Design consistent small area estimators based on M-quantile regression

Fabrizi Enrico
DISES, Università Cattolica
Via Emilia Parmense 84
29122 Piacenza,Italy
E-mail: enrico.fabrizi@unicatt.it

Pratesi Monica
DSMAE, Università di Pisa
Via Ricolfi 10
56124 Pisa, Italy
E-mail: m.pratesi@ec.unipi.it

Salvati Nicola
DSMAE, Università di Pisa
Via Ricolfi 10
56124 Pisa, Italy
E-mail: salvati@ec.unipi.it

Tzavidis Nikos
School of Social Science, University of Southampton
SO17 1BJ Southampton, United Kingdom
E-mail: n.tzavidis@soton.ac.uk

1 Introduction

In large statistical surveys, estimates of population descriptive quantities for a target variable Y are usually needed for the population as a whole and for different collections of sub-populations (domains or areas). Provided that an adequate domain-specific sample size is available, statistical agencies apply the same design-based methods used for the estimation of population level quantities to domain estimation. When available samples are not large enough to allow for reliable estimation in all or most of the domains, we have a small area estimation problem.

The application of design-based estimators, namely the Generalized Regression (GREG) estimator, to the small area setting is introduced by Särndal (1984) and reviewed in Rao (2003, Sections 2.4 and 2.5). These estimators are design consistent, approximately design unbiased for moderate area specific sample sizes, their randomization based variances may be easily estimated. Unfortunately, they typically overlook 'local effects', that is the between area variation not accounted for by the regressors. For this reason, model dependent estimators relying on mixed models became very popular (see Rao, 2003; Jiang and Lahiri, 2006a). Informally, we may classify the mixed models applied to small area estimation in two classes. In the first class, often labeled as area level models, the data provide information only at the domain level and a model is assumed on the direct survey estimates. In the second class a model is specified at the unit level. The reliability of these methods hinges on the validity of model assumptions, a criticism often raised within the design-based research tradition (Estevao and Särndal, 2004). Moreover, model-based estimators based on unit-level models such as the popular nested error regression (Battese et al., 1988) or multi-level (You and Rao, 2000) typically do not make use of the unit level survey weights; as a result the estimators are not design consistent

as the area sample sizes become large, unless the sampling design is self-weighting within areas.

Design consistency is a form of protection against model failures, at least in large domains. Kott (1989) was the first to propose a design consistent estimator based on a linear mixed model. Prasad and Rao (1999), assuming a nested error regression model, introduced a pseudo-EBLUP, demonstrated its superiority to the estimator proposed in Kott (1989) and obtained a more stable Mean Square Error (MSE) estimator. You and Rao (2002) extended Prasad and Rao (1999) and obtained a pseudo-EBLUP achieving the nice property of benchmarking. Note that, although design consistent, these predictors are 'model based' and their statistical properties such as bias and MSE are evaluated with respect to the distribution induced by the data generating process and not randomization. Jiang and Lahiri (2006b) obtained design consistent predictors also for generalized linear models and evaluated MSE with respect to the joint randomization-model distribution.

In this paper we consider small area estimators based on M-quantile regression models (Chambers and Tzavidis, 2006), a robust alternative to estimators based on linear mixed models. M-quantile regression relies on the assumption of a quantile-specific linear relationship between the target and the auxiliary variables. They are free of distributional assumptions and do not require explicit specification of the random part of the model. Moreover the recourse to M-estimation protects from presence of outliers and influential observations. However, the small area estimators based on M-quantile regression discussed by Chambers and Tzavidis (2006) and by Tzavidis et al. (2010) do not make use of sampling weights and are, in general, not design consistent.

The main goal of this paper is to obtain design consistent small area estimators based on M-quantile regression models, thereby generalizing Chambers and Tzavidis (2006) and Tzavidis et al. (2010). The proposed estimators are obtained by adopting a model-assisted approach: a working linear M-quantile regression model is assumed only to motivate the estimators but only properties with respect to the randomization distribution induced by the sample design will be considered. Along with estimators of area means and totals we consider also estimators of alternative functionals of the area-specific distribution functions, more specifically of quantiles.

We consider a general probability sampling design and and consider two different situations with respect to its ignorability: in the first, sampling design is ignorable given the auxiliary variables included in the small area model, so the sample and the population obey the same model and estimators of model parameters are model consistent regardless of the use of sampling weights. In the second situation, we assume that the design is ignorable only conditionally on the variables included in the small area model and the sampling weights; this means that some of the variables contributing to weights are not included into the small area model. This may the case in practice, where weights reflect the design and non-response corrections but variables used in this process are not available to the researcher. We show that our estimators preserve their nice design-based properties also in this second situation, while pseudo-EBLUPs, because of the unweighted estimation of variance components, are less efficient.

The paper is organized as follows. In Section 2 we propose a design consistent estimation of the M-quantile regression coefficients. Our new design consistent small area predictors are introduced in Section 3, along with estimators for their design-based MSE. In Section 4 we introduce a simulation exercise in order to comparing weighted and non-weighted M-quantile based small area estimators. The simulation is also aimed at comparing the properties of the proposed estimator of the mean to the popular psuedo-EBLUP by You and Rao (2002) and to the generalized regression (GREG) estimator for small areas. Moreover, in the simulation study the estimators of MSE of the proposed predictors are tested.

2 Design consistent estimation of the M-quantile regression coefficients

Let's suppose that a population U of size N is divided into m non overlapping subsets U_i (domains of study or areas) of size N_i, $i = 1, ..., m$. We index the population units by j and the small areas by i. The population data consist of values y_{ij} of the variable of interest, values $\mathbf{x}_{ij}$ of a vector of p auxiliary variables. We assume that $\mathbf{x}_{ij}$ contains 1 as its first component. Suppose that a sample s is drawn according to some, possibly complex, sampling design such that the inclusion probability of unit j within area i is given by π_{ij}, so that area-specific samples $s_i \subset U_i$ of size $n_i \geq 0$ are available for each area. Note that non-sample areas have $n_i = 0$, in which case s_i is the empty set. The set $r_i \subset U_i$ contains the $N_i - n_i$ indices of the non-sampled units in small area i. Values of y_{ij} $(j = 1 \dots n)$ are known only for sampled values while for the $p-$vector of auxiliary variables the area mean is known.

Since much of the development in this paper is based on the application of linear M-quantile regression, we now give a brief definition of related concepts. The q-th M-quantile for the random variable Z with distribution function $F(Z)$, Q_q, is defined as the solution of the equation

$$(1) \quad \int \psi_q\left(\frac{Z - Q_q}{\sigma_q}\right) F(dz) = 0,$$

where $\psi_q(u) = 2\psi(u)\big\{qI(u > 0)+(1-q)I(u \leq 0)\big\}$ and ψ is an influence function, that we assume to be a bounded and monotone non decreasing function over the real line with $\psi(0) = 0$. The parameter σ_q is a suitable measure of the scale of the random variable $Z-Q_q$. Note that if we relax boundedness, i.e. assuming $\psi(u) = u$ we obtain the expectile of order q, which represents a quantile-like generalization of the mean, while for $\psi(u) = sgn(u)$ we obtain the ordinary population quantiles.

Since this presentation of M-quantile models is not directly related to small area estimation, let's drop the subscript i from the notation of this Section. Ordinary linear regression is based on the idea of modelling the expected value of the dependent variable as a function of the regressors; that is, on the assumption that $E(y_j|\mathbf{x}_j) = \mathbf{x}_j^T\boldsymbol{\beta}$, $j = 1, \dots, N$. In M-quantile regression (Breckling and Chambers, 1988) it is the conditional M-quantile $Q_q(y_j|\mathbf{x}_j)$ that is assumed to be a linear function of the auxiliary information, that is a distinct (hyper-)plane, characterized by quantile-specific regression coefficients $\boldsymbol{\beta}_\psi(q)$ is assumed to have generated the data at each $q \in (0, 1)$. More specifically we may write the basic model assumption as:

$$(2) \quad Q_q(y_j|\mathbf{x}_j) = \mathbf{x}_j^T\boldsymbol{\beta}_\psi(q),$$

with $q \in (0, 1)$. Under ignorable sampling, for given q and the influence function ψ, a consistent estimate of the vector of the regression parameters $\boldsymbol{\beta}_\psi(q)$ may then be obtained by solving the following normal equations respect to $\boldsymbol{\beta}_\psi(q)$:

$$n^{-1}\sum_{j\in s} \psi_q\left(\frac{y_j - \mathbf{x}_j^T\boldsymbol{\beta}_\psi(q)}{\hat{\sigma}_q}\right)\mathbf{x}_j = \mathbf{0},$$

where $\hat{\sigma}_q$ is some robust estimate of the scale of the residuals $y_j - \mathbf{x}_j^T\boldsymbol{\beta}_\psi(q)$, e.g. $\hat{\sigma}_q = median|y_j - \mathbf{x}_j^T\boldsymbol{\beta}_\psi(q)|/0.6745$. The solution may then be obtained via iterative re-weighted least squares. M-quantile regression provides a 'quantile-like' generalization of regression based on influence functions. As such, both the quantile regression introduced by Koenker and Bassett (1978) and the expectile regression by Newey and Powell (1987) may be obtained as special cases. See Breckling and Chambers (1988) for a more detailed introduction to M-quantile regression.

When data are obtained from complex surveys, weights may be included into the estimation process to obtain design consistent estimators. Design consistency provides a protection against the

failure of (2). We now introduce a design consisten estimator $\hat{\boldsymbol{\beta}}_{w\psi}(q)$ for $\boldsymbol{\beta}_{\psi}(q)$ assuming a general sampling design characterized by inclusion probabilities π_j, allowing that it may be non-ignorable given the regressors included into model (2), as it may be the case when some of the design variables are not available or not used in the modelling stage.

For any $q \in (0,1)$ let $\mathbf{B}_{\psi}(q)$ be the solution of the Census system of equations $N^{-1}\sum_{j\in U}\psi_q(y_j - \mathbf{x}_j^T\boldsymbol{\beta}_{\psi}(q))\mathbf{x}_j = \mathbf{0}$. Kocic et al. (1997) proved that this solution is unique when $\psi(u)$ is a continuous monotone function in u.

We now introduce the following set of assumptions: *i)* $\psi(u)$ is a monotone continuous function in u and ψ is differentiable in $\boldsymbol{\beta}_{\psi}$; *ii)* The regularity conditions on the sampling design that guarantee the consistency of the Horwitz-Thompson estimator hold. They are needed to prove the following result:

Theorem 2.1 *Under assumptions i) and ii), for each* $q \in (0,1)$, $\hat{\boldsymbol{\beta}}_{w\psi}(q)$ *defined as the solution of the weighted normal equations*

$$n^{-1}\sum_{j\in s} w_j\psi_q(y_j - \mathbf{x}_j^T\boldsymbol{\beta}_{\psi}(q))\mathbf{x}_j = \mathbf{0}$$

is design consistent for $\mathbf{B}_{\psi}(q)$.

An iteratively re-weighted least squares algorithm is used to calculate the design-weighted M-quantile regression coefficients at q. The weighted least squares estimates of $\boldsymbol{\beta}_{\psi}(q)$ can be written as

$$\hat{\boldsymbol{\beta}}_{w\psi}(q) = \left(\mathbf{X}^T\mathbf{W}\mathbf{C}(q)\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{C}(q)\mathbf{y}, \tag{3}$$

where $\mathbf{X}$ and $\mathbf{y}$ are the $n \times p$ matrix of sample x values and the vector of sample y values, respectively; $\mathbf{W}$ is the diagonal sampling weight matrix of order n and $\mathbf{C}(q)$ is the diagonal weight matrix of order n that defines the estimator of the design-weighted M-quantile regression coefficient at q.

3 M-quantile regression methods applied to small area estimation

3.1 Estimation of small area characteristics based M-quantile regression

In the application of M-quantile regression to small area estimation, Chambers and Tzavidis (2006) characterize the variability across the population not accounted for by the regressors using the M-quantile coefficients of the population units. For unit j in area i, this coefficient is the value θ_{ij} such that $Q_{\theta_{ij}}(y_{ij}|\mathbf{x}_{ij}) = y_{ij}$. The authors observe that if a hierarchical structure does explain part of the variability in the population data, units within areas defined by this hierarchy are expected to have similar M-quantile coefficients. This represents an alternative to more popular recourse to area-specific random effects and has the advantage of avoiding distributional assumptions.

The small area descriptive quantities can be represented as functionals of the target variable distribution function within the area in question:

$$F_i(t) = N_i^{-1}\sum_{j\in U_i} I(y_{ij} \le t) = \Big[\sum_{j\in s_i} I(y_{ij} \le t) + \sum_{j\in r_i} I(y_{ij} \le t)\Big]$$

with $U_i = r_i \cup s_i$. Estimators of F_i under M-quantile linear models are the used to obtain M-quantile (MQ) predictors of area descriptive quantities. For instance the two authors propose

$$\hat{F}_i^{MQ} = N_i^{-1}\Big[\sum_{j\in s_i} I(y_{ij} \le t) + \sum_{j\in r_i} I(\hat{y}_{ij} \le t)\Big], \tag{4}$$

where $\hat{y}_{ij} = \mathbf{x}_{ij}^T \hat{\beta}_\psi(\bar{\theta}_i)$ and $\bar{\theta}_i = \sum_{j=1}^{n_i} \theta_{ij}$ when $n_i > 0$, while if $n_i = 0$ we set $\bar{\theta}_i = 0.5$ and estimators of F_i reduce to synthetic estimators based on M-median regression. Tzavidis et al. (2010) note that this 'naive' estimator of the distribution function implies a potentially severely biased estimator of the area means under the linear M-quantile regression model; they propose to use an alternative estimator of F_i based on a smearing argument and discussed in Chambers and Dunstan (1986):

$$(5) \quad \hat{F}_i^{MQ/CD} = N_i^{-1}\Big[\sum_{j\in s_i} I(y_{ij} \le t) + \sum_{j\in r_i}\sum_{h\in r_i} I[\hat{y}_{ij} + (y_{ih} - \hat{y}_{ih}) \le t]\Big].$$

Both (4) and (5) are unweighted model-based estimators that neglect sampling inclusion probabilities or weights. For this reason, the associated estimators of area descriptive quantities will not be, in general, consistent. For instance, if we adopt (5), the associated estimator of the small area mean is

$$(6) \quad \hat{\bar{Y}}_i^{MQ/CD} = \int t d\hat{F}_i^{MQ/CD}(t) = N_i^{-1}\Big[\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_\psi(\bar{\theta}_i) + \frac{N_i - n_i}{n_i}\sum_{j\in s_i}\{y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_\psi(\bar{\theta}_i)\}\Big],$$

that may be written in the form:

$$(7) \quad \hat{\bar{Y}}_i^{MQ/CD} = n_i^{-1}\sum_{j\in s_i} y_{ij} + \Big\{N_i^{-1}\sum_{j\in U_i}\mathbf{x}_{ij}^T - n_i^{-1}\sum_{j\in s_i}\mathbf{x}_{ij}^T\Big\}\hat{\boldsymbol{\beta}}_\psi(\bar{\theta}_i).$$

It resembles a GREG estimator of the small-area mean under the assumption of simple random sampling or some other self-weighting design, but under more general sampling designs it will not be design consistent.

3.2 Design consistent small area predictors

We may obtain design consistent estimators of area descriptive quantities using a modified Rao-Kovar-Mantel estimator of F_i (Rao, Kovar and Mantel, 1990) defined as

$$(8) \quad \hat{F}_i^{WMQ/RKM} = N_i^{-1}\Big[\sum_{j\in s_i} w_{ij} I(y_{ij} \le t) + \sum_{j\in U_i} I(\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_{w\psi}(\bar{\theta}_i) \le t) - \sum_{j\in s_i} w_{ij} I(\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_{w\psi}(\bar{\theta}_i) \le t)\Big].$$

This estimator is nearly unbiased and consistent with respect to the randomization distribution. In fact, under the mild assumption that the limit in probability $\theta_i = p\lim \bar{\theta}_i$ exists, $\hat{\boldsymbol{\beta}}_{w\psi}(\bar{\theta}_i)$ is design consistent for $\boldsymbol{\beta}_\psi(\theta_i)$. Nice design-based properties of the point estimators obtained as functionals of $\hat{F}_i^{WMQ/RKM}$ and namely, the design consistency follows. More specifically, the M-quantile regression based estimator of the area mean associated to (8) is given by:

$$(9) \quad \hat{\bar{Y}}_i^{WMQ} = \int t d\hat{F}_i^{WMQ/RKM}(t) = \frac{1}{N_i}\sum_{j\in s_i} w_{ij} y_{ij} + \Big(\frac{1}{N_i}\sum_{j\in U_i}\mathbf{x}_{ij}^T - \frac{1}{N_i}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}^T\Big)\hat{\boldsymbol{\beta}}_{w\psi}(\bar{\theta}_i).$$

We note that because an iteratively re-weighted least squares algorithm is used to calculate the design-weighted M-quantile regression fit at $\bar{\theta}_i$, it immediately follows that we may write (9) as linear combinations of the sample values of y, i.e., $\hat{\bar{Y}}_i^{WMQ} = N_i^{-1}\mathbf{w}_i^{*T}\mathbf{y}$ where $\mathbf{w}_i^* = (w_{ij}^*) = \mathbf{W1}_i + \mathbf{C}(\bar{\theta}_i)\mathbf{X}\Big(\mathbf{X}^T\mathbf{W}\mathbf{C}(\bar{\theta}_i)\mathbf{X}\Big)^{-1}\Big(\sum_{j\in U_i}\mathbf{x}_{ij}^T - \sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}^T\Big)^T$.

Here $\mathbf{1}_i$ is the n-vector with j-th component equal to one whenever the corresponding sample unit is in area i and is zero otherwise. The area-specific M-quantile coefficient θ_i may be estimated using $\tilde{\bar{\theta}}_i = \sum_{j=1}^{n_i} \breve{w}_{ij}\theta_{ij}$ instead of the unweighted average $\bar{\theta}_i$. Using the simulation exercise discussed in next Section, we found that this choice has no appreciable impact on the efficiency of estimators.

As anticipated, expression (9) has a GREG-like form and may be seen as the design-weighted version of (7). It may easily be noted that it is nearly unbiased and design consistent. Moreover, under

the usual assumptions that guarantee the application of the finite population central limit theorem to the Horwitz-Thompson estimators, we have that $(\hat{\bar{Y}}_i^{WMQ} - \bar{Y}_i)/\sqrt{V(\hat{\bar{Y}}_i^{WMQ})} \rightarrow N(0,1)$. See Breidt et al. (2005). With respect to ordinary GREG estimators used in small area literature (see Rao, 1999, Section 2.5), note that: *i*) the use of an area specific coefficient in M-quantile regression accounts for area characteristics not explained by the auxiliary variables; *ii*) the use of M-estimation makes estimator (9) robust to data points with high leverages. However, similarly to the ordinary GREG, estimator (9) is not robust to outliers in the y that are not outliers in the auxiliary variables.

Estimates of small area quantiles may be obtained straightforwardly by inverting $\hat{F}_i^{WMQ/RKM}$:

$$\hat{Q}_i^{WMQ}(q) = inf_{x\in\Re}\hat{F}_i^{WMQ/RKM}(x) \geq q = (\hat{F}_i^{MQ/RKM})^{-1}(q), \quad (10)$$

$q \in (0,1)$. Note that, while estimator of the area mean (9) needs only the means of the auxiliary variables to be known at the population level, the estimator for the quantiles does require $\mathbf{x}_{ij}$ to be known $\forall j \in r_i$.

3.3 Estimation of the design-based variance

Estimators of the design-based Mean Square Error of $\hat{\bar{Y}}_i^{WMQ}$, $\hat{Q}_i^{WMQ}(q)$ and other functionals of $\hat{F}_i^{WMQ/RKM}$ may be obtained using bootstrap. For sampling desings as general as multistage stratified design with unequal inclusion probabilities bootstrap algorighms are known in the literature. See for instance Rao (1999, Section 5).

As far as $\hat{\bar{Y}}_i^{WMQ}$ is concerned, in view of its nearly design unbiasedness we may consider the simple estimator of its variance based on Taylor linearization:

$$\hat{V}(\hat{\bar{Y}}_i^{WMQ}) = \frac{1}{N_i^2}\sum_{j\in s_i}\sum_{k\in s_i}\frac{\pi_{ijk}-\pi_{ij}\pi_{ik}}{\pi_{ijk}}\frac{(y_{ij}-\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{\psi w}(\bar{\theta}_i))}{\pi_{ij}}\frac{(y_{jk}-\mathbf{x}_{ik}^T\hat{\boldsymbol{\beta}}_{\psi w}(\bar{\theta}_i))}{\pi_{ik}}, \quad (11)$$

where π_{ijk}s are the joint order inclusion probabilities. The estimator is a first order approximation because (11) does not take into account both the variability due to the estimation of the value $\bar{\theta}_i$ and that associated to the estimation of $\hat{\boldsymbol{\beta}}_{\psi w}(\bar{\theta}_i)$. The estimator (11) underestimates the actual $MSE(\hat{\bar{Y}}_i^{WMQ})$, but if the overall sample size (on which estimation of the M-quantile model is based) is at least moderate and sampling variance of the y_{ij} and $\mathbf{x}_{ij}$ dominates that associated to the uncertainty in estimating θ_i, the underestimation is likely to be small. This issue will be considered also in the simulation exercise of next Section, when it will be shown that it enjoys good properties even with quite small n.

4 Simulation exercise

In this Section we report results from a simulation based on the *Swissmunicipalities* population, that provides information about the Swiss municipalities in 2003 (Tillé and Matei, 2009). As we are interested in design based properties of estimators, this population will be kept fix and, repeatedly sampled according to a design described below. We are interested in comparing the bias and mean square error of the discussed estimators of small area means and quantiles to those of selected alternatives and, secondly to assess the performances of the proposed MSE estimators. In this study we focused on MSE estimation for the 20th and 80th percentile using the bootstrap estimator, and for the mean using either the approximated variance estimator (11) or the bootstrap estimator.

4.1 Description of the simulation experiment

The *Swissmunicipalities* population consists in the records of 22 variables for each of the 2896 Swiss municipalities. One of this variables is `Canton` that will define areas of interest in the simulation. More

specifically, there are twenty-six of these areas whose sizes range from 3 to 400. We merge the smallest canton (of size 3) with the adjacent canton so we work with a collection of 25 areas. Our target variable is the area with buildings (`Airbat`; Y); its distribution is skewed to the right. As auxiliary information, assumed known for each unit in the population, we consider a single variable defined as the square root of the the total municipality population (`PopTot`; X). Note that $corr_U(Y, X) = 0.78$, where $corr_U(,)$ indicates the population-level linear correlation coefficient. Fitting the mixed model with canton-specific random effects and computing the Shapiro-Wilk normality test on the residuals, we obtain a p-value $< 2.2e^{-16}$ showing that the null hypothesis that the residuals follow a normal distribution is rejected. For this reason, the use of an M-quantile model that relaxes these assumptions, with a bounded influence function, seems reasonable for these data.

Samples are selected according to a fixed size, unequal probability without replacement sampling design using the Midzuno's method (Deville and Tillé , 1998) implemented in the `Sampling` package (Tillé and Matei, 2009) running under `R`. We consider a sample sizes of $n = 290$ units corresponding approximately to a 10% sampling rate and two different size variables Z: i) Area under cultivation (`Surfacescult`), ii) a Uniform variable on the interval $(1, 20)$. More specifically, in case i), we define $\pi_j = 0.2 \times z_j + 0.05$, $\forall j \in U$ to avoid that some inclusion probabilities are greater than one.

Using the first size variable the design is non-ignorable if we condition only on `PopTot` as $corr_U(\texttt{Airbat}, \texttt{Surfacescult}|\texttt{PopTot}) = 0.327$ and this correlation is signficant. Of course, the design would become ignorable if we include `Surfacescult` into the models, but we prefer not to do so, to mimic situations where not all variables relevant for the design or the non response correction are included into the model. With the second size variable we have an ignorable desing conditionally on `PopTot`. To emphasize the difference between the two scenarios, we label the desing as 'non-ignorable' when the size variable is `Surfacescult` and 'ignorable' when the Uniform is the size variable.

The estimators we are going to compare with those obtained with the illustrated Weigthed M-quantile (WMQ) method are the 'direct', the unweighted M-quantile (MQ), and only as far as the mean is concerned, the EBLUP (see Rao, 2003, Chapter 7), pseudo-EBLUP (You and Rao, 2000) and the GREG (see Rao, 2003, Section 2.5) for small areas. The 'direct' estimator of the mean is a post-stratified ratio estimator that does not make use of auxiliary information, i.e. $\hat{\bar{Y}}_i = \sum_{j \in s_i} \tilde{w}_{ij} y_{ij}$ with $\tilde{w}_{ij} = N_i(\hat{N}_i \pi_{ij})^{-1}$, $\hat{N}_i = \sum_{j \in s_i} \pi_{ij}^{-1}$. The direct estimator of quantiles is obtained by inverting the weight empirical distribution function (with weights $\tilde{w}_{ij}$).

Note that for the M-quantile models the Huber Proposal 2 influence function is used with $c = 1.345$. This value provides 95% efficiency when the errors are normal and still offers protection against outliers Huber (1981). In addition the Huber Proposal 2 satisfies the assumption i) of *Theorem 1.*

In the simulation, we use the following procedure described by Särndal et al. (1992, Chapter 11) for estimating the MSE. We illustrate it for $\hat{\bar{Y}}_i^{WMQ}$, but it applies similarly to $\hat{Q}_i^{WMQ}(q)$.

- Using the sample data to build an artificial population U^* by using the sample weights w_{ij}.
- Draw T independent samples s^* from U^* by using the same design used to draw s from U.
- For each bootstrap sample compute the replication $\hat{\bar{Y}}_{it}^{WMQ*}$ of $\hat{\bar{Y}}_i^{WMQ}$ for $i = 1, \ldots, m$ and $t = 1, \cdots, T$.

The bootstrap estimator of the MSE of $\hat{\bar{Y}}_i^{WMQ}$ for each small area can be computed as $\widehat{MSE}_i = T^{-1} \sum_{t=1}^{T} \left(\hat{\bar{Y}}_i^{WMQ*} - \hat{\bar{Y}}_i^{WMQ} \right)^2$.

The Monte-Carlo experiment consists in drawing $R = 5,000$ samples from this population and calculating small area estimators for the mean, the 20th and the 80th percentile of `Airbat`, along with bootstrap estimators of their MSE based on $T = 200$ bootstrap replicates and the approximated MSE

Table 1: Design-based simulation results using the Switzerland data. Results show across areas distribution of Absolute Relative Bias (ARB%) and Relative Root Mean Square Error (RRMSE%) over simulations.

		Summary of across areas distribution					
Predictor	Indicator	Min	Q1	Median	Mean	Q3	Max
		non-ignorable design					
MQ	ARB(%)	0.36	1.85	3.58	5.33	5.97	18.35
	RRMSE(%)	6.45	9.68	13.46	16.34	21.82	42.91
WMQ	ARB(%)	0.01	0.24	0.52	0.74	0.75	2.69
	RRMSE(%)	6.15	9.19	12.27	13.37	14.27	34.55
EBLUP	ARB(%)	0.33	7.22	12.09	15.60	20.35	46.48
	RRMSE(%)	7.01	12.32	20.64	23.34	26.04	57.15
pseudo-EBLUP	ARB(%)	0.33	6.11	9.02	11.53	15.56	37.62
	RRMSE(%)	6.60	12.07	15.04	17.65	20.15	40.81
GREG	ARB(%)	0.01	0.19	0.37	0.69	0.85	2.41
	RRMSE(%)	8.08	12.78	16.47	19.00	21.02	46.69
Direct	ARB(%)	0.08	0.88	1.37	2.35	3.49	8.89
	RRMSE(%)	19.23	24.96	31.96	36.99	41.62	101.57
		ignorable design					
MQ	ARB(%)	0.03	0.42	1.25	1.63	2.72	4.34
	RRMSE(%)	6.37	10.36	11.92	14.21	17.11	32.60
WMQ	ARB(%)	0.00	0.29	0.45	0.70	0.91	2.52
	RRMSE(%)	6.17	11.14	14.16	16.30	18.82	45.29
EBLUP	ARB(%)	0.07	3.09	6.45	8.54	12.88	32.30
	RRMSE(%)	6.65	9.84	13.12	14.69	16.94	34.80
pseudo-EBLUP	ARB(%)	0.24	4.46	7.07	9.22	12.06	31.82
	RRMSE(%)	7.66	10.88	15.90	16.55	18.91	35.24
GREG	ARB(%)	0.03	0.22	0.39	0.72	0.58	6.80
	RRMSE(%)	10.05	14.64	16.48	19.98	21.77	42.78
Direct	ARB(%)	0.05	0.21	1.05	1.33	1.92	5.74
	RRMSE(%)	20.15	30.30	36.90	43.48	50.24	115.92

for the mean estimators. The performance of the different small area estimators is evaluated with respect the absolute relative bias and the relative root mean square error of estimates of the small area parameters. The absolute relative bias for small area i is computed as

$$ARB_i = \frac{1}{T}\left|\sum_{t=1}^{T}\frac{\hat{m}_{it} - m_i}{m_i}\right| \times 100,$$

and the relative root mean square error for area i is computed as

$$RRMSE_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{(\hat{m}_{it} - m_i)^2}{m_i}} \times 100.$$

Note that the subscript t ($t = 1 \ldots T$) indexes the Monte-Carlo simulations, with m_i denoting the value of the small area i parameter while $\hat{m}_{it}$ denotes the estimate of m_i in MC replication t.

4.2 Discussion of the results

We show detailed results only for small area means estimators, because of space constraints. Tables illustrating the rest of results are available from the authors upon request. Summaries of the distributions of the ARB and RRMSE of the estimators for the mean across the areas are set out in Tables 1. It shows that the WMQ predictor has a much better performance in terms of bias and efficiency when the design is non-ignorable. In addition the relative bias and thus the RRMSE of the WMQ estimator tend to zero faster than the those values of the pseudo-EBLUP. As stated in You and Rao (2002) the pseudo-EBLUP is derived under the assumption of ignorability of the sampling design; more specifically variance components used in weighting the composite estimator's elements are unweighted, design-biased estimators. In fact, if we include `Surfacescult` into the equation of the linear mixed model, pseudo-EBLUP compares to WMQ similarly to the case of ignorable design.

Under this scenario, we have that WMQ and GREG predictors have a smaller design-bias than MQ, pseudo-EBLUP and EBLUP estimators, but WMQ, pseudo-EBLUP and GREG estimators show bigger variability than MQ and EBLUP estimators, and for this reason they loose in terms of efficiency.

In principle we may expect that $\hat{\boldsymbol{\beta}}_{w\psi}(q)$ is more biased and less variable than $\hat{\mathbf{B}}$ on which the GREG relies, because of the bias-variance trade-off typical of robust estimators. We may also expect that this is true also for small area mean estimators as they share the same structure. But, since WMQ makes use of area-specific M-quantile coefficients, the resulting estimators show biases comparable to those of the GREG, while keeping smaller variances.

Let's now summarize other results (tables not reported). As far as the estimation of quantiles, we note that WMQ is more efficient than its unweighted counterpart under both ignorable and non-ignorable design. In the first case it is characterized by larger variances, but remarkably smaller biases. We also calculated estimators of the percentiles based on the (unweighted) linear mixed models. Their performances are close to those of MQ, but the biases are slightly larger. About MSE estimation, we note that the estimators based on the bootstrap algorithm track very well the actual root MSE of mean and percentiles area by area. They show a small negative bias, especially in areas with very small (less than 10) area-specific sample sizes. For the estimation of the MSE of the WMQ estimator of small area means, we have that estimator (11) performs well, showing a small amount of underestimation.

Acknowledgements

This work is financially supported by the European Project SAMPLE "Small Area Methods for Poverty and Living Condition Estimates", European Commission 7th FP - www.sample-project.eu, and by the Research Project PRIN2007 "Inference in the presence of imperfect auxiliary information: sampling from elusive populations and estimation for unplanned domains".

REFERENCES (RÉFERENCES)

Battese G.E., Harter R.M., Fuller W.A. (1988). An error components model for prediction of county crop area 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.

Breidt F.J., Claeskens G., Opsomer J.D. (2005). Model-assisted estimation for complex surveys using penalised splines. Biometrika. **92**, 831–846.

Chambers R., 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

Deville J.C., and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. Biometrika, **85**, 89–101.

Estevao V.M., Särndal C.E (2004). Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. Journal of Official Statistics, **20**, 645–669.

Huber P.J. (1981). Robust statistics. Wiley, New York.

Jiang J., Lahiri P. (2006a) Mixed Model Prediction and Small Area Estimation (with discussion). TEST. **15**, 1–96.

Jiang J., Lahiri P. (2006b). Estimation of finite population domain means: a model-assisted empirical best prediction approach. Journal of the American Statistical Asssociation. **101**, 301–311.

Koenker R., Bassett G. (1978). Regression quantiles. Econometrica. **101**, 33–50.

Kocic P., Chambers R., Breckling J., Beare S. (1997) A measure of production performance. Journal of Business and Economics Statistics, **15**, 445–451.

Kott P. (1989). Robust small domain estimation using random effects modelling. Survey Methodology. **15**, 1–12.

Newey W.K., Powell J.L. (1987). Asymetric least squares estimation and testing. Econometrica. **55**, 819–847.

Prasad N.G.N., Rao J.N.K. (1999). On robust small area estimation using a simple random effects model. Survey Methodology. **25**, 67–72.

Rao J.N.K., Kovar J.G., Mantel H.J. (1990) On estimating distribution functions and quantiles from survey data using auxiliary information. Biometrika, **77**, 365–375.

Rao J.N.K.(1999). Some current trends ins sample survey theory and methods. Sankhyā: The Indian Journal of Statistics, Series B. **61**, 1–57.

Rao, J.N.K.(2003). Small Area Estimation. Wiley, New York.

Särndal C.E (1984). Design-consistent versus model-dependent estimation for small domains. Journal of the American Statistical Association. **79**, 624–631.

Särndal C.E., Swenssson B., Wretman J. (1992). Model Assisted Survey Sampling. Springer-Verlag, New York.

Tillé Y., Matei A. (2009) Package `Sampling` Functions for drawing and calibrating samples, dowloadable at `http://cran.r-project.org/web/packages/sampling/`.

Tzavidis N., Marchetti S., Chambers R. (2010) Robust estimation of small-area means and quantiles. The Australian and New Zealand Journal of Statistics, **52**, 167–186.

You Y., Rao J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. Survey Methodology. **26**, 173–181.

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