. Wiley and Sons

Prasad N, Rao J (1990) The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association 85:163–171

R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Rao J, Molina I (2015) Small Area Estimation. Wiley Series in Survey Methodology, Wiley, URL https://books.google.it/books?id=i1B_BwAAQBAJ

Richardson A, Welsh A (1995) Robust estimation in the mixed linear model. Biometrics 51:1429–1439

Rudemo M (1982) Empirical choice of istograms and kernel density estimators. Scandinavian Journal of Statistics 9:66–78

Ruiz-Gazen A, Marie-Sainte S, Berro A (2010) Detecting multivariate outliers using projection pursuit with particle swarm optimization. In: Lechevallier Y, Saporta G (eds) Proceedings of COMPSTAT'2010, Physica-Verlag HD, pp 89–98, DOI 10.1007/978-3-7908-2604-3_8, URL http://dx.doi.org/

10.1007/978-3-7908-2604-3_8
Sinha S, Rao J (2009) Robust small area estimation. The Canadian Journal of Statistics 37(3):381–399
Stone C (1974) Cross-validatory choice and assessment of statistical predictors (with discussion). Journal of the Royal Statistical Society 36:11–147
Stone C (1984) An asymptotically optimal window selection rule for kernel density estimates. Annals of Statistics 12:1285–1297
Torabi M, Datta G, Rao J (2009) Empirical bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates. Scandinavian Journal of Statistics (36):355–368
Tzavidis N, Marchetti S, Chambers R (2010) Robust estimation of small area means and quantiles. Australian and New Zeland Journal of Statistics 52(2):167–186
Ybarra L, Lohr S (2008) Small area estimation when auxiliary information is measured with error. Biometrika (95):919–931