# PARAMETRIC MODELLING OF M-QUANTILE REGRESSION COEFFICIENT FUNCTIONS WITH APPLICATION TO SMALL AREA ESTIMATION

Paolo Frumento and Nicola Salvati

*Karolinska Institutet, Institute of Environmental Medicine, Unit of Biostatistics*
*University of Pisa, Department of Economics and Management*

*Abstract:* We describe how M-quantile regression coefficients, $\boldsymbol{\beta}(\tau)$, can be modelled as (flexible) parametric functions of $\tau$. This approach is referred to as *M-quantile regression coefficients modelling* (MQRCM), and is implemented in the R package `Mqrcm` that accompanies the paper. We illustrate the advantages of this method and suggest a range of feasible modelling strategies. We introduce an estimator, describe its asymptotic properties, and present simulation results. The proposed approach is combined with small area estimation (SAE) techniques to estimate the average equivalised household income using the EU-SILC data.

*Key words and phrases:* Equivalised income, EU-SILC data, M-quantile function, R package `Mqrcm`.

## 1 Introduction

A central objective of the Horizon 2020 programme is the promotion of innovative and inclusive societies in Europe and worldwide. To reach this goal, it is important to investigate how social cohesion, solidarity and reconciliation of differences between social groups or individuals can be achieved. Measuring and analysing social and economic inequalities, which have multiple dimensions including income, wealth, employment, health, environment and wellbeing, is of fundamental importance to provide the necessary information to policymakers and stakeholders in general.

Information at disaggregated geographical level is of particular interest in this context, and can be obtained from data collected in national surveys. However, sample sizes within small domains or areas are frequently too small to estimates a

parameter of interest with a sufficient precision. The demand for reliable statistics for small areas has promoted the development of small area estimation (SAE) methods.

A general discussion of different techniques for small area estimation is found in Rao and Molina (2015). A commonly used model-based approach is that of fitting a linear mixed model in which between-area variation is accounted for by area-level effects. An alternative solution, described by Chambers and Tzavidis (2006), is to implement small area estimation using M-quantile regression (MQR) models. This method permits to obtain outlier-robust estimators without making parametric assumptions, using the general theory of M-estimation. The distinguishing features of this approach include the protection that a careful choice of a quantile-specific loss function offers against the effect of outliers, and the characterisation of domain heterogeneity in terms of domain-specific M-quantiles. Estimation of small area means based on M-quantile regression is considered a standard approach (Tzavidis et al., 2010; Salvati et al., 2012; Chambers et al., 2014; Rao and Molina, 2015) and has been used in numerous applications (e.g., Tzavidis et al., 2008; Pratesi et al., 2008; Giusti et al., 2012; Fabrizi et al., 2014). For a complete review of M-quantile regression models in SAE see Bianchi et al. (2018).

In this paper we introduce a new estimator of M-quantiles, in which the regression coefficients, $\boldsymbol{\beta}(\tau)$, are modelled as parametric functions of $\tau$. This approach is referred to as *M-quantile regression coefficients modelling*, and is related to the existing literature on quantile regression (Frumento and Bottai, 2016, 2017) .

The idea of describing the M-quantile function by a parametric model presents significant advantages over standard MQR, in which different M-quantiles are estimated one at a time. First, the proposed method is not grid-based, as it does not require selecting an arbitrary grid $\{\tau_1, \ldots, \tau_r\}$ at which to estimate M-quantile regression coefficients. Second, parametric models simplify summarising and interpreting the results, and typically generate more efficient estimators. Third, imposing a parametric structure can stabilise the behaviour of the estimated regression coefficients, especially in the tails, and alleviate the M-quantile crossing problem, occurring when the fitted M-quantile function is not monotonically

non-decreasing.

We apply the proposed methodology to analyse data from the 2006 European Survey on Income and Living Conditions (EU-SILC), using small area estimation techniques to combine the survey data with those of the population Census 2001 and estimate the average equivalised household income of the local labour systems (LLSs) of three large administrative regions in Italy, namely Lombardia (northern Italy), Toscana (central Italy) and Campania (southern Italy). The goal of the research is to investigate both the within-region variability, and the so called "north-south" divide characterising the Italian territory.

The paper is structured as follows. In Section 2 we review the existing methods for M-quantile regression and their use in small area estimation. In Section 3 we present the 2006 EU-SILC and Census 2001 data that are used to estimate the average equivalised household income. In Section 4 we introduce a parametric approach to M-quantile regression and provide general guidelines for model building. We describe the estimator in Section 5, and discuss its asymptotic properties in Section 6. In Section 7 we present simulation results, and Section 8 demonstrates the properties of the proposed procedure presenting the application that motivated this research. Finally, in Section 9 we summarize the main findings of the paper and discuss future research aimed at outlier robust small area inference.

# 2 An overview of M-quantile regression (MQR) models and their application to small area estimation

## 2.1 Linear M-quantile regression

Through the paper, we denote by $Y_i$ a response variable of interest, and by $\boldsymbol{x}_i$ a $q$-dimensional vector of observed covariates, $i = 1, \ldots, n$. Following standard notation, we assume that

$$M(\tau \mid \boldsymbol{x}_i) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(\tau) \tag{1}$$

is the $\tau$-th M-quantile of $Y_i$ given $\boldsymbol{x}_i$, $\tau \in (0,1)$. The $\tau$-th MQR coefficients, $\hat{\boldsymbol{\beta}}(\tau)$, minimise

$$L_n(\boldsymbol{\beta}(\tau), \sigma(\tau)) = n^{-1}\Big[n\log(\sigma(\tau)) + \sum_{i=1}^{n} \rho_\tau\left(\frac{y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau)}{\sigma(\tau)}\right)\Big], \tag{2}$$

where $y_i$ is a realisation from $Y_i$, $\rho_\tau(u) = |\tau - I(u<0)|\rho(u)$ is the tilted version of a loss function $\rho(u)$, and $\sigma(\tau)$ is a nuisance scale parameter. The quantity expressed in (2) corresponds, up to an additive constant, to the negative log-likelihood of a Generalised Asymmetric Least Informative (GALI) distribution (Bianchi et al., 2018). This model is not assumed to reflect the true data distribution, but provides a unified framework for joint estimation of $\boldsymbol{\beta}(\tau)$ and $\sigma(\tau)$. Minimising (2) with respect to $\boldsymbol{\beta}(\tau)$ and $\sigma(\tau)$ requires solving

$$\sum_{i=1}^{n} \psi_\tau\left(\frac{y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau)}{\hat{\sigma}(\tau)}\right)\boldsymbol{x}_i \;=\; \mathbf{0} \tag{3}$$

$$-\frac{n}{\sigma(\tau)} + \frac{1}{\sigma^2(\tau)}\sum_{i=1}^{n} \psi_\tau\Big(\frac{y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau)}{\sigma(\tau)}\Big)(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau)) \;=\; 0, \tag{4}$$

where $\psi_\tau(u) = \mathrm{d}\rho_\tau(u)/\mathrm{d}u$ is an influence function.

A popular choice of $\rho_\tau$ is the tilted version of Huber's loss function,

$$\rho_\tau(u) = 2\begin{cases} (c|u| - c^2/2)|\tau - I(u \le 0)| & |u| > c \\ u^2/2|\tau - I(u \le 0)| & |u| \le c, \end{cases} \tag{5}$$

where $I(\cdot)$ is an indicator function and $c$ is a cutoff constant. The Least Informative (LI) distribution described in Huber (1981, Section 4.5) is based on this loss function, with $\tau = 0.5$. Depending on the choice of $\rho(\cdot)$, MQR models may reduce to ordinary quantiles regression ($\rho(u) = |u|$) or to expectiles regression ($\rho(u) = u^2$) while other choices are also possible (Dodge and Jureckova, 2000). However, quantiles and expectiles should be treated separately due to different properties of the corresponding influence functions (Wooldridge, 2010, p. 407).

## 2.2 Use of M-quantile regression in small area estimation

Using standard notation for SAE, we denote by $U$ a population of size $N$ divided into $m$ non-overlapping subsets $U_j$ (small areas) of size $N_j$, $j = 1, ..., m$. Consistently, we denote by $y_{ij}$ and $\boldsymbol{x}_{ij}$ the response and the covariates of the $i$-th unit of area $j$, $i = 1, \ldots, N_j$.

Suppose that the quantities of interest are the area means, $\bar{y}_j = N_j^{-1} \sum_{i \in U_j} y_{ij}$, and assume to draw a random sample $s \subset U$ of population units, such that area-specific samples $s_j \subset U_j$ of size $n_j \geq 0$ are available for each area. SAE methods are used when the $y_{ij}$ values are only available for the units belonging to the set $s_j$, while the area-level means of the covariates, $\bar{\boldsymbol{x}}_j = N_j^{-1} \sum_{i \in U_j} \boldsymbol{x}_{ij}$, are known from external sources.

Estimation of small area means is implemented as follows. First, MQR is applied to the sampled units, allowing to compute the "M-quantile coefficients" $\hat{\tau}_{ij}$ such that $\boldsymbol{x}_{ij}^{\mathrm{T}} \hat{\boldsymbol{\beta}}(\tau_{ij}) = y_{ij}$, $i \in s_j$, $j = 1, \ldots, m$. If a hierarchical structure does explain part of the variability in the population data, units within areas are expected to have similar M-quantile coefficients. Then, an estimate of the mean M-quantile coefficient for area $j$ is obtained as $\hat{\bar{\tau}}_j = n_j^{-1} \sum_{i=1}^{n_j} \hat{\tau}_{ij}$. Finally, an estimator of $\bar{y}_j$ is given by

$$\hat{\bar{y}}_j^{\mathrm{MQR/Naive}} = N_j^{-1} \Big[ \sum_{i \in s_j} y_{ij} + \sum_{i \in \bar{s}_j} \boldsymbol{x}_{ij}^{\mathrm{T}} \hat{\boldsymbol{\beta}}(\hat{\bar{\tau}}_j) \Big], \tag{6}$$

where $\bar{s}_j$ is the set of non-sampled units in area $j$. The above estimator, introduced by Chambers and Tzavidis (2006), uses the linear M-quantile regression model described in (1) and assumes $\boldsymbol{\beta}(\tau)$ to be a "sufficiently smooth" function of $\tau$. If $n_j = 0$, $\hat{\bar{\tau}}_j = 0.5$ is used and (6) reduces to a synthetic estimators based on M-median regression.

Chambers et al. (2014) defined such method as *robust-projective* as it projects sample non-outlier (i.e., working model) behaviour onto the non-sampled part of the survey population. They also proposed a method that allows for contributions from representative sample outliers. Their method is said to be *robust-predictive* since it attempts to predict the contribution of the population outliers to the population quantity of interest. In the robust-predictive context, a bias-corrected version of estimator (6) is given by

$$\hat{\bar{y}}_j^{\mathrm{MQR/BC}} = N_j^{-1} \Big\{ \sum_{i \in s_j} y_{ij} + \sum_{i \in \bar{s}_j} \boldsymbol{x}_{ij}^{\mathrm{T}} \hat{\boldsymbol{\beta}}(\hat{\bar{\tau}}_j) + \frac{N_j - n_j}{n_j} \sum_{i \in s_j} \hat{\omega}_{ij} \phi \Big\{ \frac{y_{ij} - \boldsymbol{x}_{ij}^{\mathrm{T}} \hat{\boldsymbol{\beta}}(\hat{\bar{\tau}}_j)}{\hat{\omega}_{ij}} \Big\} \Big\}, \tag{7}$$

where $\hat{\omega}_{ij}$ is a robust estimator of the scale of the residual $y_{ij} - \boldsymbol{x}_{ij}^{\mathrm{T}} \hat{\boldsymbol{\beta}}(\hat{\bar{\tau}}_j)$ in area $j$. The robust influence function $\psi$, used to define $\hat{\boldsymbol{\beta}}(\hat{\bar{\tau}}_j)$, is replaced in the third addend of (7) by a new function $\phi$; such function is still bounded, but

more accommodating with respect to sample outliers, as $|\psi| \leq |\phi|$. 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 (7) treat sample outliers as non-representative (Chambers et al., 2014). A method to estimate the mean squared error (MSE) of MQR-based robust predictors of small area means under the robust-projective and robust-predictive approaches has been proposed by Chambers et al. (2014). Their approach uses first-order approximations to the variances of solutions of estimating equations to develop conditional MSE estimators for predictors (6) and (7).

## 3 Italian Census 2001 and 2006 EU-SILC data

In the EU-SILC data, the regional samples are based on a stratified two-stage sample design, in which municipalities are the Primary Sampling Units (PSUs), and households are the Secondary Sampling Units (SSUs). The PSUs are divided into strata according to their dimension in terms of population size; the SSUs are selected by means of systematic sampling in each PSU.

The goal of our study is to estimate the average equivalised income for the local labour systems (LLSs) of Lombardia, in the north, Toscana, in central Italy, and Campania, in southern Italy. LLSs refer to 611 unplanned domains obtained as clusters of municipalities where the bulk of the labour force lives and works, and are defined on a functional basis, the key criterion being the proportion of commuters who cross the LLS boundary on the way to their workplace. According to the official EU nomenclature of local units, LLSs are intermediate between LAU 1 and LAU 2 levels (Eurostat, 2016). The choice of the LLSs of these three regions, out of the 20 existing in Italy, is motivated by the geographical differences characterising the Italian territory. In particular, the selected regions can be considered as representative of Northern, Central and Southern/Insular Italy, respectively, and can be used to investigate the so-called "north-south" divide.

The data used in this paper come from the 2006 wave of EU-SILC and the 2001 Population Census. The response variable of interest is the equivalised income, which is obtained by dividing the total disposable household income by a factor that takes into account the size and composition of the household. This

factor is computed using the modified OECD scale (Hagenaars et al., 1994).

The target small areas are 172 in total: 59 in Lombardia (34 sampled and 25 out-of-sample areas), 57 in Toscana (32 sampled and 25 out-of-sample) and 56 in Campania (18 sampled and 38 out-of-sample). Among the observed LLSs, the sample size ranges between 13 and 261. The mean value is 64.3, while quartiles are 23, 35, and 59, respectively. Figure 1 shows the distribution of the LLSs by sample size.

The presence of numerous LLSs with a very small sample size makes it difficult to obtain reliable estimates at the area level, and motivates the use of SAE techniques. Covariate information at the population level is obtained from the Census 2001 data. A description of the variables from both the EU-SILC 2006 survey and the Census 2001 is reported in Table 1. Some variables refer to the head of the household (HH), while others are measured at the household level.

The use of lagged Census information may lead to bias in small area estimators. However, the variables whose totals are known from the Census have been proven to be powerful predictors of household income according to tests conducted within the SAMPLE program (Small Area Methods for Poverty and Living Conditions, `http://www.sample-project.eu/`, SAMPLE (2010)). Moreover, the impact of the time lag is limited, because the area-level means of the considered auxiliary variables evolve slowly over time.

Fabrizi et al. (2014) presented some preliminary diagnostics obtained by fitting a linear mixed model with households at the first level, LLSs at the second level, equivalised household income as the response variable, and the variables in Table 1 used as covariates. They suggested that the hypotheses of normality used by linear mixed models may not hold, and highlighted the presence of outlying values. In this situation, small area methods based on M-quantiles are likely to generate more reliable estimators.

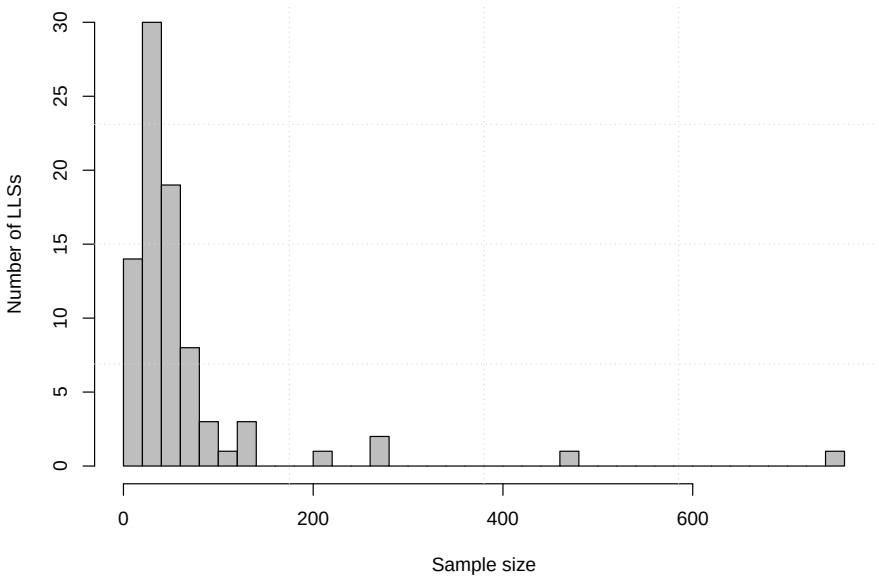


Figure 1: Distribution of the available sample size (88 areas with no observations are not represented).

# 4 Modelling M-quantile regression coefficient functions

Standard M-quantile regression is nonparametric in the sense that estimation is carried out separately for each value of $\tau$, treating $\boldsymbol{\beta}(\tau)$ and $\sigma(\tau)$ as infinite-dimensional parameters. Typically, the estimated coefficient functions $\hat{\boldsymbol{\beta}}(\tau)$ are summarised graphically. An example is provided in Figure 2, which represents the coefficients associated with gender of the HH (reference = female) and the occupational status of the HH (reference = other) in the EU-SILC data, computed by running M-quantile regression at $\tau = (0.005, 0.01, \ldots, 0.995)$ with Huber's loss function ($c = 1.345$).

Frequently, the coefficient functions have a rather simple behaviour which can be conveniently described by a mathematical formula or an expression in natural language (e.g., a straigth line, a parabola, a J-shape). For example,

Table 1: Description of variables available from EU-SILC 2006 and Population Census 2001.

| Variable Name | Description |
|---|---|
| Ownership of the house | Two levels: Owner or free accommodation (reference) / Other |
| Age of the HH | Continuous |
| Occupational status of the HH | Two levels: Working / Other (reference) |
| Gender of the HH | Two levels: Male / Female (reference) |
| Years in education of the HH | Continuous |
| Household size | Continuous |
| Region | Three levels: Campania, Lombardia, Toscana (reference) |

visual interpolation of the coefficient functions displayed in Figure 2 suggests that in both situations we could use a linear model, say $\beta(\tau \mid \boldsymbol{\theta}) = \theta_0 + \theta_1\tau$, to provide a good fit with only two parameters. Also, some volatility and very large standard errors are observed in the tails of the coefficient functions, as $\tau$ approaches 0 or 1. This is usually due to data sparsity, and could be alleviated by introducing some structural assumptions on the functional form of $\boldsymbol{\beta}(\tau)$.

We assume model (1) to hold,

$$M(\tau \mid \boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}(\tau),$$

and introduce two finite-dimensional parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ such that

$$\boldsymbol{\beta}(\tau) = \boldsymbol{\beta}(\tau \mid \boldsymbol{\theta}), \quad \sigma(\tau) = \sigma(\tau \mid \boldsymbol{\phi}). \tag{8}$$

Since the scale parameter $\sigma(\tau)$ is not of scientific interest, the main focus of the paper will be to provide a framework for parametric modelling of $\boldsymbol{\beta}(\tau)$, expanding the work of Frumento and Bottai (2016, 2017) on quantile regression. Hereafter, we will speak of *M-quantile regression coefficients modelling* and use the abbreviation MQRCM.

The existence of a parameter vector $\boldsymbol{\theta}$ such that $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta}(\tau \mid \boldsymbol{\theta})$ can hardly be thought of as an assumption, as long as mild regularity conditions are maintained on the data-generating process. As suggested later in the paper, using a parametric approach presents important advantages over the standard, "non-parametric" estimator: (a) it is generally more efficient in terms of standard errors; (b) it permits describing the coefficient functions using a finite and usually

very limited number of parameters; and (c) it facilitates controlling for quantile crossing.


Figure 2: Estimated MQR coefficients ($\tau = 0.005, 0.01, \ldots, 0.995$) associated with gender of the HH (reference = female) and occupational status of the HH (reference = other) in the EU-SILC data, computed using Huber's loss function with tuning parameter $c = 1.345$. Pointwise confidence intervals are represented by the shaded area, while the dotted line indicates the zero.

### 4.1 Model building

We briefly describe a strategy for model building, with the help of a simple example. We assume model (8) to hold, and write the conditional M-quantile function as

$$M(\tau \mid \boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}(\tau \mid \boldsymbol{\theta}). \tag{9}$$

Following Frumento and Bottai (2016, 2017), we adopt the following linear parametrisation:

$$\boldsymbol{\beta}(\tau \mid \boldsymbol{\theta}) = \boldsymbol{\theta} \boldsymbol{b}(\tau) \tag{10}$$

where $\boldsymbol{b}(\tau) = [b_1(\tau), \ldots, b_k(\tau)]^{\mathrm{T}}$ is a $k$-dimensional set of known functions. With this notation, $\boldsymbol{\theta}$ is a $q \times k$ matrix, and the M-quantile function is rewritten as

$$M(\tau \mid \boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}(\tau \mid \boldsymbol{\theta}) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\theta} \boldsymbol{b}(\tau). \tag{11}$$

This model has the advantage of being analytically tractable and relatively simple to implement. Allowing $\boldsymbol{\beta}(\tau \mid \boldsymbol{\theta})$ to be a nonlinear function of $\boldsymbol{\theta}$ is possible, but would have a high cost in terms of computation. On the other hand, the suggested linear parametrisation is very practical and does not represent a limitation in terms of flexibility.

Consider, for example, a regression model with a single covariate $x$:

$$M(\tau \mid x, \boldsymbol{\theta}) = \beta_0(\tau \mid \boldsymbol{\theta}) + \beta_1(\tau \mid \boldsymbol{\theta}) x.$$

A possible approach is to define a "sufficiently flexible" model such as

$$\beta_0(\tau \mid \boldsymbol{\theta}) = \theta_{00} + \theta_{01}\tau + \theta_{02}\tau^2 + \theta_{03}\tau^3 - \theta_{04}\log(1-\tau),$$

$$\beta_1(\tau \mid \boldsymbol{\theta}) = \theta_{10} + \theta_{11}\tau.$$

In this example, the intercept is parametrised using the quantile function of an Exponential distribution, $-\log(1-\tau)$, that determines a long right tail; and a 3rd-degree polynomial $(\tau, \tau^2, \tau^3)$ that allows for a deviation from it. The coefficient associated with $x$ is assumed to be a linear function of $\tau$. In matrix form, the

model is defined by

$$\boldsymbol{b}(\tau) = \begin{bmatrix} 1 \\ \tau \\ \tau^2 \\ \tau^3 \\ -\log(1-\tau) \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_{00} & \theta_{01} & \theta_{02} & \theta_{03} & \theta_{04} \\ \theta_{10} & \theta_{11} & 0 & 0 & 0 \end{bmatrix}.$$

Possible choices of $\boldsymbol{b}(\tau)$ include polynomials $\left[\tau, \tau^2, \tau^3, \dots\right]$, splines, piecewise linear functions, roots $\left[\tau^{1/2},\ (1-\tau)^{1/2},\ \tau^{1/3},\ (1-\tau)^{1/3}, \dots\right]$, logarithms $[\log(\tau), -\log(1-\tau)]$, trigonometric functions $[\cos(2\pi\tau), \sin(2\pi\tau)]$, quantile functions of known distribution (e.g., that of a standard Normal), and combinations of the above. Typically, unbounded functions are used to model the intercept, $\beta_0(\tau)$, while the other coefficients are commonly assumed to be bounded and, in many situations, to be linear functions of $\tau$ or to not depend on $\tau$ at all.

The described modelling approach is much more parsimonious than standard MQR, that can be seen as a special case of model (8) in which $\boldsymbol{\beta}(\tau)$ is allowed to be an arbitrarily flexible function of $\tau$. Using a low-dimensional parametric model allows to represent the coefficient functions by simple closed-form mathematical equations, making it easier to report and interpret the results.

Finally, working with a parametric M-quantile function simplifies controlling for quantile crossing, which can be diagnosed using the first derivative of $M(\tau \mid x, \boldsymbol{\theta})$ with respect to $\tau$. In some situations, it is possible to determine in advance which values of the parameters generate a well-defined M-quantile function. For example, if $M(\tau \mid x, \boldsymbol{\theta}) = \tau(\theta_0 + \theta_1 x)$ with $x \geq 0$, $M'(\tau \mid x, \boldsymbol{\theta}) > 0$ if (i) $\theta_0 > 0$ and $\theta_1 \geq 0$; or (ii) $\theta_1 < 0$ and $\theta_0 > -\theta_1 \max(x)$.

Additional examples of model building are presented in Section 8, in the papers by Frumento and Bottai (2016, 2017), and in the documentation of the `Mqrcm` and `qrcm` R packages (Frumento, 2017, 2018).

## 5 The estimator

Assume model (8) to hold,

$$\boldsymbol{\beta}(\tau) = \boldsymbol{\beta}(\tau \mid \boldsymbol{\theta}), \qquad \sigma(\tau) = \sigma(\tau \mid \boldsymbol{\phi}).$$

We propose estimating $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ by minimising

$$\bar{L}_n(\boldsymbol{\theta}, \boldsymbol{\phi}) = \int_{\tau_1}^{\tau_2} L_n(\boldsymbol{\beta}(\tau \mid \boldsymbol{\theta}), \sigma(\tau \mid \boldsymbol{\phi}))\mathrm{d}\tau, \quad 0 \leq \tau_1 < \tau_2 \leq 1, \tag{12}$$

which corresponds to the integral, with respect to $\tau$, of the loss function of standard M-quantile regression given in (2). This quantity can be thought of as an average loss function, that carries information on multiple M-quantiles at once. This approach is analogous to the estimation method introduced by Frumento and Bottai (2016) and referred to as *integrated loss minimisation* (ILM). Although in most situations $\tau_1 = 0$ and $\tau_2 = 1$, it is possible to model a subset of quantiles of interest, e.g., those below the median ($\tau_1 = 0$, $\tau_2 = 0.5$)

The exact expression for (12) depends on the selected loss function, $\rho_\tau$. The standard, "nonparametric" estimator of $\boldsymbol{\beta}(\tau)$ can be obtained in two different ways: (i) by allowing $\boldsymbol{\beta}(\tau \mid \boldsymbol{\theta})$ to be arbitrarily flexible, e.g., using a piecewise-linear function with a very large number of knots; or (ii) by setting $\tau_1 = \tau - \Delta$, $\tau_2 = \tau + \Delta$, and letting $\Delta \to 0$.

In the `Mqrcm` R package, $\bar{L}_n(\boldsymbol{\theta}, \boldsymbol{\phi})$ is evaluated numerically, and a Newton-type algorithm is used to perform minimisation. The coefficient functions are modeled as in (10),

$$\boldsymbol{\beta}(\tau \mid \boldsymbol{\theta}) = \boldsymbol{\theta}\boldsymbol{b}(\tau),$$

allowing the user to define $\boldsymbol{b}(\cdot)$. The scale parameter $\sigma(\tau \mid \boldsymbol{\phi})$ is treated as a piecewise constant function of $\tau$, which guarantees flexibility and avoids specifying a parametric model for a nuisance paramter.

## 6 Asymptotic theory

To derive asymptotic properties, we apply the standard theory of M-estimators (e.g., Amemiya, 1985; Newey and McFadden, 1994). To facilitate the notation, we write as $\boldsymbol{\xi} = (\boldsymbol{\theta}, \boldsymbol{\phi})$ the vector of all model parameters. We denote by $\hat{\boldsymbol{\xi}}_n = (\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\phi}}_n)$ the minimiser of $\bar{L}_n(\boldsymbol{\xi}) = \bar{L}_n(\boldsymbol{\theta}, \boldsymbol{\phi})$, the integrated loss function defined by (12), and by $\boldsymbol{\xi}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\phi}_0)$ the true parameter value minimising $\bar{L}_0(\boldsymbol{\xi}) = E[\bar{L}_n(\boldsymbol{\xi})]$.

**Theorem 1 (consistency)**. *Assume that (i) the parameter space $\boldsymbol{\Xi}$ is a compact set; (ii) $\bar{L}_n(\boldsymbol{\xi})$ converges uniformly in probability to $\bar{L}_0(\boldsymbol{\xi})$; (iii) $\bar{L}_0(\boldsymbol{\xi})$ is uniquely minimised at $\boldsymbol{\xi}_0$; (iv) $\bar{L}_0(\boldsymbol{\xi})$ is continuous. Then $\boldsymbol{\xi}_n \xrightarrow{p} \boldsymbol{\xi}_0$.*

The above assumptions require the following non-trivial conditions for consistency: (a) $\int_{\tau_1}^{\tau_2} \rho_\tau \left( \frac{y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(\tau)}{\sigma(\tau)} \right) \mathrm{d}\tau$ is finite; and (b) $\sigma(\tau \mid \boldsymbol{\phi}) > 0$ at all $\tau$. Condition (a) may not hold if $\boldsymbol{\beta}(\tau \mid \boldsymbol{\theta})$ is non-integrable, e.g., $\boldsymbol{\beta}(\tau \mid \boldsymbol{\theta}) = \theta \tan(\pi(\tau - 0.5))$. Condition (b) requires the distribution to be non-degenerated. A proof of Theorem 1 is given in Newey and McFadden (1994), Theorem 2.1, p. 2121.

**Theorem 2 (asymptotic normality)**. *Suppose that the conditions of Theorem 1 are satisfied, and that (i) $\boldsymbol{\xi}_0$ is an interior point of $\boldsymbol{\Xi}$; (ii) $\bar{L}_n(\boldsymbol{\xi})$ is twice continuously differentiable in a neighborhood $\mathcal{N}$ of $\boldsymbol{\xi}_0$; (iii) $\sqrt{n} \nabla_{\boldsymbol{\xi}} \bar{L}_n(\boldsymbol{\xi}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega})$; (iv) there is $\boldsymbol{H}(\boldsymbol{\xi})$ that is continuous at $\boldsymbol{\xi}_0$ and $\sup_{\boldsymbol{\xi} \in \mathcal{N}} ||\nabla_{\boldsymbol{\xi}\boldsymbol{\xi}} \bar{L}_n(\boldsymbol{\xi}) - \boldsymbol{H}(\boldsymbol{\xi})|| \xrightarrow{p} \mathbf{0}$; (v) $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{\xi}_0)$ is nonsingular. Then,*

$$\sqrt{n}(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{H}^{-1} \boldsymbol{\Omega} \boldsymbol{H}^{-1}\right).$$

Condition (i) is standard. Conditions (ii) and (iii) permit applying a central limit theorem to $\nabla_{\boldsymbol{\xi}} \bar{L}_n(\boldsymbol{\xi}_0)$, while conditions (iv) and (v) guarantee that the asymptotic covariance matrix is well-defined. For a proof of Theorem 2, we refer to Newey and McFadden (1994), Theorem 3.1, p. 2143.

An estimate of the covariance matrix of $\hat{\boldsymbol{\xi}}$ can be obtained using the sample counterparts of $\boldsymbol{H}$, which can be estimated by the Hessian matrix of $\bar{L}_n(\hat{\boldsymbol{\xi}})$, and of $\boldsymbol{\Omega}$, which is consistently estimated by the outer product of the summands of $\nabla_{\boldsymbol{\xi}} \bar{L}_n(\boldsymbol{\xi})$. Under model (10), where $\boldsymbol{\beta}(\tau \mid \boldsymbol{\theta}) = \boldsymbol{\theta}\boldsymbol{b}(\tau)$ and $\boldsymbol{\theta}$ is a $q \times k$ matrix, $\mathrm{cov}(\hat{\boldsymbol{\theta}})$ is given by the upper $qk \times qk$ block of $\boldsymbol{H}^{-1} \boldsymbol{\Omega} \boldsymbol{H}^{-1}$. At any given $\tau$, an estimate of $\mathrm{cov}(\hat{\boldsymbol{\beta}}(\tau))$ can be easily computed using the fact that $\hat{\boldsymbol{\beta}}(\tau) = \hat{\boldsymbol{\theta}}\boldsymbol{b}(\tau)$ is a linear transformation of $\hat{\boldsymbol{\theta}}$.

# 7 Simulation results

To assess the finite-sample performance of the proposed estimator and compare it with standard techniques, we designed a simulation study composed of two parts. In the first part of the simulation, we compare our method with standard M-quantile regression; in the second, the described method is applied to small area estimation.

## 7.1 Comparing MQRCM with standard MQR

We considered the M-quantile function defined by Huber's loss function with tuning constant $c = 1.345$, and parametrised as follows:

$$M(\tau \mid x_1, x_2) = \beta_0(\tau) + \beta_1(\tau)x_1 + \beta_2(\tau)x_2, \tag{13}$$

where $x_1$ is a discrete uniform variable with support $\{1, \ldots, 5\}$, and $x_2$ is binary with $P(x_2 = 1) = 0.5$.

To simulate data with a desired M-quantile function $M(\tau \mid \boldsymbol{x}) = \beta_0(\tau) + \beta_1(\tau)x_1 + \ldots$, we proceeded as follows. We defined $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \ldots\}$, and considered a "flexible" quantile function $Q_{\boldsymbol{\gamma}}(\tau \mid \boldsymbol{x}) = [\beta_0(\tau) + s(\tau)\boldsymbol{\gamma}_0] + [\beta_1(\tau) + s(\tau)\boldsymbol{\gamma}_1]x_1 + \ldots$, where $s(\tau)$ was the basis of a B-spline with knots at $\tau = (0, 0.05, 0.10, 0.25, 0.5, 0.75, 0.90, 0.95, 1)$. The regression coefficients of $Q_{\boldsymbol{\gamma}}(\tau \mid \boldsymbol{x})$ are equal to those of the desired M-quantile function, plus a deviation expressed by a spline function with parameters $\boldsymbol{\gamma}$. We defined $M_{\boldsymbol{\gamma}}(\tau \mid \boldsymbol{x})$ to be the M-quantile function obtained by applying ordinary M-quantile regression to a "perfect draw" from $Q_{\boldsymbol{\gamma}}(\tau \mid \boldsymbol{x})$. By "perfect draw" we mean that: (i) the empirical distribution of $\boldsymbol{x}$ is equal to the true one, and (ii) for each unique value of $\boldsymbol{x}$, data are "simulated" by evaluating $Q_{\boldsymbol{\gamma}}(\tau \mid \boldsymbol{x})$ at a grid of evenly spaced values of $\tau$ in $(0, 1)$. We then computed $\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \int_0^1 [M_{\boldsymbol{\gamma}}(\tau \mid \boldsymbol{x}) - M(\tau \mid \boldsymbol{x})]^2 \mathrm{d}\tau$. With this procedure, the true M-quantiles associated with $Q_{\hat{\boldsymbol{\gamma}}}(\cdot \mid \boldsymbol{x})$ were approximately equal to their target value $M(\cdot \mid \boldsymbol{x})$.

We implemented the following simulation scenarios:

*Simulation 1.* We defined $\beta_0(\tau) = \log(\tau) - 0.5\log(1 - \tau)$, $\beta_1(\tau) = 1 + 2\exp\{-5(1 - \tau)\}$, and $\beta_2(\tau) = 1 + 5\tau$.

*Simulation 2.* We defined $\beta_0(\tau) = z(\tau)$, $\beta_1(\tau) = 1 + 5(1 - (1 - \tau)^{0.25})$, and $\beta_2(\tau) = 1 + \tau + 10(\tau - 0.5)^3$, where $z(\tau)$ denotes the quantile function of a standard Normal distribution.

For each scenario, we generated $R = 1000$ simulated dataset. For each dataset, we estimated standard MQR coefficients, and applied MQRCM with two different model specifications: (i) $\text{M}^{(a)}$, the true model, and (ii) $\text{M}^{(b)}$, a model in which $\boldsymbol{\beta}(\tau \mid \boldsymbol{\theta})$ was parametrised by a B-spline basis with knots at $\tau = \{0, 0.05, 0.25, 0.5, 0.75, 0.95, 1\}$. This allowed to assess the performance of the described estimator in a situation in which a flexible model is used in absence of stronger parametric assumptions.

We measured the bias and standard errors of estimated M-quantile regression coefficients across datasets. The bias of all estimators was negligible and is not reported, while standard errors are summarised in Table 2. Results demonstrate that MQRCM estimators can be much more efficient than ordinary MQR, when a parsimonious model is used to describe $\boldsymbol{\beta}(\tau \mid \boldsymbol{\theta})$. When a flexible model is used instead, the efficiency gain tends to vanish.

## 7.2 Using MQRCM in Small Area Estimation

We carried out model-based simulations to compare small area predictors based on MQRCM with those obtained from standard MQR. We generated the population data following Chambers et al. (2014), defining $m = 40$ small areas. Samples were selected by simple random sampling without replacement within each area. The population and sample sizes were $N_j = 100$ and $n_j = 5$ for all areas, $j = 1, \ldots, 40$. We considered a single predictor $x$ generated from a lognormal distribution with mean 1 and standard deviation 0.5 on the log scale. The response was generated as $y_{ij} = 100 + 5x_{ij} + \upsilon_j + \epsilon_{ij}$, where the area-level effects, $\upsilon_j$, and the individual effects, $\epsilon_{ij}$, were independently generated according to two scenarios:

- *Scenario (0,0)*: $u \sim N(0, 3)$ and $\epsilon \sim N(0, 6)$.
- *Scenario ($\epsilon$,$\upsilon$)*: $\upsilon \sim N(0, 3)$ for areas $1 - 36$, $\upsilon \sim N(9, 20)$ for areas $37 - 40$; $\epsilon \sim \delta N(0, 6) + (1 - \delta)N(20, 150)$ where $\delta$ was a binary variable with $P(\delta = 1) = 0.97$.

Scenario (0,0) corresponds to a "well-behaved" distribution. In scenario ($\epsilon$,$\upsilon$),

the area effects $37-40$ have a nonzero mean, and a much larger variance than those of areas $1-36$. Moreover, the individual effects come from a mixture of two normal distributions that generates outliers with a probability of 3%.

Each scenario was independently simulated $R = 1000$ times. For each simulated dataset, we estimated M-quantile regression coefficients in two different ways: (i) using the traditional MQR estimator; and (ii) applying the MQRCM approach described in this paper. In both cases the influence function $\psi$ was a Huber-type function with tuning constant $c = 1.345$.

To implement MQRCM, we used the following model specification:

$$M(\tau \mid x, \boldsymbol{\theta}) = \beta_0(\tau \mid \boldsymbol{\theta}) + \beta_1(\tau \mid \boldsymbol{\theta})x$$

with

$$\beta_0(\tau \mid \boldsymbol{\theta}) = \theta_{00} + \theta_{01}\tau + \theta_{02}\tau^2 + \theta_{03}\tau^3 + \theta_{04}\tau^4 + \theta_{05}\tau^5 + \theta_{06}z(\tau),$$

$$\beta_1(\tau \mid \boldsymbol{\theta}) = \theta_{10} + \theta_{11}\tau,$$

where $z(\tau)$ denotes the quantile function of a standard Normal distribution. The intercept was defined by the combination of a Normal model and a 5-th degree polynomial, allowing for a great flexibility. The regression coefficient associated with $x$, that in the true model was a constant, was assumed to be a linear function of $\tau$.

For each dataset, we used the fitted M-quantile regression coefficients to estimate the area means using the naive formula given in (6) and its corrected version given in (7). Results are summarised in Table 3, where we report the median value (across areas) of the relative bias and the relative root mean squared error (rrMSE), which are obtained by dividing the absolute bias and the root mean squared error of each area by the true area means $\bar{y}_j$, $j = 1, \ldots, 40$.

Under the $(0, 0)$ scenario, the predictors based on MQR and on MQRCM showed similar results in terms of bias and $rr$MSE. In scenario $(\epsilon,\upsilon)$, predictors based on MQRCM had a significantly smaller bias and $rr$MSE in areas 37-40. These results suggest that when outlier values are present, small area predictors based on parsimonious parametric models can be more reliable and efficient than those based on ordinary M-quantile regression.

Table 2: Simulation results: comparing MQRCM with standard MQR

| Sim1 | | $\text{se}(\hat{\beta}_0(\tau))$ | | | $\text{se}(\hat{\beta}_1(\tau))$ | | | $\text{se}(\hat{\beta}_2(\tau))$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\tau$ | MQR | $\text{M}^{(a)}$ | $\text{M}^{(b)}$ | MQR | $\text{M}^{(a)}$ | $\text{M}^{(b)}$ | MQR | $\text{M}^{(a)}$ | $\text{M}^{(b)}$ |
| | 0.05 | .61 | .41 | .62 | .16 | .11 | .16 | .51 | .45 | .53 |
| | 0.25 | .46 | .22 | .44 | .14 | .11 | .14 | .45 | .41 | .44 |
| | 0.50 | .57 | .21 | .58 | .19 | .11 | .20 | .59 | .44 | .61 |
| | 0.75 | .72 | .34 | .73 | .27 | .14 | .28 | .75 | .56 | .75 |
| | 0.95 | .66 | .70 | .64 | .25 | .27 | .24 | .59 | .68 | .61 |
| Sim2 | | $\text{se}(\hat{\beta}_0(\tau))$ | | | $\text{se}(\hat{\beta}_1(\tau))$ | | | $\text{se}(\hat{\beta}_2(\tau))$ | | |
| | $\tau$ | MQR | $\text{M}^{(a)}$ | $\text{M}^{(b)}$ | MQR | $\text{M}^{(a)}$ | $\text{M}^{(b)}$ | MQR | $\text{M}^{(a)}$ | $\text{M}^{(b)}$ |
| | 0.05 | .22 | .32 | .23 | .08 | .11 | .08 | .23 | .26 | .23 |
| | 0.25 | .43 | .34 | .43 | .16 | .12 | .16 | .44 | .40 | .44 |
| | 0.50 | .62 | .42 | .63 | .23 | .14 | .23 | .62 | .48 | .64 |
| | 0.75 | .76 | .52 | .77 | .29 | .20 | .30 | .79 | .65 | .79 |
| | 0.95 | .77 | .69 | .77 | .34 | .32 | .34 | .84 | .84 | .86 |

Empirical standard errors of the estimated M-quantile regression coefficients across simulations. MQR denotes ordinary M-quantile regression; $\text{M}^{(a)}$ and $\text{M}^{(b)}$ denote MQRCM estimators, in which $\hat{\boldsymbol{\beta}}(\tau) = \boldsymbol{\beta}(\tau \mid \hat{\boldsymbol{\theta}})$. In $\text{M}^{(a)}$ we fitted the true model, while in $\text{M}^{(b)}$we parametrised $\boldsymbol{\beta}(\tau \mid \boldsymbol{\theta})$ by a B-spline basis with knots at $\tau = \{0, 0.05, 0.25, 0.5, 0.75, 0.95, 1\}$.

Table 3: Simulation results: performance of different predictors of small area means

| Scenario | $(0,0)$ | $(\epsilon,u)$ | $(\epsilon,u)$ |
|---|---|---|---|
| Areas | $1-40$ | $1-36$ | $37-40$ |
| *Median relative bias* | | | |
| MQR/*Naive* | -0.005 | -0.285 | -0.826 |
| MQRCM/*Naive* | -0.006 | -0.276 | -0.728 |
| MQR/BC | 0.002 | -0.228 | -0.287 |
| MQRCM/BC | 0.002 | -0.223 | -0.278 |
| *Median rr*MSE | | | |
| MQR/*Naive* | 0.835 | 0.986 | 1.486 |
| MQRCM/*Naive* | 0.834 | 0.992 | 1.310 |
| MQR/BC | 0.915 | 1.240 | 1.339 |
| MQRCM/BC | 0.916 | 1.244 | 1.222 |

Median value (across areas) of the relative bias and the relative root mean squared error (rrMSE) of naive and bias-corrected estimators of the area means obtained using ordinary M-quantile regression (MQR) and M-quantile regression coefficients modelling (MQRCM).

# 8 Analysis of EU-SILC data

We used data from the 2006 EU-SILC and the 2001 Population Census of Italy presented in Section 3 to estimate the mean equivalised income for the Local Labour Systems (LLSs) of three Italian regions: Lombardia (Northern Italy), Toscana (Central Italy) and Campania (Southern Italy). The target small areas were 172 in total. In addition to evaluating the potential dissimilarities within each region, we were also interested in understanding the so-called "north-south" divide characterising Italy in terms of poverty and living conditions.

We formulated an M-quantile regression model that included all predictors summarised in Table 1, for a total of eight coefficients, plus an intercept $\beta_0(\tau)$. To apply MQRCM, we considered a variety of parametrisations of $\boldsymbol{\beta}(\tau \mid \boldsymbol{\theta}) = \boldsymbol{\theta}\boldsymbol{b}(\tau)$, choosing different specifications of the "basis" $\boldsymbol{b}(\tau)$ that defines the parametric form of the regression coefficients. In principle, each coefficient could be described by its own parametric model. For example, $\beta_0(\tau)$ could be a combination of logarithmic functions, the coefficient associated with gender may be described

by a linear function, and so on. In practice, the same parametric form was used to describe all coefficients. This simplification is very convenient, although it requires $\boldsymbol{b}(\tau)$ to be sufficiently flexible.

Selected parametrisations are illustrated in Table 4. Model 0 uses a piecewise linear functions with discontinuity points at $\tau = \{0.05, 0.10, \ldots, 0.95\}$, and should be regarded as a "nonparametric" model which is essentially equivalent to standard M-quantile regression. Model 1 describes the regression coefficients by a third-degree polynomial, while models 2 and 3 use different roots of $\tau$ and $1-\tau$. Model 4 defines a Logistic distribution, while model 5 uses a linear combination of $\log(\tau)$ and $-\log(1-\tau)$, that corresponds to the quantile function of the asymmetric Logistic distribution. Model 6 uses the quantile function of a "double" Rayleigh distribution, and model 7 is a generalisation of it. Models 8 and 9 parametrise the coefficients by a combination of a linear trend, and a set of trigonometric functions with different periods.

For each model $h = \{0, 1, \ldots, 9\}$, we obtained an estimate of the parameters, say $\hat{\boldsymbol{\theta}}^{(h)}$, and computed $F(y_{ij} \mid \hat{\boldsymbol{\theta}}^{(h)}, \boldsymbol{x}_{ij}) = M^{-1}(y_{ij} \mid \hat{\boldsymbol{\theta}}^{(h)}, \boldsymbol{x}_{ij})$. Using model 0 as a benchmark to assess model fit, we selected the minimiser of $\sum_{i,j} |F(y_{ij} \mid \hat{\boldsymbol{\theta}}^{(h)}) - F(y_{ij} \mid \hat{\boldsymbol{\theta}}^{(0)})|$, $h = \{1, 2, \ldots, 9\}$. With this procedure, model 5 was selected as our final model. The estimated model parameters are summarised in Table 5, while selected M-quantile regression coefficients are reported in Table 6. The model is represented graphically in Figures 3 and 4, where we also report the output of standard M-quantile regression, which was estimated at a grid of quantiles, $\tau = \{0.005, 0.01, \ldots, 0.995\}$.

Based on the fitted parametric model, all estimated M-quantile regression coefficients showed a monotonic trend, and some of them had a long left or right tail. As illustrated in Figure 3, estimates based on MQRCM did not show the erratic tail behaviour that was observed using standard MQR. Additionally, as shown in Figure 4, using a parametric model reduced the variability in the tails, at the cost of a slightly increased variability at intermediate values of $\tau$. Finally, while standard MQR estimates suffered from significant M-quantile crossing (that affected approximately 1.9% of the observations), no crossing M-quantiles were found using MQRCM. Note that, under model 5, the first derivative of the M-quantile function is available in closed form and is given by $M'(\tau \mid \boldsymbol{x}, \boldsymbol{\theta}) =$

$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}\boldsymbol{b}'(\tau)$ with $\boldsymbol{b}'(\tau) = [0, \tau^{-1}, (1-\tau)^{-1}]^{\mathrm{T}}$.

A positive effect was found for the indicators of house ownership and current employment, age and male gender, and education, while the household size did not reach statistical significance. Campania had significantly lower income compared with Lombardia and Toscana.

In Figure 5 we present maps of MQRCM/BC estimates of average equivalised income for each LLS in Toscana, Lombardia and Campania. The maps obtained with the MQRCM/Naive estimator are not displayed for reasons of space, and are available upon request to the authors.

In Lombardia (northern Italy) the LLSs with higher estimated average equivalised income are concentrated in the south-western and south-eastern parts of the region. Instead, the LLSs with lower estimates are concentrated in the central and northern parts of the region. The map of Toscana (central Italy) indicates that the LLSs in the north of the region, corresponding to the LLSs of the province of Massa-Carrara and those in the northern part of the provinces of Lucca and Prato, are characterised by the the lowest estimates of the mean household equivalised income. These areas can be considered as the more critical in the region. On the other hand, the LLSs with the highest average equivalised income are concentrated in the provinces of Firenze, Siena and Arezzo, in the central-eastern part of the region. The map of Campania (southern Italy) shows that the average equivalised income in this region is more geographically differentiated. For example, the LLSs with the highest estimates are spread across the region. These results confirm that the mean household income is usually higher in the LLSs around the largest cities, and are consistent with the estimates at provincial level obtained in the SAMPLE project (SAMPLE, 2010).

Estimating the average equivalised income at the LLS level allowed us to investigate the gap in living conditions between the three considered regions. The gap between Lombardia and Toscana is not very pronounced, as shown by the relatively small differences in terms of mean equivalised income between the two regions. Instead, Campania presents a big gap with the central and northern regions. The highest estimates of average equivalised income in Campania are comparable to the lowest ones in Lombardia and Toscana. These results confirm the existence of the so-called "north-south" divide in Italy.

In Figure 6 we compare the estimates obtained with MQRCM/BC predictor with those produced by direct estimators. In the left panel of the figure, we plotted the direct estimates on the $x$-axis, and the model-based ones on the $y$-axis. The Pearson's linear correlation coefficient was $r = 0.82$, suggesting that the two estimates are somewhat consistent. In the right panel we plotted the difference between the two estimates against the available sample size, $n_j$. As expected, larger differences were found in correspondence of smaller sample sizes.

In Figure 7, we report the standard errors of the direct estimators $\hat{\bar{y}}_j^{\text{Direct}}$ and the root mean squared errors of MQRCM/BC, as a function of the area-specific sample size, $n_j$ (only for areas with $n_j > 0$). It can be seen that using SAE yields a larger gain in precision when $n_j$ is relatively small. The more stable behaviour of MSE of $\hat{\bar{y}}_j^{\text{MQRCM/BC}}$ when compared to standard error of direct estimators reflects the robustness of the method with respect to the presence of outliers, that are influential not only on point estimates but also on estimation of MSE.

To validate the reliability of the model-based small area estimates, we used a goodness-of-fit test (Brown et al., 2001) for the null hypothesis that the direct and model-based estimates are statistically equivalent. The diagnostic test is based on the following Wald statistic:

$$W = \sum_{j=1}^{m} \left\{ \frac{(\hat{\bar{y}}_j^{\text{MQRCM/BC}} - \hat{\bar{y}}_j^{\text{Direct}})^2}{\text{MSE}(\hat{\bar{y}}_j^{\text{MQRCM/BC}}) + Var(\hat{\bar{y}}_j^{\text{Direct}})} \right\}.$$

The test statistic is compared with the 95th percentile of a chi-squared distribution with $m = 172$ degrees of freedom (203.60). In our case, $W = 35.72$, showing that the model-based estimates were not statistically different from the direct estimates.

Table 4: Alternative model specifications.

| Model | $\boldsymbol{b}(\tau)$ | n. of free parameters |
|---|---|---|
| 0 | $L(\tau, \text{knots} = \{0.05, 0.10, \ldots, 0.95\})$ | $21 \times 9 = 189$ |
| 1 | $\tau, \tau^2, \tau^3$ | $4 \times 9 = 36$ |
| 2 | $\tau^{1/2}, (1-\tau)^{1/2}$ | $3 \times 9 = 27$ |
| 3 | $\tau^{1/3}, (1-\tau)^{1/3}$ | $3 \times 9 = 27$ |
| 4 | $\log\{\tau/(1-\tau)\}$ | $2 \times 9 = 18$ |
| 5 | $\log(\tau), -\log(1-\tau)$ | $3 \times 9 = 27$ |
| 6 | $[-\log(\tau)]^{1/2}, [-\log(1-\tau)]^{1/2}$ | $3 \times 9 = 27$ |
| 7 | $[-\log(\tau)]^{1/3}, [-\log(1-\tau)]^{1/3}$ | $3 \times 9 = 27$ |
| 8 | $\tau, \cos(\pi\tau), \sin(\pi\tau)$ | $4 \times 9 = 36$ |
| 9 | $\tau, \cos(2\pi\tau), \sin(2\pi\tau)$ | $4 \times 9 = 36$ |

Different parametrisations based on the set of functions that define $\boldsymbol{b}(\tau)$. An intercept $b(\tau) = 1$ was always included. In the table, $L(\tau, \text{knots})$ indicates a piecewise linear function with discontinuity points at the supplied knots. Model 5 was selected as our final model. The estimated model parameters are summarised in Table 5, while selected M-quantile regression coefficients are reported in Table 6. The model is represented graphically and compared with standard M-quantile regression in Figures 3 and 4.

Table 5: Summary of model 5.

| | 1 | $\log(p)$ | $-\log(1-p)$ | P-value |
|---|---|---|---|---|
| Intercept | 10545 (492)$^*$ | 2229 (201)$^*$ | 4938 (428)$^*$ | .000$^*$ |
| Ownership of the house | 1350 (328)$^*$ | 325 (121)$^*$ | 851 (297)$^*$ | .000$^*$ |
| Age of the HH (centered at mean) | 21 (12) | -2 (4) | 50 (12)$^*$ | .000$^*$ |
| Occupational status of the HH: working | 2183 (388)$^*$ | -651 (119)$^*$ | 46 (353) | .000$^*$ |
| Gender of the HH: male | 1047 (367)$^*$ | 94 (119) | 201 (345) | .000$^*$ |
| Years in education of the HH (centered at mean) | 322 (38)$^*$ | 48 (12)$^*$ | 409 (41)$^*$ | .000$^*$ |
| Household size (centered at mean) | 65 (125) | -44 (45) | -61 (103) | .199 |
| Region: Campania | -4052 (382)$^*$ | -531 (134)$^*$ | -425 (353) | .000$^*$ |
| Region: Lombardia | -278 (364) | -47 (135) | 758 (369)$^*$ | .158 |
| P-value | .000$^*$ | .000$^*$ | .000$^*$ | |

Estimates of $\boldsymbol{\theta}$ and asymptotic standard errors (in brackets) based on model 5, in which M-quantile regression coefficients are parametrised as $\beta_j(\tau \mid \boldsymbol{\theta}) = \theta_{j0} + \theta_{j1}\log(\tau) - \theta_{j2}\log(1-\tau)$, $j = 0, \ldots, 8$. For example, the estimated M-quantile regression coefficient associated with the number of years in education is $\widehat{\beta}_5(\tau) = 322 + 48\log(\tau) - 409\log(1-\tau)$. All estimated coefficients are represented graphically in Figure 3. The p-values in the table margins are obtained from a test of nullity of the corresponding row or column. In particular, the bottom row can be used to assess the significance of each component of $\boldsymbol{b}(\tau)$, while the last column is a significance test for the effect of the covariates. In the table, the asterisk ($^*$) denotes significance less than 0.05.

Table 6: Estimated M-quantile regression coefficients.

| | $\tau = 0.05$ | | $\tau = 0.5$ | | $\tau = 0.95$ | |
|---|---|---|---|---|---|---|
| | MQRCM | MQR | MQRCM | MQR | MQRCM | MQR |
| Intercept | 4120 (412) | 4572 (362) | 12423 (340) | 12422 (353) | 25223 (1072) | 23869 (1512) |
| Ownership of the house | 421 (236) | 515 (263) | 1715 (227) | 1762 (257) | 3882 (744) | 4051 (1098) |
| Age of the HH (centered at mean) | 30 (8) | 31 (7) | 57 (8) | 54 (7) | 171 (29) | 186 (31) |
| Occupational status of the HH: working | 4135 (265) | 4243 (266) | 2666 (277) | 2667 (260) | 2354 (885) | 2812 (1112) |
| Gender of the HH: male | 775 (264) | 748 (248) | 1122 (257) | 1092 (242) | 1646 (857) | 1713 (1037) |
| Years in education of the HH (centered at mean) | 198 (23) | 210 (20) | 572 (26) | 562 (19) | 1544 (102) | 1592 (83) |
| Household size (centered at mean) | 195 (94) | 228 (87) | 54 (87) | 74 (85) | -115 (257) | -168 (365) |
| Region: Campania | -2482 (266) | -2530 (280) | -3978 (263) | -3938 (274) | -5297 (880) | -4751 (1170) |
| Region: Lombardia | -100 (263) | -43 (234) | 279 (240) | 231 (229) | 1994 (909) | 2499 (980) |

Estimated M-quantile regression coefficients of different orders, obtained from model 5 (MQRCM), and from ordinary M-quantile regression (MQR). Estimated standard errors in brackets.

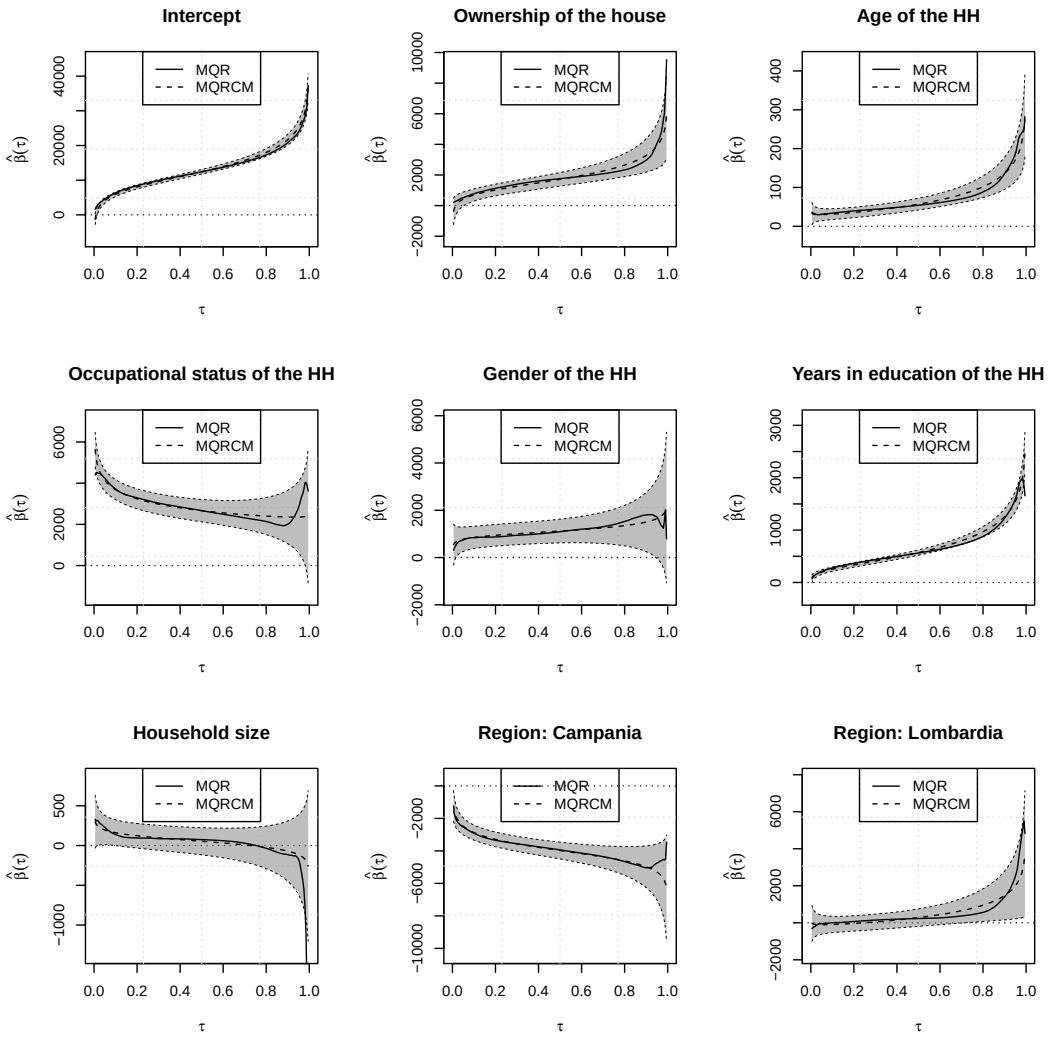


Figure 3: Estimated M-quantile regression coefficients using MQR (continuous line) and MQRCM (dashed line with associated pointwise 95% confidence interval).

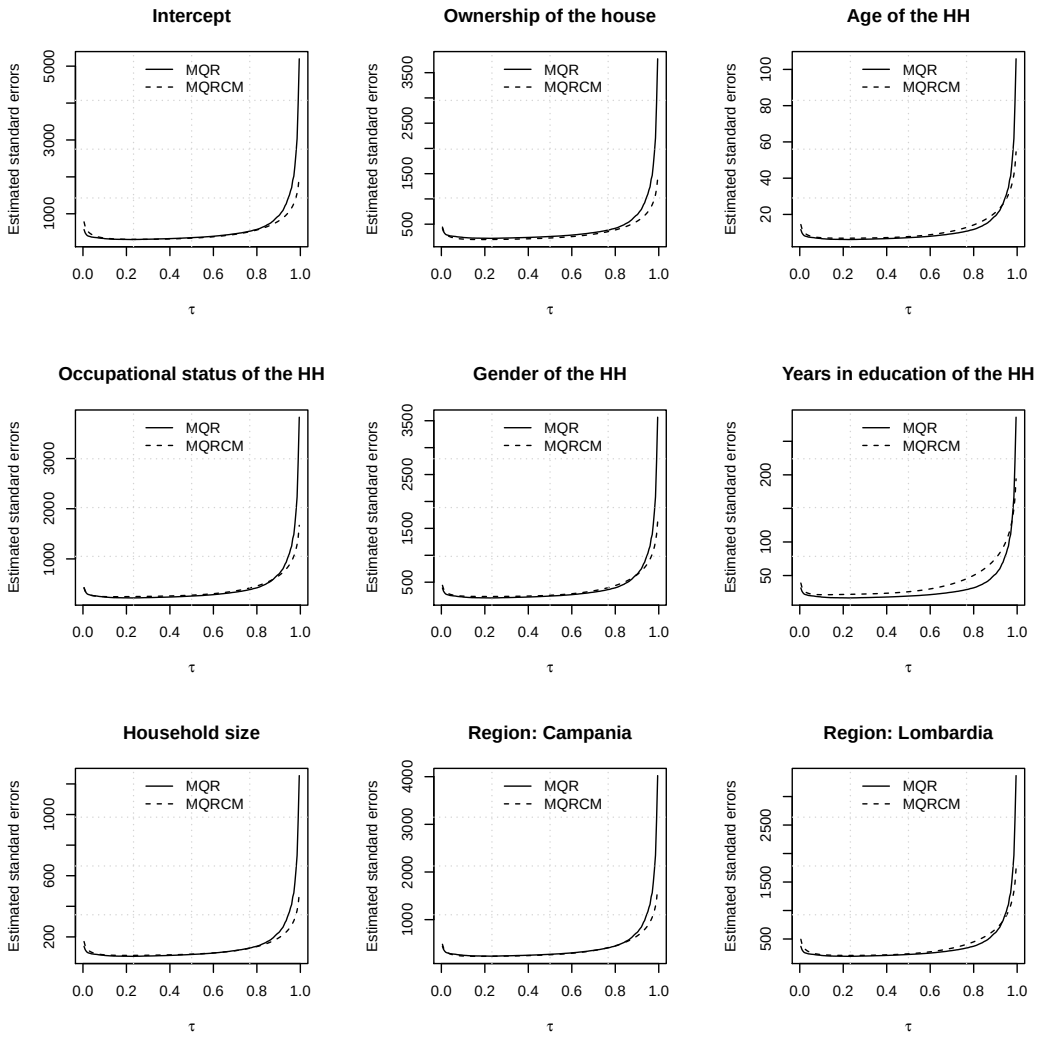


Figure 4: Estimated standard errors of M-quantile regression coefficients estimated by MQR (continuous line) and MQRCM (dashed line).

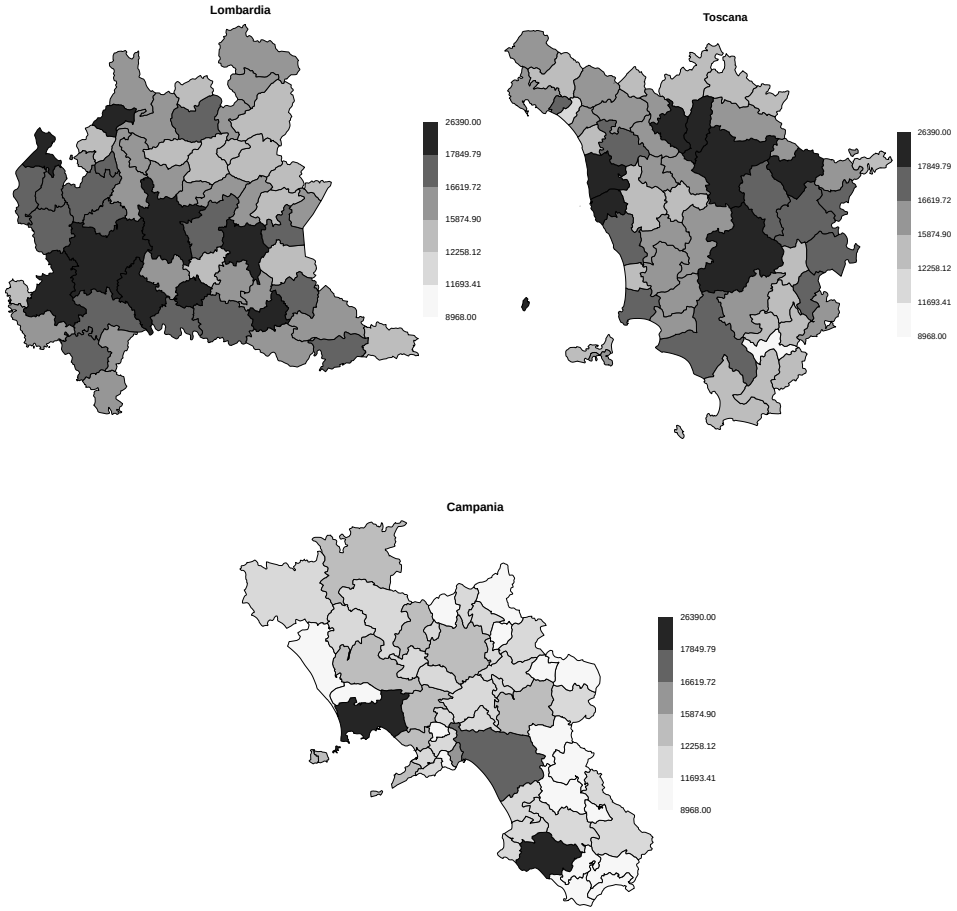


Figure 5: Estimated average equivalised income by MQRCM/BC based on EU-SILC and Census data for Lombardia, Toscana, Campania, income year 2005 .

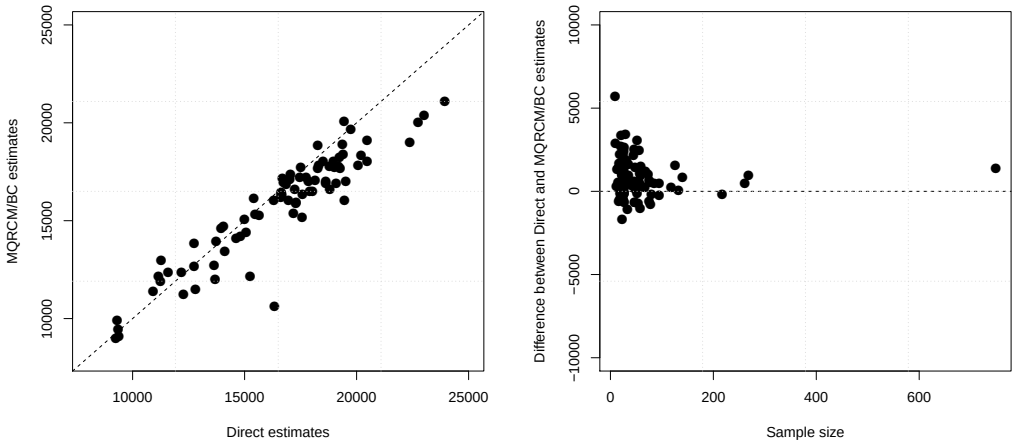


Figure 6: A comparison between the direct estimates and those obtained by MQRCM/BC.

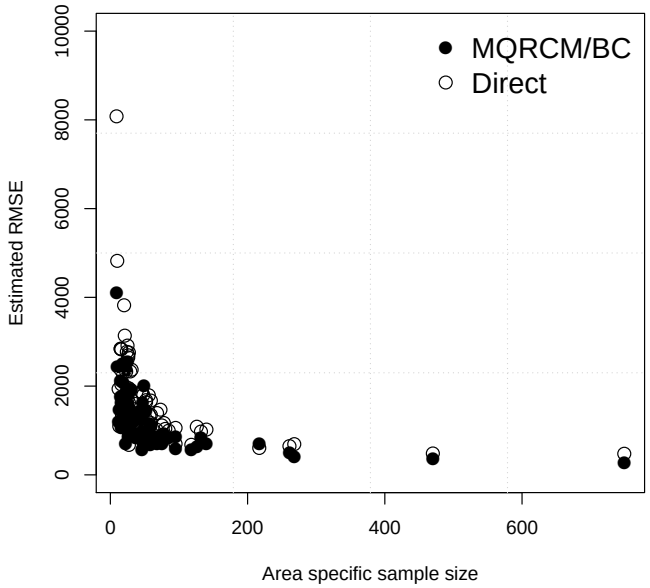


Figure 7: A comparison of root mean square error estimates for the direct and the small area estimators obtained by MQRCM/BC.

# 9 Conclusions

We introduced a new estimator of M-quantiles, in which the regression coefficients, $\boldsymbol{\beta}(\tau)$, are described by parametric functions of $\tau$. Unlike ordinary M-quantile regression, our method is not grid-based and allows to model the entire M-quantile function at once. As shown by simulations, fitting parsimonious models permits achieving significant gains in efficiency in comparison with standard methods. Moreover, using a parametric model permits stabilising the tail behaviour of the M-quantile regression coefficients, reducing the risk of M-quantile crossing and making it very easy to diagnose it. Possible future developments of the method outlined in this paper include M-quantile modelling of binary, count and multicategory outcomes.

The described approach has been applied in small area estimation to evaluate the average equivalised household income of local labour systems (LLSs) of Toscana, Lombardia and Campania administrative regions in Italy. Our estimates confirm the existence of the so-called "north-south" divide in Italy.

An efficient implementation of the proposed estimator is provided in the `Mqrcm` R package, which includes a main function `iMqr` for model fitting, and a variety of auxiliary function for summary, plotting, and prediction.

## References

Amemiya, T. (1985). *Advanced econometrics*. Cambridge, MA: Harvard University Press.

Bianchi, A., Fabrizi, E., Salvati, N., Tzavidis, N. (2018). "Estimation and testing in M-quantile regression with application to small area estimation", *International Statistical Review*, **0**(0), 1–30, doi:10.1111/insr.12267.

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. Statistics Canada Symp.Achieving Data Quality in a Statistical Agency: a Methodological Perspective. Hull: Statistics Canada.

Chambers, R. and Tzavidis, N. (2006). "M-quantile Models for Small Area Estimation", *Biometrika* **93**, 255-268.

Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2014). "Outlier robust small area estimation". *Journal of the Royal Statistical Society, series B*, 76, 47–69.

Dodge, Y., Jureckova, J. (2000). *Adaptive Regression*. New York: Springer.

Eurostat (2016). *NUTS - Nomenclature of territorial units for statistics*. https://ec.europa.eu/eurostat/web/nuts/background.

Fabrizi, E., Giusti, C., Salvati, N. and Tzavidis, N. (2014). "Mapping average equivalized income using robust small area methods", *Papers in Regional Science*, **93**(3), 685–702, doi:10.1111/pirs.12015.

Frumento, P. (2017). `qrcm`: *Quantile Regression Coefficients Modeling*. R package version 2.1, url: http://CRAN.R-project.org/package=qrcm

Frumento, P. (2018). `Mqrcm`: *M-Quantile Regression Coefficients Modeling*. R package version 1.0, url: http://CRAN.R-project.org/package=Mqrcm

Frumento, P., and Bottai, M. (2016). "Parametric modeling of quantile regression coefficient functions", *Biometrics*, **72**(1), 74–84, doi: 10.1111/biom.12410.

Frumento, P., and Bottai, M. (2017). "Parametric modeling of quantile regression coefficient functions with censored and truncated data", *Biometrics*, **73**(4), 1179–1188, doi: 10.1111/biom.12675.

Gilchrist, W. (2000). *"Statistical modeling with Quantile Functions"*, Chapman & Hall, ISBN 1-58488-174-7.

Giusti C., Marchetti S., Pratesi M. and Salvati N. (2012). "Robust Small Area Estimation and Oversampling in the Estimation of Poverty Indicators", *Survey Research Methods*, **3**(4), 155–163.

Hagenaars, A., de Vos, K. and Zaidi, M.A. (1994) *"Poverty statistics in the late 1980s: Research based on micro-data"*, Office for Official Publications of the European Communities. Luxembourg

Huber, P. J. (1981). *"Robust Statistics"*, John Wiley and Sons, New York.

Koenker, R. (2005). *"Quantile regression"*, Econometric Society Monograph Series, Cambridge: Cambridge University Press.

Koenker, R. and Bassett, G., Jr. (1978). "Regression Quantiles", *Econometrica* **46**, 33–50.

Newey, W. K. and McFadden, D. (1994). *Large Sample Estimation and Hypothesis Testing*. in R. F. Engle and D. L. McFadden (eds.), *Handbook of Econometrics* **4**, Ch. 36, 2111–2245, Handbooks in Econometrics, 2, North-Holland, Amsterdam.

Pratesi, M., Ranalli, M. G. and Salvati, N. (2008). "Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the northeastern US", *Environmetrics* **19**, 687–701.

Rao, J.N.K. and Molina, I. *Small Area Estimation*. 2nd ed. New York: Wiley.

Salvati, N., Tzavidis, N., Pratesi, M. and Chambers, R. (2012). "Small area estimation via M-quantile geographically weighted regression", *TEST* **21**, 1–28.

SAMPLE project Deliverable 17, (2010). "Pilot Applications", Available on the project website www.sample-project.eu.

Tzavidis, N., Salvati, N., Pratesi,M. and Chambers, R. (2008). "M-quantile models with application to poverty mapping", *Statistical Methods & Applications* **17**, 393–411.

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

Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge (Mass.): The MIT Press.