Robust Bayesian small area estimation based on quantile regression

Robust Bayesian small area estimation based on quantile regression

Enrico Fabrizi

Università Cattolica del S. Cuore. Address: Via Emilia Parmense 84, 29122 Piacenza, Italy

Nicola Salvati

Dipartimento di Economia e Management, Università di Pisa

Carlo Trivisano

Dipartimento di Scienze Statistiche 'P. Fortunati', Università di Bologna

Abstract

Quantile and M-quantile regression have been applied successfully to small area estimation within the frequentist approach. Quantile regression is applied in the same context but from a Bayesian perspective. Joint modelling of the quantile function is considered, adopting a non parametric assumption on the data generating process that nonetheless explicitly includes the normal distribution as a special case. A specification of the random part of the model that is simple and consistent with the predictive aim of small area estimation is proposed. Although the main output of the method is the estimation of the whole quantile function, estimators of the small area means based on the integration of the quantile function are proposed and discussed. A simulation exercise is used to assess the frequentist properties of these proposed predictors, that result at least as efficient as frenquentist small area estimators based on quantile regression in scenarios characterized by the presence of outliers. The proposed method is illustrated using data from the European survey on Income and Living Conditions (EU-SILC).

Keywords: Bayesian non-parametrics, Dirichlet process priors, quantile function, survey data, frequentist properties of Bayesian methods
2010 MSC: 62D05, 62G08, 62P25

1. Introduction

Small area estimation is aimed at predicting finite population descriptive quantities (such as mean, totals, quantiles, concentration indexes) for sub-populations for which specific sample sizes are, in all or most cases, not large enough to allow for reliable inference using ordinary, *direct* design based estimation methods. These sub-populations are usually labelled as small areas. More precise predictions can be obtained when auxiliary information is available at both the sample and the census level.

Email address: enrico.fabrizi@unicatt.it (Enrico Fabrizi)

Linking auxiliary information to the target variable under consideration requires some form of modelling. When auxiliary information is available for each statistical unit a *unit level* model can be specified. Linear mixed models are the most popular class of models used for this purpose. A review of the literature on the application of linear mixed models to small area estimation from both the frequentist and Bayesian perspective can be found in Rao and Molina (2015, chapter 7).

Chambers and Tzavidis (2006) introduce quantile and M-quantile regression as an alternative class of models for the same purpose relaxing some of the conventional modeling assumptions such as the normality of the random components and obtaining estimators robust to outliers. Several developments of their proposal and applications to various research problems appeared since then; see Bianchi et al. (2018) for a review. All these applications of quantile regression to small area estimation consider a frequentist approach to inference and are based on separate fitting of regression planes indexed on different quantiles. Moreover they are focused on the estimation of point population descriptive quantities and namely the mean.

In this paper, we adopt a Bayesian approach to estimation; more importantly, we consider simultaneous quantile regression, that amounts to specify a model for the quantile function. This allows us to widen the scope of estimation to get estimates of the whole quantile function at the area level, a richer picture that we can get from the estimation of single descriptive quantities.

Early Bayesian papers on quantile regression are based on the idea of estimating regression planes indexed on different quantiles separately, using the Asymmetric Laplace or some generalization of it as pseudo-likelihood for the data. See Yu and Moyeed (2001) and Kottas and Krnjajic (2009), among others. Following a different line of research (see Reich et al., 2010; Tokdar and Kadane, 2012; Yang and Tokdar, 2017), we consider simultaneous linear quantile regression, in which the estimation of the quantile function is directly targeted.

Specifically, we consider an approach to linear quantile regression due to Reich (2012), Reich and Smith (2013) and adapt it to the purpose of small area estimation by introducing random effects into the model to describe between area variation not accounted for by the regressors. Random effects are introduced also in Smith et al. (2015) for the same model; nonetheless our specification is different as we are interested in inference conditional on the random effects and in prediction rather than assessing the marginal effects of the covariates.

The considered approach has several advantages. First, although the specification for the quantile function is non-parametric, it is centered around a parametric baseline that can be recovered as a special case. Second, area-specific parameters determine not only the location but also the scale and shape of the distribution, so area level features can be accounted for. Third, as under mild conditions a closed form for the likelihood exists, standard MCMC algorithms, such as Gibbs sampling with Metropolis-Hastings blocks can be used to explore posterior distributions. As a consequence, not only the quantile functions but also posterior distributions of their functionals can easily explored. Fourth, prediction of small area level quantile functions, and their functionals, although non linear, require only averages of auxiliary variables to be known at the area level and not individual out of sample values.

We will devote a special attention to the estimation of small area means, characterized as functional of the quantile function. This choice is motivated mainly by comparability purposes: as most small area literature is focused on mean estimation, we investigate how our approach compares to others in the estimation of means, despite its purpose is more general. We prove that, under the considered model and mild conditions on the chosen baseline, small area means can be expressed as a linear combination of the parameters characterizing the quantile function. The posterior distribution of these functions of parameters can be explored using MCMC algorithm and their posterior means proposed as point predictors.

We are interested in assessing frequentist properties of various estimators of the small area means based on our model. We do this by means of a simulation exercise, in which they are compared to selected frequentist alternatives including the M-quantile methods proposed by Chambers and Tzavidis (2006) and Chambers et al. (2014), the Empirical Best Linear Unbiased Predictor (EBLUP) (Rao and Molina, 2015, chapter 5) based on the assumption of normality and its robust version (REBLUP) by Sinha and Rao (2009). The simulation setting is the same considered in Chambers et al. (2014) and it is focused on the study of the robustness and the efficiency of the estimators. We also compare our estimators to alternatives from the Bayesian literature. Specifically, we consider the one based on the Bayesian analysis of the nested error regression model, known as Battese-Harter-Fuller model in the small area literature, the way it is proposed by Datta and Ghosh (1991). Quite surprisingly, although the Bayesian literature is rich of robust extensions of the Fay-Herriot area-level model, there are few proposals of robust alternatives to the unit-level nested error regression model. We can mention Chakraborty et al. (2018); nonetheless these authors allow for non-normality of the individual level errors, but keep the normality for the area level random effects. As far as we know, there are no applications of Bayesian quantile regression to small area estimation.

We apply the proposed method to the analysis of a real data set. We consider a sample of households from the survey on Income and Living Conditions (SILC) conducted in Italy as in most European countries with consistent methodology. We focus on households with at least one person aged 60 or more and receiving old age benefits (pension earner) and we target the quantile functions and the mean of the equivalized household income. We are interested in households with pension earners as they represent a subset of the population particularly exposed to the risk of poverty and social exclusion. We target equivalized household income as poverty and inequality measures currently adopted in the EU are based on this variable.

In summary, this research contributes to the literature in two ways: *i*) it introduces Bayesian quantile regression based on the joint estimation of quantiles in small area estimation literature with the aim of directly targeting the estimation of the quantile function at the area level; *ii*) it provides the basis for estimating any functional of the quantile function at the small area level.

The paper is organized as follows. In Section 2 we set some notation and review frequentist approach to quantile regression applied to small area estimation. In Section 3 we introduce the Reich's approach to quantile regression and our extension of it, while in Section 4, we prove our result on the representation of small area means. In Section 5 we present the simulation exercise and analyze the results, while the application to real data is presented in Section 6. Concluding remarks and possible extensions are discussed in the Section 7.

2. Review of quantile regression methods in small area estimation

Let's consider a population P of size N and a partition of it into m small areas $\{P_1, \ldots, P_h, \ldots, P_m\}$ of size N_h, $\sum_{h=1}^{m} N_h = N$. Suppose we target a variable y. Several quantities that describe the distribution of y in the small areas can be of interest. Most small area estimation focuses on the estimation of area level descriptive quantities; in most cases the mean:

$$\bar{Y}_h = N_h^{-1} \sum_{i \in P_h} y_i.$$

Sometimes also area-specific τ-quantiles: $Q_h(\tau) = inf\{t | N_h^{-1} \sum_{i \in P_h} \Delta(t - y_i) \geq \tau\}$ with $\Delta(t) = 1$ when $t \geq 0$ and $\Delta(t) = 0$ elsewhere, are of interest.

Let's assume that an overall sample s of size n is drawn from the population and that sampling is non-informative. The sample is partitioned into area-specific samples s_h of size n_h ($n_h \geq 0$). We also assume that a $p+1$ vector of auxiliary information $\mathbf{x}_i$ is known for each unit in the sample (i.e $i \in s$); moreover $x_{i1} = 1$. Area-level means of the auxiliary variables $\bar{\mathbf{x}}_h = N_h^{-1} \sum_{i \in P_h} \mathbf{x}_i$ are assumed to be accurately known and so are $\bar{\mathbf{x}}_{rh} = (N_h \bar{\mathbf{x}}_h - \sum_{i \in s_h} \mathbf{x}_i)/(N_h - n_h)$.

We exploit the relationship between the target and the auxiliary variables to improve the estimation of $\bar{Y}_h$ by assuming a super-population model, as we are interested in problems where n_h are so small that estimators of $\bar{Y}_h$ based on area-specific sampling are not precise enough. We focus on predictors in the form:

$$\hat{\bar{Y}}_h^P = \frac{1}{N_h}\Big\{ \sum_{i \in s_h} y_i + (N_h - n_h)\hat{\bar{Y}}_{rh} \Big\}. \tag{1}$$

where $\bar{Y}_{rh} = \sum_{i \in (P_h - s_h)} y_i$ and $\hat{\bar{Y}}_{rh}$ its predictor.

The nested error regression model proposed by Battese et al. (1988) can be written as:

$$y_i = \mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \mathbf{v} + e_i, \tag{2}$$

where $\mathbf{v}$ is an m vector with independent components for which we assume $v_h \sim N(0, \sigma_v^2)$ while $e_i \sim N(0, \sigma_e^2)$; $\mathbf{z}_i$ is an m vector such that $z_{ih} = 1$ when $i \in s_h$ and 0 otherwise.

Under this model $\hat{\bar{Y}}_{rh} = \bar{\mathbf{x}}_{rh}^t \tilde{\boldsymbol{\beta}} + \tilde{v}_h$, where $\tilde{\boldsymbol{\beta}}$ is an estimate of the regression parameter and $\tilde{v}_h$ a prediction of the area-specific random effect. Model (2) is the most popular unit level model used in small area estimation and inference and is discussed in several papers, under both the frequentist and the Bayesian approach, (Rao and Molina, 2015, chapter 6). The popular Empirical Best Linear Unbiased predictor (EBLUP) of $\bar{Y}_h$ associated to (2) lacks robustness when normality of e_i, v_h fails. Sinha and Rao (2009) introduce M-estimation ideas for the variance components leading to robustified predictors based on the Battese-Harter-Fuller model.

The recourse to quantile regression offers an alternative to linear mixed models that is particularly useful when distributional assumption are difficult to specify. Let' assume the linear quantile regression model for the variable y associated to unit i in the sample:

$$q(\tau | \mathbf{x}_i) = \mathbf{x}_i^t \boldsymbol{\beta}_\tau,$$

where $\boldsymbol{\beta}_\tau$ defined as the minimizer of

$$\min_{\beta} E\left[|\tau - I(u < 0)| \rho(u) \right],$$

with $u = (y - \mathbf{x}^t \beta)/\sigma_\tau$, σ_τ being a scale parameter. The loss function $\rho(u)$ is given by $\rho(u) = |u|$ in ordinary quantile regression while it is generalized using M-estimation ideas in M-quantile regression (Breckling and Chambers, 1988), using for instance the popular Huber loss function:

$$\rho(u) = \begin{cases} (c|u| - c^2/2) & |u| > c \\ u^2/2 & |u| \leq c. \end{cases}$$

The constant $c > 0$ can be fixed to compromise between efficiency and robustness (a common choice is $c = 1.345$). Assuming ρ is (a.e.) continuously differentiable and convex, an estimator of $\boldsymbol{\beta}_\tau$, $\hat{\boldsymbol{\beta}}_\tau$, can be obtained as the solution of the following system of equations

$$\sum_{i \in s} \psi_\tau \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\hat{\sigma}_\tau} \right) \mathbf{x}_i = \mathbf{0},$$

where $\hat{\sigma}_\tau$ is a consistent estimator of σ_τ and $\psi_\tau(u) = d\rho_\tau(u)/du = |\tau - I(u < 0)|\psi(u)$, with $\psi(u) = d\rho(u)/du$. An iterative method is needed here to obtain a solution, like an iteratively re-weighted least squares algorithm or the Newton-Raphson algorithm.

The application of M-quantile regression to small area estimation introduced by Chambers and Tzavidis (2006) is based on the idea of area-characterizing quantile τ_h (M-quantile coefficient). For unit i the (unique) M-quantile coefficient τ_i is the one for which $y_i = \mathbf{x}_i^t \boldsymbol{\beta}_{\tau_i}$ hold exactly. The area-characterizing quantile will be defined as an average of τ_i for units within the same area: $\tau_h = N_h^{-1} \sum_{i \in P_h} \tau_i$. The value $\hat{\tau}_h$ will be a sample based estimator of τ_h. 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 approach to estimating area random effects without the need for parametric assumptions. Chambers and Tzavidis (2006) propose a predictor in the form

$$\hat{\bar{Y}}_h^{MQ} = N_h^{-1}\Big\{ \sum_{i \in s_h} y_i + (N_h - n_h)\hat{\bar{Y}}_{rh} \Big\}, \tag{3}$$

where, in this case, $\hat{\bar{Y}}_{rh} = \bar{\mathbf{x}}_{rh}^t \hat{\boldsymbol{\beta}}_{\hat{\tau}_h}$ with $\hat{\boldsymbol{\beta}}_{\hat{\tau}_h}$ estimating $\boldsymbol{\beta}_{\tau_h}$ in $q_{\tau_h}(y_i|\mathbf{x}_i) = \mathbf{x}_i^t \boldsymbol{\beta}_{\boldsymbol{\tau_h}}$.

This idea proved to be effective and led to many applications and theoretical developments. See Bianchi et al. (2018) for a review. As predictors in the form (3) tend to be biased, Chambers et al. (2014) propose a bias-corrected predictor in the form:

$$\hat{\bar{Y}}_h^{MQ-BC} = N_h^{-1}\Big\{ \sum_{i \in s_h} y_i + (N_h - n_h)\bar{\mathbf{x}}_{rh}^t \hat{\boldsymbol{\beta}}_{\tau_h} + \sum_{i \in s_h} w_i^{\psi} \xi\Big(\frac{y_i - \hat{y}_i^{\psi}}{w_i^{\psi}}\Big) \Big\}, \tag{4}$$

where ξ, ψ, such that $|\xi| > |\psi|$, are influence functions associated to the loss function (3) defined for different values of the tuning constant c. The second influence function, ξ, is still bounded but 'less restrictive' than ψ, and its purpose is to define an adjustment for the bias caused by the fact that the first two terms on the right-hand side of equation (4) treat sample outliers as self-representing; w_i^{ψ} is a robust estimator of scale of the residuals $y_i - \hat{y}_i^{\psi}$ with $\hat{y}_i^{\psi} = \mathbf{x}_i^t \hat{\boldsymbol{\beta}}_{\tau_h}$ (see Chambers et al., 2014, for details).

We note that, although relying on quantile regression, Chambers and Tzavidis (2006) and subsequent developments are in most cases concerned with the estimation of $\bar{Y}_h$ and do not consider the estimation of the quantile function at the area level as an explicit target.

3. Bayesian quantile regression and the Reich's model

In the Bayesian literature, quantile regression has been studied using different approaches. Although an accurate review of this literature is beyond the scope of this article, we shortly discuss what we need to motivate our choice. Some contributions adopt parametric *pseudo-likelihoods* such as the asymmetric Laplace (Yu and Moyeed , 2001; Yue and Rue, 2011) or more flexible alternatives (Taddy and Kottas, 2010; Wichitaksorn et al., 2014) and estimate regression planes separately for each quantile of interest. These methods are computationally simple but the use of *pseudo-likelihoods* has been criticized (see Yang et al., 2016, and the contributions discussing this paper). Separate estimation of quantile regression planes is also a limitation as quantile plane crossing can be a problem; moreover inference of functionals of different quantiles can be difficult as the relationship between different estimated quantiles would be overlooked.

Among the methods considering joint estimation of all quantile regression planes (see Tokdar and Kadane, 2012; Yang and Tokdar, 2017) we consider an approach due to Reich (2012) and Reich and Smith (2013). The reasons for this choice are that, *i*) although non-parametric the model for the quantile function is centered on a parametric baseline that can then be obtained as a special case; *ii*) if the baseline has a density in closed form, a closed form for the likelihood can be obtained; *iii*) the model is analytically tractable and a closed form expression for the population mean can be obtained.

To describe this method let's consider a linear quantile regression model in which we assume that both location and scale of y depend linearly on the auxiliary variables $\mathbf{x}$:

$$\begin{aligned} q(\tau|\mathbf{x}_i) &= \mathbf{x}_i^t\boldsymbol{\alpha}_0 + \mathbf{x}_i^t\boldsymbol{\alpha}_1 q_0(\tau) = \mathbf{x}_i^t\boldsymbol{\beta}(\tau) \\ &= \sum_{j=1}^{p+1} x_{ij}\alpha_{0j} + \sum_{j=1}^{p+1} x_{ij}\alpha_{1j}q_0(\tau) = \sum_{j=1}^{p+1} x_{ij}\{\alpha_{0j} + \alpha_{1j}q_0(\tau)\} = \sum_{j=1}^{p+1} x_{ij}\beta_j(\tau). \end{aligned} \tag{5}$$

The expression $q_0(\tau)$ is assumed to be the quantile function of a continuous variable with location equal to 0 and scale equal to 1. If the choice $q_0(\tau) = \Phi^{-1}(\tau)$ is taken, where $\Phi^{-1}(\tau)$ is the quantile function of a $N(0,1)$ random variable, it would imply $y_i|\mathbf{x}_i \sim N(\mathbf{x}_i^t\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1^t\mathbf{x}_i\mathbf{x}_i^t\boldsymbol{\alpha}_1)$, i.e. an heteroskedastic normal regression model. The impact of the j-th regressor on the scale of the distribution changes with the quantile whenever $\alpha_{1j} \neq 0$ for some $j > 1$, otherwise we would have a simpler homoskedastic model where the auxiliary variables impact only on the location.

Reich and Smith (2013) propose the specification of a more flexible model generalizing the derivative of $q(\tau|\mathbf{x}_i)$ in τ:

$$\frac{dq(\tau|\mathbf{x}_i)}{d\tau} = \mathbf{x}_i\boldsymbol{\alpha}_1\frac{d}{d\tau}q_0(\tau) = \sum_{j=1}^{p+1} x_{ij}\alpha_{1j}\frac{d}{d\tau}q_0(\tau), \tag{6}$$

assuming $\frac{d}{d\tau}q_0(\tau)$ exists. Specifically they propose to replace (6) with a piece-wise derivative function

$$\frac{dq(\tau|\mathbf{x}_i)}{d\tau} = \sum_{\ell=1}^{L+1} I(\kappa_{\ell-1} < \tau \leq \kappa_\ell)\mathbf{x}_i^t\boldsymbol{\alpha}_\ell\frac{d}{d\tau}q_0(\tau), \tag{7}$$

where the $L+1$ piece-wise intervals are separated by breakpoints $0 = \kappa_0 < \kappa_1 < \cdots < \kappa_L < \kappa_{L+1} = 1$. The idea is to obtain a quantile function whose shape is locally that of the baseline $q_0(\tau)$ but where each piece is characterized by a different *local* parameter. Reich and Smith (2013) show that the continuous quantile function corresponding to (7) is given by:

$$q(\tau|\mathbf{x}_i) = \mathbf{x}_i^t\boldsymbol{\alpha}_0 + \sum_{\ell=1}^{L+1} \mathbf{x}_i^t\boldsymbol{\alpha}_\ell B_\ell(\tau). \tag{8}$$

The $B_\ell(\tau)$ result from a piecewise decomposition of $q_0(\tau)$, such that $\sum_{\ell=1}^{L+1} B_\ell(\tau) = q_0(\tau)$; we report a detailed definition of $B_\ell(\tau)$ in the Appendix. In order to have a monotonically increasing quantile function we must have $\mathbf{x}_i^t\boldsymbol{\alpha}_\ell > 0\ \forall \ell, i$, so $\boldsymbol{\alpha}_\ell$ parameters must be constrained in the estimation process. The larger L, the more flexible is the model (and richer in terms of parameters). According to Reich and Smith (2013) L between 3 and 5 already introduces a reasonable amount of flexibility, a finding that our result will confirm.

The normal baseline, i.e. $q_0(\tau) = \Phi^{-1}(\tau)$, deserves special attention in small area estimation. In fact the baseline model becomes the actual model whenever $\boldsymbol{\alpha}_\ell = \boldsymbol{\alpha}$ ($\ell = 1, \ldots, L$) and

differences in the (estimated) $\boldsymbol{\alpha}_\ell$ can be consistently interpreted as departures from the baseline model. The normality assumption is considered in the Battese-Harter-Fuller model (2) and plays a reference role in small area estimation.

Other choices of the baseline are possible provided that its location is set to 0 and its scale to 1 to guarantee identification (see Reich and Smith, 2013). Additional shape parameters can in principle be included and treated as additional model parameters. Possible alternatives to $\Phi^{-1}(\tau)$ include t, exponential power, generalized hyperbolic, Fréchet distributions.

Note that, in (8), the term $\mathbf{x}_i^t\boldsymbol{\alpha}_0$ does not depend on τ, so $\boldsymbol{\alpha}_0$ influences only the location of the conditional distribution. The interpretation of the $\boldsymbol{\alpha}_\ell$, $\ell \geq 1$ is that of *local scale parameters* that, as such, influence the shape of the distribution (i.e. the tails).

There exists a closed form expression for the density associated to (8) that we report here as our notation is slightly different from that of Reich and Smith (2013):

$$f(y|\mathbf{x}_i, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_L) = \sum_{\ell=1}^{L+1} \left\{ \frac{I[q(\kappa_{\ell-1}|\mathbf{x}_i) < y \leq q(\kappa_\ell|\mathbf{x}_i)]}{\mathbf{x}_i^t\boldsymbol{\alpha}_\ell} f_0\left(\frac{y - q(\kappa_{\ell-1}|\mathbf{x}_i) + \mathbf{x}_i^t\boldsymbol{\alpha}_\ell q_0(\kappa_\ell|\mathbf{x}_i)}{\mathbf{x}_i^t\boldsymbol{\alpha}_\ell} \right) \right\}, \quad (9)$$

where f_0 is the density of the baseline model (that we suppose is well defined). The availability of (9) is of great importance for computation as it allows the implementation of Gibbs sampling type of algorithms in the approximation of the posterior distribution. For this reason the choice of a baseline $q_0(\tau)$ that admits an associated density function is advisable. The density (9) has discontinuities at interior breakpoints. Discontinuous densities are rather common in Bayesian non-parametrics since Ferguson (1973).

4. Extension to small area estimation

In this section we propose the extension of the approach due to Reich (2012), Reich and Smith (2013) to small area estimation. Section 4.1 presents how the quantile function can become area-specific by introducing random effects; in Section 4.2 we propose a Bayes point predictor for small area means under quadratic loss and in Section 4.3 we discuss prior specification for the model parameters.

4.1. Area-specific quantile functions

We propose to generalize (8) by making the $\boldsymbol{\alpha}_0$ parameters area specific, i.e. $\boldsymbol{\alpha}_0$ becomes $\boldsymbol{\alpha}_{0h}$:

$$q_h(\tau|\mathbf{x}_i) = \mathbf{x}_i^t\boldsymbol{\alpha}_{0h} + \sum_{\ell=1}^{L+1} \mathbf{x}_i^t\boldsymbol{\alpha}_\ell B_\ell(\tau). \quad (10)$$

We do not allow parameters $\boldsymbol{\alpha}_\ell$ to be area-specific as this would lead to a too richly parametrized model for which it is very likely not to have enough information for reliable estimation.

Specifically, we can express $\boldsymbol{\alpha}_{0h}$ as $\boldsymbol{\alpha}_{0h} = \boldsymbol{\alpha}_0 + \mathbf{v}_{0h}$ where $\mathbf{v}_{0h}$ are 0 meaned random variables. Note that these random effects are not quantile-specific in line with Koenker (2004), Reich et al. (2010), Smith et al. (2015), but differently from Geraci and Bottai (2014). The random effects $\mathbf{v}_{0h}$ are actually vector valued: not only the general intercept α_{01} is allowed to vary across areas, but also slopes α_{0j}, $j = 1, \ldots, p$ are assumed to be area-specific. A simplified model, more in

line with usual assumptions in small area estimation (i.e. only the intercepts are assumed to be random) can be easily obtained by setting

$$\mathbf{x}_i^t \boldsymbol{\alpha}_{0h} = \alpha_{01h} + \sum_{j=2}^{p+1} x_{ij}\alpha_{0j}. \tag{11}$$

As anticipated in the introduction, our specification of the random effects is different, and simpler, with respect to Smith et al. (2015). The reason is that we are interested in inference conditional on the random effects and in prediction rather than assessing the marginal effects of the covariates.

4.2. Prediction of small area quantiles and means

The linear quantile function (10) can be straightforwardly used to predict a quantile conditional on a given $\mathbf{x}_i$, $i \in P_h$. If we average over all population units in area P_h, i.e. we integrate out the $\mathbf{x}_i$ (assuming a constant point mass probability on each population unit), we obtain

$$\hat{q}_h(\tau) = \frac{1}{N_h}\sum_{i\in P_h} q_h(\tau|\mathbf{x}_i) = \bar{\mathbf{x}}_h^t \boldsymbol{\alpha}_{0h} + \sum_{\ell=1}^{L+1} \bar{\mathbf{x}}_h^t \boldsymbol{\alpha}_\ell B_\ell(\tau), \tag{12}$$

i.e. the marginal τ-quantile of y in area h. The hat over q in $\hat{q}_h(\tau)$ emphasizes that its form depends on the assumed model. As it is function of the auxiliary variables only through $\bar{\mathbf{x}}_h$ its estimation does not require knowledge of individual $\mathbf{x}_i$ for units not in the sample. Note that, although the local scale parameters $\boldsymbol{\alpha}_\ell$ are common to all areas, the scale of the distribution depend on area-specific information through $\bar{\mathbf{x}}_h$, so $\hat{q}_h(\tau)$ are area-specific in both location and shape.

In most small area applications, estimation of small area means is of interest. For the super-population mean θ_h we can obtain, consistently with (12), a useful representation based on the well known relationship between the expectation of a random variable and its quantile function:

$$\hat{\theta}_h = \int_0^1 \hat{q}_h(\tau)d\tau,$$

that leads to the statement of the following result.

Theorem 1. *If we assume that* $\int_{\kappa_\ell}^{\kappa_{\ell+1}} q_0(\tau)d\tau$ *exists and can be expressed as* $\int_{\kappa_\ell}^{\kappa_{\ell+1}} q_0(\tau)d\tau = g\{q_0(\kappa_\ell)\} - g\{q_0(\kappa_{\ell+1})\}$ *for some non-negative function* g *such that* $g\{q_0(0)\} = g\{q_0(1)\} = 0$ *then*

$$\hat{\theta}_h = \int_0^1 \hat{q}_h(\tau)d\tau = \bar{\mathbf{x}}_h^t \boldsymbol{\alpha}_{0h} + \bar{\mathbf{x}}_h^t \Big[\sum_{\ell=1}^{L} (\boldsymbol{\alpha}_{\ell+1} - \boldsymbol{\alpha}_\ell)\Big\{g(q_0(\kappa_\ell)) - (1-\kappa_\ell)q_0(\kappa_\ell)\Big\}\Big]. \tag{13}$$

Proof. Let's start from

$$\int_0^1 \hat{q}_h(\tau)d\tau = \bar{\mathbf{x}}_h^t \Big[\boldsymbol{\alpha}_{0h} + \int_0^1 \sum_{\ell=1}^{L+1} \boldsymbol{\alpha}_\ell B_\ell(\tau)d\tau\Big],$$

and note that $\boldsymbol{\alpha}_\ell$ are vectors of size p, but $B_\ell(\tau)$, see the Appendix for a detailed definition, is a scalar function of τ that multiplies each component of $\boldsymbol{\alpha}_\ell$. For this reason we can work on the

integral above component-wise. For simplicity we will denote α_ℓ a generic component of $\boldsymbol{\alpha}_\ell$. Let $\tilde{q}_B(\tau) = \sum_{\ell=1}^{L+1} \alpha_\ell B_\ell(\tau)$, then

$$\int_0^1 \tilde{q}_B(\tau)d\tau = \int_{\kappa_0}^{\kappa_1} \tilde{q}_B(\tau)d\tau + \int_{\kappa_1}^{\kappa_2} \tilde{q}_B(\tau)d\tau + \cdots + \int_{\kappa_\ell}^{\kappa_{\ell+1}} \tilde{q}_B(\tau)d\tau,$$

where $\kappa_0 = 0$ and $\kappa_{\ell+1} = 1$. If $\tau < \kappa_1$ then $\tilde{q}_B(\tau) = \alpha_1 q_0(\tau)$ so $\int_{\kappa_0}^{\kappa_1} \tilde{q}_B(\tau)d\tau = -\alpha_1 g\{q_0(\kappa_1)\}$. If $\kappa_1 < \tau \leq \kappa_2$ then $\tilde{q}_B(\tau) = \alpha_1 q_0(\kappa_1) + \alpha_2\{q_0(\tau) - q_0(\kappa_1)\} = (\alpha_1 - \alpha_2)q_0(\kappa_1) + \alpha_2 q_0(\tau)$ and $\int_{\kappa_1}^{\kappa_2} \tilde{q}_B(\tau)d\tau = (\alpha_1 - \alpha_2)(\kappa_2 - \kappa_1)q_0(\kappa_1) + \alpha_2[g\{q_0(\kappa_1)\} - g\{q_0(\kappa_2)\}]$. As a consequence

$$\int_{\kappa_0}^{\kappa_2} \tilde{q}_B(\tau)d\tau = (\alpha_1 - \alpha_2)(\kappa_2 - \kappa_1)q_0(\kappa_1) - (\alpha_1 - \alpha_2)g\{q_0(\kappa_1)\} - \alpha_2 g\{q_0(\kappa_2)\}.$$

More generally, if $\kappa_\ell < \tau \leq \kappa_{\ell+1}$ then

$$\tilde{q}_B(\tau) = \sum_{j=1}^{\ell} (\alpha_j - \alpha_{j+1})q_0(\kappa_j) + \alpha_{\ell+1} q_0(\tau),$$

that implies

$$\int_{\kappa_\ell}^{\kappa_{\ell+1}} \tilde{q}_B(\tau)d\tau = (\kappa_{\ell+1} - \kappa_\ell) \sum_{j=1}^{\ell} (\alpha_j - \alpha_{j+1})q_0(\kappa_j) + \alpha_{\ell+1}[g\{q_0(\kappa_\ell)\} - g\{q_0(\kappa_{\ell+1})\}]$$

In view of $\int_{\kappa_L}^{\kappa_{L+1}} q_0(\tau)d\tau = \alpha_{L+1} g\{q_0(\kappa_L)\}$ formula (13) follows from simple algebra. □

The conditions in the statement of Theorem 1 are satisfied, for instance, by $\Phi^{-1}(\tau)$ for which $\int_{\kappa_\ell}^{\kappa_{\ell+1}} \Phi^{-1}(\tau)d\tau = \phi(\kappa_\ell) - \phi(\kappa_{\ell+1})$ along with other distributions such as the t or the Frechet distributions.

Equation (13) allows to express θ_h as a linear function of the $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\ell)$ parameters. Note that $g\{q_0(\kappa_\ell)\}$ are known constants that depend on the chosen knots $\kappa_1, \ldots, \kappa_L$ and the baseline $q_0(\tau)$. In fact the terms in (13) can be re-arranged in order to express $\hat{\theta}_h$ directly as a linear combination of the $\boldsymbol{\alpha}$:

$$\hat{\theta}_h = \bar{\mathbf{x}}_h^t \boldsymbol{\alpha}_{0h} + \bar{\mathbf{x}}_h^t \Big\{ \sum_{\ell=1}^{L+1} \boldsymbol{\alpha}_\ell b_\ell \Big\} = \bar{\mathbf{x}}_h^t \Big\{ \boldsymbol{\alpha}_{0h} + \sum_{\ell=1}^{L+1} b_\ell \boldsymbol{\alpha}_\ell \Big\}, \tag{14}$$

with $b_\ell = g\{q_0(\kappa_{\ell-1})\} - g\{q_0(\kappa_\ell)\} + (1 - \kappa_\ell)q_0(\kappa_\ell) - (1 - \kappa_{\ell-1})q_0(\kappa_{\ell-1})$ provided that conventionally we assume $q_0(\kappa_{L+1}) = q_0(\kappa_0) = 0$.

A Bayes point predictor under quadratic loss for the small area mean $\bar{Y}_h$ can then be expressed as:

$$\begin{aligned} \hat{\bar{Y}}_h^{QR} &= E\Big[N_h^{-1} \Big[\sum_{i \in s_h} y_i + (N_h - n_h)\bar{\mathbf{x}}_{rh}^t \Big\{ \boldsymbol{\alpha}_{0h} + \sum_{\ell=1}^{L+1} b_\ell \boldsymbol{\alpha}_\ell \Big\} \Big] | \mathbf{d} \Big] \\ &= E\Big[N_h^{-1} \Big[\sum_{i \in s_h} y_i + (N_h - n_h)\hat{\bar{Y}}_{rh} \Big] | \mathbf{d} \Big], \end{aligned} \tag{15}$$

where $\mathbf{d}$ is a shortcut notation for the data, that is $\mathbf{d} = (\mathbf{y}_s, \mathbf{X}_s, \bar{\mathbf{x}}_h)$. Posterior standard deviation $sd(\hat{\bar{Y}}_h^{QR}|\mathbf{d})$ can consistently be used to evaluate uncertainty around $\hat{\bar{Y}}_h^{QR}$.

If the normal baseline is assumed, i.e. $q_0(\tau) = \Phi^{-1}(\tau)$ then (13) specializes to

$$\hat{\theta}_h|q_0 = \bar{\mathbf{x}}_h^t\boldsymbol{\alpha}_{0h} + \bar{\mathbf{x}}_h^t\Big[\sum_{\ell=1}^{L}(\boldsymbol{\alpha}_{\ell+1} - \boldsymbol{\alpha}_\ell)\Big\{\phi(\Phi^{-1}(\kappa_\ell)) - (1-\kappa_\ell)\Phi^{-1}(\kappa_\ell)\Big\}\Big], \tag{16}$$

as the function g coincides with the density ϕ in this case. If $\alpha_1 = \cdots = \alpha_{L+1}$, then $\theta_h = \bar{\mathbf{x}}_h^t\boldsymbol{\alpha}_{0h}$, that is the expectation does depend only on the location of the distribution and not on its shape. In this sense our representation of θ_h generalizes normal random effects models that are popular in small area estimation: if we have no evidence of departures from normality the model reduces to the nested error regression model (2).

4.3. Prior specification for the random effects and other hyperparameters

In the specification of the distribution of the random effects $\boldsymbol{\alpha}_{0h}$ we consider three alternatives of increasing complexity. The first is simply to assume independent normal distributions:

$$\alpha_{0hj} \sim N(\alpha_{0j}, \zeta_j^{-1}) \tag{17}$$

where ζ_j^{-1} are precision parameters. The second alternative is represented by

$$\alpha_{0hj} \sim t(\alpha_{0j}, g_j, \zeta_j^{-1}), \tag{18}$$

i.e. independent t-distributions with g_j degrees of freedom and ζ_j as before. We assume $g_j \sim Exp(0.1)$. In both (17) and (18) we assume $\zeta_j \sim Gamma(.01, .01)$, a standard assumption for the precision motivated by computational convenience. Diffuse independent normal priors are considered for parameters α_{0j}. The rationale for (18) is that of introducing some flexibility in the prior of α_{0hj} with respect to (17) while keeping computational convenience. Both these priors are not flexible enough to accommodate general configurations of the random effects and are not consistent with the non-parametric specification of the rest of the model.

Our third alternative consists in a non-parametric prior based on the Dirichlet process. Actually we specify a Dirichlet process prior, truncated for implementation purposes. We follow the lines of Ohlssen et al. (2007) and we adopt also their notation and their strategy in the specification of the priors for the hyperparameters:

$$\begin{aligned} \alpha_{0hj} &\sim TDP(\gamma_j, G_{0j}, T_{cj}) \\ G_{0j} &= N(\mu_{0j}, \sigma_{0j}^2) \end{aligned} \tag{19}$$

The acronym TDP stays for Truncated Dirichlet process. Ohlssen et al. (2007) discuss the choice of T_{cj} which is related to the prior chosen for γ_j. Consistently with their proposal we set $T_{cj} = 50$ that results adequate also in terms of measure of approximation to the full Dirichlet process and $\gamma_j \sim Uniform(0.3, 10)$. The hyper-location parameters μ_{0j} are given a diffuse normal prior: $\mu_{0j} \sim N(0, 100)$, while we assume $\sigma_{0j} \sim Uniform(0, 50)$. The prior (19) offers the advantage of being very flexible and it is in line with the non-parametric specification of the model. The main drawback is that is computationally demanding and its implementation in practical MCMC algorithm can slow down computations. In all cases priors for different α_{0hj} are assumed to be independent.

The consideration of alternative distributions for the random effects will allow us to make comparisons and checking the sensitivity of relevant posterior summaries to alternative specifications of the distribution assumed for the random effects.

For the $\boldsymbol{\alpha}_\ell$ ($\ell \geq 1$) we adopt the same latent variable approach suggested by Reich and Smith (2013) to guarantee the monotonicity of the quantile function. Specifically, the derivative of the quantile fuction $q_h(\tau|\mathbf{x}_i)$ is positive if and only if $\mathbf{x}_i^t\boldsymbol{\alpha}_\ell > 0, \forall i, \forall \ell \geq 1$. We assume that covariates are scaled so that $x_{ij} \in [-1, 1]$, $\forall i,\ j = 1, \ldots, p$; the quantity $wc(\boldsymbol{\alpha}_\ell) = \boldsymbol{\alpha}_{\ell 1} - \sum_{j=2}^{p+1} |\alpha_{\ell j}|$ defines the *worst case* for the positiveness of $\mathbf{x}_i^t\boldsymbol{\alpha}_\ell$, reached when either $x_{ij} = -1$ and $\alpha_{\ell j} > 0$ or $x_{ij} = 1$ and $\alpha_{\ell j} < 0$, $j > 1$. In line with Reich and Smith (2013) we build the prior by introducing a set of latent unconstrained parameters $\boldsymbol{\alpha}^\star = (\alpha^\star_{\ell 1}, \ldots, \alpha^\star_{\ell p})$ and set

$$\alpha_{\ell j} = \begin{cases} \alpha^\star_{\ell j} & wc(\boldsymbol{\alpha}^\star_\ell) > 0, \\ \\ \varepsilon \mathbf{1}(j = 1) & wc(\boldsymbol{\alpha}\star_\ell) \leq \varepsilon, \end{cases}$$

where $\varepsilon > 0$ is a small constant. Theorem 1 in Reich and Smith (2013) shows how these restrictions on $\boldsymbol{\alpha}_\ell$ are note severe and allow for a flexible prior. Differently from Reich and Smith (2013) we consider a simpler specification for $\alpha^\star_{\ell j}$: we select independent normal priors instead of multivariate normal priors with an auto-regressive variance-covariance structure controlled by an additional parameter ρ_j.

5. Simulation exercise

We now present a simulation exercise aimed at assessing the frequentist properties of the Bayesian predictors of the small area means discussed in Section 4. Quantile regression has been introduced in small area estimation as a distribution-free robust alternative to linear mixed models, thereby we focus on the ability of quantile regression based predictors to deal with outliers.

We consider the same simulation exercise presented in Chambers et al. (2014), Section 5, simply adding new predictors to those they study.

For each Monte Carlo iteration, population data are generated within area h ($h = 1, \ldots, m = 40$) according to the following model:

$$y_i = 100 + 5x_i + v_h + \epsilon_i, \quad i \in P_h \tag{20}$$

where sub-populations P_h have size $N_h = 100$. Samples are selected by simple random sampling without replacement within each area with $n_h = 5$. With respect to Chambers et al. (2014) the case $n_h = 15$ has not been considered. The auxiliary variable values are generated according to a log-normal distribution, i.e. $x_i \sim LogN(\mu = 1, \sigma = 0.5)$.

The random components v_h and e_i are generated according to four different scenarios:

Scenario $[0, 0]$, no outliers: $v \sim N(0, 3)$ and $\epsilon \sim N(0, 6)$;
Scenario $[e, 0]$, individual outliers only: $v \sim N(0, 3)$, $\epsilon \sim \delta N(0, 6) + (1 - \delta)N(20, 150)$ where $\delta \overset{ind}{\sim} Bernoulli(0.97)$;
Scenario $[0, v]$, area outliers only: $v \sim N(0, 3)$ for areas $h = 1, \ldots, 36$ and $v \sim N(9, 20)$ for areas $h = 37, \ldots, 40$; $\epsilon \sim N(0, 6)$;
Scenario $[e, v]$, outliers in both area and individual effects: $v \sim N(0, 3)$ for areas $h = 1, \ldots, 36$

and $v \sim N(9, 20)$ for areas $h = 37, \ldots, 40$; $\epsilon \sim \delta N(0, 6) + (1 - \delta)N(20, 150)$ where $\delta \overset{ind}{\sim}$ *Bernoulli*(0.97).

Note that when outliers are present at the individual level they contaminate each area; the outlier perturbation of area effects always concerns the same group of areas. This makes possible comparing the behaviour of estimators in area that are outliers and in those that are not. The outliers are introduced by means of additive perturbation of residuals. This choice is motivated by the creation of residuals' distribution skewed and heavy tailed, and not simply heavy tailed as it would result from scale mixtures.

As our methods are based on MCMC algorithm and therefore computationally demanding, we limit the number of Monte Carlo replications to $R = 250$. All codes are written in R, while for MCMC we use our own codes in `jags` called through the R package `rjags` (Plummer et al., 2016).

We consider models with $q_0(\tau) = \Phi^{-1}(\tau)$, $L = 3$ and $\kappa_\ell = 0.25, 0.5, 0.75$, which is also the default in the `BSquare` package (Smith and Reich, 2013). More complex models with $L = 5$ and $\kappa_\ell = 0.1, 0.25, 0.5, 0.75, 0.9$ will only be considered to illustrate how performances can be improved in situations where the residuals distribution exhibits heavy tails.

Posterior summaries are based on 10,000 samples, while the length of burn-in in set to 5,000. TPD priors (19) are assumed for the random effects, unless in cases where the use of normal or t priors (18) is explicitly mentioned. To speed up convergence we choose the initial values in the simulation in this way: we generate one sample from the simulation design, that we label 'iteration 0' as it is not considered in the subsequent analysis. For these data we run the model without the random effects using the `BSquare` package, (Smith and Reich, 2013) in order to get initial values for the α_ℓ, $\ell \geq 1$. We use these and randomly generated values for the remaining parameters to run our model. The posterior expectations obtained after convergence are then used to initialize the models for all the samples generated in the simulation.

We consider various predictors for comparison. In the first place, we consider the standard EBLUP predictor associated to the nested error regression model (2) that can play the role of benchmark because of its popularity, but that is known to be sensitive to outliers because of the normality assumption on the random components. We also consider the posterior mean of θ_h obtained from the Bayesian analysis of model (2) as it is implemented in `hbsae` package (Boonstra, 2012), with default choices for the priors. This implementation is essentially based on Datta and Ghosh (1991). We denote this predictor as HB-NER. We consider also the REBLUP associated to model (2) and defined according to the methodology of Sinha and Rao (2009), the M-quantile (MQ) and the bias-corrected M-quantile (MQ-BC) defined in (3) and (4), respectively. Other Bayesian predictors, and namely the one proposed by Chakraborty et al. (2018) quoted in the introduction, are not considered, as comparisons would be unfair: they do not allow for non-normal random effects and their assumption on the individual level errors do not encompass asymmetric contamination of scenarios $[e, 0]$ and $[e, v]$.

A sample code, along with data and initial values, written in the `R` language is available in the supplementary material associated to the paper.

We compare the predictors in terms of median values (across areas) of relative bias (RB) and

of relative root mean square error (RRMSE) defined for specific areas as

$$RB_h = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{\bar{Y}}^*_{hr} - \bar{Y}_{hr}}{\bar{Y}_{hr}}$$

$$RRMSE_h = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\bar{Y}}^*_{hr} - \bar{Y}_{hr})^2}}{\frac{1}{R}\sum_{r=1}^{R}\bar{Y}_{hr}},$$

where $\hat{\bar{Y}}^*_{hr}$ denotes the generic predictor being considered. Median values will be denoted as mRB and $mRRMSE$.

Table 1: Model-based simulation results: performances of predictors of small area means in scenarios [0,0] and [e,0].

Scenario	$[0,0]$		$[e,0]$	
Predictor	mRB	mRRMSE	mRB	mRRMSE
EBLUP	0.02	0.81	-0.02	1.22
HB-NER	0.01	0.82	-0.03	1.22
REBLUP	0.03	0.82	-0.39	1.01
MQ	0.02	0.82	-0.43	1.03
MQ-BC	0.02	0.91	-0.28	1.24
$\hat{\bar{Y}}^{QR}$	0.02	0.80	0.19	1.18
$\hat{\bar{Y}}^{QR}\|L=5$	0.02	0.83	0.02	0.98

Results in Table 1 show that in the $[0,0]$ scenario, all predictors perform similarly (with the only MQ-BC being slightly less efficient). Probability intervals calculated as the range between the 0.025 and 0.975 quantile of $\hat{\bar{Y}}^{QR}$ posterior distributions show a frequentist coverage between 0.94 and 0.96. We note how EBLUP and HB-NER perform very closely. This is not true only for this scenario, but also for all the remaining ones, so we will not insist on this point.

When outliers are introduced in the unit-level residuals (scenario $[e,0]$), all robust frequentist predictors are negatively biased. This could be expected, as they downweight outliers that are, according to the design of scenario $[e,0]$, all positive, thus improving efficiency at the price of some negative bias (see also Chambers et al., 2014). On the contrary, $\hat{\bar{Y}}^{QR}$ exhibits a positive bias: we have a scale parameter ruling the right tail of the distribution and the chosen threshold with $L = 3$ is the quantile 0.75 so that all the last quarter of the distribution is influenced by the outliers. To confine the impact of outliers only to the extreme of the right tail we can explore a model with more nodes. In fact, if we set $L = 5$ (and $\kappa_\ell = 0.1, 0.25, 0.5, 0.75, 0.9$) only the last decile is influenced by outliers and as result we have a much smaller bias and a predictor more efficient in terms of mRRMSE. As far as the frequentist coverage of posterior intervals in scenario $[e,0]$ are concerned it is close to 0.95 for $\hat{\bar{Y}}^{QR}|L = 5$, while between 0.88 and 0.92 when $L = 3$. We note that when $L = 5$ the Bayes predictor we propose is more efficient than all alternatives, with the exception of REBLUP (that is nonetheless more biased), so it seems that more nodes would lead to more efficient estimators. It has a price in terms of computational burden that may be unnecessary in scenarios less extreme than those considered in this simulation. A large L may also lead to poor estimation of shape parameters ruling the tails unless the sample is very large; in this case the computational burden can be heavy. In this exercise we do not explore the choice $L > 5$.

Table 2: Model-based simulation results: performances of predictors of small area means in scenarios $[0, v]$. (N), (t), (TDP) means that a Normal. a t or a Dirichlet process prior has been specified for the random effects, see Section 4.3 for details.

Areas	$[1-36]$	$[37-40]$	$[1-36]$	$[37-40]$
Predictor	mRB		$mRRMSE$	
EBLUP	0.10	-0.54	0.85	0.97
HB-NER	0.10	-0.50	0.88	1.00
REBLUP	0.11	-0.47	0.84	1.02
MQ	0.09	-0.94	0.83	1.46
MQ-BC	0.03	-0.07	0.92	0.86
$\hat{\bar{Y}}^{QR}$ (N)	0.09	-0.59	0.87	1.04
$\hat{\bar{Y}}^{QR}$ (t)	0.06	-0.33	0.83	0.94
$\hat{\bar{Y}}^{QR}$ (TDP)	0.07	-0.27	0.82	0.91

When we introduce perturbed random effects (scenario $[0, v]$, Table 2), we observe negative biases in the outlying areas as modified random effects have a positive mean and estimators are only partially able to accommodate it. For the Bayesian predictors it means that, despite the assumed prior is flexible, it is not flexible enough to fully account for the extreme outliers considered in the $[0, v]$ scenario. We also observe a positive median bias for the non-outlying areas because of the over-estimation of variance components due to a variance inflation effects caused by the presence of outliers. Only MQ-BC, that follows a different logic and is explicitly bias-adjusted, is very close to unbiasedeness (in median) in both groups of areas. For the Bayesian predictors, the amount of bias depends on the flexibility of the specified prior. In table 2 we compare Normal, t and TDP priors; clearly the more flexible the prior the smaller is the median bias observed in the outlying areas. Note that the improvement from Normal to t looks bigger than that from t to TDP; this is interesting especially in view of the fact that Normal and t have similar levels of computational complexity in MCMC algorithms, while TDP is computationally more demanding. In terms of $mRRMSE$, $\hat{\bar{Y}}^{QR}$ endowed with the TDP prior is more efficient than MQ-BC in non-outlying areas, while it is slightly less performing in the outlying ones. Frequentist coverage of posterior intervals reaches 0.95 for areas [1-36] and lies between 0.9 and 0.95 for the outlying ones. We observe that this slight undercoverage is quite usual when the posterior distribution is not centered on the actual value of the parameter.

Results for scenario $[e, v]$ are presented in Table 3. This scenario is quite extreme as we have big outliers at both the unit and the area level. For this complex data structure we found that predictors based on $L = 5$ perform significantly better than those based on $L = 3$ and is in this case the best predictor among those considered. In summary, the method we propose has the potential to improve the efficiency of many popular predictors in a variety of situations, with and without outliers contaminating the data.

6. An application to Italian income survey data

In this section we apply our methodology to the analysis of data from the 2013 Italian sample of the EU-SILC survey. This survey is conducted yearly across many European countries by the relevant National Institutes of Statistics using harmonized questionnaires and survey methodolo-

Table 3: Model-based simulation results: performances of predictors of small area means in scenarios $[e, v]$.

Areas	$[1-36]$	$[37-40]$	$[1-36]$	$[37-40]$
Predictor	*mRB*		*mRRMSE*	
EBLUP	0.17	-1.59	1.37	2.39
HB-NER	0.15	-1.51	1.38	2.33
REBLUP	-0.36	-1.00	0.99	1.44
MQ	-0.32	-0.99	1.01	1.57
MQ-BC	-0.26	-0.30	1.26	1.49
$\hat{\bar{Y}}^{QR}$	0.46	0.02	1.31	1.33
$\hat{\bar{Y}}^{QR}\|L=5$	0.34	-0.26	1.07	1.19

gies (Atkinson and Marlier, 2010, chapter 2). We consider a subset of these data focusing on people aged 60 or more and receiving old age or survivor benefits (in short, pension earners).

The target variable is the equivalized income, defined as the total disposable household income divided by an equivalence factor, in this case the so called modified-OECD equivalence factor (see Fusco et al., 2010, for more details). The goal of our estimation exercise is to estimate the average equivalized income of pension earners at the level of *local labour systems* also known as travel to work areas. We limit our attention to central Italy (i.e. region `ITE` according to the Nomenclature of Territorial Units for Statistics).

Our target population is divided into 113 areas. The overall sample size is 2,779 individuals. Area-specific sample sizes range from 0 (39 cases) to 496. For the areas with at least one sample household, the median sample size is equal to 19, so many of them are not large enough to provide estimates of average equivalized income of pension earners with an adequate precision. To evaluate the precision of the direct estimators, we follow the suggestions of Statistics Canada (2007) and we count the number of small areas with values of the coefficient of variation (CV) less than 16.6% (reliable for general use), between 16.6% and 33.3% (to be accompanied by a warning to users) and over 33.3% (unreliable). We have that only 32 areas are endowed with reliable estimates, 31 with estimates publishable with a warning and 11 completely unreliable. Moreover we have 39 areas for which we do not have any sample.

Pension income is a component of total individual income, thereby of total household income used in the definition of our target variable. The relationship between pension income and equivalized income for units in our sample is described in Figure 1. This relationship can be exploited in small area estimation as total (and average) pension income which is known at the municipal and thereby at the *local labour system* level from administrative (fiscal) archives. Nonetheless, from Figure 1 we get that this relationship is characterized by the presence of outlying observations and heteroskedastic residuals, a situation that can be dealt with using quantile regression.

To model the quantile function we consider a normal baseline, $q_0(\tau) = \Phi^{-1}(\tau)$ and compare models based on $L = 3$ ($\kappa = 0.25, 0.5, 0.75$) and $L = 5$ ($\kappa = 0.1, 0.25, 0.5, 0.75, 0.9$), along with different assumptions on the random effects.

We fit all models using the MCMC sampler `jags` (Plummer et al., 2016). Initial values for regression parameters are obtained by running the model (without random effects) using the `BSquare` package (Smith and Reich, 2013). We run a conservative burn-in of 10,000 draws and base our posterior summaries on a 40,000 MCMC sample.

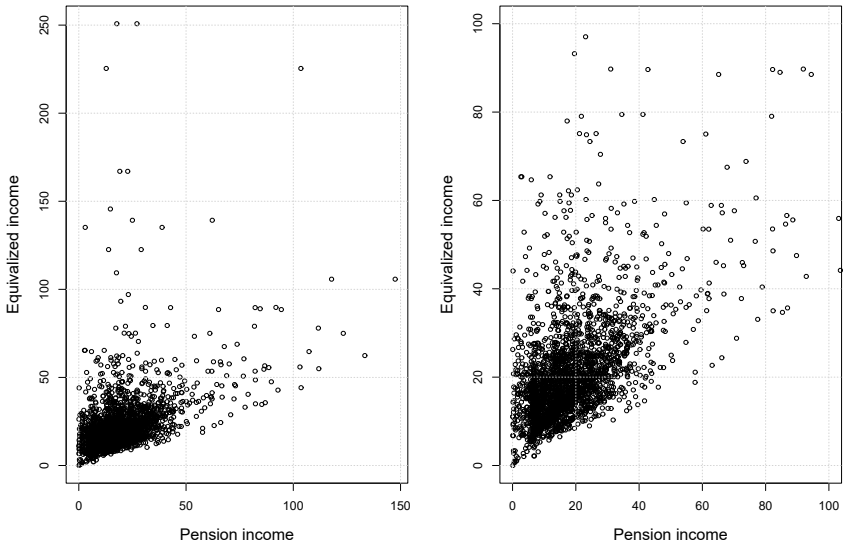

Figure 1: Pension income vs Equivalized income, values in thousend euros (left panel). Focus on values less than 100 on both axes (right panel).

We start with a TDP prior for the random effects. Regardless of the chosen L, the posterior distribution of the hyperparameter γ introduced in (19) is negatively skewed with a peak close to the upper limit of its domain. This indicates that the random effects distribution is close to the base distribution assumed in the TDP specification, in our case the normal (see Ohlssen et al., 2007). We compare models with either normal or t distribution for the random effects in terms of DIC (Spiegelhalter et al., 2002); it provides a slight preference in favour of the model with normal random effects. This provides evidence that there are no outlying random effects and assuming their normality is a tenable assumption. We also considered simplified models where only the intercepts are area-specific while the slopes are not, as described in (10), in both the $L = 3$ and $L = 5$ cases models with random effects on both intercepts and slopes exhibit a lower DIC.

Eventually, the model with $L = 5$ is to be preferred in terms of DIC to that with $L = 3$. This is in line with the evidence provided by the simulation exercise of Section 5 for the $[e, 0]$ scenario, characterized by unit level but not area level outliers, similarly to the data currently analyzed.

A limited sensitivity exercise on the priors for the hyper-parameters has been made. Specifically for the degrees of freedom of the t distribution we considered beside the exponential prior quoted in section 4.3, a $Gamma(2, 0.01)$ prior discussed in Juárez and Steel (2010), finding no appreciable impact on the posterior distributions of parameters relevant to this analysis and namely the α parameters. Similarly we modified the parameters of the Gamma distributions used for precisions considering $Gamma(1, 1)$, $Gamma(1, 0.1)$ and $Gamma(.1, .1)$ alternatives without finding these choices to have any impact on the posterior of the αs. Eventually, we did a limited sensitivity exercise for γ_j parameter in TPP prior specification. Also in this case we did not find any relevant effect of these modifications. In Figure 2 (left panel) the posterior expectation of

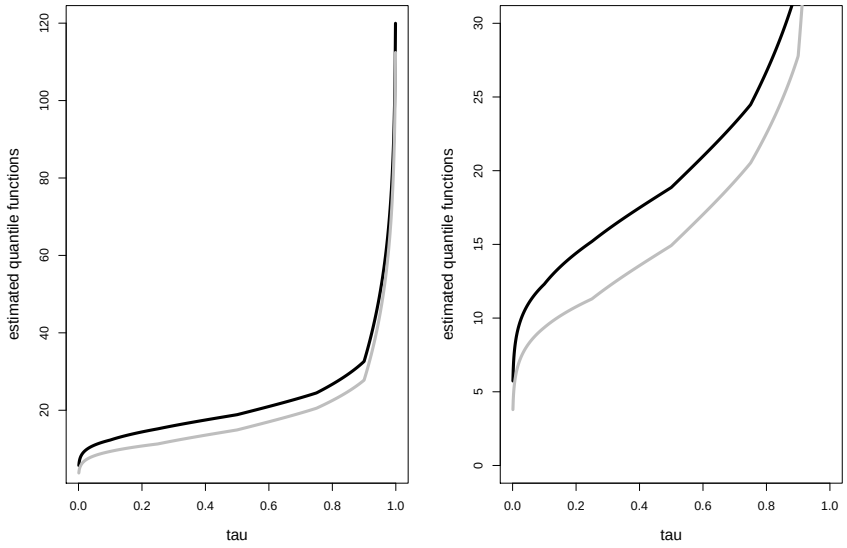

Figure 2: Left panel: Estimated quantile functions for Rome (black) and Porto Sant'Elpidio (red). Right panel: same plot but range on y axis limited.

quantile function $E(\hat{q}_h(\tau)|\mathbf{d})$ is plotted for two different areas: Rome (which is the richest on average according to direct estimates) and Porto Sant'Elpidio (a rural area near the Adriatic coast), the poorest. It is apparent how the model is flexible enough to accommodate the skew distribution of the data despite the adopted normal baseline. Area-specific intercepts move the curve for Rome up to stochastically dominate the quantile function estimated for Porto Sant'Elpidio. The differences between the two curves can be better evaluated from the right panel, where the influence of the $\bar{\mathbf{x}}_h$ on the $\boldsymbol{\alpha}$ parameters lead to a steeper curve for Porto Sant'Elpidio right of $\tau = 0.75$, as we can expect for an area where pensions are more concentrated on lower quantiles.

Small area estimators can be often thought as composite estimators that complement the information from the area-specific sample with that provided by all other areas in the sample. This is true also for the method that we propose. In Figure 3 we compare the empirical quantile functions of two areas with the quantile function $E(\hat{q}_h(\tau)|\mathbf{d})$ estimated from the model. In the left panel we show the situation for the Local Labour system of Florence. Although the area-specific sample is adequate ($n_h = 127$) the empirical quantile function exhibits a short tail; the observed maximum equivalized income in Florence is less than half of the maximum in Rome, which does not make much sense in view of our prior knowledge of the problem and can be attributed to a sample effect. The model-based estimated right tail of the distribution is thereby more realistic. In the right panel of Figure 3 we show the results for the Local Labour system of Jesi (a middle sized town in the Marche administrative region). The sample here is smaller ($n_h = 79$) but the presence of an observation in the highest percentile of the (overall) sample distribution has an impact on the empirical quantile function. The model based estimated quantile function moderates the impact of the outlier.

We now turn our attention to the estimation of small area means. Although integrating the

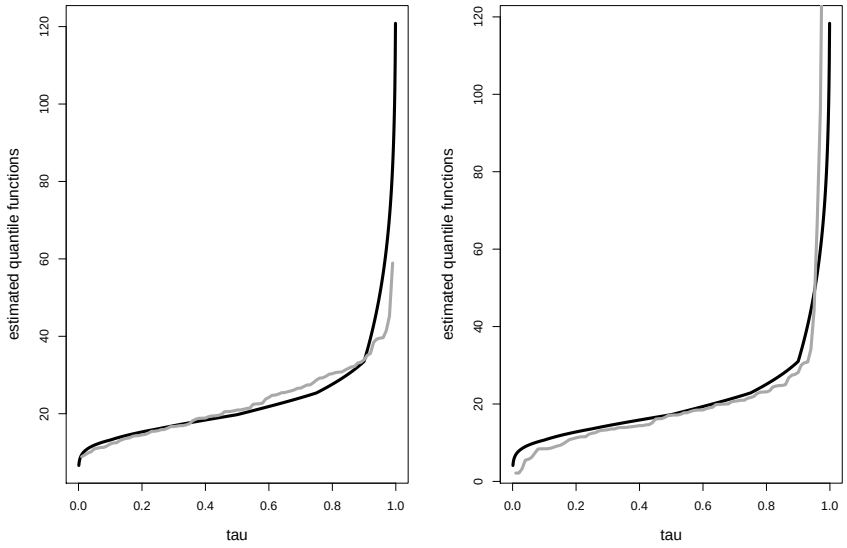

Figure 3: Empirical quantile function (grey) versus model based estimated quantile function (black). Left panel: Florence. Right panel: Jesi

whole quantile function is liable to outlier influence, we note that these shape parameters $\boldsymbol{\alpha}_\ell$ are estimated on the whole sample; moreover the influence of outliers is limited to just one of the shape parameters. Figure 4 shows the difference between direct and model based estimators, that diminishes as the sample size grows. We limited the x-axis to 150 to show more clearly what happens with small and medium sample sizes.

If we compare the posterior standard deviations $sd(\hat{\bar{Y}}_h^{QR}|\mathbf{d})$, to the standard deviations of direct estimators (i.e. sample means) we find that they are much smaller: 66% smaller in median and 67% in mean; we find similar numbers when comparing coefficient of variations. The posterior standard deviations $sd(\hat{\bar{Y}}_h^{QR}|\mathbf{d})$ are also much smaller than those associated to the HB-NER estimator (see section 5) that reduce the coefficient of variation by 20% in median; a poor performance due to the failure of the normality assumption that lead to over-estimation of the random effects variance. Estimates for out-of-sample areas can be straightforwardly be obtained using the formulas of our predictors. With reference to the Statistics Canada (2007) guidelines all estimates obtained using the small area method can be labelled as reliable.

We finally test whether the model-based and direct estimates proposed have the same expected value using the approach by Brown et al. (2001) who considered model-based estimates as unbiased predictors of direct estimates and proposed a goodness-of-fit diagnostic based on the null hypothesis that the model-based estimates are equal to the expected values of the direct estimates. In particular, the following Wald statistic, distributed according to a $\chi^2(m)$ under the null hypothesis is proposed:

$$W = \sum_{h=1}^{H} \frac{(\hat{\bar{Y}}_h - \hat{\bar{Y}}_h^{QR})^2}{Var(\hat{\bar{Y}}_h) + V(\hat{\bar{Y}}_h^{QR}|\mathbf{d})} \tag{21}$$

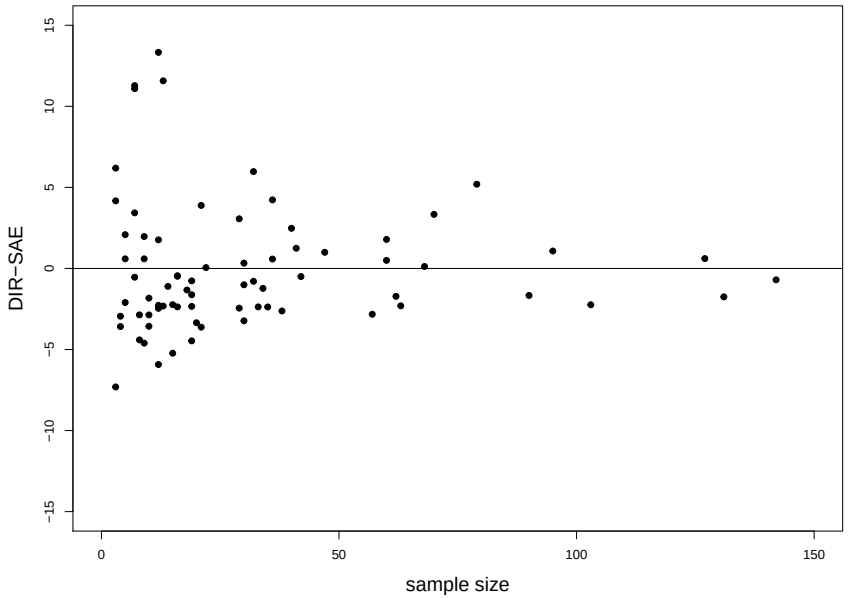

Figure 4: Difference between the direct estimator and $\hat{\bar{Y}}_h^{QR}$ plotted against the sample size (up to 150).

where $Var(\hat{\bar{Y}}_h)$ denotes the sampling variance of $\hat{\bar{Y}}_h$. The p-value is in our case 0.139, so we do not have evidence against the null hypothesis. As the procedure proposed by Brown et al. (2001) assumes that the sample sizes in the small areas are sufficient to justify central limit assumptions, we repeat the calculation on the 36 areas with sample size at least 20: the p-value is in this case 0.331.

7. Conclusions

In this article we introduce Bayesian analysis of quantile regression models in the context of small area estimation. Our distributional assumptions are very flexible but we keep normality, that often plays a central role in small area estimation, as a special case.

With the proposed method, we obtain an estimate of the whole quantile function at the area level. The method can be applied to the estimation of quantities defined as functionals of the quantiles. In this research we focused on the mean, that is the most common target of small area estimation. The estimator of the mean assumes a simple, interpretable form. We compare the predictors we propose with those based on quantile regression under a frequentist approach and we find that ours are at least as performing in terms of bias and mean square error under different scenarios. We also find that probability intervals based on posterior quantiles have frequentist coverages close to their nominal levels. Our research can be extended to other functionals of the quantiles as concentration indexes, headcount ratios, quintile share ratios that are used in the analysis of the income distributions.

Posterior distributions are explored by means of MCMC algorithms that can be implemented using standard softwares, such as `jags` we consider in our application. Nonetheless it is computationally demanding, especially when the sample size is large. For this reason we considered only solution based on relatively simple specifications, based on a limited number of nodes. More efficient, faster algorithms represent an area for possible future research.

Acknowledgements

The work of Nicola Salvati has been carried out with the support of project InGRID 2 (grant agreement 730998, EU) and of project PRA2018-9 ('Fromsurvey-based to register-based statistics: a paradigm shift using latent variable models').

References

Atkinson, A.B., Marlier, E. (2010), *Income and living conditions in Europe*, Eurostat Statistical books, Publication Office of the European Union, Luxembourg.

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

Boonstra, H.J. (2012), `hbsae`: Hierarchical Bayesian Small Area Estimation, R package version 1.0., `URL: https://CRAN.R-project.org/package=hbsae`.

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

Bianchi, A., Fabrizi, E., Salvati, N., Tzavidis, N.(2018), Estimation and Testing in M-quantile regression with applications to small area estimation, *International Statistical Review*, 86, 541–570.

Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001) Evaluation of small area estimation methods. An application to unemployment estimates from the UK LFS. In *Proc. Symp.Achieving Data Quality in a Statistical Agency: a Methodological Perspective*. Ottawa: Statistics Canada.

Chakraborty, A., Datta, G.S., Mandal, A. (2018), Robust hierarchical Bayes small area estimation for nested error linear egression model, `arXiv:1702.05832 [stat.ME]`

Chambers, R., Tzavidis, N. (2006), M-quantile models for small area estimation, *Biometrika*, 76, 47–69.

Chambers, R., Chandra, H., Salvati, N., Tzavidis, N. (2014), Outlier robust small area estimation, *Journal of the Royal Statistical Society*, B 93, 255–268.

Datta, G.S., Ghosh, M. (1991), Bayesian prediction in linear models: applications to small area estimation, *Annals of Statistics*, 19, 1748–1770.

Ferguson, T. S. (1973), A Bayesian analysis of some nonparametric problems, *The Annals of Statistics*, 1, 209–230.

Fusco, A., Guio, A.C., Marlier, E. (2010), Chracterizing the income poor and the materially deprived in European countries, in *Eurostat Statistics books: Income and living conditions in Europe,* (Atkinson A.B. and Marlier E. (eds.), Publication Office of the European Union, Luxembourg.

Geraci, M., Bottai, M. (2014) Linear Quantile Mixed Models, *Statistics and Computing*, 24, 461–479.

Koenker, R. (2004), Quantile regression for longitudinal data, *Journal of Multivariate Analysis*, 91, 74–89.

Juárez, M.A., Steel, M.F.J. (2010) Model-Based Clustering of Non-Gaussian Panel Data Based on Skew-t Distributions, *Journal of Business and Economic Statistics*, 28, 52-66.

Ohlssen, D.I., Sharples, L.D. and Spiegelhalter, D.J. (2007) Flexible random-effects models using Bayesian semiparametric models: application to institutional comparisons, *Statistics in Medicine*, 26, 2088-2112.

Kottas, A., Krnjajic, M. (2009) Bayesian semiparametric modelling in quantile regression, *Scandinavian Journal of Statistics*, 36, 297–319.

Plummer, M., Stukalov, A., Denwood, M. (2016). `rjags` (R package version 4-6, pp. 1–19). Vienna, Austria: The Comprehensive R Archive Network

Rao, J.N.K., Molina, I. (2015), Small area estimation, Wiley Series in Survey Methodology, New York.

Reich, B. J., Bondel, H. D. and Wang, H. J. (2010), Flexible Bayesian quantile regression for independent and clustered data, *Biostatistics*, 11, 337–352.

Reich, B. J., Fuentes, M. and Dunson D. B. (2011), Bayesian spatial quantile regression, *Journal of the American Statistical Association*, 106, 6–20.

Reich, B. J. (2012), Spatiotemporal quantile regression for detecting distributional changes in environmental processes, Journal of the Royal Statistical Society: Series C 64, 535–553.

Reich, B.J., Smith, L.B. (2013), Bayesian quantile regression for censored data, *Biometrics*, 69, 651–660.

Sinha, S. K. and Rao, J. N. K. (2009) Robust small area estimation, *Canadian Journal of Statistics*, 37, 381–399.

Smith, L.B., Reich, B.J. (2013) `BSquare`: An `R` package for Bayesian simultaneous quantile regression, North Carolina State University; 2013. Available from: `http://www4.stat.ncsu.edu/ reich/QR/BSquare.pdf`

Smith, L.B., Fuentes, M., Gordon-Larsen, P., Reich B.J., (2015), Quantile regression for mixed models with application to examine blood pressure trends in China, *Annals of Applied Statistics*, 9, 1226–1246.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society*, Series B 64, 583–639.

Statistics Canada (2007) 2005 Survey of Financial Security - Public Use Microdata File, User Guide. Published by authority of the Minister responsible for Statistics Canada, `http://www.statcan.gc.ca/pub/13f0026m/13f0026m2007001-eng.htm`.

Taddy, M.A., Kottas, A. (2010), A Bayesian Nonparametric Approach to Inference for Quantile Regression, *Journal of the American Statistical Association*, 28, 357–369.

Tokdar, S. T., Kadane, J.B. (2012), Simultaneous linear quantile regression: a semiparametric Bayesian approach, *Bayesian Analysis*, **7**, 51–72.

Wichitaksorn, N., Choy, S.T.B., Gerlach, R. (2014), A generalized class of skew distributions and associated robust quantile regression models, *The Canadian Journal of Statistics*, 42, 579–596.

Yang, Y., Wang, H.J. and He, X. (2015), Posterior Inference in Bayesian Quantile Regression with Asymmetric Laplace Likelihood, *International Statistical Review*, 84, 327–344.

Yang, Y., Tokdar, S.(2017), Joint estimation of quantile planes over arbitrary predictor space, Cornell University Library *Journal of the American Statistical Association*, 112, 1107–1120.

Yue, Y.R., Rue, H. (2011), Bayesian inference for additive mixed quantile regression model, *Computational Statistics and Data Analysis*, 55, 84–96.

Yu, K. and Moyeed, R.A. (2001), Bayesian quantile regression, Statistics and Probability Letters, 54, 437–447.

Appendix

We report the definition $B_\ell(\tau)$ introduced in (8). It is presented also in Reich and Smith (2013) but our notation is slightly different. We restrict to the case $L = 3$ (generalization to larger L is trivial):

- if $\tau \leq \kappa_1$ then

$$\begin{aligned} B_1(\tau) &= q_0(\tau) \\ B_{\ell>1}(\tau) &= 0 \end{aligned}$$

- if $\kappa_1 < \tau \leq \kappa_2$ then

$$\begin{aligned} B_1(\tau) &= q_0(\kappa_1) \\ B_2(\tau) &= q_0(\tau) - q_0(\kappa_1) \\ B_{\ell>2}(\tau) &= 0 \end{aligned}$$

- if $\kappa_2 < \tau \leq \kappa_3$ then

$$\begin{aligned} B_1(\tau) &= q_0(\kappa_1) \\ B_2(\tau) &= q_0(\kappa_2) - q_0(\kappa_1) \\ B_3(\tau) &= q_0(\tau) - q_0(\kappa_2) \end{aligned}$$

- if $\tau > \kappa_3$ then

$$\begin{aligned} B_1(\tau) &= q_0(\kappa_1) \\ B_2(\tau) &= q_0(\kappa_2) - q_0(\kappa_1) \\ B_3(\tau) &= q_0(\tau) + \{q_0(\kappa_3) - q_0(\kappa_2)\} \end{aligned}$$