SAPIENZA
UNIVERSITÀ DI ROMA

Sapienza, University of Rome
Department of Statistical Sciences
Ph.D. in Methodological Statistic (XXXI cycle)

Mixed effect quantile and M-quantile regression for spatial data

Candidate:
Alberto Sabbi

Thesis advisor:
Prof. Marco Alfò

December 2020

Acknowledgments

I would mainly like to thank my Supervisor Prof. Marco Alfò for the support and advice he gave me on this journey. Many thanks also go to the reviewers, Prof. M. Giovanna Ranalli and Prof. Dr. Timo Schmid, for their useful suggestions on this thesis. I want to thank my travel friends too, without them this journey for me would not have been the same; among these, a special thanks goes to a friend who I had the pleasure to meet at the beginning of this path, Diego Chianella, with whom I shared the joys and sorrows of this experience.

A special thanks goes to Stefano Mugnoli, Ganluigi Salvucci, Damiano Abbatini, Stefania Lucchetti, Giovanni Lombardo, Fabio Lipizzi and Paola Giordano too, because they supported me during the whole period.

A further thanks I would like to address to Luca Salvati, historical friend, huge and inexhaustible source of knowledge always ready to help me in the most complicated moments.

I would also like to thank my mother for the constant support and affection she's daily shown to me.

Finally and not by order of importance I would like to thank the love of my life Francesca for the support and courage she gave me and gives me every day in this journey and our beautiful children Matilde and Giorgio, who make my life special.

Contents

1 Introduction

In several empirical studies, attention is focused on the analysis of the conditional mean of a response variable as a function of observed covariates/factors; this can be the base for the specification of an appropriate regression model. However, the use of such an approach could provide an incomplete analysis only and it may lead to unreliable conclusions when the assumptions for the linear regression model are not met.

Just to give an example, standard regression models may be influenced by the presence of outliers in the data, which could affect model parameter estimates.

The quantile regression model (QR) in the following, introduced by Koenker and Bassett (1978), was originally considered for its robustness features; it has raised great interest in the literature and success in various application areas since it can offer a more detailed picture of the response when compared to standard linear regression models.

The great advantage of quantile regression is the possibility to estimate the entire conditional distribution of the response variable and to study the influence of explanatory variables on the form of the distribution of Y.

This model has become highly consolidated thanks to its properties in terms of robustness against outliers, efficiency for a wide range of error distributions and equivariance to monotonic transformations.

To give more details, let us briefly remind the features of a standard linear regression model. The model aims at identifying the mean of Y conditional are a set of explanatory variables X which can be either quantitative (covariates) or qualitative (factors). On the other end, the quantile regression aims at identifying the quantile of Y at level $\tau \in (0,1)$, conditional are X = x.

Generally, a linear model is specified as follows:

$$y = X\beta + \varepsilon$$

where y is an (n,1) vector of observed values for the response, X is an (n,p) *design* matrix and β is a (p,1) vector of effects that need to be estimated, last, ε is an (n,1) measurement error vector.

Sometimes, the matrix X can be considered as a stochastic matrix including random data drawn from a distribution, but it is generally preferred to assume X values are fixed and

exogenous. As a starting point, it is assumed that the vector ε is a Gaussian rv with independent and identically distributed elements having mean 0 and constant variance σ^2.

Quantile regression (Koenker and Bassett, 1978; Koenker, 2005) represents a useful generalization of median regression whenever the interest is not limited to the estimation of a location parameter at the centre of the conditional distribution of the response variable but extends to location parameters (quantiles) at other parts of this conditional distribution.

Similarly, expectile regression (Newey and Powell, 1987) generalizes least squares regression at the centre of a distribution to estimation of location parameters at other parts of the target conditional distribution namely, expectiles.

Breckling and Chambers (1988) introduced M-quantile regression that extends the ideas of M-estimation (Huber, 1964; Huber and Ronchetti, 2009) to a different set of location parameters for that lie between quantiles and expectiles the response conditional distribution. M-quantile regression could be considered as a quantile-like generalisation of mean regression based on influence functions, or a combination of quantile and expectile regression (Newey and Powell, 1987). In fact, M-quantiles aim at combining the robustness properties of quantiles with the efficiency properties of expectiles.

The independence assumption, between the units used in the classic models, can not be always satisfied; for example, in the case of multilevel data where the observed sample is made of lower level units (pupils, time occasions) nested within higher level units (classrooms, individuals), referred to as clusters. In that context, in fact, we may think at lower level units (eg. pupils) from the same higher level unit (eg. class) as more similar to each other than to units from a different higher level unit.

In the case where, for example, the subjects have different means, but very similar within variability, we can think that the response y is obtained by adding to the linear predictor random intercepts that describe unobserved features specific to higher level units. In this case the intercept changes across subjects and it represents the effects of unobserved covariates specific to higher level units.

The use of cluster-specific effects may help us introduce a simple structure of association between observations from the same group/ cluster.

The notation should be slightly modified to account for a potential hierarchical structure; let y_{ij} i=1,…,m, j =1, …,n_i represent the value of the response variable for the j-th lower level unit within the i-th cluster unit. The corresponding design vector variable is denoted by x_{ij}.

Generalizing the previous argument, we may consider cluster-specific random effects b_i within a generalized linear model structure:

$$g\{E(y_i|x_i, b_i)\} = x_i'\beta + w_i'b_i$$

where w_i represents a subset of variables whose effects may vary across clusters according to a distribution $f_b(\cdot)$. It is usually assumed that $E(b|X) = 0$ to ensure identifiability of the corresponding elements in β, and $f_b(b_i|X_i) = f_b(b_i)$ implying exogeneity of observed covariates.

In this case, the likelihood function can not be analytically computed. The resulting integral, may be calculated by either Gaussian Quadrature (GQ), see Abramowitz and Stegun (1964) and Press et al. (2007), or adaptive Gaussian Quadrature (aGQ), see Liu and Pierce (1994), Pinheiro and Bates (1995); However, in both cases with a high computational effort. Monte Carlo and simulated ML approaches have been discussed as potential alternatives, see Geyer and Thompson (1992), McCulloch (1994) and Munkin and Trivedi (1999).

Rather than using a parametric distribution for the random effects, we may leave $f_b(\cdot)$ unspecified and approximate it using a discrete distribution on G <m locations $\{\xi_1, \dots, \xi_G\}$. The number of locations is bounded from above by the number of different higher level units profiles (see Aitkin, 1999).

This approach may be considered as a model-based clustering approach, where the population of interest is assumed to be divided into G homogeneous sub-populations which differ for the values of the regression parameter vector; this approach is therefore less parsimonious than a fully parametric one.

Observed data are frequently characterized by a spatial dependence; that is the observed values can be influenced by the "geographical" position. In such a context it is possible to assume that the values observed in a given area are similar to those recorded in neighboring areas. Such data is frequently referred to as spatial data and they are frequently met in epidemiological, environmental and social studies, for a discussion see Haining, (1990). Spatial data can be multilevel, with samples being composed of lower level units (population, buildings) nested within higher level units (census tracts, municipalities, regions) in a geographical area.

Green and Richardson (2002) proposed a general approach to modelling spatial data based on finite mixtures with spatial constraints, where the prior probabilities are modelled through a Markov Random Field (MRF) via a Potts representation (Kindermann and Snell, 1999, Strauss, 1977). This model was defined in a Bayesian context, assuming that the interaction parameter for the Potts model is fixed over the entire analyzed region. Geman

and Geman (1984) have shown that this class process can be modelled by a Markov Random Field (MRF). As proved by the Hammersley-Clifford theorem, modelling the process through a MRF is equivalent to using a Gibbs distribution for the membership vector. In other words, the spatial dependence between component indicators is captured by a Gibbs distribution, using a representation similar to the Potts model discussed by Strauss (1977).

In this work, a Gibbs distribution, with a component specific intercept and a constant interaction parameter, as in Green and Richardson (2002), is proposed to model effect of neighboring areas.

This formulation allows to have a parameter specific to each component and a constant spatial dependence in the whole area, extending to quantile and m-quantile regression the proposed by Alfò et al. (2009) who suggested to have both intercept and interaction parameters depending on the mixture component, allowing for different prior probability and varying strength of spatial dependence.

We propose, in the current dissertation to adopt this prior distribution to define a Finite mixture of quantile regression model (FMQRSP) and a Finite mixture of M-quantile regression model (FMMQSP), for spatial data.

The dissertation is structured into eight chapters. After an introduction to the thesis content the second chapter outlines the standard linear quantile regression model, discussing general properties such as robustness and equivariance and detailing the estimation method proposed by Koenker and Bassett (1978). The asymmetric Laplace distribution (ALD) is introduced to cast parameter estimation a maximum likelihood (ML). The third chapter describes the use of individual-specific random effects to model association between multi levels observations. Parametric Quantile and M-quantile regression models with random effects, proposed by Geraci and Botai (2007) and Tzavidis et al. (2016) respectively, are presented in this and the following chapter. Non-parametric models, specifically finite mixtures of quantile and M-quantile model are also presented.

The chapter 5 contains the main thesis proposal. The spatial dependence is included in the Quantile and the M-quantile model via the approach discussed by Green and Richardson (2002). This model uses finite mixture models with spatial constraints defined by Markov random field for the prior probabilities of component membership. Such an approach has been already used by Alfò et al. (2009) in the context of generalized linear models. We extend their proposal by applying this procedure to quantile and M-quantile regression models to give a more complete representation of the (conditional) response distribution.

The proposed methodology is evaluated in a Monte-Carlo simulation study in chapter 6. We compare different regression models in a series of different scenarios. While in the chapter 7 an application of the proposed models to a study on real data is described. Finally in chapter 8 we provide conclusions and suggestions for further research.

2 Standard Quantile Regression

In empirical studies attention is often focused on the analysis of the conditional mean for a response variable as a function of a set of observed covariates/factors in a linear regression framework. The use of such an approach could provide only an incomplete analysis and lead to unreliable conclusions when the assumptions of the linear regression model are not met. Standard regression models may also be influenced by the presence of outliers in the data, as these could heavily affect the estimation of model parameters.

The goal of standard regression models is to describe how the conditional expectation of a response variable changes as a function of a set of explanatory variables. However, considering the expected value as the only parameter of interest could not guarantee a reliable description of the impact of explanatory variables on the response; as we may be interested in features of such a distribution which could fruitfully described by appropriate quantiles.

In some empirical situations the major interest focuses on the tail of the distribution; for example, this is true in environmental context, where the limit value of a variable of interest (eg. Radon, PCBs, dioxins, heavy metals, etc.) may be regulated by law.

Quantile regression model (QR), introduced by Koenker and Bassett (1978), was originally appreciated for its robustness features, and it has been recently raising great interest in the literature in several application areas since it can offer a more detailed picture of the response when compared to standard regression model. The estimation of model parameters is pursued via a quite standard optimization problem. To formalize, let us introduce some notation; similarly to what happens for the calculation of the sample mean, defined as the solution to the problem of minimizing the sum of squared errors, we can define the quantile at level $\tau \in (0,1)$ as the solution of a simple minimization problem.

2.1 Definition of Quantile

The (conditional) quantile function is a basic concept which is worth to be introduced. Let us consider a (discrete or continuous) random variable, for a given level $\tau \in (0, 1)$, the τ-th quantile of X can be defined as the value $\gamma_\tau \in R$ such that

$$F^{-}(\gamma_\tau) \leq \tau \leq F(\gamma_\tau)$$

If X is continuous, the previous inequalities hold exactly and the quantile is univoquely defined. When if X is discrete, the above inequalities define a closed interval, and this implies that the quantile is not unique. To univoquely identify the quantile, we establish by convention that this is always the smallest element in the set of possible solutions. This definition can be formalized as follows

$$\gamma_\tau = min\{\gamma: F(\gamma) \geq \tau\}$$

The τ-quantile of X is a real number such that $Pr(X \leq x_\tau) \geq \tau$ and that $Pr(X \geq x_\tau) \geq 1 - \tau$. There is only one τ-quantile if the equation $F_x(x) = \tau$ has at most one solution, where $F_x(x) = Pr\ (X \leq x\ )$. corresponds to the cdf of X. We define the quantile function of X as an application $\tau \rightarrow F_X^{-1}(\tau)$ that associates a suitable τ-quantile of the random variable X to a number in the unit line $\tau \in [0, 1]$. Thus, the quantile function may be defined as follows :

$$Q_X(\tau) = F_X^{-1}(\tau) = min\{x \in \mathbb{R} | F_X(x) \geq \tau, 0 < \tau < 1\}$$

where $F_X(\cdot)$ represents the cdf of X.

This function provides the unconditional *τ-th* quantile of X, defined as the smallest value in the set of possible values that give a value of the cdf not lower that τ.

In other words, the quantile function expresses, for each level $\tau \in [0, 1]$, the value of the random variable x_τ such that $F_X(x_\tau) \geq \tau$ and $P(X > x_\tau) \leq 1 - \tau$. The quantile function is therefore defined as the inverse of the cdf:

$$\begin{cases} Q_X(x_\tau) = F_X^{-1}(\tau) \\ F_X(x_\tau) = Q_X^{-1}(x_\tau) = \tau \end{cases}$$

whenever the appropriate inverses exist. Koenker and Bassett (1978) proposed to estimate the quantiles by solving the optimization problem based on the following loss function:

$$L(\xi_\tau) = \sum_{i \in \{l | X_l \geq \xi_\tau\}} \tau |X_i - \xi_\tau| + \sum_{i \in \{l | X_l < \xi_\tau\}} (1 - \tau) |X_i - \xi_\tau| \qquad (2.1)$$

when the solution satisfies

$$\hat{\xi}_\tau = arg \min_{\xi_\tau \in R} L(\xi_\tau) \qquad (2.2)$$

The absolute differences between the observations and the optimal unknown value ξ_τ are weighted by τ for $X_i \geq \xi_\tau$, and by $(1 - \tau)$ for $X_i < \xi_\tau$ see Koenker (2005). If $\tau = 0.5$, the median is defined. The previously introduced equation can be defined, in a compact form through the introduction of the *check function*, defined as follows:

$$\rho_\tau(u) = u\big(\tau - \mathrm{I}(u < 0)\big) \quad 0 < \tau < 1$$

where $I(\cdot)$ represents the indicator function. The check function can be represented graphically as in Figure 2.1 below:

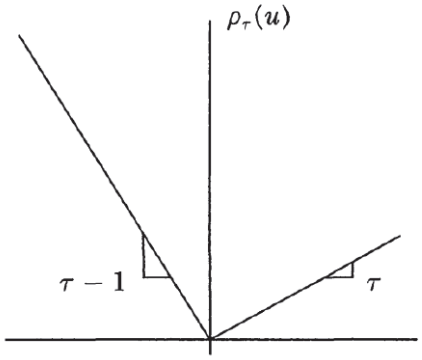

Figure 2.1 Check function

Based on such a definition, the loss function in equation (2.2) can be rewritten as follows

$$L(\xi_\tau) = \sum_i \rho_\tau(X_i - \xi_\tau) \tag{2.3}$$

and the conditional quantile is estimated as

$$\hat{\xi}_\tau = arg \min_{\xi_\tau \in R} L(\xi_\tau) \tag{2.4}$$

Considering the loss function in equation (2.1), we may notice that the solution may not be unique if nτ is not an integer (Koenker, 2005). The minimization of the loss function allows to identify the empirical quantile at any τ level. In the case X is a continuous variable, the expected loss function can be defined as:

$$E\big(\rho_\tau(X - \xi_\tau)\big) = \tau \int_{\xi_\tau}^{\infty} (x - \xi_\tau) dF_X(x) - (1 - \tau) \int_{-\infty}^{\xi_\tau} (x - \xi_\tau) dF_X(x) \tag{2.5}$$

The expected loss function can be minimized by setting equal to zero the first derivative; the first order condition is

$$\begin{aligned} \frac{dE\big(\rho_\tau(X - \xi_\tau)\big)}{d\xi_\tau} &= \tau \frac{\int_{\xi_\tau}^{\infty} (x - \xi_\tau) dF_X(x)}{d\xi_\tau} - (1 - \tau) \frac{\int_{-\infty}^{\xi_\tau} (x - \xi_\tau) dF_X(x)}{d\xi_\tau} \\ &= -\tau \int_{\xi_\tau}^{\infty} dF_X(x) + (1 - \tau) \int_{-\infty}^{\xi_\tau} dF_X(x) \\ &= -\tau\big(1 - F(\xi_\tau)\big) + (1 - \tau) F(\xi_\tau) \\ &= F_X(\xi_\tau) - \tau = 0 \end{aligned}$$

The second derivative is positive as $F_x(x)$ is non decreasing with $x \in R$; therefore the expected loss function is convex and it is minimized if and only if $F(\xi_\tau) = \tau$, i.e. $\xi_\tau = F^{-1}(\tau)$, as in equation (2.2).

The observation developed in this Section show that the quantiles may be expressed as the solution to a simple optimization problem; this leads to more general methods of estimating parameters via conditional quantile functions.

2.2 Quantile regression

The standard regression model specifies a model for the mean of the response Y conditional on a set of explanatory variables X. On the other hand, the quantile regression specifies the conditional distribution of Y, at each quantile $\tau \in (0,1)$ as a function of X.

As it as been previously noticed, the quantile regression provides a more accurate description of the conditional distribution through the evaluation of conditional quantiles. Let us assume that the quantile regression model is

$$Y_i = Q_\tau(Y_i|x_i) + \varepsilon_{i\tau} \quad \mathrm{i} = 1, \dots, \mathrm{n} \tag{2.6}$$

where Y_i denotes the response variable, x_i represent a vector of p explanatory variables and a constant term and ε_τ is an error term whose distribution varies with the quantile $\tau \in (0,1)$. The fundamental assumption of such regression model is that the τ- th conditional quantile of the error term is $Q_\tau(\varepsilon_{i,\tau}|x_i) = 0$

Therefore, based on such hypothesis, the τ- th quantile of the response Y_i conditional on the value x_i can be written as

$$Q_\tau(Y_i|x_i) = x_i'\beta_\tau \tag{2.7}$$

The parameter values β_τ may vary with τ, that is the quantile regression model is a model with varying parameters and this means that the effect of each term in the design vector may not be constant over the conditional response distribution Y|X. In this sense, a variable with no impact at the centre may well be thought of as having a substantial impact on the tails of $f_{Y|X}$. Given $\tau \in (0, 1)$ and considering equations (2.5), (2.6) and (2.7) the parameter vector β_τ can be estimated by solving the following problem

$$\hat{\beta}_\tau = \underset{\beta_\tau \in R^p}{\operatorname{argmin}} \left(\sum_{i \in \{i: y_i \geq x_i' \beta_\tau\}} \tau \left| y_i - x_i' \beta_\tau \right| + \sum_{i \in \{i: y_i < x_i' \beta_\tau\}} (1-\tau) \left| y_i - x_i' \beta_\tau \right| \right)$$

$$= \underset{\beta_\tau \in R^p}{\operatorname{argmin}} \sum_i \rho_\tau(y_i - x_i'\beta_\tau) \tag{2.8}$$

Median regression is a particular case of quantile regression with τ = 0.5 (Koenker, 2005).

Unlike least squares estimation, equation (2.8) does not lead to closed form solutions since the *check function* is not derivable at the origin. Rather the problem in equation (2.8) can be considered as a linear programming problem that can be solved numerically. As it regards the median regression problem, Barrodale and Roberts (1974) proposed an efficient simplex algorithm that was subsequently generalized for any conditional quantile by Koenker and D'Orey (1987). Such algorithms require a computational effort that make them rarely used. In fact, this algorithm is quite slow for data sets with a large number of observations (n>100'000). Another algorithm, the interior point algorithm, which also solves general linear programming problems, was introduced in this context by Portnoy and Koenker (1997). This algorithm shows great advantages in computational efficiency over the simplex algorithm for data sets with a large number of observations.

2.2.1 Parameter estimation

Let us consider a sample of n observations $(y_i, \underline{x}_i), i = 1, \dots, n;$ this can be rewritten in matrix form, where

$$\underline{Y}=[y_1,\dots,y_n]'$$

represents the response vector (with mutually independent terms), thought as composed by iid random variables Y_i, i=1,...,n.

The matrix of explanatory variables is

$$\underset{(nxp)}{X} = \begin{bmatrix} x_1^T \\ \cdot \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

The quantile regression model is based on the following hypothesis:

$$Y_i = Q_\tau(Y_i|x_i) + \varepsilon_{i\tau}$$
$$Q_\tau(Y_i) = x_i^T \beta_\tau \;\; i = 1, .., n$$
$$\varepsilon_{i\tau} \sim F \; iid:$$
$$Q\left(\varepsilon_{i\tau} | \, \underline{x}_i\right) = Q(\varepsilon_{i\tau}) = 0$$

The estimates for β_τ are obtained by solving the minimization problem:

$$\min_{\beta_\tau \in R^p} \left[\sum_{i \in \{i: y_i \geq x_i' \beta_\tau\}} \tau |y_i - x_i' \beta_\tau| + \sum_{i \in \{i: y_i < x_i' \beta_\tau\}} (1-\tau) |y_i - x_i' \beta_\tau| \right] \qquad (2.9)$$

As we have already mentioned, the solution $\hat{\beta}_\tau$ can not be written in a closed form expression as with the least squares estimate. However, the solutions to the problem equation (2.9) have some nice algebraic properties, as shown by Koenker and Bassett (1978).

First, a solution to the minimization problem above does always exist but it is not generally unique; for this purpose, let us denote by B_τ the set of solutions, and let:

- $\mathcal{L} = \{1, \dots, n\}$ be the set of indices for the sample observations;
- $\mathcal{H}_p$ be the set of all $\binom{n}{p}$ possible subsets $h = \{j_1, \dots, j_p\}$ of elements in $\mathcal{L}$ that can be obtained by selecting p different elements.

Let us consider the generic element $h \in \mathcal{H}_p$, and denote by:

$$\bar{h} = \mathcal{L} \setminus h$$

the complementary set to h, i.e. the set made by the (n-p) elements in $\mathcal{L}$ that do not belong to h. Let $X_{[h]}$ be the square matrix of order p obtained by selecting, the rows corresponding to indices $\{j_1, \dots, j_p\}$ from X and $Y_{[h]}$ the response vector for the same set of indices.

Last, denote by:

$$H_p = \{h \in \mathcal{H}_p : rank(X_{[h]}) = p\} \subseteq \mathcal{H}_p$$

the subset of $\mathcal{H}_p$ containing elements $h \in \mathcal{H}_p$ such that the matrix $X_{[h]}$ is non-singular. We can prove the following theorem

Theorem 1 **-**If the matrix X has full column rank, then the set B_τ of the possible solutions to the minimization problem in equation (2.9) has at least one element of the form:

$$\hat{\beta}_\tau = \left(X_{[h^*]}\right)^{-1} y_{[h^*]}, \qquad h^* \in \mathcal{H}_p$$

and B_τ is the convex hull of all the solutions having this form.

Theorem 1 states that, conditional on a set of hypotheses, there is an element $h^* \in H_p$ such that:

$$y_{[h^*]} = X_{[h^*]} \hat{\beta}_\tau$$

Therefore, from a graphical point of view, the quantile regression can be considered as a hyperplane passing exactly through (at least) p points out of the n observed.

The fact that the hyperplane equation for quantile regression is univocally determined by a subset of p observations, out of the n sample units, has raised some criticisms see eg. Koenker (2005), mainly regarding two aspects:

1. the estimator somewhat ignores a portion of the sample information;
2. if we consider the median regression (τ = 1/2), since the estimator is a linear function of a subset of sample units, under the conditions of the Gauss-Markov theorem, it cannot improve LS estimator in terms of efficiency.

Koneke and Bassett (1978), however, state that all sampling observations are used in the estimation process to determine the "optional" subset h^* containing the p points that determine the hyperplane equation for the quantile regression model. The estimator $\hat{\beta}_\tau$ is intrinsically non-linear if we consider the selection of the element $h^* \in H_p$.

Let us rearrange the rows of X and y in such a way that the first p rows correspond to the observations in the set h^* used to determine the hyperplane equation for the quantile regression model. We may therefore write the following equality:

$$\begin{bmatrix} y_{[h^*]} \\ y_{[\bar{h}^*]} \end{bmatrix} = \begin{bmatrix} X_{[h^*]} & 0 \\ X_{[\bar{h}^*]} & I \end{bmatrix} \begin{bmatrix} \hat{\beta}_\tau \\ \hat{\varepsilon}_{[\bar{h}^*]} \end{bmatrix} = \begin{bmatrix} X_{[h^*]}\hat{\beta}_\tau \\ X_{[\bar{h}^*]}\hat{\beta}_\tau + \hat{\varepsilon}_{[\bar{h}^*]} \end{bmatrix}$$

where $\hat{\varepsilon}_{[\bar{h}^*]}$ is the vector of residuals corresponding to the n-p observations of in the set $\bar{h}^*$, defined by:

$$\hat{\varepsilon}_{[\bar{h}^*]} = y_{[\bar{h}^*]} - X_{[\bar{h}^*]}\hat{\beta}_\tau = y_{[\bar{h}^*]} - X_{[\bar{h}^*]}X_{[h^*]}^{-1}y_{[h^*]}$$

The matrix 0 is a matrix contain null elements with size (p,n-p), while I denotes the identity matrix of order (n-p).

Comparing the estimator obtained via the median regression (τ = 1/2) with the least squares estimator, we have:

$$\hat{\beta}_{0.5} = \left(X_{[h^*]}\right)^{-1} y_{[h^*]} \qquad \text{Median Regression}$$

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y \qquad \text{Ordinary Least Squares Regression}$$

The two estimators show substantial differences:

1. the estimator for the parameter of the median regression model involves a subset of p observations drawn out of the n sample units. All the sample observations are implicitly used to establish which points determine the hyperplane equation. The ordinary least squares estimator is an explicit function of all the observations.

2. the set h^* of the observations characterizing the solution $\hat{\beta}_\tau$ varies, in general, with varying realizations of the error vector ε.

From a geometrical point of view, the ordinary least squares estimators is based on taking into consideration a linear projection $\hat{y} = X\hat{\beta}$ and minimizing the Euclidean distance $\|y - \hat{y}\|$. To characterize the estimator for the parameter vector in a quantile regression model, Koenker (2005) proposes to imagine inflating a ball centered at y until it touches the subspace spanned by X. The quantile regression $\rho_\tau(\cdot)$ dissimilarity measure

$$d_\tau(y, \hat{y}) = \sum_{i=1}^{n} \rho_\tau(y_i - \hat{y}_i)$$

can be compared to the shape of a diamond. Replacing Euclidean balls with polyhedral diamonds raises some new problems, but many nice features still persist. Expression (2.9) can be rewritten as follows:

$$\psi(\beta_\tau, \tau, y, X) = \sum_{[i:y_i \geq x_i'\beta_\tau]} \tau(y_i - x_i'\beta_\tau) + \sum_{[i:y_i < x_i'\beta_\tau]} (1-\tau)(y_i - x_i'\beta_\tau)$$

$$= \sum_{i=1}^{n} \left[\tau - \frac{1}{2} + \frac{1}{2} sgn(y_i - x_i'\beta_\tau)\right] [y_i - x_i'\beta_\tau]$$

where sgn denotes the sign function:

$$sgn(u) = \begin{cases} +1 & u > 0 \\ 0 & u = 0 \\ -1 & u < 0 \end{cases}, \qquad u \in \mathbb{R}^p$$

The minimization problem can therefore be characterized by the following theorem.

Theorem 2 (Bassett and Koenker, 1978***) -*** The value $\hat{\beta}_\tau = \left(X_{[h^*]}\right)^{-1} y_{[h^*]}$ is a unique solution to the Problem (2.9) if and only if:

$$(\tau - 1)\mathbb{1}_p \leq \sum_{i \in \bar{h}^*} \left[\frac{1}{2} - \frac{1}{2} sgn\left(y_i - x_i^T \hat{\beta}_\tau\right) - \tau\right] x_i^T \left(X_{[h^*]}\right)^{-1} \leq \tau \mathbb{1}_p$$

where $\mathbb{1}_p$ denotes a p dimensional column vector with unit elements.

Considering the directional derivative of the function ψ towards the direction w, it is possible to show that the aforesaid theorem (Bassett and Koenker, 1978) is proved.

2.2.2 Properties of solutions

Quantile regression should not only be considered as a useful tool for a more detailed description of the (conditional) distribution of a response variable; it has additional properties when compared to standard linear regression, such as the equivariance to monotone transformation of the dependent variable, the robustness to outlying values and efficiency (Koenker, 2005).

These properties had already been introduced in the original paper by Koenker and Bassett (1978) and we discuss them in details below.

2.2.2.1 Equivariance

Let $B_\tau = B(\tau; y, X)$ represent the set of solutions to the minimization problem (2.9) and $\hat{\beta}_\tau = \hat{\beta}(\tau; y, X)$ be an estimate for the parameters of the regression hyperplane for the τ-th conditional quantile, $\tau \in [0,1]$ then we have that:

i) $\hat{\beta}(\tau; \lambda y, X) = \lambda \hat{\beta}(\tau; y, X),$ $\lambda > 0$

ii) $\hat{\beta}(\tau; \lambda y, X) = \lambda \hat{\beta}(1-\tau; y, X),$ $\lambda < 0$

iii) $\hat{\beta}(\tau; y + X\gamma, X) = \hat{\beta}(\tau; y, X) + \gamma,$ $\gamma \in \mathbb{R}^p$

iv) $\hat{\beta}(\tau; y, XA) = A^{-1}\hat{\beta}(\tau; y, X)$ A (n,p) non singular matrix

The first property states that when all the response values are multiplied by a quantity $\lambda > 0$, then the solution will be subject to the same transformation. The second property shows that when if the response values are multiplied by a quantity $\lambda < 0$, then the parameter estimation for the regression hyperplane of order τ correspond to the coefficients of the regression hyperplane of order 1- τ, for the original values y where sign is changed due to multiplying by λ.

According to the third property, when a linear combination of the design matrix with coefficient γ is added to the vector of responses y, the solution corresponds to the sum of the solution for vector y plus γ.

The last property is called the equivariance to reparameterization of design and derives from the effect of a non-singular matrix A (pxp) introduced in the model. The recourse to reparameterization is quite common in regression analysis (von Eye and Schuster 1998) when the matrix of the explanatory variables is not of full column rank.

2.2.2.2 Equivariance under monotonic transformations

Another important feature is equivariance under monotonic transformations; let us remind the definition of the quantile function $Q_y(\tau)$, of a random variable Y, with distribution function $F_{Y|X}(\cdot)$

$$Q_y(\tau) = F_Y^{-1}(\tau) = inf\{y|F_Y(y) \geq \tau\} \quad 0 < \tau < 1$$

If $g(\cdot)$ is a strictly monotone increasing, continuous from the left function, we have that:

$$\tau = P[Y \leq Q_Y(\tau)] = P\big[g(Y) \leq g\big(Q_Y(\tau)\big)\big] \quad 0 < \tau < 1$$

In other words, the quantile function of the r.v. $g(Y)$, obtained by applying an increasing monotone function $g(\cdot)$ to Y, is given by $g\big(Q_Y(\tau)\big)$.

Similarly, for the conditional quantile function we have:

$$Q_{g(Y)|X}(\tau|x) = g\left(Q_{Y|X}(\tau|x)\right), \quad 0 < \tau < 1$$

This property is peculiar to the quantiles since, for example, it does not hold for the (conditional) mean

$$\mathbb{E}(g(Y)|x) \neq g\big(\mathbb{E}(Y|x)\big)$$

as equality holds only for specific forms of $g(\cdot)$, for example when $g(\cdot)$ is linear.

2.2.2.3 Distribution of the sign of residuals

There is a relation between the number of positive, negative and null residuals. Given a solution $\hat{\beta}_\tau \in B_\tau$ to the minimization problem in equation (2.9), the corresponding vector of regression residuals is defined by:

$$\hat{\varepsilon} = \mathrm{y} - \mathrm{X}\hat{\beta}_\tau$$

Let us consider a partition of the set of indices $\{1, \dots, n\}$ based on the sign of the corresponding residuals. Then let

- $Z_{\widehat{\beta}_\tau} = \{i: \hat{\varepsilon}_i = 0\}$ denote the set of indexes corresponding to points with a zero residual, with cardinality, n_0.
- $N_{\widehat{\beta}_\tau} = \{i: \hat{\varepsilon}_i < 0\}$ denote the set of indexes corresponding to points with a negative residual, with cardinality, n_-.
- $P_{\widehat{\beta}_\tau} = \{i: \hat{\varepsilon}_i > 0\}$ denote the set of indexes corresponding to points with a positive residual; with cardinality, n_+.

The three sets are disjoint and we can prove the following theorem.

Theorem 3 - If the design matrix X contains a column of 1's (i.e. the regression hyperplane contains the intercept), then:

$$n_- \leq n\tau \leq n - n_+ = n_- + n_0$$

If $\hat{\beta}_\tau$ is the unique solution to the minimization in equation (2.9) then the inequality holds strictly.

2.2.2.4 Robustness

As we have previously mentioned, one of advantages of the quantile regression model when compared to standard regression is robustness to outliers. The robustness of solution to outlying values can be characterized by the following theorem.

Theorem 4 - If $\hat{\beta} \in B(\tau; y, X)$, where $B(\tau; y, X)$ represent the set of solutions to the minimization problem (2.9), then $\hat{\beta} \in B\left(\tau; X\hat{\beta} + D\hat{\varepsilon}, X\right)$ where D is a (nxn) diagonal matrix with non-negative elements and $\hat{\varepsilon} = y - X\hat{\beta}(\tau; y, X)$.

The theorem states that a perturbation in y leaving the sign of the residuals unchanged leaves also the solution to the minimization problem unchanged.

In geometrical terms, this property means that the solution does not change if you "move" the points by varying the corresponding Y values until these ones remain on the same side of the hyperplane, i.e. the sign of the corresponding residuals does not change.

2.2.3 Parameter estimation using the Asymmetric Laplace distribution

A method to cast estimation for quantile regression in a parametric context, eg. ML approach is based on adopting the asymmetric Laplace distribution (see Geraci and Bottai, 2007). In this case, the optimization problem in equation (2.9) is equivalent to estimate parameter via optimizing the likelihood function based on the ALD.

Let us consider data in the form (x_i, y_i), i = 1,…,n , where y_i are independent scalar observations of a continuous response variable with common cdf $F_y(\cdot\,|x)$, whose shape is not exactly known, and x_i are design vectors X. Linear conditional quantile functions are defined by:

$$Q(\tau|x_i) = x_i^T\beta, \quad i = 1, \ldots., N \qquad (2.10)$$

where $\tau \in (0,1)$, $Q(\cdot) \equiv F_{y_i}^{-1}(\cdot)$, $\beta_\tau \in \mathbb{R}^P$ is an unknown vector of parameters. As we have previously remarked, the parameter estimates in τ-quantile regression, represent the solution to the following minimization problem:

$$L_\tau(\beta) = \min_{\beta \in R^p} \left\{ \sum_{i \in (i: y_i \geq x_i^T \beta)}^{n} \tau \left| y_i - x_i^T \beta \right| + \sum_{i \in (i: y_i < x_i^T \beta)}^{n} (1-\tau) \left| y_i - x_i^T \beta \right| \right\} \qquad (2.11)$$

and, the estimate $\hat{\beta}$ will clearly depend on the value τ.

Koenker and Machado (1999) and Yu and Moyeed (2001) have introduced the asymmetric Laplace density (ALD), using it as a parametric distribution that help recast the minimization of the sum of absolute deviations into a maximum likelihood into framework.

Given a response Y with an ALD (μ, σ, τ) density, $Y \sim ALD(\mu, \sigma, \tau)$, the individual contribution to the likelihood function is given by

$$f(y|\beta, \sigma) = \frac{\tau(1-\tau)}{\sigma} exp\left\{ -\rho_\tau \left(\frac{y - \mu_\tau}{\sigma} \right) \right\} \qquad (2.12)$$

where $\rho_\tau(\upsilon) = \upsilon\big(\tau - I(\upsilon \leq 0)\big)$ is the loss function, I (.) denotes the indicator function, τ∈ (0, 1) is the asymmetry parameter (skewness), σ> 0 and $\mu_\tau \in \mathbb{R}$ are the scale and the location parameters. The loss function $\rho(\cdot)$ assigns weight τ or 1 –τ to observations that are respectively, higher or lower than μ_τ, with by $Pr(y \leq \mu_\tau) = \tau$. Therefore, the distribution of Y is divided by the location parameter μ_τ into two parts, one on the left associated to a weight τ and one on the right with (1-τ), see Yu and Zhang (2005).

Let us set $\mu_{\tau i} = x_i^T \beta_\tau$ and $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_\mathrm{n})$. Assuming that $y_i \sim ALD(\mu_i, \sigma, \tau)$ the likelihood from a sample of n independent observations is

$$L(\beta, \sigma; y, \tau) \propto \sigma^{-1} exp\left\{ -\sum_{i=1}^{n} \rho_\tau \left(\frac{y_i - \mu_{i\tau}}{\sigma} \right) \right\}$$

If we consider σ as nuisance parameter, the maximization of the above mentioned likelihood $L(\beta_\tau, \sigma; y, \tau)$ with respect to parameter β_τ is equivalent to the minimization of the objective function $L_\tau(\beta)$.

Thus, the ALD is useful as a bridge between the likelihood and the non parametric framework for estimation of model parameters in a linear quantile regression model.

2.3 M-quantile regression

The M-quantile regression model integrates the expectile regression (a generalization of the standard regression see Newey and Powell, 1987) and the quantile regression by Koenker and Bassett (1978) into a single framework. The integration of these two modelling approaches help define a quantile-type generalization of robust regression

estimated via influence functions (M-quantile regression). M-estimation is a method, based on the use of influence functions, introduced by Huber (1973) to guarantee robustness of parameter estimates to outliers. It controls the effect of outliers by limiting the effect of those points with a residual greater than a given threshold c.

The M-Quantile regression (MQ) of order τ for a response with conditional density $f(y|x)$, introduced by Breckling and Chambers (1988), is defined as the solution to the following estimating equation:

$$\int \psi_\tau \left(y - MQ_\tau(y|x;\psi)\right) f(y|x) dy = 0$$

where ψ_τ is a (potentially asymmetric) influence function, corresponding to the first derivative of a (potentially asymmetric) loss function ρ_τ, $\tau \in (0, 1)$. In the case of a linear M-Quantile regression model, we assume that:

$$MQ_\tau(y|x;\psi) = x_i'\beta_\tau \qquad (2.13)$$

which correspond to the assumption $MQ_\tau(\varepsilon|x;\psi) = MQ(\varepsilon|\psi) = 0$, where ε is the measurement error. The estimates of β_τ are obtained by minimizing

$$\sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta_\tau) \qquad (2.14)$$

By specifying the form of the asymmetric loss function $\rho_\tau(\cdot)$, it is possible to obtain the standard regression model (ρ_τ quadratic and τ=0.5), the linear expectile regression model (ρ_τ quadratic and $\tau \neq 0.5$, see Newey and Powell, 1987), and the quantile regression model, if the loss function introduced by Koenker and Bassett (1978) is used.

For M quantile regression, we choose to adopt the Huber loss function (Breckling and Chambers, 1988) defined by:

$$\rho_\tau(u) = \begin{cases} 2c|u| - c^2\{\tau I(u > 0) + (1-\tau) I(u \leq 0)\}, & if\ |u| > c \\ u^2\{\tau I(u > 0) + (1-\tau) I(u \leq 0)\}, & if |u| \leq c \end{cases}$$

where $I(\cdot)$ represents the indicator function and $c \in \mathbb{R}_+$ denotes a tuning constant. The value assumed by this constant plays a fundamental role in the estimation process; in fact, this value can be used to balance robustness and efficiency in the MQ regression model. With c tending to zero, robustness increases, but efficiency decreases (in this case we are moving towards quantile regression); with large and positive c robustness decreases and efficiency increases (as we are moving towards *expectile* regression).

Setting the first derivative of (2.14) equal to zero leads to the following estimating equations:

$$\sum_{i=1}^{n} \psi_\tau\,(r_{i\tau})x_i = 0$$

where $r_{i\tau} = y_i - MQ_\tau(y_i|x_i;\psi)$ is the residual, and:

$$\psi_\tau(r_{i\tau}) = 2\psi(s^{-1}r_{i\tau})\{\tau I(r_{i\tau} > 0) + (1-\tau)I(r_{i\tau} \le 0)\}$$

Here, s>0 represents the scale parameter. In the case of robust regression, it is often estimated by $\hat{s} = median|r_{i\tau}|/0.6745$. Since the focus of this paper is on M-type estimation, we use as influence function the so called Huber Proposal 2:

$$\psi(u) = uI(-c \le u \le c) + c\,sgn(u)I(|u| > c)$$

Provided that the tuning constant c is strictly greater than zero, the estimates of $\beta\tau$ can be obtained using an iterative by weighted least squares algorithm, IWLS (Kokic et al, 1987).

Quantiles have a more intuitive interpretation than M-quantiles, even if they both target essentially the same part of the distribution of interest (Jones, 1994). It should be emphasized that M-quantile estimation offers some advantages:

i. it easily allows to robustly estimate regression parameters;
ii. it can trade robustness for efficiency in inference by varying the choice of the *tuning* constant c in the influence function;
iii. it offers computational stability as we can use a wide range of continuous influence functions instead of L_1 norm used in the quantile regression context (Tzavidis et al., 2016).

The asymptotic theory for M-quantile regression with i.i.d. errors and fixed regressors can be derived from the results in Huber (1973), see Breckling and Chambers (1988). Bianchi and Salvati (2015) prove the consistency of the estimator of β_τ, provide its asymptotic covariance matrix when regressors are stochastic and propose a variance estimator for the M-quantile regression coefficients based on a sandwich approach.

3 Quantile regression with random effects

Generally, a linear regression model with constant effects is defined as follows:

$$y = X\beta + \varepsilon \tag{3.1}$$

where y is an $(n,1)$ vector of observed response values, $X \in M(n,p)$ is a *design* matrix, β is the corresponding vector of parameters to be estimated and ε is an $(n,1)$ measurement error vector.

While this is not strictly necessary, it is often assumed that the vector ε is normally distributed with mean 0 and constant variance σ^2, with independent and identically distributed iid elements. The choice for a parametric distribution makes the theory, for the sample distribution of parameter estimates, more easily to be applied; the only really necessary hypothesis are $E(\varepsilon|x) = E(\varepsilon|x) = 0$ and $var(\varepsilon|x) = var(\varepsilon) = \sigma^2$.

The independence assumption cannot be always satisfied, as in the case of multilevel or hierarchically structured data where the observed sample is composed by lower level units (pupils, temporal occasions, results) nested within higher level units (classrooms, individuals, questionnaires), usually referred to as clusters.

For example, let us consider a test carried out on patients who have been administered a given treatment with the aim at evaluating how this can affect patients' symptoms. By taking measurements on each subject at different times, corresponding to the days following treatment administration, we implicitly use the patient as the sample unit with measurements corresponding to time occasion nested within patients.

From a model specification perspective, the measurements referring to the same subject cannot be assumed to be independent from each other, while subjects can be still considered independent.

When the subjects have different mean, but a very similar variability, we can think that the response y is obtained by adding to the linear predictor individual-specific intercepts that describe the individual characteristics at the study start or at a baseline measurement. The intercept changes across subjects and help us explain the differences between subjects that cannot be explained by observed covariates. In this case, model (3.1) can be rewritten as follows:

$$Y = D\alpha + X\beta + \varepsilon$$

where $\alpha = (\alpha_1, \dots, \alpha_n)$ represents the vector of individual specific intercepts, referring to individual-specific deviations from the overall intercept term in $X\beta$, $D = I_n \otimes \mathbb{1}_T$, and individual and time indexes are denoted by i=1,...,n and t=1,...,T.

3.1 Linear and Generalized Linear Mixed Models

A number of study designs, such as those derived by multilevel, longitudinal and cluster sampling, typically require the application of ad hoc statistical methods to take into account the association between observations belonging to the same unit or cluster. To analyse this type of complex data, we can use popular and flexible models referred to as mixed-effect models. By means of cluster-specific effects, they account for variability between clusters, while fixed effects are usually included to account for variability in the response within clusters.

By generalizing the standard regression models, we assume that a set of observations grouped into m clusters have been recorded for the response and the design vector. Let us denote the size of group i by n_i (i = 1, ..., m). The linear mixed models (LMMs) see eg. McCulloch and Searle (2000), Searle et al. (1992), Verbeke and Molenberghs (2000) is defined by

$$y_{ij} = x_{ij}^T \beta + w_{ij}^T b_i + \varepsilon_{ij} \tag{3.2}$$

where y_{ij} is the response value for the j-th (lower level) unit in the i-th clusters j=1,…, n_i, i=1 ,…, m; x_{ij} is the corresponding p-dimensional design vector β is the fixed parameter vector and w_{ij} is a q< p-dimensional vector. Usually $w_{ij} \subseteq x_{ij}$ so that b_i can be thought of as individual-specific effects measuring, for a given covariate, the individual specific deviation from the corresponding element in the fixed parameter vector β. Last, ε_{ij} is the measurement error vector. Last, b_i is the individual specific effect, and we assume that $b \sim N(0, \sigma_b^2)$, $\epsilon_{ij} \sim N(0, \sigma^2)$ $b_i \perp \varepsilon_{ij}$, we may estimate parameters in (3.2) via maximum likelihood. Assuming normality for the error components, $cov(b_j, b_{j'}) = 0$ $j \neq j'$, so if we assume, for simplicity sake, that $b_i \sim MVN_q(0, \Sigma_b)$, we derive for the marginal likelihood the following expression (Harville, 1979).

$$l(\beta, \sigma_b^2, \sigma^2) = -\frac{1}{2} log|V| - \frac{1}{2}(y - X\beta)' V^{-1}(y - X\beta) \tag{3.3}$$

where y is the response vector, $V = \Sigma \ + W\Sigma_b W'$, $\Sigma \ = \sigma^2 I_n$, $\Sigma_b = \sigma_b^2 I_m$, W is an (n,m) matrix of known positive constants. Parameter estimates are obtained by solving the

estimating equations obtained by differentiating the log-likelihood with respect to the parameters and setting these derivatives equal to zero (Goldstein, 2003).

The sensitivity of ML parameter estimates to assumptions upon the random effects distribution has been the focus of a huge literature, see Rizopoulos et all. (2009), McCulloch and Neuhaus (2005) among others.

We notice that the poss function in equation (3.3) is quadratic. This loss function is based on Gaussian assumptions; however, the presence of outliers may produce inefficient and biased parameters estimates (Richardson and Welsh, 1995).

One approach to robustifying the mixed effects model against departures from normality is to use an alternative loss function, growing at slower rate than quadratic.

A robust estimation method has been followed by Huggins (1993), Huggins and Loesch (1998), Richardson and Welsh (1995), and Welsh and Richardson (1997). This approach consists in replacing, in the log-likelihood function, the quadratic loss function by a new function that grows with the residuals but at a slower rate.

The log-likelihood function, for such a model, becomes:

$$l(\beta, \sigma_b^2, \sigma^2) = -\frac{K_1}{2} log|V| - \rho(r)$$

where r denotes a scaled residual defined by $\mathrm{r} = \mathrm{V}^{-\frac{1}{2}}(\mathrm{y} - \mathrm{X}\beta)$, $\rho(\cdot)$ is a continuous loss function with derivative denoted by $\psi(\cdot)$, $\mathrm{K}_1 = \mathrm{E}[\epsilon\psi(\epsilon)']$ is a correction factor for consistency, $\varepsilon \sim \mathrm{N}(0, \mathrm{I}_\mathrm{n})$, and the terms $\psi(\mathrm{r})$ and $\mathrm{r}^{\mathrm{T}}\psi(\mathrm{r})$ are assumed to be limited. This is the robust maximum likelihood proposal I by Richardson and Welsh (1995).

Richardson and Welsh (1995) proposed a further alternative, suggesting to solve the estimating equation for σ_b^2, σ^2 derived in the context of robust maximum likelihood estimation

$$\frac{1}{2}\psi(r')V^{-\frac{1}{2}}WW'V^{-\frac{1}{2}}\psi(r) - \frac{K_2}{2}tr(V^{-1}WW') = 0 \qquad (3.4)$$

where $K_2 = E[\psi(\varepsilon)\psi(\varepsilon)']$, $\varepsilon \sim N(0, I_n)$. Richardson and Welsh (1995) called this robust maximum likelihood proposal II. It can be viewed as a generalization of Huber's proposal II (Huber, 1981).

3.1.1 The semiparametric case

As it has been noted above, in real life problems, observations are often organized in the form of hierarchical data; the potential association between dependent observations should be considered to provide valid and efficient inferences. The use of individual-specific random effects may help us introduce a simple structure of association between observations.

Let us consider a set of individual-specific random effects b_i in a generalized linear model:

$$y_{ij}|x_{ij}, b_i \sim LEF(\theta_{ij})$$

$$\theta_{ij} = g\{E(y_{ij}|x_{ij}, b_i)\} = x'_{ij}\beta + w'_i b_i$$

where the assumption already discussed before hold.

Based on the local independence assumption and, if needed after a Mundlak (1978) type connection, the likelihood function may be written as follows:

$$L(\Phi) = \prod_{i=1}^{m} \left\{ \int_{\mathcal{B}} \prod_{j=1}^{n_i} f(y_{ij}|x_{ij}, b_i)\, f_b(b_i|X_i) db_i \right\} = \prod_{i=1}^{m} \left\{ \int_{\mathcal{B}} \prod_{j=1}^{n_i} f(y_{ij}|x_{ij}, b_i)\, f_b(b_i) db_i \right\}$$

where Φ is the global parameters vector. The terms b_i i=1,...,m account for individual-specific heterogeneity common to each lower-level units within the same i-th cluster.

Generally, the likelihood function does not have a closed form and, to calculate the previous integral, we need to use either Gaussian Quadrature (GQ), Abramowitz and Stegun (1964) and Press et al. (2007), adaptive Gaussian Quadrature (aGQ) Liu and Pierce (1994), Pinheiro and Bates (1995), or other approximation approaches. Monte Carlo and simulated ML approaches are potential alternatives, even if the cost could be even higher.

In the case of finite samples and for reduced individual sequences, these methods may either not provide a good approximation or be inefficient as the impact of missing information (ie. b_i i=1,...,m) increase.

3.1.2 Non parametric random effects

Rather than using a parametric distribution for the random effects, we may leave $f_b(\cdot)$ unspecified, and approximate it using a discrete distribution on G≤m locations $\{\xi_1, \dots, \xi_G\}$:

$$\pi_k = Pr(b_i = \xi_k)\,; \;\; b_i \sim \sum_{k=1}^{G} \pi_k \delta_{\xi_k} \qquad i = 1, \dots, m\,; k = 1, \dots, G \tag{3.5}$$

where $\delta_Q(\cdot)$ is a function that puts a unit of mass at Q. Using such an approximation, the likelihood function can be rewritten as

$$L(\Phi) = \prod_{i=1}^{m}\left\{\sum_{k=1}^{G}\prod_{j} f\left(y_{ij}\middle|x_{ij},\xi_k\right)\pi_k\right\} =: \prod_{i=1}^{m}\left\{\sum_{k=1}^{G}\prod_{j} f_{ijk}\pi_k\right\} \qquad (3.6)$$

where $\Phi = \{\beta, \xi_1, \dots, \xi_G, \pi_1, \dots, \pi_G\}$ is the global parameter vector and $f_{ijk} = f\left(y_{ij}\middle|x_{ij},\xi_k\right)$ is the distribution of the response for the j-th measurement in the i-th cluster drawn from the k-th component of the finite mixture, k = 1, ..., G.

While the above distribution is based on Non Parametric Maximum Likelihood (NPML) theory by Laird (1978), Simar (1976), Böhning (1982) and Lindsay (1983 a, b), the previous equation (3.6) can be also motivated by a model-based clustering approach, where the population of interest is (assumed to be) divided into G homogeneous sub-populations which differ for the values of the regression parameters only. The number of unknown parameters to be estimated is higher than in the corresponding parametric model. In fact, both $\xi_k$and π_k k=1,...,G are unknown parameters, and G itself is unknown, even if it is usually considered as fixed and estimated through appropriate penalized likelihood criteria. The seminal papers by Aitkin (1996, 1999) establish a connection between mixed-effect models and finite mixtures.

When using discrete individual-specific coefficient, the regression model can be expressed, in the k-th component of the mixture as follows:

$$g\{E\left(y_{ij}\middle|x_{ij},\xi_k\right)\} = x'_{ij}\beta + w'_{ij}\xi_k$$

The score function is:

$$S(\Phi) = \frac{\partial \log[L(\Phi)]}{\partial \Phi} = \frac{\partial \ell(\Phi)}{\partial \Phi} = \sum_{i=1}^{m}\sum_{k=1}^{G}\left(\frac{f_{ik}\pi_k}{\sum_l f_{il}\pi_l}\right)\sum_j \frac{\partial \log f_{ijk}}{\partial \Phi} =: \sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik}\sum_j \frac{\partial \log f_{ijk}}{\partial \Phi}$$

where the weights

$$\omega_{ik} = \frac{\prod_j f_{ijk}\pi_k}{\sum_l \prod_j f_{ijl}\pi_l} \; i = 1, \dots, m, K = 1, \dots, G$$

represent the posterior probability of component membership. The score function is just a sum of the likelihood equations for a standard GLM with weights ω_{ik}. The log-likelihood function can be directly maximized, or indirectly maximized through and EM-type algorithms. The basic EM algorithm is defined by solving equations for a given set of the weights, and updating the weights according to the current parameter estimates see Aitkin (1999) for details.

3.2 Linear quantile mixed models

Extension of standard QR models to multilevel or hierarchical data has led to several distinct approaches based on individual specific effects. These can be classified into two types: parametric and non-parametric.

The latter family includes fixed effect (Koenker 2004, Lamarche 2010) and weighted effect (Lipsitz et al., 1997) models. The former is based on the use of the asymmetric Laplace density (ALD) with individual-specific effects having a Gussian (Geraci and Bottai 2007, Liu and Bottai 2009; Yuan and Yin 2010; Lee and Neocleous 2010; Farcomeni 2012) or other parametric distributions (Reich et al. 2010). The two families are not mutually exclusive; just to give an example, the penalization method suggested by Koenker (2004) may, as noted by Geraci and Bottai (2007), be considered as based on the asymmetric Laplace density with individual-specific effects.

Let us consider multilevel data in the form (y_{ij}, x_{ij}) i=1,...,m, j=1,…,n_i, where x_{ij} denotes a p-dimensional design vector andy_{ij} is the j-th value of a continuous random variable measured on subject (cluster) i.

In a fixed effect framework, Koenker (2004) proposed to consider the following optimization problem with a penalty term λ

$$\min_{\alpha,\beta} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \omega \rho_\tau \left(y_{ij} - x_{ij}^T \beta - b_i\right) + \lambda \sum_{i=1}^{m} |b_i| \qquad (3.7)$$

where λ is the penalization parameter, ω is the weight regulating the influence of individual-specific effects b_i on the τ-th quantile. $\beta\tau$ summarizes the impact of the observed covariates on the τ-th quantile for an individual whose baseline level is equal to b_i. The dependence between observations from the same individual is not accounted, even if the term $\lambda \sum_{i=1}^{m} |b_i|$ resembles a log-density.

The penalization parameter λ must be arbitrarily set and its choice may heavily influence inference on β_τ; therefore, this choice is a fundamental issue. Using a penalized approach, estimating a τ-distributional individual effect would be impracticable (Geraci and Bottai, 2007).

Since the results achieved by adopting this method depend on the choice of the parameter λ, we need to select a suitable value. Lamarche (2010) proposed a method to select the

penalty term; as this term influences asymptotic variance, it can be selected by minimizing the trace of the estimated asymptotic covariance matrix.

Geraci and Bottai (2007) proposed to avoid these issue by using random individual-specific effects, defining the linear quantile mixed model (LQMM). Let b_i represent a q-dimensional vector of individual-specific parameters; the quantile function depends on x_{ij} as follows (see Yu Zhu et al, 2016, Geraci and Bottai, 2007, 2014 and Liu and Bottai, 2009):

$$Q_{y_{ij}|b_i}(\tau|x_{ij}, b_i) = x'_{ij}\beta + w'_i b_i, \qquad j = 1, \dots, n_i, \quad i = 1, \dots, m$$

where w_i represents a subset of x_{ij} associated to individual-specific effects b_i varying across individuals according to a distribution $f_b(\cdot)$.

If we assume that, conditional on b_i, y_{ij} i = 1, ..., m and j = 1,…,n_i are independently random variables with Asymmetric Laplace density, we obtain:

$$f(y_{ij}|\beta, b_i, \sigma) = \frac{\tau(1-\tau)}{\sigma} exp\left\{-\rho_\tau\left(\frac{y_{ij} - \mu_{ij}}{\sigma}\right)\right\}$$

where $\mu_{ij} = Q_{y_{ij}|b_i}(\tau|x_{ij}, b_i)$ represents the location parameter for the τ-th quantile (Geraci e Bottai, 2007), and $\tau \in (0,1)$ is a fixed known value.

Dependence between observations from the same subject (cluster) is introduced in the model by the individual specific considered as iid random variables. To complete assumptions on individual effects we will denote the corresponding density by f_b, indexed by a parameter ϕ_τ which may depend on τ. Finally, we assume that ε_{ij} and b_i are independent on X and each other. If b_i and x_i are dependent a Mundlak -type approach can be used.

Given the individual sequence y_i=(y_{i1},…,y_{ini}) and assuming local independence, the conditional (on b_i) density for the joint individual sequence is

$$f(y_i|\beta, b_i, \sigma) = \prod_{j=1}^{n_i} f(y_{ij}|\beta, b_i, \sigma)$$

The density for the complete data (y_i, b_i) can be rewritten as follows:

$$f_{y,b}(y_i, b_i|\Phi) = f_{y|b}(y_i|\beta, b_i, \sigma) f_b(b_i|\sigma_b) \qquad (3.8)$$

i=1,…,m, where $f_b(b_i|\sigma_b)$ denotes the density for the individual-specific effects b_i and $\Phi = (\beta, \sigma, \sigma_b)$ is the "global" vector of parameters. The marginal density based on the assumption of independence of higher-level cluster is:

$$f(y, b|\Phi) = \prod_{i=1}^{m} f_Y(y_i|\beta, b_i) f_b(b_i|x_i)$$

3.2.1 Parameter estimation

By integrating out the random effects, based on the exogeneity assumption $f_b(b_i|X_i) = f_b(b_i)$ or of the controlling for the linear effect of x_i on b_i , we obtain the marginal distribution for the individual sample of the response :

$$f(y_i|\Phi) = \int_{R^N} f_{Y,b}(y_i, b_i|\Phi)\, db_i \tag{3.9}$$

Inference on the parameter vector $\Phi = (\beta, \sigma, \sigma_b)$ is based on the marginal likelihood, defined by

$$l(\eta; y) = \sum_{i}^{m} log\ f(y_i|\Phi).$$

In general, however, the integral in equation (3.9) that does not have a closed form.

Therefore, Geraci and Bottai (2007) propose to estimate model parameters through a Monte Carlo EM algorithm, often applied in the context of LMMs (Meng and Van Dyk, 1998; Booth and Hobert, 1999).

Let us consider a generic individual; the E-step at the (t + 1) th iteration can be used to calculate the (conditional on observed data) expectation of the complete data log-likelihood:

$$\begin{aligned} Q_i\left(\Phi \middle| \Phi^{(t)}\right) &= E\left\{l_c(\Phi; y_i, b_i) \middle| y_i; \Phi^{(t)}\right\} \\ &= \int \{\log f(y_i|\beta, b_i, \sigma) + \log f(b_i|\sigma_b)\} f\left(b_i \middle| y_i, \Phi^{(t)}\right) db_i \end{aligned} \tag{3.10}$$

where $\Phi^{(t)} = \left\{\beta^{(t)}, \sigma^{(t)}, \sigma_b^{(t)}\right\}$ is the vector of current parameter estimates and $l_c(\cdot)$ is the complete data log-likelihood. The expected value is expressed with respect to the missing data distribution conditional on the observed data y and the current parameter estimates. The approximation of the expected value of the complete data log-likelihood can be carried out using a suitable Monte Carlo approach.

If we consider the random effects as known, the LQMM model is specified by

$$\hat{L}_{\tau i}(\beta) = \sum_{j \in \left(j:\hat{y}_{ij} \geq x_{ij}^T \beta\right)}^{n_i} \tau \left|\hat{y}_{ij} - x_{ij}^T \beta\right| + \sum_{j \in \left(j:\hat{y}_{ij} < x_{ij}^T \beta\right)}^{n_i} (1-\tau) \left|\hat{y}_{ij} - x_{ij}^T \beta\right|$$

where $\tilde{y}_{ij} = y_{ij} - b_i$. By applying the usual linear programming algorithm, we may update β by the following minimization

$$\min_{\beta \in R^p} \tilde{L}_\tau(\beta), \quad \tilde{L}_\tau(\beta) = \sum_{i=1}^{m} \tilde{L}_{\tau i}(\beta)$$

Since the random effects are not observed, Geraci and Bottai (2007) propose to take a sample $b_i = (b_{i1}, \ldots, b_{iM})$ of size M form the posterior density of the random effects conditional on the observed data

$$f(b_i|y_i, \Phi) \propto f(y_i|\beta, b_i, \sigma) f(b_i|\sigma_b)$$

and to approximate the expected value by

$$Q_i^*\left(\Phi \middle| \Phi^{(t)}\right) = \frac{1}{M} \sum_{k=1}^{M} l_c(\Phi; y_i, b_{ik})$$

An iterative procedure can then be applied to obtain the estimates for Φ, with fixed τ. The standard errors of parameter estimates, can be obtained through a bootstrap approach by considering the matrix (yi,Xi) as the basic resampling unit (Lipsitz et al., 1997), as it is usual for longitudinal data.

3.2.2 The choice of the random individual-specific distribution

The choice of the distribution for the individual-specific random effects has been largely debated. The use of a Gaussian distribution in the linear mixed models framework considerably simplifies the analytical form of the expectation of the complete data log-likelihood function that is needed when using the EM algorithm. We may also note that, for a fixed covariance matrix, β can be estimated via GLS. This simplification allows to reduce the time to convergence.

Assuming that the individual-specific effects are independent and identically distributed random variates, the joint individual density can be rewritten as:

$$f(y_i, b_i|\Phi) = f(b_i|\sigma_b)\prod_{j=1}^{n_i} f\left(y_{ij}\middle|\beta, b_i, \sigma\right)$$

$$= \left\{\frac{\tau(1-\tau)}{\sigma}\right\}^{n_i} \frac{1}{\sqrt{2\pi\sigma_b}} exp\left[-\sum_{j=1}^{n_i}\left\{\rho_\tau\left(\frac{y_{ij}-\mu_{ij}}{\sigma}\right)\right\} - \frac{1}{2\sigma_b} b_i^2\right]$$

In the case of a strong asymmetry we may assume that $b_i \sim ALD(0, \sigma_b, \tau)$, where τ defines the degree of asymmetry of the individual-specific random effect distribution, and the joint density for the i-th unit is equal to

$$f(y_i, b_i|\Phi) = f(b_i|\sigma_b)\prod_{j=1}^{n_i} f\left(y_{ij}\middle|\beta, b_i, \sigma\right)$$

$$= \left\{\frac{\tau(1-\tau)}{\sigma}\right\}^{n_i} \frac{1}{4\sigma_b} exp\left[-\sum_{j=1}^{n_i}\left\{\rho_\tau\left(\frac{y_{ij}-\mu_{ij}}{\sigma}\right)\right\} - \frac{1}{2\sigma_b}|b_i|\right]$$

In the case of a median regression, fixing τ=0.5 and λ = σ / σ_b, the joint density can be rewritten as:

$$f(y_i, b_i|\Phi) = \frac{1}{(4\sigma_b)^{n_i+1}\lambda^{n_i}} exp\left[-\frac{1}{2\sigma}\left\{\sum_{j=1}^{n_i}\left(\left|y_{ij}-\mu_{ij\tau}\right|\right) + \lambda|b_i|\right\}\right]$$

It is worth noting that the exponential argument has a form similar to that of the penalized quantile regression (PQR) proposed by Koenker (2004) for quantile regression applied to multilevel data, based on a fixed effect type estimator.

3.2.3 Fimite Mixtures of Quantile Regression

Let us consider hierarchical data with repeated measurements in the form $\left(y_{ij}, x'_{ij}\right)$ for j=1,...,n_i, and i= 1,...,m, where y_{ij} is the j-th value of a continuous on response the i-th subject, x'_{ij} j are row vectors of a known design matrix and β is a (p,1) vector of fixed regression coefficients. We follow the notation in the linear-mixed model and define the linear mixed-effects quantile function for response y_{ij} as:

$$y_{ij} = x'_{ij}\beta + w'_{ij}b_i + \varepsilon_{ij}$$

where w_{ij} is a (q,1) subset of x_{ij} with individual-specific effects; b_i is a (q,1) vector of random regression coefficients; the error term ε_{ij} j=1,...,n_i, and i= 1,...,m, is assumed to be independently distributed as ALD. We also assume independence between b_i and ε_{ij} and

between the random regression coefficients b_i and the explanatory variables x'_{ij}, see Mundlak (1978).. The conditional density function of $y_{ij}|b_i$ can be written as:

$$f(y_{ij}|b_i, x_{ij}; \beta, \sigma) = \frac{\tau(1-\tau)}{\sigma} exp\left\{-\rho_\tau\left(\frac{y_{ij} - \mu_{ij\tau}}{\sigma}\right)\right\}$$

where $\mu_{ij\tau} = x'_{ij}\beta_\tau + w'_{ij}b_i$ is the linear predictor for the τ-th quantile function, and $\tau \in (0,1)$ is fixed and known.
Let

$$f(y_i|b_i, x_i; \beta, \sigma) = \prod_{j=1}^{n_i} f(y_{ij}|b_i, x_{ij}; \beta, \sigma)$$

be the density for the i-th subject conditional on the random effect b_i. The complete data density (y, b) is

$$f(y|b, x; \Phi) = \prod_{i=1}^{m}\prod_{j=1}^{n_i} f(y_{ij}|b_i, x_{ij}; \Phi) f_b(b_i|x_i; \Phi) \qquad (3.11)$$

where we assume that $f_b(b_i|x_i; \Phi)$ the is density of b_i, $\Phi = \{\beta, \sigma, \Sigma\}$ is the vector of parameters of interest.

We obtain the estimates for Φ by maximising the marginal density $f(y|x; \Phi)$, which is calculated by integrating out the random effects b in equation (3.11). That is,

$$L(\Phi) = \int f(y|b, x; \Phi)\, f_b(b_i|x_i; \Phi) db.$$

As discussed above this integral is frequently intractable. Then, using a non parametric distribution for the random effects, we may approximate it using a discrete distribution on G <m locations $\{\xi_1, \ldots, \xi_G\}$ as we have showen for GLLMs in section 3.1.2.

In this case, as we have commented before, the likelihood function is

$$L(\Phi) = \prod_{i=1}^{m}\left\{\sum_{k=1}^{G}\prod_j f(y_{ij}|x_{ij}, \xi_k)\pi_k\right\} =: \prod_{i=1}^{m}\left\{\sum_{k=1}^{G}\prod_j f_{ijk}\pi_k\right\}$$

$$= \prod_{i=1}^{m}\sum_{k=1}^{G}\prod_j \pi_k \frac{\tau(1-\tau)}{\sigma_k} exp\left\{-\rho_\tau\left(\frac{y_{ij} - \mu_{ijk}}{\sigma_k}\right)\right\}$$

where $\mu_{ijk} = x'_{ij}\beta_\tau + w'_{ij}\xi_k$ and $\Phi = \{\beta, \xi_1, \ldots, \xi_G, \pi_1, \ldots, \pi_G\}$ is the global parameter vector, f_{ijk} is the AL distribution for the j-th measurement in the i-th cluster drawn from the k-th component of the finite mixture, k = 1, ..., G.

We discuss the EM the algorithm for ML estimation in the general case of multilevel data, with j = 1,..,n_i measurements corresponding to i=1,…,m upper-level units, and a given quantile $\tau \in (0,1)$.

Let us denote by $z_{ik,\tau}$, the variable indicating whether the i-th unit comes from the k-th component of the mixture when the τ-th quantile is considered. Each component of the mixture is characterized by a different vector of regression parameters, $\xi_{k,\tau}$, k = 1, …, G. Therefore, we may write:

$$\pi_k = Pr\left(Z_{ik,\tau} = 1\right) = Pr\left(b_{i,\tau} = \xi_{k,\tau}\right)$$

while the remaining parameters β_τ and σ_τ are constant across the components. In a complete data setting, we would have observed the couple $(y_i, z_{i,\tau})$ where $z_{i,\tau}=(z_{i1,\tau},\ldots,z_{iG,\tau})$, and the log-likelihood function would therefore be equal to:

$$\ell_c(\Phi_\tau) = \sum_{i=1}^{m}\sum_{k=1}^{G} z_{ik,\tau}\left\{\log f_\tau\left(y_i\middle|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right) + \log\left(\pi_{k,\tau}\right)\right\} =$$

$$= \sum_{i=1}^{m}\sum_{k=1}^{G} z_{ik,\tau}\left\{\log\left[\prod_{j=1}^{n_i} f_\tau\left(y_{ij}\middle|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right)\right] + \log\left(\pi_{k,\tau}\right)\right\} =$$

$$= \sum_{i=1}^{m}\sum_{k=1}^{G} z_{ik,\tau}\left\{\sum_{j}^{n_i}\log\left[f_\tau\left(y_{ij}\middle|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right)\right] + \log\left(\pi_{k,\tau}\right)\right\} \qquad (3.12)$$

where, as used, $\Phi = \{\beta, \xi_1, \ldots, \xi_G, \pi_1, \ldots, \pi_G\}$ represents the global of parameters vector for $\tau \in (0,1)$.

As we have remarked above, in the E-step of the EM algorithm, at the (t+1)th iteration, we compute the expected value of the complete data log likelihood over the indicator vector $z_{ik,\tau}$, conditional on the observed data y_i and the current parameter estimates $\widehat{\Phi}_\tau^{(t)}$. Therefore, at the t+1 iteration of the algorithm we compute

$$\omega_{ik,\tau}^{(t+1)} = \frac{\pi_{k,\tau}^{(t)} f_{ik,\tau}\left(\widehat{\Phi}_\tau^{(t)}\right)}{\sum_l \pi_{l,\tau}^{(t)} f_{il,\tau}\left(\widehat{\Phi}_\tau^{(t)}\right)}$$

i=1,…,m and k=1,…,G. The conditional expectation of the complete data log-likelihood for multilevel data, given the observed data y_i and the current parameters $\widehat{\Phi}_\tau^{(t)}$, is:

$$Q\left(\Phi_\tau\middle|\widehat{\Phi}_\tau^{(t)}\right) = E_{\widehat{\Phi}_\tau^{(t)}}[\ell_c(\Phi_\tau)|y_i] = \sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t+1)}\left\{\sum_{j}^{n_i}\log\left[f_\tau\left(y_{ij}\middle|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right)\right] + \log\left(\pi_{k,\tau}\right)\right\}$$

Maximizing the function Q (.) with respect to Φ_τ, we obtain updated maximum likelihood estimates $\widehat{\Phi}_\tau^{(t+1)}$ given the posterior probability $\omega_{ik;\tau}^{(t+1)}$, through the likelihood equations:

$$\frac{\partial Q\left(\Phi_\tau \middle| \widehat{\Phi}_\tau^{(t)}\right)}{\partial \Phi_\tau} = 0$$

From the likelihood point of view, the Fimite Mixtures of Quantile Regression (FMQR) model is equivalent to a finite mixture of regression models with errors having on asymmetric Laplace distribution (ALD), which can be regarded as an extension of the traditional mixtures of regression models.

4 M-quantile regression with random effects

Tzavidis et al. (2010, 2016) extended M-quantile regression to multilevel data by considering cluster-specific effects to account for the hierarchical structure of observed data. In the simplest case, an individual-specific intercept b_i is added to the M-quantile model:

$$MQ_{y_{ij}}(\tau|x_{ij}, b_i; \psi) = x'_{ij}\beta_\tau + w'_{ij}b_i \tag{4.1}$$

Tzavidis et al. (2016) propose the use of estimating equations based on asymmetric loss functions to estimate model parameters, thus defining MQRE regression.

To obtain the estimating equations, robust maximum likelihood estimation is applied by asymmetrically weighting the residuals (Sinha and Rao, 2009), using Huber proposal II:

$$\begin{gathered} X'V_\tau^{-1}B_\tau^{\frac{1}{2}}\psi_\tau(r_\tau) = 0 \\ \frac{1}{2}\psi_\tau(r_\tau)'B_\tau^{\frac{1}{2}}V_\tau^{-1}WW^TV_\tau^{-1}B_\tau^{\frac{1}{2}}\psi_\tau(r_\tau) - \frac{K_{2\tau}}{2}tr(V_\tau^{-1}WW') = 0 \\ \frac{1}{2}\psi_\tau(r_\tau)'B_\tau^{\frac{1}{2}}V_\tau^{-1}V_\tau^{-1}B_\tau^{\frac{1}{2}}\psi_\tau(r_\tau) - \frac{K_{2\tau}}{2}tr(V_\tau^{-1}) = 0 \end{gathered} \tag{4.2}$$

where $r_\tau = B_\tau^{-\frac{1}{2}}(y - X\beta_\tau)$ is the vector of scaled residuals with components $r_{ij\tau}$, $B_\tau = diag(V_\tau)$ and

$$V_\tau = \Sigma_\tau + W\Sigma_{b\tau}W', \Sigma_\tau = \sigma_\tau^2 I_n, \Sigma_{b\tau} = \sigma_{b\tau}^2 I_m$$

where σ_τ^2 and $\sigma_{b\tau}^2$ are the MQ specific variance parameters and

$$K_2 = E[\psi_\tau(\varepsilon)\psi_\tau(\varepsilon)'],\ \varepsilon \sim N(0, I_n).$$

To solve these equations, Tzavidis et al. (2016) proposed the joint use of a Newton-Raphson algorithm and an iterative fixed point method (Anderson, 1973). Fixed effects are estimated by using the Newton-Raphson algorithm, while variance parameters are estimated by using the fixed point method. This approach allows to prevent convergence problems during the estimation of variance components, which are frequently observed with the use of Newton-Raphson algorithms.

4.1 Non parametric individual-specific effects

Starting from the work of Tzavidis et al. (2016) Borgoni et al. (2019) introduce a M-quantile model with bivariate splines used to model spatial dependence, representing a non parametric alternative to usual individual-specific effects.

In the case of data whose behaviour is non linear, we may add a flexible component to the linear predictor of the M-quantile model as proposed by Borgoni et al (2019). In the form of a penalization, penalized splines are effective tools for a number of reasons. First, they are reasonably simple to implement, as they represent a relatively straightforward extension of a linear term. Second, their flexibility allow to include wide range of modelling features, eg. to account, as in the present case, for spatial dependence.

Splines refer to piecewise polynomials defined using a set of basis functions to handle non-linear structures in the data; for example the spatial pattern of a variable of interest can be explained as a function of the location of the points, represented by their coordinates. A bivariate smoothing spline can be included in the model specification to represent spatial dependence in terms of a set of bivariate basis functions. Following Ruppert et al. (2003), Pratesi et al. (2009) suggested the use of radial basis functions to derive low-rank thin plate splines.

In this context, the M-quantile model in equation *(4.1)* can be rewritten:

$$MQ_y(\tau|X, b, ; \psi) = X\beta_\tau + Wb_\tau + W_{sp}\gamma_\tau \qquad (4.3)$$

where X is a design matrix, β_τ is the vector of fixed parameters, b_τ is an individual-specific parameter vector (cluster-specific), W is a matrix containing elements in x associated to varying effects; W_{sp} is a spline matrix and K is the number of spline knots. More specifically (Opsomer et al. 2008),

$$W_{sp} = \left[C\left(u_i - k_j\right)\right]_{1\le i\le K}^{1\le i\le n}\left[C(k_i - k_k)\right]_{1\le i,k\le K}^{-1/2}$$

where k_j and k_i , j=1,...,K, i=1,...,K, are two-dimensional vectors representing the cartographic coordinates of knots i and j. u_i is a two-dimensional vector representing the cartographic coordinates of sampling location i and $C(s) = \|s\|_2^2 log\|s\|_2$ where $s \in \mathbb{R}^2$ and $\|s\|_2$ is the Euclidean norm of s in $\mathbb{R}^2$.

Model fit is a practical advantage of the representation offered by spline based models. The usual penalised spline-fitting criterion requires estimating a penalising or smoothing parameter prior to model estimation. Cross-validation is generally a way to tackle this problem. Using mixed models avoids this step as the model can be directly estimated using the appropriate routines for linear mixed models. Moreover, including random

coefficients for the spline basis components allows to account for the bias due to omitted variables. Treating the coefficients of the knots as random leads to a smoother representation of the estimated effect, compared to using fixed effects specification, and avoids data overfitting (Ruppert et al., 2003).

The estimating equations are based on extending the idea of asymmetric weighting of the residuals in equation *(4.2)* (Tzavidis et al. 2016; Borgoni et al. 2018):

$$X'V_\tau^{-1}B_\tau^{\frac{1}{2}}\psi_\tau(r_\tau) = 0 \qquad (4.4)$$

$$\frac{1}{2}\psi_\tau(r_\tau)'B_\tau^{\frac{1}{2}}V_\tau^{-1}WW^TV_\tau^{-1}B_\tau^{\frac{1}{2}}\psi_\tau(r_\tau) - \frac{K_{2\tau}}{2}tr(V_\tau^{-1}WW') = 0$$

$$\frac{1}{2}\psi_\tau(r_\tau)'B_\tau^{\frac{1}{2}}V_\tau^{-1}V_\tau^{-1}B_\tau^{\frac{1}{2}}\psi_\tau(r_\tau) - \frac{K_{2\tau}}{2}tr(V_\tau^{-1}) = 0$$

where $r_\tau = B_\tau^{-\frac{1}{2}}(y - X\beta_\tau)$ is the vector of scaled residues with components $r_{ij\tau}$, B_τ is the diagonal matrix with main diagonal elements $b_{ij\tau}$ equal to the components placed on the diagonal of covariance, $\psi_\tau(r_\tau)$ the derivative of a loss function ρ_τ

$$V_\tau = \Sigma_{\varepsilon_\tau} + W\Sigma_{b_\tau}W' + W_{sp}\Sigma_{\gamma_\tau}W_{sp}', \Sigma_{\varepsilon_\tau} = \sigma_{\varepsilon_\tau}^2 I_n, \Sigma_{\gamma_\tau} = \sigma_{\gamma_\tau}^2 I_K, \Sigma_{b_\tau} = \sigma_{b_\tau}^2 I_m$$

where $\sigma_{\varepsilon_\tau}^2$, $\sigma_{\gamma_\tau}^2$ and $\sigma_{b_\tau}^2$ are the MQ specific variance parameters and

$$K_2 = E[\psi_\tau(\varepsilon)\psi_\tau(\varepsilon)'],\ \varepsilon \sim N(0, I_n).$$

To solve equations (4.4) Tzavidis et al. (2016) propose to use a Newton-Raphson algorithm and an iterative fixed point method (Anderson, 1973). In particular, fixed effects are estimated by using the Newton-Raphson algorithm, while variance parameters are estimated by using the fixed point method. This approach allows to prevent convergence problems during the estimation of variance components, which are frequently observed with the use of the Newton-Raphson algorithm.

4.2 Finite mixtures of M-quantile regressions

Alfò et al. (2017) extended the M-quantile models with individual-specific effects starting from the semiparametric approach we have discussed in section 2.3. Where a discrete distribution for the effects is considered.

Yu and Moyeed (2001) show the relationship between the loss function for quantile regression and the maximization of a likelihood function based on a Asymmetric Laplace density.

Bianchi et al. (2015) propose a similar alternative for the M-quantile regression. This proposal is based on the assumption that the loss functions belong to the large class of continuously differentiable convex functions, with special attention the Huber to loss function. It is possible to show that minimizing the Huber loss function (Huber, 1964) is equivalent to maximizing a likelihood function based on the Asymmetric Least Informative (ALID) density, see Bianchi et al. (2015). A further extension has been proposed by Alfò et al. (2017),where the ALID is used and finite mixtures of M-quantile models are fitted using the EM algorithm.

The ALID distribution (Asymmetric Least Informative Density) proposed by Bianchi et al. (2015) has the following form:

$$\mathrm{f}_\tau(\cdot) = \frac{1}{\mathrm{B}_\tau(\sigma_\tau)} \exp\{-\rho_\tau(\cdot)\} \qquad (4.5)$$

where $B_\tau(\sigma_\tau)$ is a normalizing constant

$$B_\tau(\sigma_\tau) = \sigma_\tau\sqrt{\frac{\pi}{\tau}}\left[\Phi\left(c\sqrt{2\tau}\right) - \frac{1}{2}\right] + \sigma_\tau\sqrt{\frac{\pi}{1-\tau}}\left[\Phi\left(c\sqrt{2(1-\tau)}\right) - \frac{1}{2}\right] + \frac{\sigma_\tau}{2c\tau} exp\{-c^2\tau\} + \frac{\sigma_\tau}{2c(1-\tau)} exp\{-c^2(1-\tau)\}$$

In the M-quantile model

$$MQ_\tau\left(y_{ij}|x_{ij}, \xi_{k,\tau}; \psi\right) = x'_{ij}\beta_\tau + w'_{ij}\xi_{k,\tau}$$

The distribution of b_i may vary with quantiles. As previously stated, we assume that, without loss of generality, the following equality holds:

$$\sum\nolimits_k \pi_k \xi_{k,\tau} = 0 \; \forall\tau$$

which mimic $E(b_i|x_i) = E(b_i) = 0$

To estimate parameters β_τ e ξ_k, we maximize the log-likelihood function:

$$\ell(\Phi) = \sum_{i=1}^{m} \log\left\{\sum_{k=1}^{G}\prod_{j} f_\tau(y_{ij}|x_{ij},\xi_k)\pi_k\right\} =: \sum_{i=1}^{m} \log\left\{\sum_{k=1}^{G}\prod_{j} f_{ijk}\pi_k\right\}$$

where $f^{\tau}_{Y|b}$ denotes the ALID density. The score function components are:

$$S(\beta_\tau)\sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}\sum_{j=1}^{n_i}\psi_\tau\left(y_{ij}-x'_{ij}\beta_\tau-w'_{ij}\xi_{k,\tau}\right)x_{ij}=0$$

$$S\left(\xi_{k,\tau}\right)\sum_{i=1}^{m}\omega_{ik,\tau}\sum_{j=1}^{n_i}\psi_\tau\left(y_{ij}-x'_{ij}\beta_\tau-w'_{ij}\xi_{k,\tau}\right)w_{ij}=0$$

where

$$\omega_{ik,\tau}=\frac{\pi_{k,\tau}f_{ik,\tau}}{\sum_l \pi_{l,\tau}f_{il,\tau}}=\frac{\pi_{k,\tau}\prod_j f_{ijk,\tau}}{\sum_l \pi_{l,\tau}\prod_j f_{ijl,\tau}}$$

can be interpreted as the posterior probability for the i-th higher level unit (cluster) from the k-th component of the mixture at quantile τ. Here $f_{ijk,\tau}$ represents the ALID distribution coming for the k-th component and the τ-th quantile.

We discuss the EM the algorithm for ML estimation in the general case of multilevel data, with j = 1,..,n_i measurements corresponding to i=1,…,m upper-level units, and a given quantile τ∈ (0,1).

Let us denote by $z_{ik,\tau}$, the variable indicating whether the i-th unit comes from the k-th component of the mixture when the τ-th quantile is considered. Each component of the mixture is characterized by a different vector of regression parameters, $\xi_{k,\tau}$, k = 1, …, G. Therefore, we may write:

$$\pi_k = Pr\left(Z_{ik,\tau}=1\right)=Pr\left(b_{i,\tau}=\xi_{k,\tau}\right)$$

while the remaining parameters β_τ and σ_τ are constant across the components. In a complete data framework, for each higher level unit we would have observed the couple $(y_i,z_{i,\tau})$ where $z_{i,\tau}=(z_{i1,\tau},\ldots,z_{iG,\tau})$, and the log-likelihood function would be equal to:

$$\ell_c(\Phi_\tau)=\sum_{i=1}^{m}\sum_{k=1}^{G} z_{ik,\tau}\left\{\log f_\tau\left(y_i|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right)+\log\left(\pi_{k,\tau}\right)\right\}=$$

$$=\sum_{i=1}^{m}\sum_{k=1}^{G} z_{ik,\tau}\left\{\log\left[\prod_{j=1}^{n_i} f_\tau\left(y_{ij}|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right)\right]+\log\left(\pi_{k,\tau}\right)\right\}= \tag{4.6}$$

$$= \sum_{i=1}^{m}\sum_{k=1}^{G} z_{ik,\tau}\left\{\sum_{j}^{n_i}\log\left[f_\tau\left(y_{ij}\middle|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right)\right]+\log\left(\pi_{k,\tau}\right)\right\}$$

where $\Phi = \{\beta, \xi_1, \ldots, \xi_G, \pi_1, \ldots, \pi_G\}$ represents the global of parameters vector for $\tau \in (0,1)$.

As we have remarked above, in the E-step of the EM algorithm, at the (t+1)th iteration we compute the expected value of the complete data log likelihood over the indicator vector $z_{ik,\tau}$, conditional on the observed data y_i and the current parameter estimates $\widehat{\Phi}_\tau^{(t)}$. Therefore, at the t+1 iteration of the algorithm we compute

$$\omega_{ik,\tau}^{(t+1)} = \frac{\pi_{k,\tau}^{(t)} f_{ik,\tau}\left(\widehat{\Phi}_\tau^{(t)}\right)}{\sum_l \pi_{l,\tau}^{(t)} f_{il,\tau}\left(\widehat{\Phi}_\tau^{(t)}\right)}$$

i=1,…,m and k=1,…,G. The conditional expectation of the complete data log-likelihood for multilevel data, given the observed data y_i and the current parameters $\widehat{\Phi}_\tau^{(t)}$, is:

$$Q\left(\Phi_\tau\middle|\widehat{\Phi}_\tau^{(t)}\right) = E_{\widehat{\Phi}_\tau^{(t)}}[\ell_c(\Phi_\tau)|y_i] = \sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t+1)}\left\{\sum_{j}^{n_i}\log\left[f_\tau\left(y_{ij}\middle|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right)\right]+\log\left(\pi_{k,\tau}\right)\right\}$$

Maximizing the function Q (.) with respect to Φ_τ, we obtain updated maximum likelihood estimates $\widehat{\Phi}_\tau^{(t+1)}$ given the posterior probability $\omega_{ik;\tau}^{(t+1)}$:

$$\frac{\partial Q\left(\Phi_\tau\middle|\widehat{\Phi}_\tau^{(t)}\right)}{\partial \Phi_\tau} = 0$$

the updates are obtained by solving likelihood equation that can be written separately for each element in Φ:

$$S(\beta_\tau) = \sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t)}\sum_{j}^{n_i}\frac{\partial}{\partial\beta_\tau}\log\left[f_\tau\left(y_{ij}\middle|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right)\right]$$

$$= -\sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t)}\sum_{j}^{n_i}\frac{\partial\rho_\tau\left(\left(y_{ij}-x_{ij}'\beta_\tau-w_{ij}'\xi_{k,\tau}\right)/\sigma_\tau\right)}{\partial\beta_\tau}$$

$$= \sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t)}\sum_{j}^{n_i}\psi_\tau\left(\frac{y_{ij}-x_{ij}'\beta_\tau-w_{ij}'\xi_{k,\tau}}{\sigma_\tau}\right)\frac{x_{ij}}{\sigma_\tau}$$

$$S\left(\xi_{k,\tau}\right) = \sum_{i=1}^{m}\omega_{ik,\tau}^{(t)}\sum_{j}^{n_i}\psi_\tau\left(\frac{y_{ij}-x_{ij}'\beta_\tau-w_{ij}'\xi_{k,\tau}}{\sigma_\tau}\right)\frac{w_{ij}}{\sigma_\tau}$$

$$S(\sigma_\tau) = \sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t)}\sum_{j}^{n_i}\frac{\partial}{\partial\sigma_\tau}\log\left[f_\tau\left(y_{ij}\middle|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right)\right] =$$

$$= \sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t)}\sum_{j}^{n_i}\frac{\partial}{\partial\sigma_\tau}\left\{-\log B_\tau(\sigma_\tau) + \log\left[f_\tau\left(y_{ij}\middle|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right)\right]\right\} =$$

$$= \sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t)}\sum_{j}^{n_i}\left[-\frac{1}{\sigma_\tau} + \psi_\tau\left(\frac{y_{ij}-x_{ij}'\beta_\tau - w_{ij}'\xi_{k,\tau}}{\sigma_\tau}\right)\left(\frac{y_{ij}-x_{ij}'\beta_\tau - w_{ij}'\xi_{k,\tau}}{\sigma_\tau^2}\right)\right]$$

$$S(\pi_k) = \sum_{i=1}^{m}\left[\frac{\omega_{ik,\tau}^{(t)}}{\pi_k} - \frac{\omega_{iG,\tau}^{(t)}}{\pi_G}\right] = \sum_{i=1}^{m}\left[\frac{\omega_{ik,\tau}^{(t)}}{\pi_k} - \frac{1-\sum_{l=1}^{G-1}\omega_{il,\tau}^{(t)}}{1-\sum_{l=1}^{G-1}\pi_l}\right], k = 1,\dots,G-1$$

Regardless of the (conditional) density adopted for the response, the expression in $S(\pi_k)$ leads to the standard updates:

$$\hat{\pi}_k^{(t)} = \frac{\omega_{ik,\tau}^{(t)}}{m} \quad k = 1,\dots,G-1$$

The updated estimates for the remaining model parameters depend on the specific conditional density form we choose.

In the case where $f_\tau(\cdot)$ corresponds to the asymmetric least informative distribution (ALID), the M-step updates are calculated by an IWLS algorithm, which corresponds to a weighted version of the standard algorithm for M-quantile regression.

4.2.1 The provision of standard errors

To provide the standard error for model parameter estimates, we need first to compute the Hessian for Φ:

$$He(\beta_\tau,\beta_\tau) = -\sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t)}\sum_{j}^{n_i}\psi_\tau'\left(\frac{y_{ij}-x_{ij}'\beta_\tau - w_{ij}'\xi_{k,\tau}}{\sigma_\tau}\right)\frac{x_{ij}}{\sigma_\tau}\frac{x_{ij}'}{\sigma_\tau}$$

$$He\left(\beta_\tau,\xi_{k,\tau}\right) = -\sum_{i=1}^{m}\omega_{ik,\tau}^{(t)}\sum_{j}^{n_i}\psi_\tau'\left(\frac{y_{ij}-x_{ij}'\beta_\tau - w_{ij}'\xi_{k,\tau}}{\sigma_\tau}\right)\frac{x_{ij}}{\sigma_\tau}\frac{w_{ij}'}{\sigma_\tau}$$

$$He(\beta_\tau,\sigma_\tau) = -\sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t)}\left[\sum_{j}^{n_i}\psi_\tau'\left(\frac{y_{ij}-x_{ij}'\beta_\tau - w_{ij}'\xi_{k,\tau}}{\sigma_\tau}\right)\left(\frac{y_{ij}-x_{ij}'\beta_\tau - w_{ij}'\xi_{k,\tau}}{\sigma_\tau^2}\right)\frac{x_{ij}}{\sigma_\tau}\right.$$

$$\left. + \psi_\tau\left(\frac{y_{ij}-x_{ij}'\beta_\tau + w_{ij}'b_{k,\tau}}{\sigma_\tau}\right)\frac{x_{ij}}{\sigma_\tau^2}\right]$$

$$He\left(\xi_{k,\tau},\xi_{k,\tau}\right)=-\sum_{i=1}^{m}\omega_{ik,\tau}^{(t)}\sum_{j}^{n_i}\psi_\tau'\left(\frac{y_{ij}-x_{ij}'\beta_\tau-w_{ij}'\xi_{k,\tau}}{\sigma_\tau}\right)\frac{w_{ij}}{\sigma_\tau}\frac{w_{ij}'}{\sigma_\tau}$$

$$He\left(\xi_{k,\tau},\sigma_\tau\right)=-\sum_{i=1}^{m}\omega_{ik,\tau}^{(t)}\left[\sum_{j}^{n_i}\psi_\tau'\left(\frac{y_{ij}-x_{ij}'\beta_\tau-w_{ij}'\xi_{k,\tau}}{\sigma_\tau}\right)\left(\frac{y_{ij}-x_{ij}'\beta_\tau-w_{ij}'\xi_{k,\tau}}{\sigma_\tau^2}\right)\frac{w_{ij}}{\sigma_\tau}\right.$$

$$\left.+\psi_\tau\left(\frac{y_{ij}-x_{ij}'\beta_\tau+w_{ij}'\xi_{k,\tau}}{\sigma_\tau}\right)\frac{w_{ij}}{\sigma_\tau^2}\right]$$

$$He(\sigma_\tau,\sigma_\tau)=\sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t)}\sum_{j}^{n_i}\left[\frac{1}{\sigma_\tau^2}-\psi_\tau'\left(\frac{y_{ij}-x_{ij}'\beta_\tau-w_{ij}'\xi_{k,\tau}}{\sigma_\tau}\right)\left(\frac{y_{ij}-x_{ij}'\beta_\tau-w_{ij}'\xi_{k,\tau}}{\sigma_\tau^2}\right)^2\right.$$

$$\left.-2\psi_\tau\left(\frac{y_{ij}-x_{ij}'\beta_\tau-w_{ij}'\xi_{k,\tau}}{\sigma_\tau}\right)\left(\frac{y_{ij}-x_{ij}'\beta_\tau-w_{ij}'\xi_{k,\tau}}{\sigma_\tau}\right)\right]$$

$$He(\pi_k,\pi_l)=\sum_{i=1}^{m}\left[-\frac{\omega_{ik,\tau}^{(t)}}{\pi_k^2}1(l=k)-\frac{1-\sum_{g=1}^{G-1}\omega_{ig,\tau}^{(t)}}{\left(1-\sum_{g=1}^{G-1}\pi_g\right)^2}\right],\qquad k,l=1,\dots,G,$$

where

$$\psi_\tau'(u)=\begin{cases}2(1-\tau) & -c\le u<0\\ 2\tau & 0\le u\le c\\ 0 & |u|>c\end{cases}$$

and the term $1(A)$ denotes the indicator function for condition A to hold, while $He\left(\beta_\tau,\pi_{k,\tau}\right)=He\left(\xi_{k,\tau},\pi_{k,\tau}\right)=He\left(\sigma_\tau,\pi_{k,\tau}\right)=0$ due to parameter distinctiveness.

In the M-step in the EM algorithm, the weights $\omega_{ik,\tau}^{(t)}$ are kept fixed and model parameters are updated; in the E-step parameters are kept fixed and weights updated. However, the EM algorithm does not provide standard error estimates for model parameters, because it does not consider the portion of information which is missing due to unavailable information on z_{ik} i=1,…,m e k=1,…,G. In the case of quantile regression, standard errors are usually computed using a non-parametric bootstrap approach, see Geraci and Bottai (2014). In the current M-quantile context we may rely on the approach discussed by Louis (1982) and, for practical purposes, on the formula by Oakes (1999).

Let us denote by $I(\Phi_\tau)$ the observed information for Φ_τ where:

$$I(\Phi_\tau)=-\frac{\partial^2\ell(\Phi_\tau)}{\partial\Phi_\tau\partial\Phi_\tau'}$$

and by $\widehat{\Phi}_\tau^{(t)}$ the parameters current estimates. The Oakes (1999) identity is:

$$I\left(\widehat{\Phi}_\tau\right)=-\left\{\left.\frac{\partial Q\left(\Phi_\tau\middle|\widehat{\Phi}_\tau\right)}{\partial\Phi_\tau\partial\Phi_\tau'}\right|_{\Phi_\tau=\widehat{\Phi}_\tau}+\left.\frac{\partial^2 Q\left(\Phi_\tau\middle|\widehat{\Phi}_\tau\right)}{\partial\Phi_\tau'\partial\widehat{\Phi}_\tau}\right|_{\Phi_\tau=\widehat{\Phi}_\tau}\right\}$$

The right hand side consists of two arguments. The first is given by the second derivative of the conditional expected value of the log-likelihood given the observed data, and represent the expected information. This component is simple to get from the EM algorithm. The second component is the first derivative of the conditional expected value of the complete data score function with respect to the current parameter estimates, that is the posterior weights are not considered as fixed but as functions of the current parameter estimates, and this represents the impact of missing information.

Once the observed information has been calculated at the ML estimate $\widehat{\Phi}_\tau$, we calculate the sandwich estimator (see e.g. Huber, 1967; White, 1980):

$$\widehat{Cov}\left(\widehat{\Phi}_\tau\right) = I\left(\widehat{\Phi}_\tau\right)^{-1} V\left(\widehat{\Phi}_\tau\right) I\left(\widehat{\Phi}_\tau\right)^{-1}$$

where $V\left(\widehat{\Phi}_\tau\right) = \sum_{i=1}^{m} S_i\left(\widehat{\Phi}_\tau\right) S_i\left(\widehat{\Phi}_\tau\right)'$ is the estimate for cov(S). This approach has been discussed by Alfò et al. (2017) and dates back to the procedure described by Friedl and Kauermann (2000). This estimator allows to obtain a standard error estimate robustified to model misspecification, which may arise due to the assumption on response distribution. Therefore, using this estimate for the covariance matrix, we can approximate standard errors for model parameter estimates.

Standard conditions for identifiability in regression models do not apply directly to finite mixtures of regression models, see De Sarbo and Cron (1988). Identifiability needs the covariance matrix to be of full rank; However non-identifiability may also occur when the full rank condition is met, see Hennig (2000) and Wang et al. (1996). Given that the M-quantile model is a specific linear model estimated by IWLS, we may speculate that the regression parameters are identifiable if and only if the number of components is not higher than the number of distinct (p-1)-dimensional hyperplanes in the sample.

5 Models for spatial data

Observed data frequently show a spatial pattern and this may induce some sort of "spatial" dependence; that is the observed values can be influenced by statistical unit geographical position. This type of data is called spatial data and they are frequently met in epidemiological, environmental and social studies (Haining, 1990).

Spatial dependence, arises because of a principle described in the Tobler's first law, which states that:

> *…everything is related to everything else, but near things are more related than distant things. (Tobler , 1970)*

For example, it is possible to assume that the values observed in a given area are similar to those recorded in the neighboring areas.

This spatial relationship is very evident if we consider the environmental data from the PMetro project. In this project, between 2012 and 2015, various space-time measurements of aerosols and gases were collected using instruments integrated in one of the Minimetro's cabins, a public transport in the city of Perugia (Italy), see Del Sarto et al. (2019). The variable being studied by Del Sarto et al. (2019) is the logarithm of the concentration of particles, which can be spatially represented (see Figure 5.1); as it can be noticed closer shows how similar values are quite similar.

Figure 5.1 (Pmetro project) cartography of the (log) particle concentration referred Perugia

Spatial data can be multilevel, with a sample composed by lower level units (population, buildings) nested within higher level geographical units (census tracts, municipalities, regions).

In the following, we will consider neighboring areas, as those areas that are adjacent according to the Queen criterion. For two zones to be adjacent in the sense of the Queen's contiguity, they only need to share a common limit point. This name comes from the chess game, in the case of a regular grid, where the Queen has freedom of movement.

5.1 The Potts Model

Let us consider a finite mixture model, with marginal density described by

$$f(y_i) = \sum_k \pi_k f(y_{ij} | z_{ik} = 1)$$

If spatial dependence is present, we may assume according to Green and Richardson (2002) that the prior probability for a certain area to belong to a given component is influenced by the component the neighboring areas belong to.

Green and Richardson (2002) proposed a general approach based on finite mixture models with spatial constraints, where the prior probabilities are modelled through a Markov

Random Field (MRF) by adopting a Potts representation (Kindermann and Snell 1999, Strauss and Ikeda 1990).

The proposed Gibbs distribution is characterized by the use of a component specific intercept and a constant interaction parameter to model the effect of neighboring areas, under the assumption that spatial dependence is constant over the whole analyzed region. According to the work by Geman and Geman (1984), the class process z_{ik} i=1,...,m k=1,...,G , can be modelled by a MRF.

Let us denote by $z_{ik,\tau}$ the binary variable indicating whether the i-th unit comes from the k-th component of the finite mixture. We may assume that the prior probability for a given cluster to belong to a given component is influenced by the component the neighboring clusters belong to. Geman and Geman (1984) have shown that such a class process can be modelled by a Markov Random Field (MRF). According to the Hammersley-Clifford theorem (Hammersley and Clifford, 1971), modelling the process through a MRF is equivalent to using a Gibbs distribution for the component indicator. In other words, the spatial dependence between component labels is described by a Gibbs distribution, using a representation similar to the one proposed by Strauss (1977). That is we should no longer consider the priors as

$$\pi_{ik} = \Pr(z_{ik} = 1) = \frac{\exp(\alpha_k)}{\sum_k \exp(\alpha_k)}, \qquad (5.1)$$

but, rather, the same term defined as conditional on the neighboring observations

$$\zeta_{ik} = \Pr(z_i = k | z_l, l \in N_i) = \Pr(z_{ik} = 1 | z_{lk}, l \in N_i). \qquad (5.2)$$

where, as already discussed, z_{ik} denotes the vector of cluster indicators for cluster i = 1, ..., m, N_i, is a subset of S describing the set of units neighboring to the i-th, for witch we have: $i \notin N_i \quad \forall i \in S,\ j \in N_i \Leftrightarrow i \in N_j$ and S is a regular lattice. In this case, the Gibbs distribution is defined by:

$$\zeta_{ik} = \Pr(z_{ik} = 1 | z_{lk}, l \in N_i) \propto \exp\left\{-\vartheta \sum_{l \in N_i} V(z_{ik}, z_{lk})\right\} \qquad (5.3)$$

i=1,...,m, k=1,...,G. The parameter ϑ represents the interaction (regularization) parameter describing the strength of spatial dependence, which is supposed to be constant over the analyzed region. The term V(·) denotes the potential function:

$$V(z_{ik}, z_{lk}) = \begin{cases} -1 & \text{if } z_{ik} z_{lk} = 1 \\ 0 & \text{else} \end{cases} \qquad (5.4)$$

A potential alternative to this standard Gibbs distribution can be defined by using cluster-specific parameters as proposed by Alfò et al. (2009):

$$\zeta_{ik} = \Pr(z_{ik} = 1 | z_{lk}, l \in N_i) \propto \exp\left\{\alpha_k - \vartheta_k \sum_{l \in N_i} V(z_{ik}, z_{lk})\right\} \quad (5.5)$$

where, the potential function $V(\cdot)$ has been modified to:

$$V(z_{ik}, z_{lk}) = \begin{cases} -1 & \text{if } z_{ik} z_{lk} = 1 \\ +1 & \text{else} \end{cases} \quad (5.6)$$

In this way, the probability of being a member of the k-th component increases (if θ>0) with the number of neighbours coming from that component and decreases with the number of neighbours belonging to different component. Given the k-th component, the conditional distribution is:

$$\log(\zeta_{ik}) \propto \left[\alpha_k - \vartheta_k \sum_{l \in N_i} V(z_{ik}, z_{lk})\right] = [\alpha_k - \vartheta_k (n_{ik} - n_{i\bar{k}})] \quad (5.7)$$

k = 1, ..., G, where the values n_{ik} and $n_{i\bar{k}}$ represent the member of areas in N_i that do or do not belong to the k-th component. Equation (5.7) defines a multinomial logit or, more precisely, a Strauss automodel (Strauss and Ikeda, 1990) for the prior (conditional) probability of component membership.

The parameter α_k is a component-specific intercept, and $\log(\alpha_k)$ is proportional to the prior probability of the corresponding component when $n_{ik} = n_{i\bar{k}}$ that is in the case of maximum uncertainty or when $\theta = c$ that is independence holds . Finally, the parameter ϑ_k provides information on the strength of association as it describes how belonging to a component is influenced by neighboring areas membership. In the present work, we adopt a different form of the Gibbs distribution:

$$\zeta_{ik} = \Pr(z_{ik} = 1 | z_{lk}, l \in N_i) \propto \exp\left\{\alpha_k - \vartheta \sum_{l \in N_i} V(z_{ik}, z_{lk})\right\} \quad (5.8)$$

This distribution is characterized by the use of a cluster-specific intercept α_k which refer to the log prior when spatial dependence is absent, or when maximum uncertainty $n_{ik} = n_{i\bar{k}}$ holds and a constant interaction parameter ϑ. This formulation allows to have a parameter specific to each component and a constant spatial dependence in the whole area, while the potential function is defined as in Alfò et al. (2009):

$$V(z_{ik}, z_{lk}) = \begin{cases} -1 & \text{if } z_{ik} z_{lk} = 1 \\ +1 & \text{else} \end{cases} \quad (5.9)$$

In this function, the number of concordant areas in N_i is penalized by the number of discordant areas to account for a different size of the set of neighbours N_i i=1,…,m.

The previous expression **(5.8)** does no longer describe a multidimensional logit, but rather a conditional logit, McFadden (1973). In fact, the multidimensional logit model focuses on the individual as a unit of analysis and considers the individual features as explanatory variables, while the conditional logit model focuses on the set of alternatives for each individual and the explanatory variables are specific to each alternative (Hoffman and Duncan, 1988).

5.2 Finite Mixtures model for spatial data, estimation

Let us consider a study area defined on a regular lattice S with cells indexed by i =1,...,m. Given cell $i \in S$, y_i represents the response value observed at that position. The set N_i, is a subset of S including the units neighboring to the i-th one, for witch we have that: $i \notin N_i \forall i \in S,\ j \in N_i \Leftrightarrow i \in N_j$ hold.

Using such definition for neighbouring units (cells), and considering the prior conditional probability of component membership ζ_{ik} decribed in equation (5.8), the likelihood function can be rewritten

$$L(\theta) = \prod_{i=1}^{m} \sum_{k} f(y_i|\theta_k)\zeta_{ik} \tag{5.10}$$

and the log likelihood function is:

$$l(\theta) = \sum_{i=1}^{m} log \sum_{k} f(y_i|\theta_k)\zeta_{ik} \tag{5.11}$$

As usual, the component labels are unobserved and, therefore, they must be considered as missing data; this leads to the use of the EM algorithm. The hypothetical space of the complete data is given by $(y_i; z_{ik})$ i = 1, ..., m and k = 1, ..., G. The complete data log pseudo-likelihood (PL) function is:

$$l_c(\Phi_q) = \log\left\{\prod_{i=1}^{m}\prod_{k}[f(y_i|\theta_k)\zeta_{ik}]^{z_{ik}}\right\} =$$

$$= \sum_{i}\sum_{k} z_{ik}[\log f(y_i|\theta_k) + \log \zeta_{ik}]$$

which should be maximized under the constraint $\sum_k \zeta_{ik} = 1$. This represents an approximation of the "true" likelihood, which cannot be calculated, at least not in a simple way. In fact, the approximation of the log pseudo-likelihood consists in considering the areas as independent and it is appropriate when the spatial dependence is weak. In cases of heavy spatial dependence, MPL estimates may present a strong bias, see van Duijn et al.(2007).

5.3 Finite Mixtures with multilevel spatial data

We may also consider spatial dependence in the case where each unit of the lattice includes n_i elements, with i = 1, ..., m, j=1,…,n_i. That is, we may consider multilevel spatial data with lower-level units included within higher level clusters corresponding to cells of a regular lattice. In this case, the pseudo-likelihood function can be rewritten:

$$L(\theta) = \prod_{i=1}^{m} \sum_{\mathrm{k}=1}^{\mathrm{G}} \prod_{j=1}^{n_i} f(y_{ij}|\theta_k)\zeta_{ik} \tag{5.12}$$

The hypothetical space of the complete data is given by $(y_{ij}; z_{ijk})$ i=1,…,m, j=1,…,n_i and k=1,…,G, and the pseudo-likelihood (PL) for complete data can be written as:

$$l_c(\Phi) = \log\left\{\prod_{i=1}^{m}\prod_{k}\prod_{j=1}^{n_i}[f(y_{ij}|\theta_k)\zeta_{ik}]^{z_{ijk,\tau}}\right\} =$$

$$= \sum_i \sum_k \sum_j z_{ijk,\tau}[\log f(y_{ij}|\theta_k) + \log \zeta_{ik}]$$

The score function is:

$$\begin{aligned} S(\Phi) = \frac{\partial \log[L(\Phi)]}{\partial \Phi} = \frac{\partial \ell(\Phi)}{\partial \Phi} &= \sum_{i=1}^{m}\sum_{k=1}^{G}\left(\frac{f_{ik}\zeta_{ik}}{\sum_l f_{il}\zeta_{il}}\right)\sum_j \frac{\partial \log f_{ijk}}{\partial \Phi} \\ &=: \sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik}\sum_j \frac{\partial \log f_{ijk}}{\partial \Phi} \end{aligned} \tag{5.13}$$

where:

$$\omega_{ik} = \frac{\zeta_{ik}\prod_j f_{ijk}}{\sum_l \zeta_{il}\prod_j f_{ijl}} i = 1, \dots, m; k = 1, \dots, G$$

represent the posterior (conditional) probability of component membership. As we have previously seen, the score function represents a sum of the score functions for component-specific regression models with weights ω_{ik}. The log likelihood function can be maximized indirectly, by EM-type algorithms.

The basic EM algorithm is defined by solving equations for a given set of weights and updating the weights according to the current parameter estimates .

5.4 Finite Mixtures of (M-) Quantile regression for spatial data

We may consider spatial dependence also in the (M-) quantile regression model with individual- specific effects. The hypothetical space of the complete data is given by $(y_{ij}; z_{ijk,\tau})$ i=1,…,m, j=1,…,n_i and k=1,…,G. The pseudo-likelihood (PL) for the complete data can be written as:

$$l_c(\Phi_\tau) \propto \sum_{i=1}^{m}\sum_{k=1}^{G} z_{ik,\tau}\left\{\sum_{j=1}^{n_i}\log\left[f_\tau\left(y_{ij}\middle|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right)\right]+\log\left(\zeta_{k,\tau}\right)\right\} \tag{5.14}$$

where $f_\tau(\cdot)$ is the ALD/ALID in expressions(2.12)and for (4.5). In the E-step of the EM algorithm, we define the expected log pseudo-likelihood by taking the expectation of the log pseudo-likelihood for complete data over the component indicators conditional on the observed data and the current parameter, estimates replacing z_{ik} by:

$$\omega_{ik,\tau}^{(t+1)} = \frac{\zeta_{ik}^{(t)} f_{ik,\tau}\left(\widehat{\Phi}_\tau^{(t)}\right)}{\sum_l \zeta_{il}^{(t)} f_{il,\tau}\left(\widehat{\Phi}_\tau^{(t)}\right)} \tag{5.15}$$

where $\omega_{ik,\tau}^{(t+1)}$ represents the (conditional) posterior probability for the i-th cluster to belong to the k-th component of the analyzed area, given the observed data, the membership of neighboring units and the quantile level.

The conditional expectation of the complete data log pseudo-likelihood function is:

$$\begin{aligned} Q\left(\Phi_\tau\middle|\widehat{\Phi}_\tau^{(t)}\right) &= E_{\widehat{\Phi}_\tau^{(t)}}\{l_c(\Phi_\tau)|y_i\} \\ &\propto \sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t+1)}\left\{\sum_{j=1}^{n_i}\log\left[f_\tau\left(y_{ij}\middle|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right)\right]+\log\left(\zeta_k\right)\right\} \\ &\propto \sum_{i=1}^{m}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t+1)}\left\{\sum_{j=1}^{n_i}\log\left[f_\tau\left(y_{ij}\middle|\beta_\tau,\xi_{k,\tau},\sigma_\tau\right)\right]\right\} \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{G}\omega_{ik,\tau}^{(t+1)}\left\{\log\left(\frac{exp\left[\alpha_{k,\tau}-\vartheta_\tau V\left(z_{ik,\tau},z_{lk,\tau}\right)\right]}{\sum_{h=1}^{G} exp\left[\alpha_{h,\tau}-\vartheta_\tau\sum_{l\in N_i} V\left(z_{ih,\tau},z_{lh,\tau}\right)\right]}\right)\right\} \end{aligned} \tag{5.16}$$

where ζ_k is defined in equation (5.8) and $V(\cdot)$ is a potential function in equation (5.6). The maximization of the function $Q\left(\cdot\middle|\widehat{\Phi}_\tau^{(t)}\right)$ with respect to parameters Φ_τ={β_τ, σ_τ, $\xi_{1;\tau}$,..., $\xi_{G;\tau}$, $\alpha_{1;\tau}$, …, $\alpha_{G;\tau}$, ϑ_τ} for $\tau\in$ (0,1) can be divided into different sub-step. The estimates are obtained through the score equation:

$$\frac{\partial Q\left(\Phi_\tau\middle|\widehat{\Phi}_\tau^{(t)}\right)}{\partial\Phi_\tau} = 0$$

In the first sub-step only the parameters β_τ, σ_τ and $\xi_{k;\tau}$ are taken into account, this is equivalent to considering the estimates for a model with a known spatial dependence structure.

In the second sub-step the estimates of parameters ($\alpha_{k;\tau}$, ϑ_τ) of the Gibbs distribution are updated using the following pseudo score equation (5.16):

$$\frac{\partial Q\left(\Phi_\tau \middle| \widehat{\Phi}_\tau^{(t+1)}\right)}{\partial(\alpha_{k,\tau}, \vartheta_\tau)} = \frac{\partial}{\partial(\alpha_{k,\tau}, \vartheta_\tau)} \sum_{i=1}^{m} \sum_{k=1}^{G} \omega_{ik,\tau}^{(t+1)} \left\{ \log\left(\frac{exp\left[\alpha_{k,\tau} - \vartheta_\tau V\left(z_{ik,\tau}, z_{lk,\tau}\right)\right]}{\sum_{h=1}^{G} exp\left[\alpha_{h,\tau} - \vartheta_\tau \sum_{l \in N_i} V\left(z_{ih,\tau}, z_{lh,\tau}\right)\right]} \right) \right\} = 0 \quad (5.17)$$

which is the weighted sum of the pseudo-likelihood equations for a Strauss automodel with weights $\omega_{ik,\tau}^{(t+1)}$.

At each iteration of the M step, the membership is assumed to be known and a conditional logit model is estimated. This model is implemented using $\omega_{ik,\tau}^{(t+1)}$ as a response variable, while we use G counting variables describing the potential function as covariates.

A Newton-type algorithm can be used to provide pseudo ML estimates for $\alpha_{k;\tau}$ and ϑ_τ, k=1,...,G-1.

So, updates for model parameters are defined as

$$\left\{\beta_\tau, \sigma_\tau, \xi_{k;\tau}\right\} = argmax_{\left\{\beta_\tau, \sigma_\tau, \xi_{k;\tau}\right\}} \sum_{i=1}^{m} \sum_{k=1}^{G} \omega_{ik,\tau}^{(t+1)} \left\{ \sum_{j=1}^{n_i} \log\left[f_\tau\left(y_{ij} \middle| \beta_\tau, \xi_{k;\tau}, \sigma_\tau\right)\right] \right\}$$

while for the Strauss model

$$\left\{\alpha_{k;\tau}, \vartheta_\tau\right\} = argmax_{\left\{\alpha_{k;\tau}, \vartheta_\tau\right\}} \sum_{i=1}^{m} \sum_{k=1}^{G} \omega_{ik,\tau}^{(t+1)} \left\{ \log\left(\frac{exp\left[\alpha_{k,\tau} - \vartheta_\tau V\left(z_{ik,\tau}, z_{lk,\tau}\right)\right]}{\sum_{h=1}^{G} exp\left[\alpha_{h,\tau} - \vartheta_\tau \sum_{l \in N_i} V\left(z_{ih,\tau}, z_{lh,\tau}\right)\right]} \right) \right\}$$

The E- and M-steps are alternated until convergence, obtained with a sequence of non decreasing (pseudo) likelihood values which is bounded from above. The algorithm can be simply programmed and it is quite efficient compared to other techniques. As regards the choice of the number of components it is possible to use the technique based on Model comparison with composite likelihood information criteria. We perform the estimation algorithm for an increasing number of components of the mixture, starting from the number of groups equal to 2 and stopping as soon as the Composite likelihood Bayesian information criteria (CLBIC) begins to increase. The Bayesian information criterion (BIC) can be considered a CLBIC assessment because, otherwise, calculating the Hessian matrix and the Jacobian matrix would be very complicated (Ng and Joe 2014).

6 Simulation study

In this section we describe a small scale simulation study to evaluate the performance of the proposed method, for three different quantile levels, τ = 0.25, τ = 0.5 and τ = 0.75. We have considered a regular lattice of size 20X20, with m = 400 cells; within each cell, we have n_i=n units. Two covariates are considered, normally $X_{1i} \sim EXP_m(m, \lambda)$ assumes the same value for all the units in a cell, while $X_{2ij} \sim N(0,1)$, i=1,...,m, j=1,...,n, is varying with units in a cell.

We fix G = 4, with two possible spatial structure scenarios, as depicted in Figure 6.1.

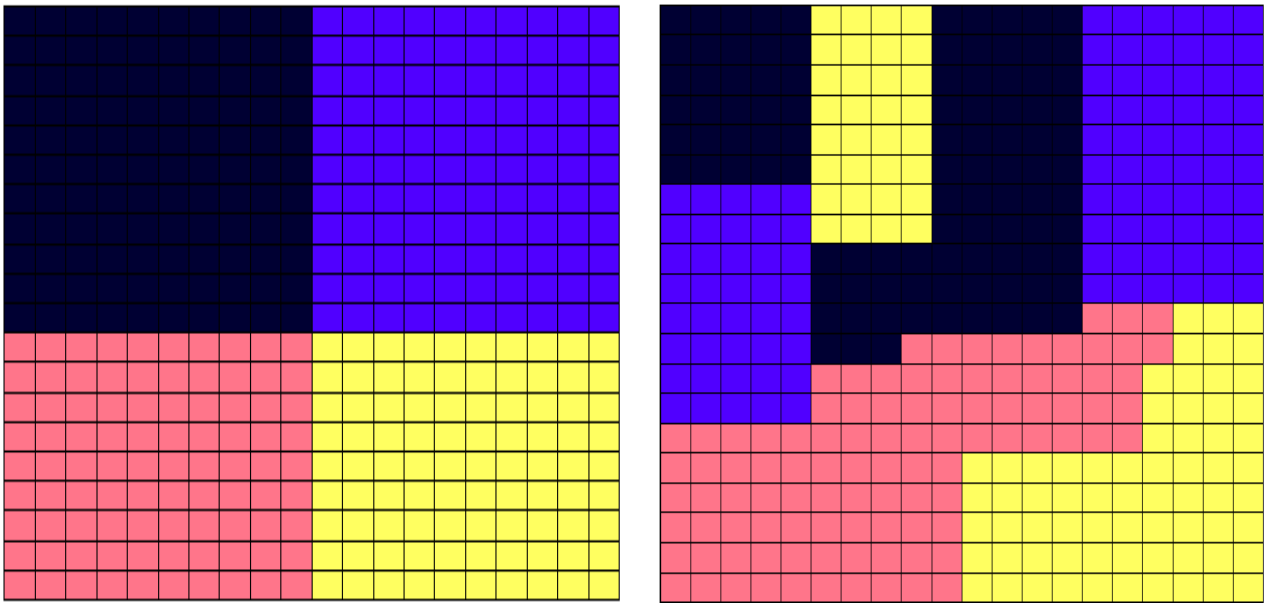

Figure 6.1 Lattice divided into G=4 groups: with a regular structure (left), and a less-regular structure (right)

The response values y_{ij} have been generated according to the following linear model:

$$y_{ij} = \sum_{k=1}^{G} \beta_{0k} I(z_i = k) + \beta_1 X_{1i} + \sum_{k=1}^{G} \beta_{2k} I(z_i = g)\, X_{2ij} + \varepsilon_{ij} \quad \text{i} = 1, \ldots, 400, \text{j} = 1, \ldots, 20$$

where $\varepsilon_{ij} \sim N(0,1)$, $\beta_0 = (-5, -3, 3, -5)$, $\beta_1 = (2)$ e $\beta_2 = (-1, 2, -3, 0.5)$. The coefficients with β_0 and β_2 vary with the component while β_1 is constant over the lattice. As, it can be noticed the spatial dependence between units is only due to the adjacency, as neighbouring units are more likely to belong to the some component.. This scenario has been replicated for B=1'000 samples.

The aim of this simulation study is to evaluate the behavior of finite mixtures of quantile (FMQR) and M quantile (FMMQ) regression, when considering geographically located units. In this way, the estimates obtained from the FMQRSP and FMMQSP models will be compared to those from FMQR and FMMQ, respectively.

The simulation study has been implemented using the R software. The estimates for FMMQ and FMQR have been obtained using the R functions developed by Alfò et al. (2017). An EM algorithm for maximum (pseudo) likelihood estimation has been implemented for FMMQSP and FMQRSP. Using a non optimized R code on a 2 GHz Intel Core i7 laptop with 16 Gb RAM, the average time for each scenario is about 17 hours for all the B=1'000 samples and all the considered models.

As we have previously seen, the tuning constant c can take on different values for the both the FMMQ and the FMMQSP models. In those cases where the data do not present outlier values, the best choice for c is a value that tends to ∞. Cantoni and Ronchetti (2001) suggest using c = 1.2, while Street et al. (1988) suggests using c = 1.25. Huber (1981) suggests c = 1.345, because it provides a reasonably high efficiency in the Gaussian case, it produces a 95% efficiency when errors are Gaussian and it still offers protection against outliers.

Wang et al. (2007) suggested a data-based approach for the definition of the tuning constant. In this way , the choice of the tuning constant can be performed automatically, within the step M of the EM algorithm. In this phase the constant can be estimated simultaneously with other parameters maximizing the log-likelihood function. However, no closed form solution is available. The estimation technique is based on the log-likelihood function, and it can be performed by defining a value in a grid defined over an appropriate segment (0.01, 100). The calculation of the log-likelihood will then be performed for a fixed c value in the grid. The c to be selected is the value from the grid that maximizes the log-likelihood function. This procedure is computationally complex, and, for this reason, we have considered two different values: c = 1.345 and c = 100 only.

$FMMQ_R$ and $FMMQSP_R$ (c = 1.345) are more robust to outliers and, when heterogeneity in the form of model parameters varying with the spatial cluster is present, we may guess that they should perform better than the single level M-quantile model (MQ). Also, when outliers are present, we expect that $FMMQ_R$ and $FMMQSP_R$ will perform better than their expectile counterparts, $FMMQ_E$ and $FMMQSP_E$ (c = 100), and the linear random effect model (Mixed).

A significant problem in the definition of spatial models is the initial allocation of the m cells to the component; in fact, once G is fixed, the initial choice of the component could be

an issue. In this simulation study we chose to assign the units to group by a simple k-means algorithm available through the library "stats", function *kmeans*.

At each iteration of algorithms discussed in section 5.4 the units are allocated according to a MAP rule. The procedure is thus repeated until convergence and if at least one group is empty, the procedure is interrupted.

To describe component membership given the neighboring units, we have used model in equation (5.8). This can be estimated via a conditional logit model by McFadden (1973), which provides a good approximation to the Potts model where covariates are random.

From a computational point of view, the conditional logit model needs for each individual a row for each category of the variable of interest. That is in the current case, for each geographical unit, we have 4 rows; in our case, leading to mXG=400*4 rows in the dataset. The conditional probability that the geographical unit i is assigned to component k=1,...,G is given by:

$$p_{ik} = \frac{exp\left(\alpha_k - \vartheta \sum_{j \in N_i} V\left(z_{ik}, z_{jk}\right)\right)}{\sum_{l=1}^{G} exp\left(\alpha_l - \vartheta \sum_{j \in N_i} V\left(z_{il}, z_{jl}\right)\right)}$$

and the R *mlogit* function from *library mlogit* was used to estimate parameters (α_k,ϑ) k=1,..,G.

To compare the performance of the different model specifications, the following measures have been used in addition to the overall mean estimates and the Mean squared error (MSE) estimate:

- for model parameters efficiency (EFF), this indicator defines the efficiency of the standard model compared to the spatial corresponding one:

$$EFF\left(\hat{\beta}\right) = \frac{\widehat{\mathrm{MSE}\left(\hat{\beta}\right)}}{\widehat{\mathrm{MSE}\left(\hat{\beta}_{sp}\right)}} \qquad (6.1)$$

- adjusted Rand index, a measure of similarity between two partitions, to understand whether the partition produced by the model agrees with the true one.

6.1 Simulation results

The first measure to evaluate and compare the models we have discussed so far is the average value of parameter estimates over the B=1000 simulation samples in the two scenarios we have considered (see Figure 6.1).

From the analysis of the distribution of estimates in the two scenarios, we notice that the estimates moderate bias with respect to the true parameter value. The results are however quite reliable but for a few cases and models .

Models without spatial effects show a greater bias only in the case of β_{21} and quantile 0.25.

Table 6.1 scenario 1 (regular lattice): Mean values of parameter estimation

Model	Quantile	fixed coefficients	component-specific slope parameter			
		β_1	β_{21}	β_{22}	β_{23}	β_{24}
True value	-	2.00	-1.00	2.00	-3.00	0.50
FMQR	25	2.00	-0.55	1.98	-2.98	0.50
	50	2.00	-0.93	2.00	-2.98	0.50
	75	2.00	-0.99	2.00	-2.78	0.48
FMMQR	25	1.99	-0.39	1.99	-2.98	0.50
	50	1.99	-0.90	2.00	-2.98	0.50
	75	2.00	-1.00	2.00	-2.83	0.48
FMMQE	25	2.00	-0.35	1.99	-3.00	0.50
	50	2.00	-0.93	2.00	-2.97	0.50
	75	2.00	-0.99	2.00	-2.81	0.48
FMQRSP	25	1.94	-0.81	1.97	-2.93	0.46
	50	1.96	-0.99	1.93	-3.00	0.44
	75	1.93	-0.91	1.89	-2.99	0.44
FMMQSPR	25	1.96	-0.80	1.97	-2.97	0.46
	50	1.96	-0.99	1.99	-3.00	0.50
	75	1.95	-0.99	1.93	-3.03	0.46
FMMQSPE	25	1.96	-0.85	1.96	-2.96	0.46
	50	1.96	-1.00	1.99	-2.99	0.47
	75	1.94	-0.99	1.98	-2.99	0.49

Table 6.2 scenario 2 (no regular lattice): Mean values of parameter estimation

Model	Quantile	fixed coefficients	component-specific slope parameter			
		β_1	β_{21}	β_{22}	β_{23}	β_{24}
True value	-	2.00	-1.00	2.00	-3.00	0.50
FMQR	25	2.00	-0.55	1.99	-2.98	0.50
	50	2.00	-0.93	2.00	-2.97	0.50
	75	2.00	-0.99	2.00	-2.81	0.48
FMMQR	25	1.99	-0.42	1.99	-2.98	0.50
	50	2.00	-0.94	2.00	-2.99	0.50
	75	2.00	-1.00	2.00	-2.98	0.48
FMMQE	25	1.99	-0.32	1.99	-3.00	0.50
	50	2.00	-0.94	2.00	-2.99	0.50
	75	2.00	-1.00	2.00	-2.84	0.48
FMQRSP	25	1.94	-0.78	1.97	-2.92	0.45
	50	1.94	-0.99	1.94	-3.00	0.44
	75	1.91	-0.89	1.88	-2.98	0.45
FMMQSPR	25	1.95	-0.79	1.97	-2.96	0.46
	50	1.94	-1.00	2.00	-3.00	0.50
	75	1.93	-0.99	1.94	-3.02	0.46
FMMQSPE	25	1.95	-0.82	1.97	-2.94	0.47
	50	1.94	-1.00	1.99	-2.98	0.48
	75	1.63	-1.00	1.98	-2.97	0.49

When we move from scenario 1 to scenario 2 we observe a (relatively large) bias arising for τ=0.25 in a few cases.

In Figure A.1-Figure A.10, A.Appendix, we report the distribution, across simulated samples, of estimates for β_1 and β_{2g} and the two scenarios

We may notice that in the second scenario, as far as the estimation of the parameters β_1 and β_{2g} is concerned, the increase of their variability is mainly characterized by the increase of anomalous values and therefore distant from the true value of the parameters.

To better analyze the behaviour of parameter estimates, we can use an estimate for the mean squared error (MSE).

The MSE analysis reveals that, in both scenarios (see Table 6.3 and Table 6.4), the discrepancy between the estimated values and the true values is higher for the models with spatial effects when compared to those not accounting for spatial dependence. This difference is essentially due to the higher variability of estimates obtained when a spatial estimators prior is considered.

Table 6.3 scenario 1 (regular lattice): mean squared error (MSE)

Model	Quantile	fixed coefficients	component-specific slope parameter			
		β_1	β_{21}	β_{22}	β_{23}	β_{24}
FMQR	25	0.0005	0.3337	0.0067	0.0126	0.0010
	50	0.0003	0.0920	0.0020	0.0206	0.0018
	75	0.0004	0.0112	0.0046	0.3223	0.0085
FMMQR	25	0.0061	0.7279	0.0057	0.0168	0.0034
	50	0.0038	0.1357	0.0019	0.0302	0.0020
	75	0.0011	0.0065	0.0007	0.2865	0.0069
FMMQE	25	0.0014	0.7735	0.0053	0.0042	0.0006
	50	0.0006	0.0980	0.0011	0.0429	0.0010
	75	0.0010	0.0069	0.0007	0.2477	0.0074
FMQRSP	25	0.0079	0.2124	0.0104	0.1070	0.0148
	50	0.0063	0.0048	0.0259	0.0025	0.0170
	75	0.0112	0.0960	0.0336	0.0623	0.0132
FMMQSPR	25	0.0063	0.2104	0.0117	0.0827	0.0154
	50	0.0063	0.0031	0.0020	0.0520	0.0026
	75	0.0077	0.0030	0.0280	0.0201	0.0092
FMMQSPE	25	0.0059	0.1613	0.0139	0.1117	0.0147
	50	0.0072	0.0023	0.0033	0.0496	0.0085
	75	0.0085	0.0028	0.0044	0.0518	0.0044

Table 6.4 scenario 2 (no regular lattice): mean squared error (MSE)

Model	Quantile	fixed coefficients	component-specific slope parameter			
		β_1	β_{21}	β_{22}	β_{23}	β_{24}
FMQR	25	0.000	0.333	0.006	0.012	0.001
	50	0.000	0.089	0.002	0.033	0.002
	75	0.000	0.008	0.004	0.260	0.007
FMMQR	25	0.005	0.697	0.004	0.035	0.001
	50	0.002	0.089	0.002	0.010	0.001
	75	0.001	0.003	0.001	0.046	0.006
FMMQE	25	0.003	0.815	0.005	0.001	0.001
	50	0.000	0.083	0.001	0.021	0.001
	75	0.001	0.005	0.001	0.238	0.006
FMQRSP	25	0.013	0.245	0.010	0.119	0.017
	50	0.016	0.003	0.022	0.002	0.017
	75	0.021	0.127	0.036	0.066	0.011
FMMQSPR	25	0.013	0.234	0.012	0.089	0.013
	50	0.014	0.002	0.002	0.043	0.002
	75	0.019	0.002	0.023	0.019	0.010
FMMQSPE	25	0.012	0.196	0.012	0.110	0.011
	50	0.016	0.002	0.004	0.055	0.006
	75	0.209	0.002	0.004	0.051	0.004

If we look at the efficiency of estimates through the EFF measure defined in equation (6.1), see Table 6.5 and Table 6.6, in both scenarios the estimates made using models without spatial effects seem to produce (slightly) higher values of MSE, even if the behaviour is not monotone.

For the first component-specific parameter, models with spatial effects seem to be more efficient, but this is quite an exception to the above mentioned phenomenon.

Table 6.5 scenario 1 (regular lattice): Efficiency (EFF)

Model	Quantile	fixed coefficients	component-specific slope parameter			
		β_1	β_{21}	β_{22}	β_{23}	β_{24}
FMQR/FMQRSP	25	0.06	1.57	0.65	0.12	0.06
	50	0.05	19.26	0.08	8.08	0.10
	75	0.03	0.12	0.14	5.17	0.65
FMMQE/FMMQSPE	25	0.24	4.79	0.38	0.04	0.04
	50	0.08	41.83	0.33	0.87	0.12
	75	0.12	2.44	0.17	4.78	1.71
FMMQR/FMMQSPR	25	0.96	3.46	0.49	0.20	0.22
	50	0.61	43.18	0.96	0.58	0.77
	75	0.14	2.16	0.03	14.23	0.75

Table 6.6 scenario 2 (no regular lattice): Efficiency (EFF)

Model	Quantile	fixed coefficients	component-specific slope parameter			
		β_1	β_{21}	β_{22}	β_{23}	β_{24}
FMQR/FMQRSP	25	0.034	1.359	0.578	0.104	0.067
	50	0.018	31.736	0.084	13.447	0.098
	75	0.018	0.061	0.118	3.922	0.659
FMMQE/FMMQSPE	25	0.252	4.149	0.428	0.012	0.054
	50	0.018	34.645	0.362	0.381	0.166
	75	0.005	2.065	0.322	4.698	1.616
FMMQR/FMMQSPR	25	0.386	2.977	0.321	0.399	0.113
	50	0.132	38.677	1.113	0.225	0.716
	75	0.072	1.389	0.057	2.467	0.581

The standard error estimates in the current context, is calculated using the bootstrap. This is a statistical technique based on resampling with replacement, which can be used to approximate the sampling distribution of given estimator.

The technique has been used, among others by Alfò et al. (2017) to evaluate finite mixtures models. In the present dissertation a spatial dataset composed by geographical units is used. The adoption of a bootstrap technique in such a context involves the creation of m samples of size n composed solely by a portion of the original geographical space. These re-samples would lead to a completely different geography and, therefore, may bias the observed spatial association.

To avoid these drawbacks, in the simulation study we have discussed, we decided to evaluate the "quality" of model estimates by using a high number of replications (B=1'000) in order to provided an estimated of the standard deviation of the model parameter estimates simulations across and compare it with the one obtained by the bootstrap approach.

To analyse the standard deviation behaviour in the two considered scenarios, the standard deviation obtained by examining just the first 20, 50, 500 was replicates considered in addition to the one obtained by considering all the replicates (see Table 6.7- Table 6.18).

Analysing the first scenario and the first 20 iterations, the standard error does seem to be almost non-convergent, as in the subsequent iterations a slight decrease in the estimates can be clearly seen. In the case of the second scenario, the estimates are slightly more variable, but they reach a certain stability already of the 50 replicates.

In this context, it is interesting to note that in both scenarios, for all the analyzed models, the behaviour is similar when the number of replicates increases.

Table 6.7 scenario 1 (regular lattice): standard deviation analysis of the model FMQR (20, 50, 500 and 1000 replicates)

Replicates	fixed coefficients	component-specific slope parameter			
	β_1	β_{21}	β_{22}	β_{23}	β_{24}
		$\tau = 0.25$			
20	0.02270	0.36312	0.10243	0.16695	0.03344
50	0.02270	0.36312	0.10243	0.15115	0.03344
500	0.02140	0.36769	0.07993	0.12629	0.03019
1000	0.02122	0.36447	0.07993	0.12629	0.03095
		$\tau = 0.50$			
20	0.01934	0.40434	0.05907	0.22159	0.05466
50	0.01826	0.40434	0.05907	0.20845	0.04571
500	0.01826	0.28132	0.04659	0.19438	0.04319
1000	0.01736	0.29583	0.04500	0.17954	0.04189
		$\tau = 0.75$			
20	0.02148	0.11163	0.07151	0.56214	0.09628
50	0.01951	0.11163	0.07151	0.56214	0.09628
500	0.01951	0.10542	0.06644	0.46780	0.08742
1000	0.01950	0.10579	0.06756	0.47587	0.08933

Table 6.8 scenario 1 (regular lattice): standard deviation analysis of the model FMMQ$_E$ (20, 50, 500 and 1000 replicates)

Replicates	fixed coefficients	component-specific slope parameter			
	β_1	β_{21}	β_{22}	β_{23}	β_{24}
		$\tau = 0.25$			
20	0.06681	0.59836	0.08908	0.22675	0.02544
50	0.06681	0.59836	0.08908	0.22675	0.02544
500	0.03762	0.59607	0.07571	0.08026	0.02459
1000	0.03762	0.59743	0.07122	0.07822	0.02434
		$\tau = 0.50$			
20	0.02756	0.32880	0.05174	0.22663	0.05119
50	0.02756	0.32880	0.04486	0.22663	0.05119
500	0.02337	0.29908	0.03663	0.15910	0.02957
1000	0.02374	0.30515	0.03301	0.15261	0.03221
		$\tau = 0.75$			
20	0.04935	0.12453	0.03904	0.52393	0.12436
50	0.03990	0.12453	0.02951	0.52393	0.09510
500	0.03642	0.09539	0.02952	0.50780	0.08681
1000	0.03160	0.08310	0.02733	0.49938	0.08403

Table 6.9 scenario 1 (regular lattice): standard deviation analysis of the model FMMQ$_R$ (20, 50, 500 and 1000 replicates)

Replicates	fixed coefficients	component-specific slope parameter			
	β_1	β_{21}	β_{22}	β_{23}	β_{24}
		$\tau = 0.25$			
20	0.13938	0.83738	0.09271	0.13758	0.08237
50	0.13054	0.59813	0.09271	0.13758	0.08237
500	0.12447	0.59809	0.08241	0.11383	0.05935
1000	0.07771	0.59834	0.07452	0.11398	0.05834
		$\tau = 0.50$			
20	0.09317	0.48312	0.04858	0.16092	0.05770
50	0.07931	0.35900	0.04858	0.14168	0.05022
500	0.07240	0.35900	0.04719	0.13448	0.04560
1000	0.06141	0.35546	0.04363	0.13251	0.04485
		$\tau = 0.75$			
20	0.03889	0.09340	0.04024	0.61890	0.10168
50	0.03889	0.09299	0.03181	0.50533	0.10168
500	0.03350	0.08866	0.02891	0.47964	0.08300
1000	0.03311	0.08027	0.02728	0.47964	0.08142

Table 6.10 scenario 1 (regular lattice): standard deviation analysis of the model FMQRSP (20, 50, 500 and 1000 replicates)

Replicates	fixed coefficients	component-specific slope parameter			
	β_1	β_{21}	β_{22}	β_{23}	β_{24}
		$\tau = 0.25$			
20	0.06430	0.60005	0.14101	0.36448	0.12615
50	0.06430	0.44237	0.10752	0.36448	0.12615
500	0.06666	0.41967	0.09815	0.31840	0.11470
1000	0.06660	0.41967	0.09810	0.30086	0.11464
		$\tau = 0.50$			
20	0.07084	0.09323	0.20348	0.07476	0.14343
50	0.07084	0.08407	0.14571	0.05635	0.14343
500	0.06813	0.06796	0.14535	0.05144	0.11739
1000	0.06773	0.06796	0.14535	0.05109	0.11618
		$\tau = 0.75$			
20	0.08400	0.26692	0.15874	0.26922	0.09787
50	0.08400	0.26692	0.15946	0.26922	0.09787
500	0.07938	0.29618	0.15128	0.24147	0.09835
1000	0.07938	0.29618	0.14273	0.24620	0.09907

Table 6.11 scenario 1 (regular lattice): standard deviation analysis of the model FMMQSP$_R$ (20, 50, 500 and 1000 replicates)

Replicates	fixed coefficients	component-specific slope parameter			
	β_1	β_{21}	β_{22}	β_{23}	β_{24}
		$\tau = 0.25$			
20	0.07640	0.59634	0.12208	0.28751	0.13397
50	0.07640	0.43963	0.12208	0.28751	0.13397
500	0.06883	0.41357	0.10040	0.28102	0.11651
1000	0.06879	0.41353	0.10368	0.28228	0.11553
		$\tau = 0.50$			
20	0.07042	0.05614	0.04500	0.22042	0.07847
50	0.07042	0.05614	0.04500	0.22042	0.06830
500	0.06651	0.05405	0.04587	0.22989	0.06453
1000	0.06820	0.05534	0.04511	0.22823	0.05012
		$\tau = 0.75$			
20	0.10036	0.07042	0.14864	0.13209	0.08996
50	0.07543	0.06429	0.14864	0.13209	0.08996
500	0.07128	0.05750	0.14943	0.13533	0.09065
1000	0.07237	0.05424	0.15179	0.13453	0.08893

Table 6.12 scenario 1 (regular lattice): standard deviation analysis of the model FMMQSP$_E$ (20, 50, 500 and 1000 replicates)

Replicates	fixed coefficients	component-specific slope parameter			
	β_1	β_{21}	β_{22}	β_{23}	β_{24}
		$\tau = 0.25$			
20	0.06640	0.45574	0.10952	0.29753	0.16061
50	0.06640	0.45574	0.10952	0.29753	0.12355
500	0.06414	0.34752	0.10974	0.27048	0.11780
1000	0.06680	0.37217	0.11141	0.27996	0.11435
		$\tau = 0.50$			
20	0.10594	0.07382	0.06222	0.23574	0.10552
50	0.08248	0.07382	0.06222	0.23574	0.10552
500	0.07529	0.04537	0.05436	0.23475	0.09776
1000	0.07241	0.04736	0.05665	0.22560	0.08888
		$\tau = 0.75$			
20	0.07591	0.04064	0.09649	0.25289	0.06325
50	0.07591	0.04064	0.07467	0.25289	0.06325
500	0.07370	0.05559	0.06816	0.23108	0.06732
1000	0.07370	0.05280	0.06508	0.22881	0.06418

*Table 6.13 scenario 2 (no regular lattice): standard deviation analysis of the model **FMQR** (20, 50, 500 and 1000 replicates)*

Replicates	fixed coefficients	component-specific slope parameter			
	β_1	β_{21}	β_{22}	β_{23}	β_{24}
		$\tau = 0.25$			
20	0.053	0.918	0.248	0.330	0.085
50	0.021	0.375	0.094	0.135	0.035
500	0.021	0.369	0.076	0.110	0.034
1000	0.021	0.367	0.076	0.110	0.034
		$\tau = 0.50$			
20	0.043	0.663	0.067	0.830	0.137
50	0.019	0.251	0.025	0.371	0.052
500	0.017	0.290	0.038	0.183	0.047
1000	0.017	0.290	0.043	0.179	0.041
		$\tau = 0.75$			
20	0.067	0.548	0.138	0.627	0.239
50	0.024	0.245	0.069	0.528	0.097
500	0.020	0.088	0.069	0.501	0.088
1000	0.020	0.088	0.065	0.472	0.082

*Table 6.14 scenario 2 (no regular lattice): standard deviation analysis of the model **FMMQ$_R$** (20, 50, 500 and 1000 replicates)*

Replicates	fixed coefficients	component-specific slope parameter			
	β_1	β_{21}	β_{22}	β_{23}	β_{24}
		$\tau = 0.25$			
20	0.316	3.155	0.252	0.634	0.086
50	0.119	1.192	0.089	0.240	0.038
500	0.095	0.847	0.064	0.194	0.038
1000	0.071	0.601	0.062	0.187	0.038
		$\tau = 0.50$			
20	0.150	1.359	0.202	0.505	0.134
50	0.053	0.514	0.091	0.191	0.050
500	0.050	0.317	0.045	0.098	0.047
1000	0.043	0.291	0.046	0.098	0.038
		$\tau = 0.75$			
20	0.099	0.646	0.075	0.522	0.212
50	0.049	0.228	0.030	0.241	0.095
500	0.039	0.075	0.034	0.223	0.079
1000	0.037	0.058	0.037	0.212	0.076

*Table 6.15 scenario 2 (no regular lattice): standard deviation analysis of the model **FMMQ$_E$** (20, 50, 500 and 1000 replicates)*

Replicates	fixed coefficients	component-specific slope parameter			
	β_1	β_{21}	β_{22}	β_{23}	β_{24}
		$\tau = 0.25$			
20	0.232	1.694	0.295	0.110	0.061
50	0.095	0.599	0.104	0.042	0.023
500	0.057	0.597	0.077	0.036	0.023
1000	0.055	0.597	0.071	0.037	0.024
		$\tau = 0.50$			
20	0.040	1109	0.196	0.391	0.133
50	0.018	0.419	0.074	0.195	0.066
500	0.017	0.279	0.039	0.152	0.040
1000	0.017	0.281	0.036	0.143	0.032
		$\tau = 0.75$			
20	0.099	0.390	0.117	1.243	0.176
50	0.049	0.195	0.052	0.507	0.062
500	0.035	0.070	0.047	0.489	0.073
1000	0.032	0.070	0.038	0.462	0.077

*Table 6.16 scenario 2 (no regular lattice): standard deviation analysis of the model **FMQRSP** (20, 50, 500 and 1000 replicates)*

Replicates	fixed coefficients	component-specific slope parameter			
	β_1	β_{21}	β_{22}	β_{23}	β_{24}
		$\tau = 0.25$			
20	0.257	0.973	0.260	1.020	0.376
50	0.105	0.397	0.116	0.416	0.133
500	0.102	0.427	0.102	0.356	0.122
1000	0.096	0.444	0.098	0.334	0.121
		$\tau = 0.50$			
20	0.326	0.313	0.415	0.140	0.341
50	0.163	0.110	0.157	0.053	0.152
500	0.125	0.061	0.137	0.050	0.118
1000	0.113	0.052	0.135	0.049	0.115
		$\tau = 0.75$			
20	0.397	0.755	0.331	0.490	0.270
50	0.140	0.338	0.148	0.245	0.096
500	0.113	0.338	0.141	0.245	0.094
1000	0.113	0.338	0.144	0.256	0.091

Table 6.17 scenario 2 (no regular lattice): standard deviation analysis of the model $FMMQSP_R$ (20, 50, 500 and 1000 replicates)

Replicates	fixed coefficients	component-specific slope parameter			
	β_1	β_{21}	β_{22}	β_{23}	β_{24}
		$\tau = 0.25$			
20	0.296	1.051	0.236	0.743	0.253
50	0.105	0.429	0.096	0.332	0.113
500	0.103	0.430	0.097	0.311	0.108
1000	0.103	0.438	0.105	0.295	0.108
		$\tau = 0.50$			
20	0.392	0.071	0.118	0.411	0.108
50	0.139	0.035	0.059	0.184	0.048
500	0.115	0.050	0.046	0.205	0.044
1000	0.105	0.048	0.043	0.207	0.045
		$\tau = 0.75$			
20	0.316	0.203	0.334	0.313	0.157
50	0.119	0.083	0.149	0.140	0.079
500	0.119	0.049	0.147	0.135	0.092
1000	0.119	0.049	0.141	0.135	0.092

Table 6.18 scenario 2 (no regular lattice): standard deviation analysis of the model $FMMQSP_E$ (20, 50, 500 and 1000 replicates)

Replicates	fixed coefficients	component-specificslopeparameter			
	β_1	β_{21}	β_{22}	β_{23}	β_{24}
		$\tau = 0.25$			
20	0.373	1.037	0.288	0.989	0.296
50	0.132	0.423	0.109	0.374	0.105
500	0.113	0.417	0.108	0.330	0.099
1000	0.097	0.406	0.105	0.326	0.098
		$\tau = 0.50$			
20	0.420	0.198	0.117	0.625	0.276
50	0.148	0.088	0.058	0.236	0.113
500	0.120	0.053	0.062	0.237	0.083
1000	0.114	0.049	0.059	0.233	0.077
		$\tau = 0.75$			
20	0.795	0.197	0.214	0.628	0.168
50	0.281	0.070	0.081	0.237	0.069
500	0.267	0.052	0.070	0.227	0.068
1000	0.267	0.049	0.065	0.223	0.061

Another interesting element to analyze is the estimated partitions. Considering the Adjusted Rand Index (ARI see Hubert and Arabie (1985)) we can see, in both the proposed scenarios, the models with spatial effects are associated to a higher average value when compared to the non-spatial models (see Table 6.19 and Table 6.20). The distribution of the ARI indicator in models with spatial effects is concentrated around the median value, ie it assumes a value that smoothly varies between 0.71 and 0.84 for the different quantiles. For the models without spatial effect, the value varies between 0.41 and 0.70.

Table 6.19 scenario 1 (regular lattice): Adjusted Rand index

	FMQR			$FMMQ_R$			$FMMQ_E$		
	25	50	75	25	50	75	25	50	75
Min.	0.43	0.43	0.43	0.00	0.00	0.00	0.00	0.00	0.00
1st Qu.	0.46	0.80	0.48	0.00	0.80	0.51	0.51	0.80	0.51
Median.	0.51	0.80	0.51	0.51	0.80	0.51	0.51	0.80	0.51
Mean	0.60	0.76	0.62	0.46	0.70	0.55	0.60	0.75	0.63
3rd Qu.	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
Max.	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
	FMQRSP			$FMMQSP_R$			$FMMQSP_E$		
	25	50	75	25	50	75	25	50	75
Min.	0.37	0.69	0.72	0.37	0.37	0.37	0.37	0.37	0.37
1st Qu.	0.80	0.82	0.82	0.81	0.82	0.83	0.81	0.82	0.83
Median.	0.82	0.83	0.83	0.82	0.83	0.84	0.82	0.84	0.84
Mean	0.81	0.83	0.84	0.80	0.85	0.86	0.80	0.84	0.86
3rd Qu.	0.83	0.84	0.88	0.82	0.87	0.91	0.82	0.88	0.91
Max.	0.97	0.97	0.97	0.92	1.00	1.00	0.89	1.00	1.00

Table 6.20 scenario 2 (no regular lattice): Adjusted Rand index

	FMQR			$FMMQ_R$			$FMMQ_E$		
	25	50	75	25	50	75	25	50	75
Min.	0.33	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.00
1st Qu.	0.36	0.70	0.38	0.00	0.70	0.41	0.41	0.70	0.41
Median.	0.41	0.70	0.41	0.41	0.70	0.41	0.41	0.70	0.70
Mean	0.50	0.66	0.52	0.39	0.62	0.46	0.51	0.65	0.55
3rd Qu.	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
Max.	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
	FMQRSP			$FMMQSP_R$			$FMMQSP_E$		
	25	50	75	25	50	75	25	50	75
Min.	0.27	0.63	0.63	0.27	0.27	0.27	0.27	0.27	0.27
1st Qu.	0.70	0.71	0.71	0.71	0.72	0.73	0.71	0.72	0.73
Median.	0.71	0.73	0.73	0.72	0.73	0.74	0.72	0.73	0.74
Mean	0.71	0.74	0.74	0.70	0.75	0.76	0.70	0.74	0.76
3rd Qu.	0.72	0.74	0.78	0.72	0.77	0.81	0.72	0.76	0.81
Max.	0.90	0.87	0.88	0.84	0.88	0.89	0.81	0.90	0.90

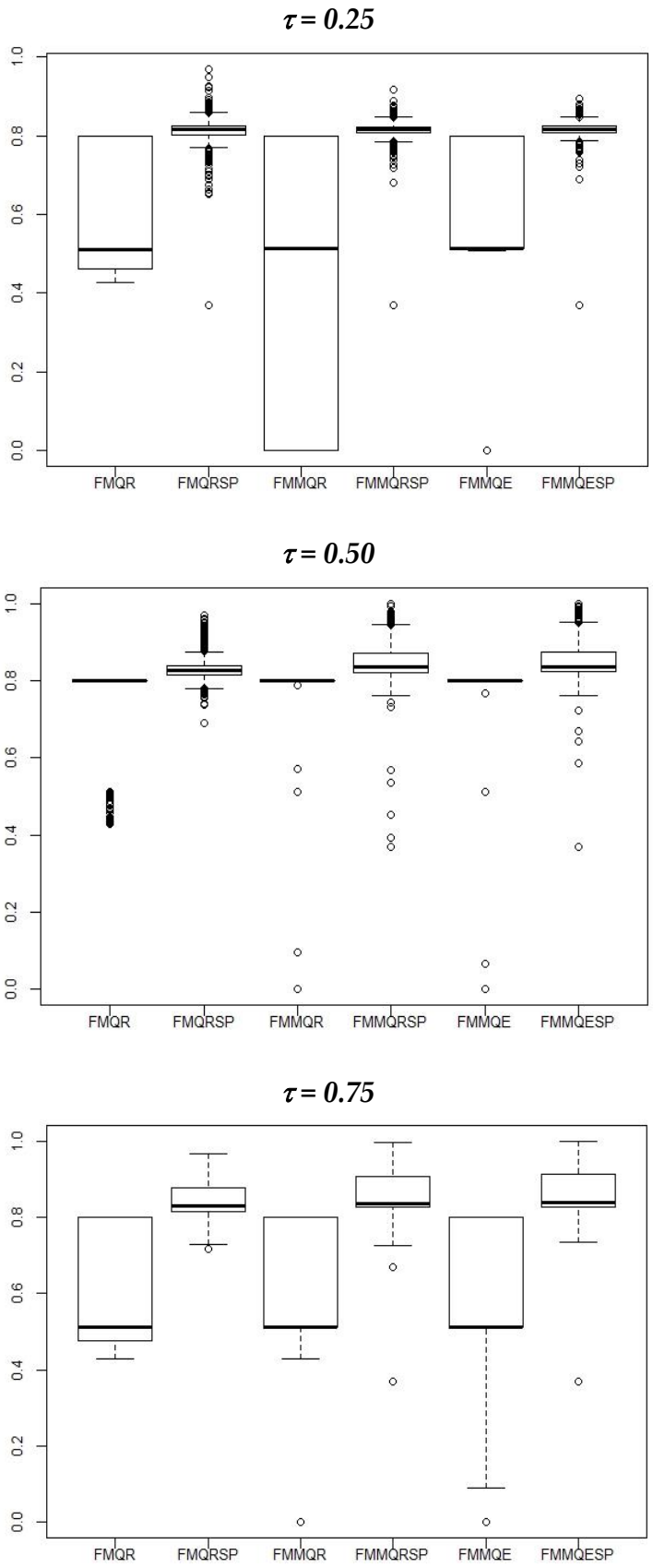

Figure 6.2 scenario 1 (regular lattice): Adjusted Rand index, boxplot

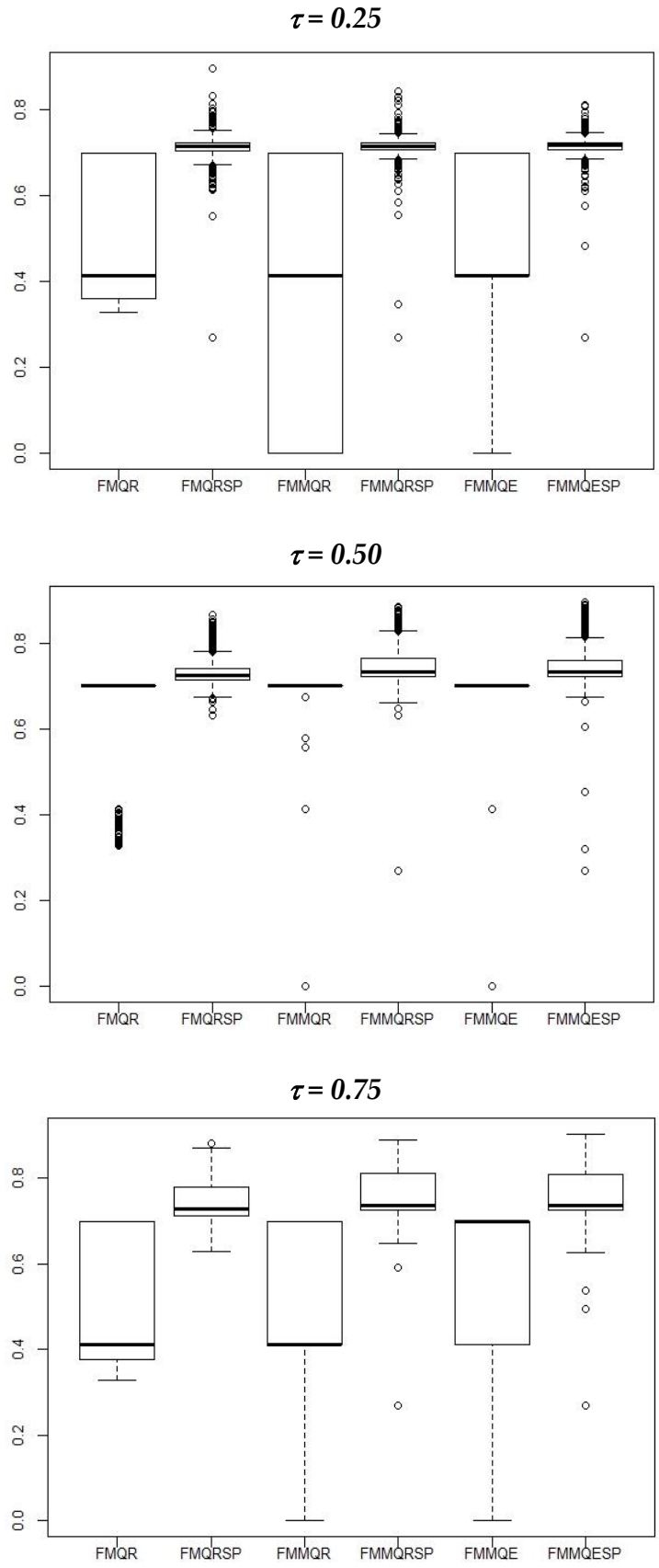

Figure 6.3 scenario 2 (no regular lattice): Adjusted Rand index, boxplot

7 Empirical application: housing prices in Rome

To better illustrate the features of the proposed approach, it was applied to the analysis of the relationship between property sales in Rome in 2001 and socio-economic characteristics of the area where the property is located. The source for the analysed data on sale of new or renovated properties is provided by the real estate agency of the Territory Agency. The prices refer to the various urban sectors of the municipality of Rome better said toponymy subdivisions. The city of Rome in 2001 was composed by 115 toponomastic subdivisions (Figure 7.1):

- 22 "Rioni" defining the historic center, established in the Middle Ages on the basis of the 14 Augustan regions and expanded at the end of the 19th century, all within the Aurelian Walls but Borgo and Prati
- 35 "Quartieri" surrounding the historic center outside the Aurelian Walls, including the three "marine quarters" Ostia is divided in
- 6 "Suburbi", territories beyond the district
- 52 "Zone" slightly populated across Grande Raccordo Anulare (GRA)

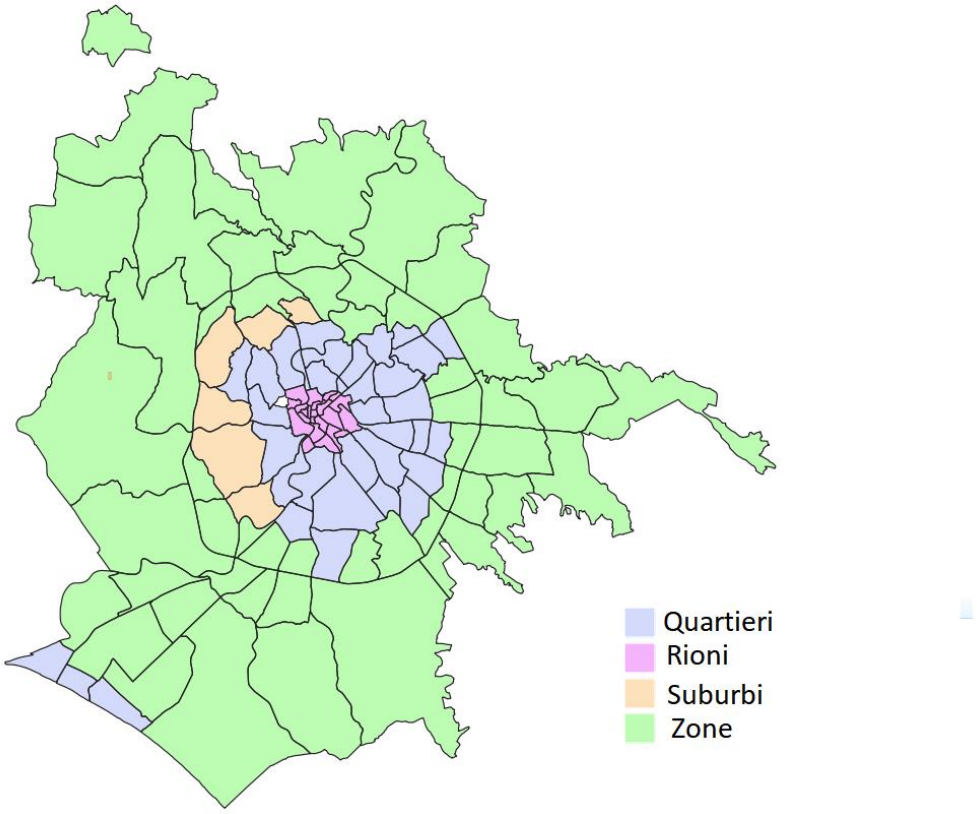

Figure 7.1 Toponomastic divisions for the city of Rome in 2001

The available dataset combines the information of the trades occurred in the 6 suburbs with the homonymous and neighboring quarters, thus leading to the partition in 109 territorial units are represented below (Figure 7.2).

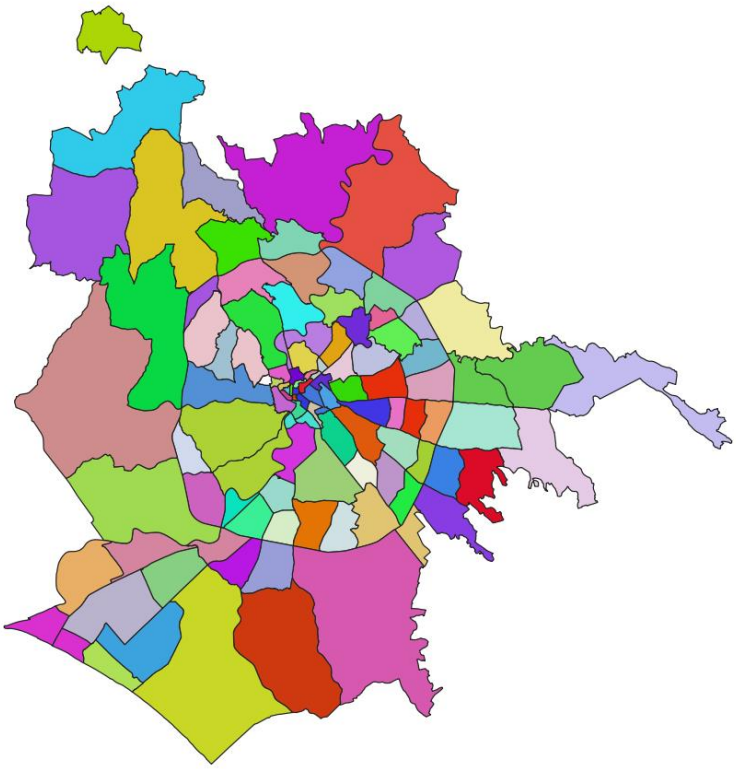

Figure 7.2 The city of Rome in 2001 is composed of 109 territorial units under study

In 2001 new or renovated homes price is equal to € 3,344,954 / sqm, with a range that varies from € 2,000 / sqm to € 5,600 / sqm (Table 7.1). Analyzing the observed price distribution in the municipality we can see that it is bimodal, with two modes, corresponding to 2'500 € / sqm, and 4'700 € / sqm (Figure 7.3).

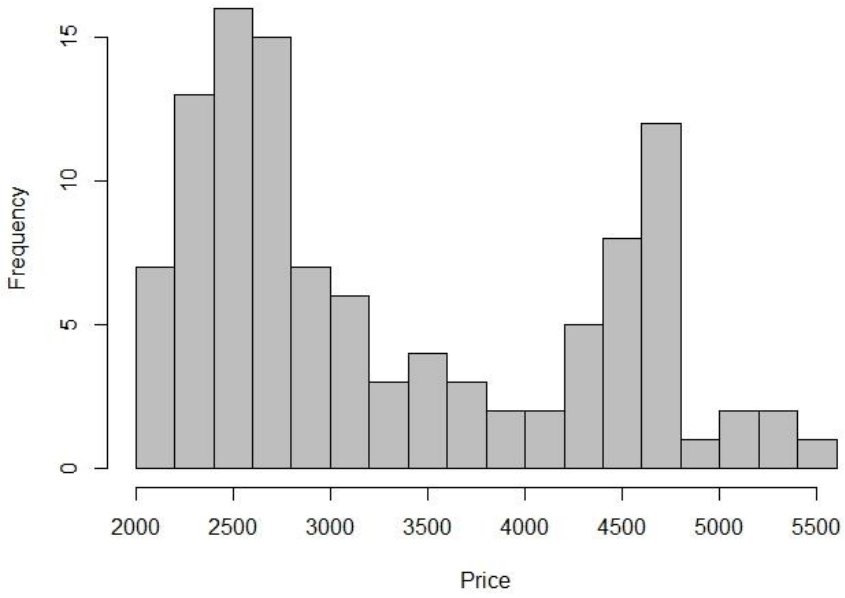

Figure 7.3 Distribution of price by sqm for new or renovated houses

Analyzing the selling price from a geographical point of view, we can see how the corresponding building position seem to greatly affect the housing price. In particular, the value decreases as the distance from the center of Rome increases(Figure 7.4).

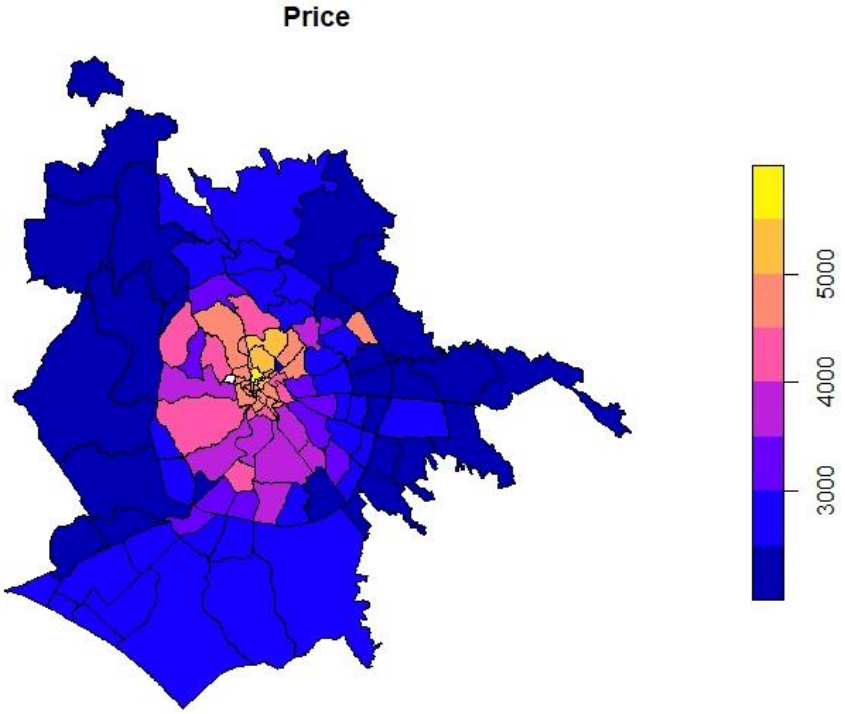

Figure 7.4 Geographical representation of the mean price of houses (€ / sqm)

Regarding the socio-economic characteristics of the territorial units/buildings, 5 indicators have been considered, based on the Population and Housing Census for year 2001. The analyzed indicators are:

1) Population percentage over 65 years

$$P_{65} = \frac{Pop_{65+}}{Pop} * 100$$

where Pop_{65+} indicates the resident population aged 65 and over, while Pop represents the resident population.

2) Average number of rooms:

$$N_{rooms} = \frac{N_{rooms\ tot}}{N_{\text{homes}}}$$

where $N_{rooms\ tot}$ and N_{homes} respectively correspond to the number of total rooms and the number of total dwellings.

3) Percentage of employees

$$P_{\text{empl}} = \frac{N_{\text{employees}}}{N_{\text{employment}}} * 100$$

where $N_{\text{employees}}$ represents the number of employees while $N_{\text{employment}}$ indicates the number of persons employed.

4) Percentage of employees in the Industry sector

$$P_{\text{ind}} = \frac{N_{\text{emp ind}}}{N_{\text{employment}}} * 100$$

where $N_{\text{emp ind}}$ indicates the number of employees in the industrial sector, while $N_{\text{employment}}$ indicates the total number of persons employed.

5) Percentage of employees in the Commerce sector

$$P_{\text{com}} = \frac{N_{\text{emp com}}}{N_{\text{employment}}} * 100$$

where $N_{\text{emp com}}$ is the number of employees in the Commerce sector and the total number is of persons employed denoted by $N_{\text{employment}}$.

Table 7.1 Descriptive statistics of the data

Variable	Label	Range	Mean	SD
Buying and selling New or renovated homes (€/sqm)	Price	(2'000; 5'600)	3'345.0	997.3
Average number of rooms	N_{rooms}	(5,57; 26,57)	17.4	5.5
Population percentage over 65 years	P_{65}	(3,09; 5,22)	4.0	0.4
Percentage of employees	P_{empl}	(53,41; 86,62)	74.7	7.9
Percentage of employees in the Industry sector	P_{ind}	(8,31; 66,67)	17.8	8.4
Percentage of employees in the Commerce sector	P_{com}	(0,00 25,94)	17.9	4.7

The choice of these socio-economic indicators is motivated by the high correlation with the phenomenon under study. Indeed analyzing the Table 7.2 we can see that the correlation between the price variable and the covariates, for N_{rooms}, assumes absolute values higher than 0.29. We also note, interestingly, that the P_{65} variable is highly correlated, 0.68, with the price variable.

In the analysis under study it was preferred to standardize the variables as these have different variability and size.

Table 7.2 correlation matrix

	Price	P_{65}	N_{rooms}	P_{empl}	P_{ind}	P_{com}
Price	1.00	0.68	0.29	-0.64	-0.51	-0.45
P_{65}	0.68	1.00	0.19	-0.33	-0.41	-0.45
N_{rooms}	0.29	0.19	1.00	-0.49	-0.05	-0.45
P_{empl}	-0.64	-0.33	-0.49	1.00	0.08	0.31
P_{ind}	-0.51	-0.41	-0.05	0.08	1.00	-0.10
P_{com}	-0.45	-0.45	-0.45	0.31	-0.10	1.00

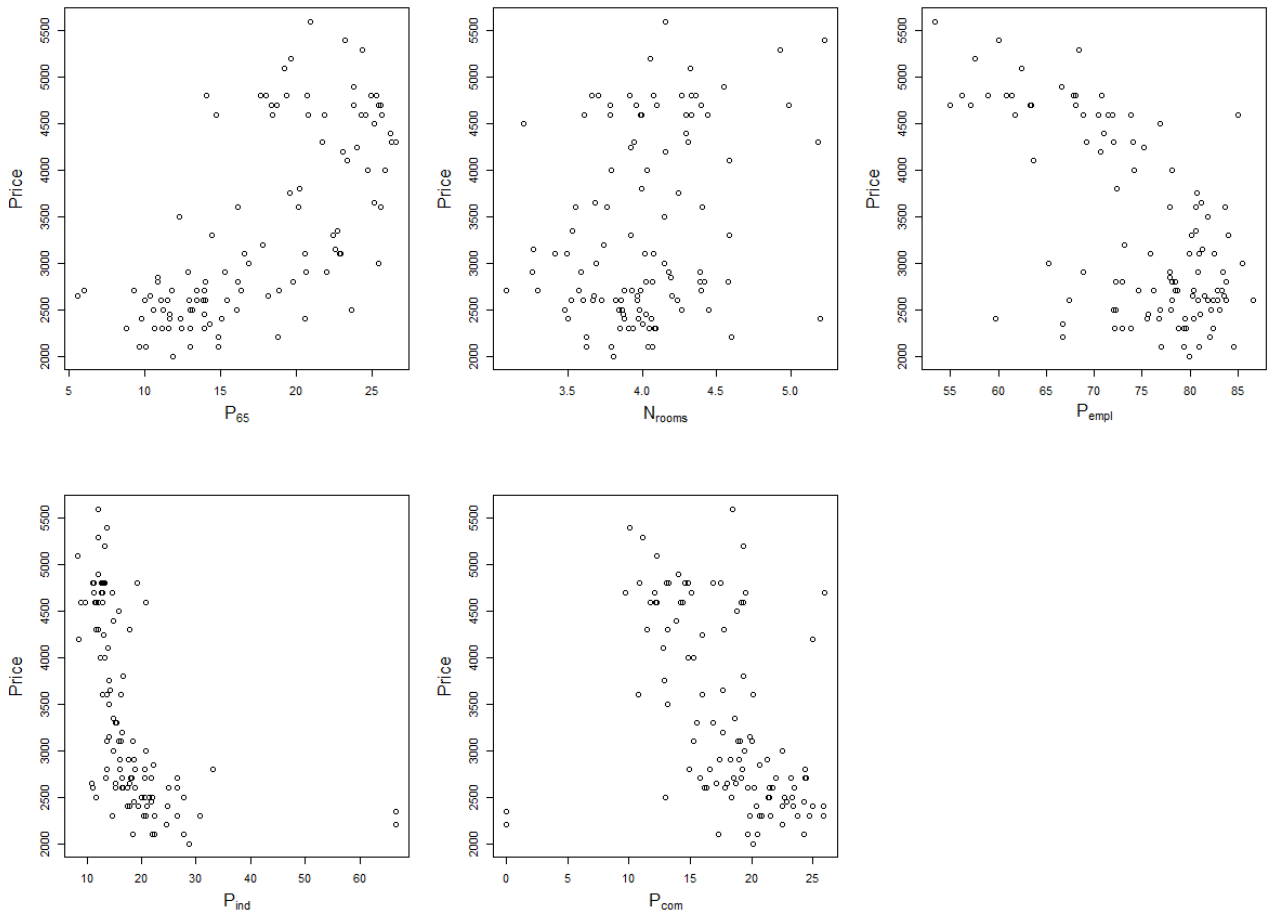

Table 7.5 Plot of price by covariates.

We extend the exploratory analysis considering the FMQR, FMMQ, FMQRSP and FMMQSP models. As before, M-quantile and M-quantile models with spatial effects are estimated using the Huber influence function with c = 1.345 ($FMMQ_R$ and $FMMQSP_R$) and c = 100 ($FMMQ_E$ and $FMMQSP_E$).

The Price variable is the dependent variable (response), the N_{rooms} variable is associated to a fixed effect, while the others namely P_{65}, P_{empl}, P_{ind} and P_{com}, are all associated to component specific effects. This model choice is motivated by the fact that the price in square meters for houses is influenced by the house size, while the other variables play the role of characterizing the socio-economic context where the house is located.

We have considered three quantiles 0.25, 0.50 and 0.75: for each, we estimated the models for an increasing number of components of the mixture, starting from G=2 and stopping as soon as the BIC begins to increase. The BIC can be considered as CLBIC assessment because, otherwise, calculating the Hessian matrix and the Jacobian matrix would be very complicated (Ng and Joe 2014).

According to this approach, the optimal number of components both for models with spatial and non-spatial effects is always G=2, but for three FMMQSP$_R$ cases of quantile 0.25, FMMQ$_E$ and FMMQ$_R$ of quantile 0.75, where G=3. (see Table 7.3).

Table 7.3 BIC values for varying number of mixture components K and varying quantile (the optimal models are reported in boldface)

Group	FMQRSP		
K	0.25	0.50	0.75
2	**169.4**	**139.2**	**162.8**
3	292.7	313.6	376.2
4	-	-	-

Group	FMQR		
K	0.25	0.50	0.75
2	**146.5**	**166.5**	**164.5**
3	160.6	168.1	199.6
4	202.0	182.2	-

Group	FMMQSP$_R$		
K	0.25	0.50	0.75
2	4'583.4	**153.2**	**166.4**
3	**206.8**	184.0	191.5
4	4649	-	4649

Group	FMMQ$_R$		
K	0.25	0.50	0.75
2	**150.2**	**151.7**	169.6
3	166.9	161.1	**165.3**
4	-	-	179.5

Group	FMMQSP$_E$		
K	0.25	0.50	0.75
2	**191.9**	**182.3**	**193.2**
3	224.0	212.6	215.5
4	-	-	-

Group	FMMQ$_E$		
K	0.25	0.50	0.75
2	**155.9**	**147.9**	4574.0
3	157.3	4'602.1	**725.2**
4	4'630.3	181.6	4'768.013

The models we have considered are characterized by two groups. Looking at the intercept parameter estimates we can deduce that the territorial units in the first group are characterized by a lower economic value.

The behaviour of covariate estimated effects is similar in all the models we have considered. In fact, if we look at the component-specific parameters the variables P_{empl}, P_{ind} and P_{com} have a negative effect, while for the variable P_{65} the parameter takes a positive value (see Table 7.4 and

Table 7.5). In the first component the values (in absolute) are more contained.

Also the N_{rooms} coefficient assumes a value that is basically negative or close to zero, but for the FMMQ$_E$ model at quantile 0.75.

Table 7.4 Parameter estimates, for FMQR, FMMQ$_R$ and FMMQ$_E$ at dipendent quantiles

	0.25		0.50		0.75		
			FMQR				
			intercept parameter				
	1 Group	*2 Group*	*1 Group*	*2 Group*	*1 Group*	*2 Group*	
	-0.80	*-0.11*	*-0.10*	*0.14*	*0.21*	0.32	
			component-specific slope parameter				
	1 Group	*2 Group*	*1 Group*	*2 Group*	*1 Group*	*2 Group*	
P$_{65}$	0.14	0.28	0.27	0.29	0.47	0.12	
P$_{empl}$	0.03	-0.56	-0.48	-0.55	-0.34	-0.66	
P$_{ind}$	-0.08	-0.41	-0.43	-0.40	-0.42	-0.24	
P$_{com}$	-0.04	-0.22	-0.25	-0.21	-0.31	-0.44	
			Fixed coefficients				
N$_{rooms}$	-0.05		-0.02		-0.10		
			FMMQ$_R$				
			intercept parameter				
	1 Group	*2 Group*	*1 Group*	*2 Group*	*1 Group*	*2 Group*	*3 Group*
	-0.66	*-0.06*	*-0.47*	*0.01*	*-0.58*	0.13	0.81
			component-specific slope parameter				
	1 Group	*2 Group*	*1 Group*	*2 Group*	*1 Group*	*2 Group*	*3 Group*
P$_{65}$	0.13	0.28	0.08	0.31	0.06	0.25	-0.04
P$_{empl}$	0.01	-0.54	0.26	-0.56	-0.01	-0.56	0.69
P$_{ind}$	-0.06	-0.44	0.01	-0.42	0.03	-0.47	-0.91
P$_{com}$	-0.08	-0.27	-0.26	-0.24	-0.15	-0.29	-0.80
			Fixed coefficients				
N$_{rooms}$	-0.07		-0.06		-0.04		
			FMMQ$_E$				
			intercept parameter				
	1 Group	*2 Group*	*1 Group*	*2 Group*	*1 Group*	*2 Group*	*3 Group*
	-0.67	*-0.09*	*0.02*	*0.02*	*-1.36*	1.93	4.11
			component-specific slope parameter				
	1 Group	*2 Group*	*1 Group*	*2 Group*	*1 Group*	*2 Group*	*3 Group*
P$_{65}$	0.11	0.30	0.43	0.28	-0.70	1.48	1.94
P$_{empl}$	0.22	-0.52	0.73	-0.54	5.13	1.13	7.67
P$_{ind}$	-0.02	-0.42	0.40	-0.40	0.85	0.32	-0.31
P$_{com}$	-0.05	-0.25	0.73	-0.24	2.00	1.84	1.26
			Fixed coefficients				
N$_{rooms}$	-0.06		-0.06		5.69		

Table 7.5 Parameter estimates, for FMQRSP, FMMQSP$_R$ and FMMQSP$_E$ at dipendent quantiles

	0.25			0.50		0.75	
				FMQRSP			
				intercept parameter			
	1 Group	*2 Group*	*1 Group*	*2 Group*	*1 Group*	*2 Group*	
	-0.52401	*0.1603706*	*-0.32*	*0.22*	*-0.25*	*0.39*	
				component-specific slope parameter			
	1 Group	*2 Group*	*1 Group*	*2 Group*	*1 Group*	*2 Group*	
P_{65}	0.21	0.05	0.27	0.16	0.28	0.15	
P_{empl}	-0.16	-0.41	-0.17	-0.44	-0.06	-0.51	
P_{ind}	-0.23	-0.61	-0.20	-0.54	-0.18	-0.56	
P_{com}	-0.16	-0.16	-0.13	-0.14	-0.11	-0.05	
				Fixed coefficients			
N_{rooms}		-0.06		0.03		0.07	
				FMMQSP$_R$			
				intercept parameter			
	1 Group	*2 Group*	*3 Group*	*1 Group*	*2 Group*	*1 Group*	*2 Group*
	-0.45	-0.02	-0.04	*-0.40*	*0.15*	*-0.31*	*0.27*
				component-specific slope parameter			
	1 Group	*2 Group*	*3 Group*	*1 Group*	*2 Group*	*1 Group*	*2 Group*
P_{65}	0.18	0.28	0.23	0.16	0.21	0.13	0.25
P_{empl}	-0.28	-0.38	-0.52	-0.16	-0.51	-0.08	-0.54
P_{ind}	-0.09	-0.48	-0.45	-0.22	-0.49	-0.22	-0.38
P_{com}	-0.30	-0.42	-0.29	-0.18	-0.22	-0.21	-0.25
				Fixed coefficients			
N_{rooms}		-0.23		-0.11		-0.12	
				FMMQSP$_E$			
				intercept parameter			
	1 Group	*2 Group*	*1 Group*	*2 Group*	*1 Group*	*2 Group*	
	-0.51	*-0.15*	*-0.43*	*0.02*	*-0.30*	*0.20*	
				component-specific slope parameter			
	1 Group	*2 Group*	*1 Group*	*2 Group*	*1 Group*	*2 Group*	
P_{65}	0.15	0.18	0.15	0.22	0.14	0.22	
P_{empl}	-0.22	-0.51	-0.12	-0.52	-0.03	-0.53	
P_{ind}	-0.22	-0.60	-0.21	-0.57	-0.21	-0.50	
P_{com}	-0.19	-0.28	-0.19	-0.25	-0.21	-0.26	
				Fixed coefficients			
N_{rooms}		-0.21		-0.14		-0.11	

To evaluate the fit of spatial and non-spatial regression models, we show in Figure 7.6 and Figure 7.7 the expected values for different Quantiles and M-quantiles.

Each panel shows that the model captures the relationship between the response variable and the covariate: the residuals for each Quantile are well dispersed, but for the FMMQ$_E$ model with quantile 0.75, where the model fails to well represent the phenomenon.

Therefore, except for this case, these graphs show a good fit of the regression models to the data.

From this figure it is possible to notice the different colors used to represent the residuals. This represents the learning group of the single unit, the components are well separated in the models with spatial effects, while for the models without spatial effects we do not have a clear distinction between the components.

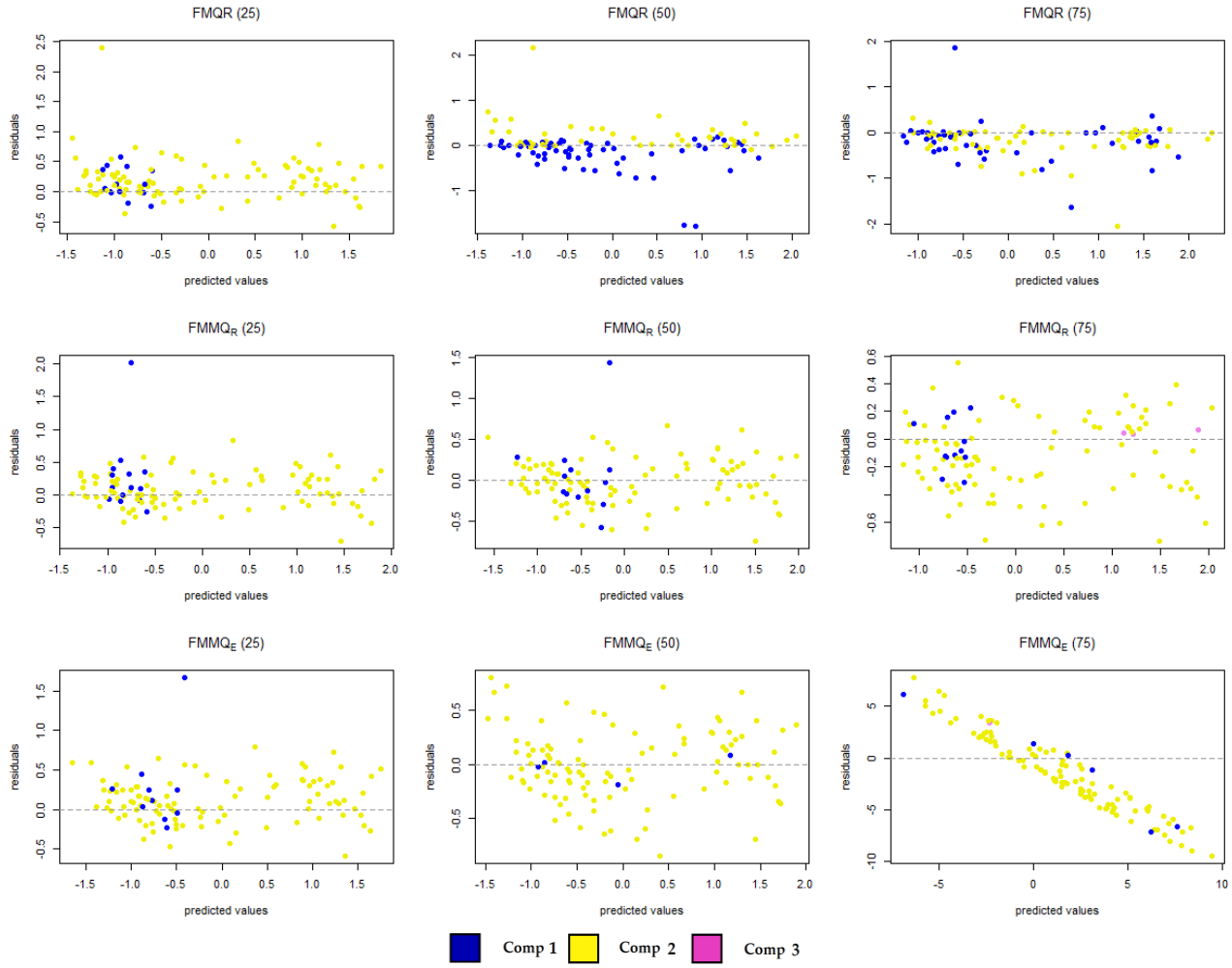

Figure 7.6 Predicted values and residuals for FMQR, FMMQ$_R$ and FMMQ$_E$ at different quantiles

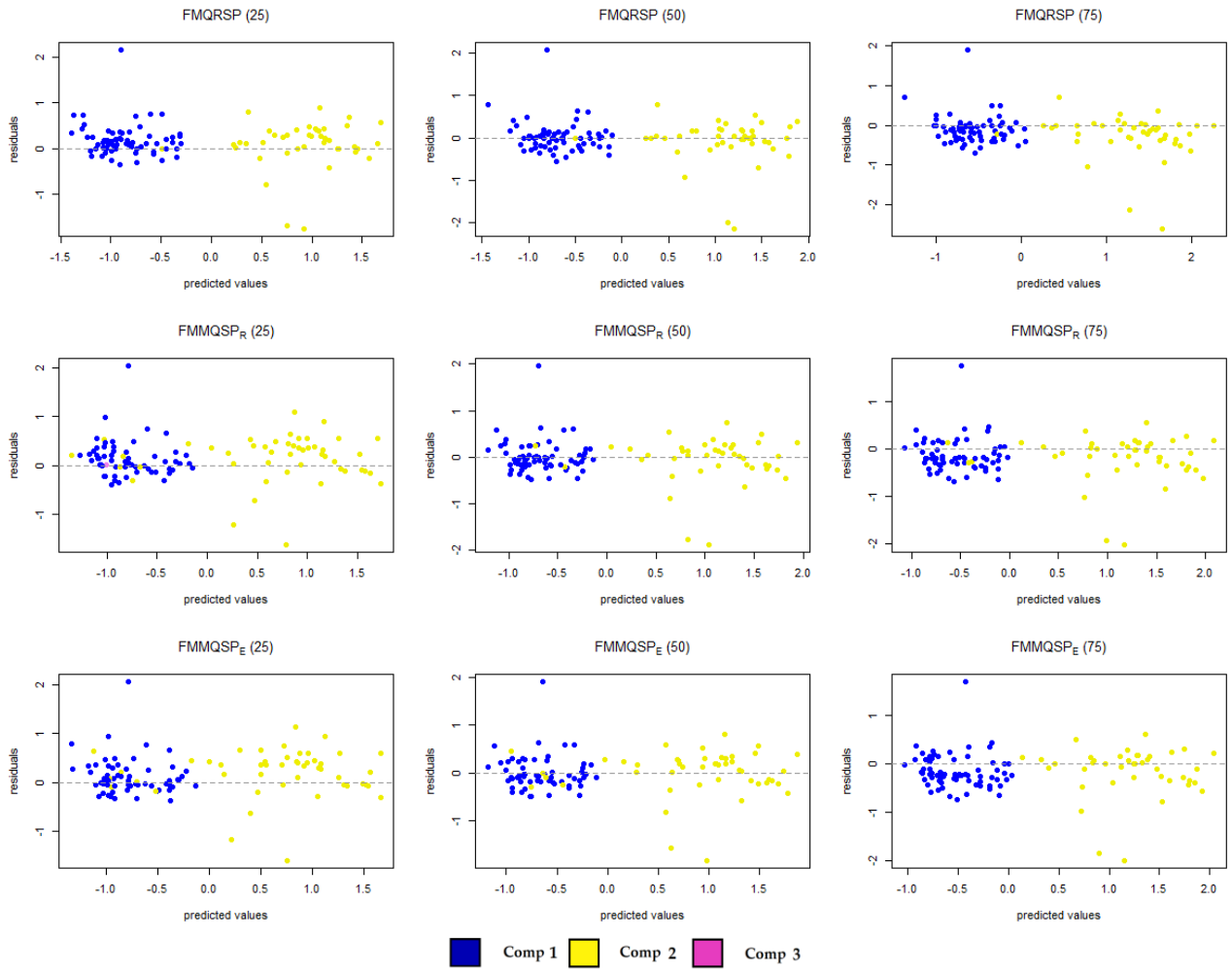

Figure 7.7 Predicted values and conditional residuals for the FMQRSP, FMMQSP$_R$ and FMMQSP$_E$ at different quantiles

Analyzing the data from a spatial point of view, the classical models (FMQR, FMMQ$_E$ and FMMQ$_R$) show a more dispersed spatial pattern which may have a difficult interpretation, and have a strong concentration of units into a single group (see Figure 7.8). The models do not seem to capture the difference between houses near and far from the center of Rome, respectively.

Instead, we can see how the models with spatial effects (FMQRSP, FMMQSP$_E$ and FMMQSP$_R$) allocate the units into component that are more compact and coherent with the nature of the phenomenon (see Figure 7.9). In fact, in the analyzed models at the different quantiles, two clusters of units are generally visualized; the first group includes all the units in the "rioni" and almost all the "quartieri" located to the north, north-west and south, while the other includes the remaining areas of Rome. Despite being in the suburbs it is interesting to notice that the "Torricola" and "Castel di Leva" areas are inserted in the pattern described above. This partition is quite stable across quantiles for FMQRSP.

Regarding the FMMQSP$_E$ model, and the quantiles 0.25 and 0.50, several zones in the north-east and south borders beyond the GRA are included, with respect to the situation we have previously seen. As for quantile 0.75, the situation is more compact with only the territorial units within in the GRA belonging to the first cluster.

Finally, as for the FMMQSP$_R$ at the 0.25 quantile, as previously seen, there are 3 groups, the units of the center of Rome are joined to different areas on the north-east and south borders beyond the GRA. As far as quantiles 0.50 and 0.75 are concerned, the situation is more compact with units located in the center of Rome associated only to areas in the southern borders beyond the GRA.

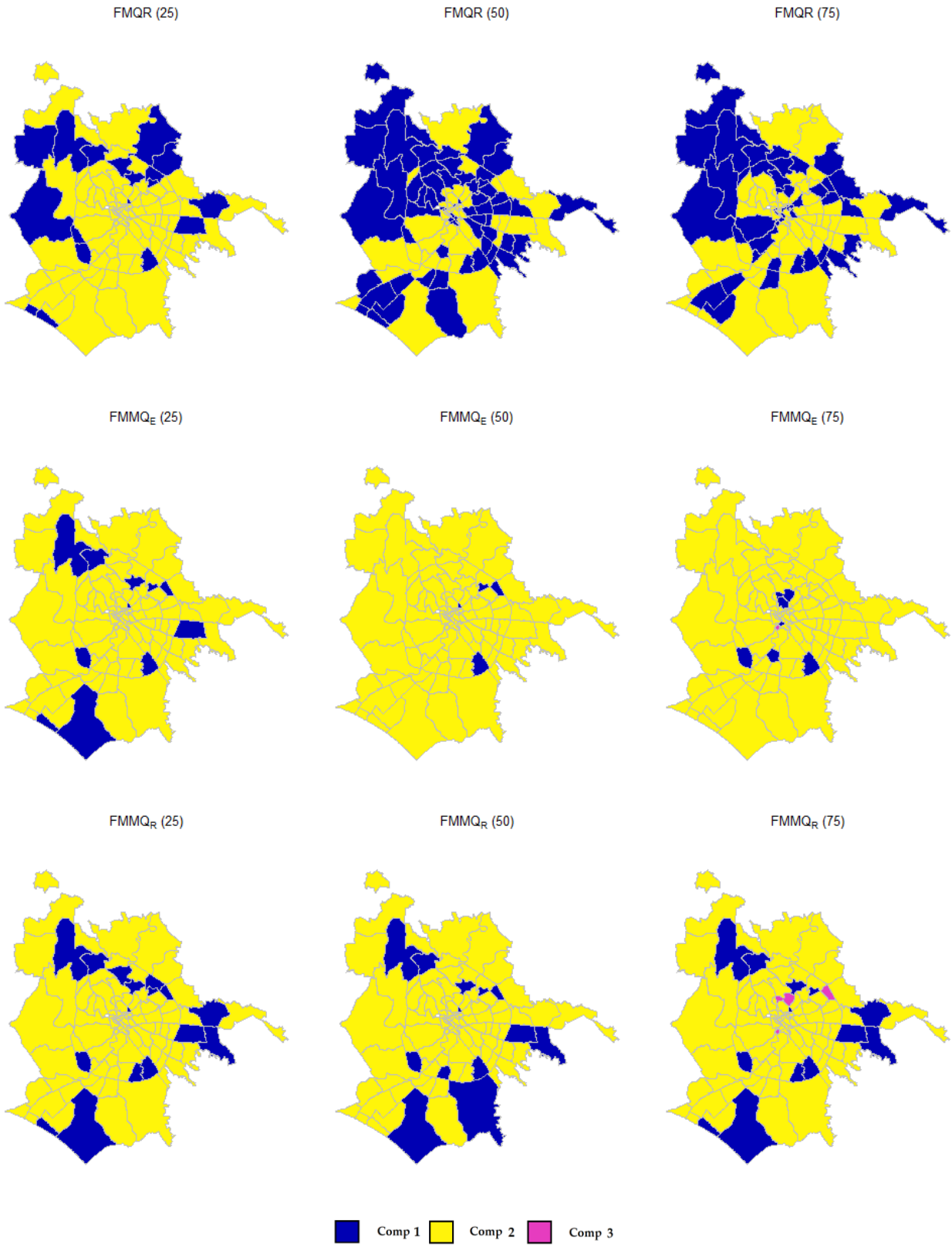

Figure 7.8 Partition of the territorial units by FMQR, FMMQ$_E$ and FMMQ$_R$ models at different quantiles

Figure 7.9 Partition of the territorial units by FMQRSP, FMMQSP$_E$ and FMMQSP$_R$ at different quantiles

From the analysis of the averages and the standard deviations of the analyzed variables for the estimated clusters, we may notice that the units clusterized by the spatial models are more homogeneous than the ones clusterized by the models without spatial effect (see Table 7.6 and

Table 7.7). Going into the details of the analysis, we observe that two profiles are created in which in the inhabited center areas, the dwellings have the highest house prices, are larger, the population is older, there is less occupied population as dependent, both in the industry sector and in the commerce sector.

Table 7.6 Table of means and standard deviation of the variables under study in the groups created by the FMQR, $FMMQ_E$ and $FMMQ_R$ models with respect to the 0.25, 0.50 and 0.75 quantiles

Quantile	Comp	Buying and selling New or renovated homes		Average number of rooms		Population percentage over 65 years		Percentage of employees		Percentage of employees in the Industry sector		Percentage of employees in the Commerce sector	
		Price		N_{rooms}		P_{65}		P_{empl}		P_{ind}		P_{com}	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
FMQR													
25	1	2'569.4	252.7	4.0	0.4	14.9	4.6	74.3	6.2	19.5	5.8	20.4	3.8
	2	3'498.4	1'018.5	4.0	0.4	17.9	5.5	74.8	8.2	17.5	8.8	17.4	4.8
50	1	3'039.5	842.0	4.0	0.4	17.3	5.7	75.0	7.1	17.8	7.7	18.1	4.5
	2	3'747.9	1'050.4	4.0	0.4	17.5	5.3	74.3	9.0	17.9	9.3	17.7	5.1
75	1	3'211.5	1'031.9	4.1	0.4	15.5	5.2	73.4	7.6	19.3	11.0	17.7	5.4
	2	3'450.0	964.8	4.0	0.4	18.8	5.3	75.7	8.1	16.6	5.2	18.1	4.2
$FMMQ_R$													
25	1	2'768.8	541.3	4.0	0.5	14.6	5.0	75.7	6.6	19.9	7.1	19.2	3.6
	2	3'444.1	1'025.8	4.0	0.4	17.9	5.5	74.5	8.1	17.5	8.5	17.7	4.9
50	1	2'888.5	575.2	4.1	0.5	14.9	5.3	76.2	6.8	20.4	6.6	18.8	3.5
	2	3'406.8	1'027.9	4.0	0.4	17.7	5.5	74.5	8.1	17.5	8.5	17.8	4.9
75	1	2'646.2	218.4	4.1	0.5	14.5	5.6	75.5	6.0	20.1	7.8	18.8	3.8
	2	3'395.7	1'008.3	4.0	0.4	17.6	5.4	74.5	8.2	17.5	8.5	17.9	4.9
	3	4'800.0	435.9	3.9	0.9	21.4	5.8	76.8	8.3	16.2	4.4	16.4	4.6
$FMMQ_E$													
25	1	2'863.6	610.4	4.1	0.6	15.6	5.5	75.6	7.1	19.1	7.1	18.3	3.4
	2	3'399.0	1'019.7	4.0	0.4	17.6	5.5	74.6	8.0	17.7	8.5	17.9	4.9
50	1	3'150.0	1'014.9	4.3	0.7	20.5	4.0	73.2	10.5	15.8	4.1	17.0	3.5
	2	3'352.4	1'000.8	4.0	0.4	17.3	5.5	74.8	7.9	17.9	8.5	18.0	4.8
75	1	4'133.3	1'294.1	4.8	0.8	20.8	7.5	66.6	7.2	16.4	5.7	14.5	6.2
	2	3'287.3	962.9	4.0	0.3	17.1	5.3	75.2	7.8	17.9	8.5	18.1	4.6
	3	4'500.0	-	3.2	-	25.1	-	76.9	-	15.7	-	18.8	-

Table 7.7 Table of means and standard deviation of the variables under study in the groups created by the FMQRSP, FMMQSP$_E$ and FMMQSP$_R$ models with respect to the 0.25, 0.50 and 0.75 quantiles

Quantile	Group	Buying and selling New or renovated homes		Average number of rooms		Population percentage over 65 years		Percentage of employees		Percentage of employees in the Industry sector		Percentage of employees in the Commerce sector	
		Price		N_{rooms}		P_{65}		P_{empl}		P_{ind}		P_{com}	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
FMQRSP													
25	1	2'678.8	428.6	3.9	0.3	14.3	4.2	78.8	4.8	20.9	9.4	19.7	4.5
	2	4'367.4	711.0	4.2	0.4	22.1	3.6	68.5	7.7	13.1	2.6	15.3	3.7
50	1	2'678.8	428.6	3.9	0.3	14.3	4.2	78.8	4.8	20.9	9.4	19.7	4.5
	2	4'367.4	711.0	4.2	0.4	22.1	3.6	68.5	7.7	13.1	2.6	15.3	3.7
75	1	2'678.8	428.6	3.9	0.3	14.3	4.2	78.8	4.8	20.9	9.4	19.7	4.5
	2	4'367.4	711.0	4.2	0.4	22.1	3.6	68.5	7.7	13.1	2.6	15.3	3.7
FMMQSP$_R$													
25	1	2'703.4	442.7	3.9	0.3	14.6	4.2	79.3	4.4	19.5	4.6	20.3	2.9
	2	4'137.8	912.6	4.2	0.4	20.8	5.0	69.3	7.7	14.8	8.3	15.5	4.6
	3	2'350.0	-	4.0	-	14.3	-	66.7	-	66.7	-	0.0	-
50	1	2'678.5	431.9	3.9	0.3	14.3	4.2	78.8	4.9	20.9	9.4	19.6	4.6
	2	4'329.5	746.3	4.2	0.4	21.9	3.9	68.6	7.6	13.2	2.7	15.4	3.9
75	1	2'678.9	435.3	3.9	0.3	14.5	4.1	78.8	4.9	21.1	9.4	19.7	4.6
	2	4'292.2	779.1	4.2	0.4	21.5	4.5	68.9	7.8	13.2	2.7	15.5	3.8
FMMQSP$_E$													
25	1	2'689.3	442.3	3.9	0.3	14.6	4.1	78.9	4.9	21.0	9.6	19.6	4.6
	2	4'178.1	876.9	4.2	0.4	20.9	5.0	69.3	7.8	13.7	3.5	15.8	4.0
50	1	2'689.3	442.3	3.9	0.3	14.6	4.1	78.9	4.9	21.0	9.6	19.6	4.6
	2	4'178.1	876.9	4.2	0.4	20.9	5.0	69.3	7.8	13.7	3.5	15.8	4.0
75	1	2'681.3	425.9	3.9	0.3	14.3	4.2	78.8	4.8	20.9	9.3	19.7	4.5
	2	4'403.6	678.5	4.2	0.4	22.4	3.1	68.2	7.6	12.9	2.3	15.1	3.7

8 Concluding remarks

When the values assumed by a response are influenced by geographical position of the units, the assumption of independence between observations can lead to substantial bias in the parameter estimates.

A potential way to solve this issue is through the use of finite mixture models with spatial constraints, where priors are modelled through a Markov Random Field (MRF) using a Potts representation, see Green and Richardson (2002).

According to the Hammersley-Clifford theorem, modelling the process through an MRF is equivalent to using a Gibbs distribution.

In this work, a form of the Gibbs distribution is proposed, characterized by a component-specific intercept and a constant interaction parameter. This formulation allows to have, at the same time, a specific benchmark probability for each component and a similar spatial dependence in the whole area. The parameters in the prior distribution can be estimated by using a conditional logit model.

This specification has been applied to finite mixtures of quantile (FMQR) and M-quantile (FMMQ) models, in order to account for spatial dependence, and to define finite mixture of quantile (FMQRSP) and M-quantile regression (FMMQSP) with spatial effects.

From a small scale simulation study we can see that the inclusion of spatial effects in the FMQR and FMMQ models does produce some additional variability in the estimates without a clear improvement is bias. In fact, in both cases there is a slight increase in the MSE values. This issue is however mitigated if we look at the partition obtained if we insert spatial dependence.

In fact, if we consider the application of the models to the empirical case concerning the price of housing in Rome, we can see that the classical models (FMQR, FMMQE and FMMQR) produce a more dispersed (and less meaningful) spatial pattern, with a strong concentration of territorial units into a single group. The models do not seem to capture the difference existing between houses near and far from the center of Rome.

Instead, analyzing the models with spatial effects (FMQRSP, FMMQSP_{E} and FMMQSP_{R}), the partitions of the units assumes a shape which is more compact and coherent with the nature of the analyzed phenomenon.

As previously seen, we have only taken into consideration the geographical units that are actually adjacent to each other. It would be interestingto apply the proposal taking into account not only the adjacent units, but also their distance. That is considering a continuous rather then a binary measure of adjacency.

A. Appendix

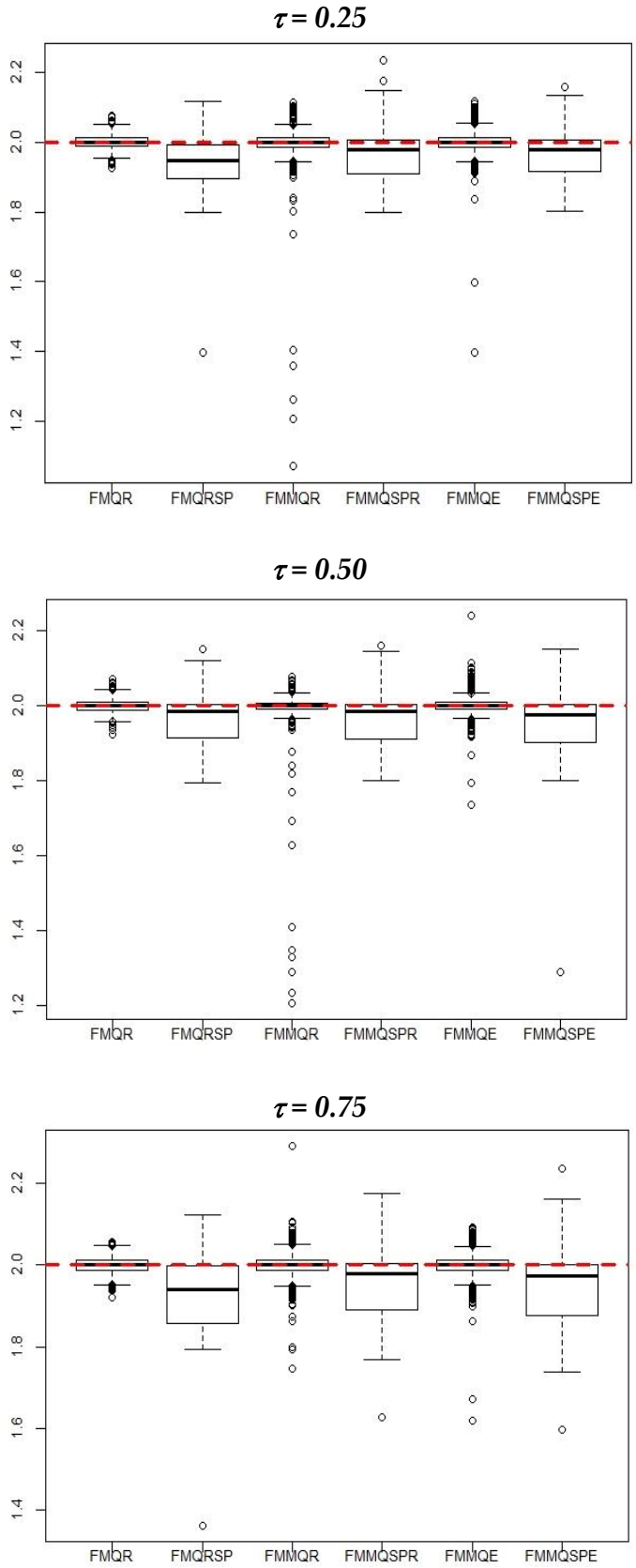

Figure A.1 Boxplot for the didtribution of the fixed parameter estimates (scenario 1)

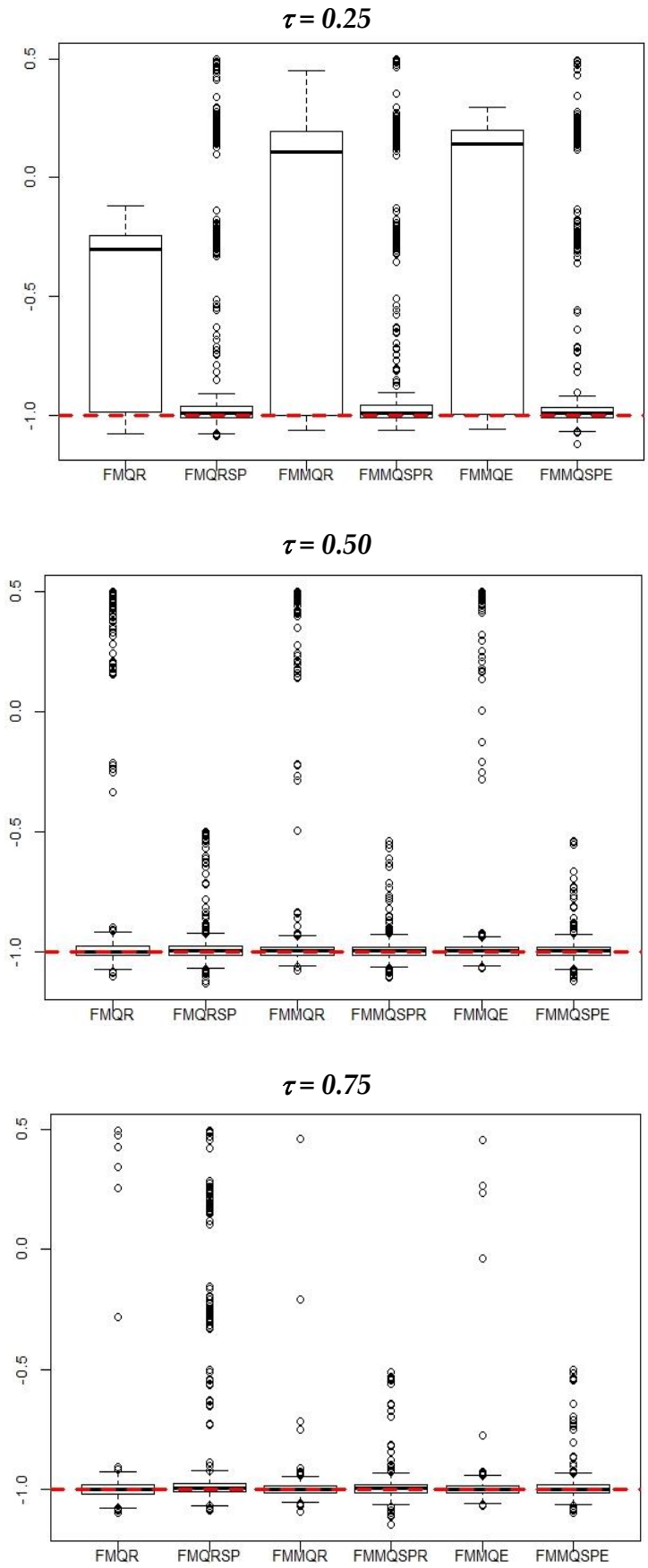

Figure A.2 Boxplot for distribution of β_{21} estimates (scenario 1)

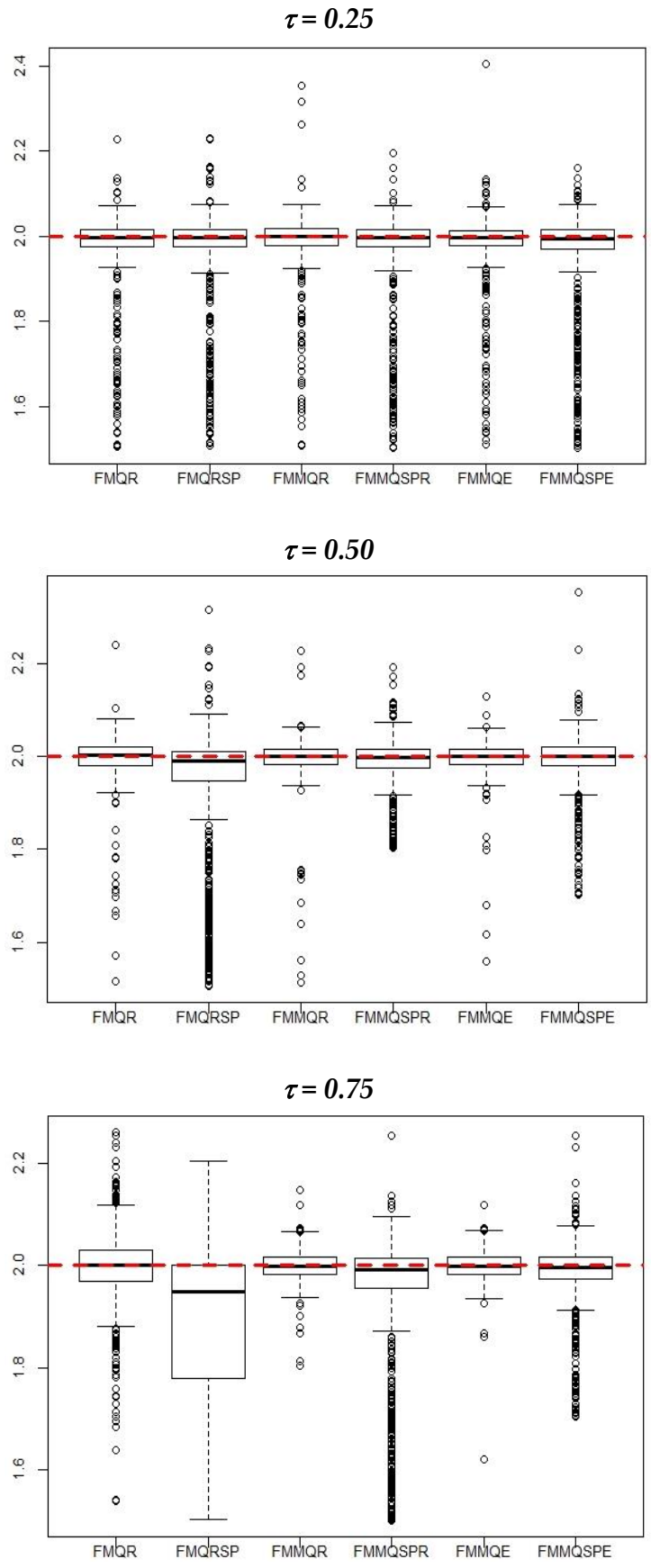

Figure A.3 Boxplot for distribution of β_{22} estimates (scenario 1)

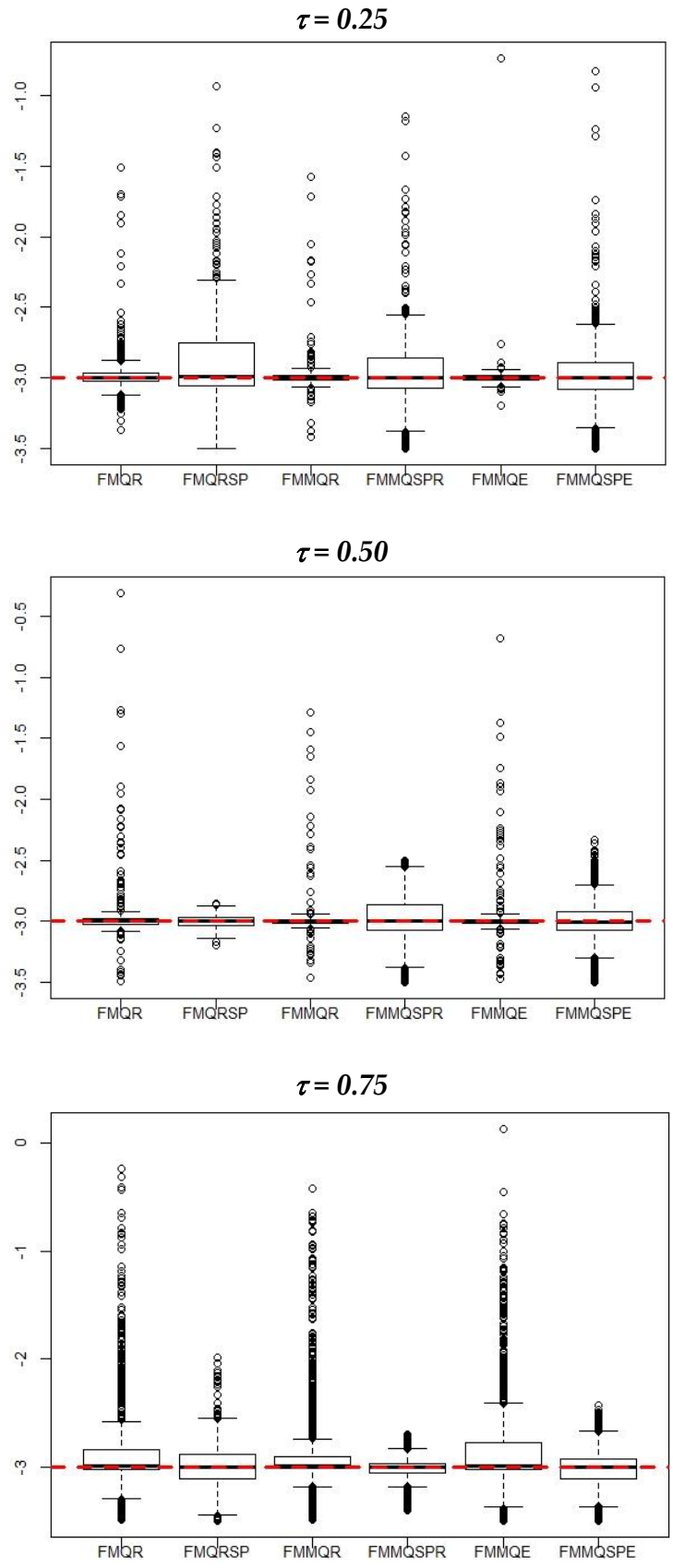

Figure A.4 Boxplot for distribution of β_{23} estimates (scenario 1)

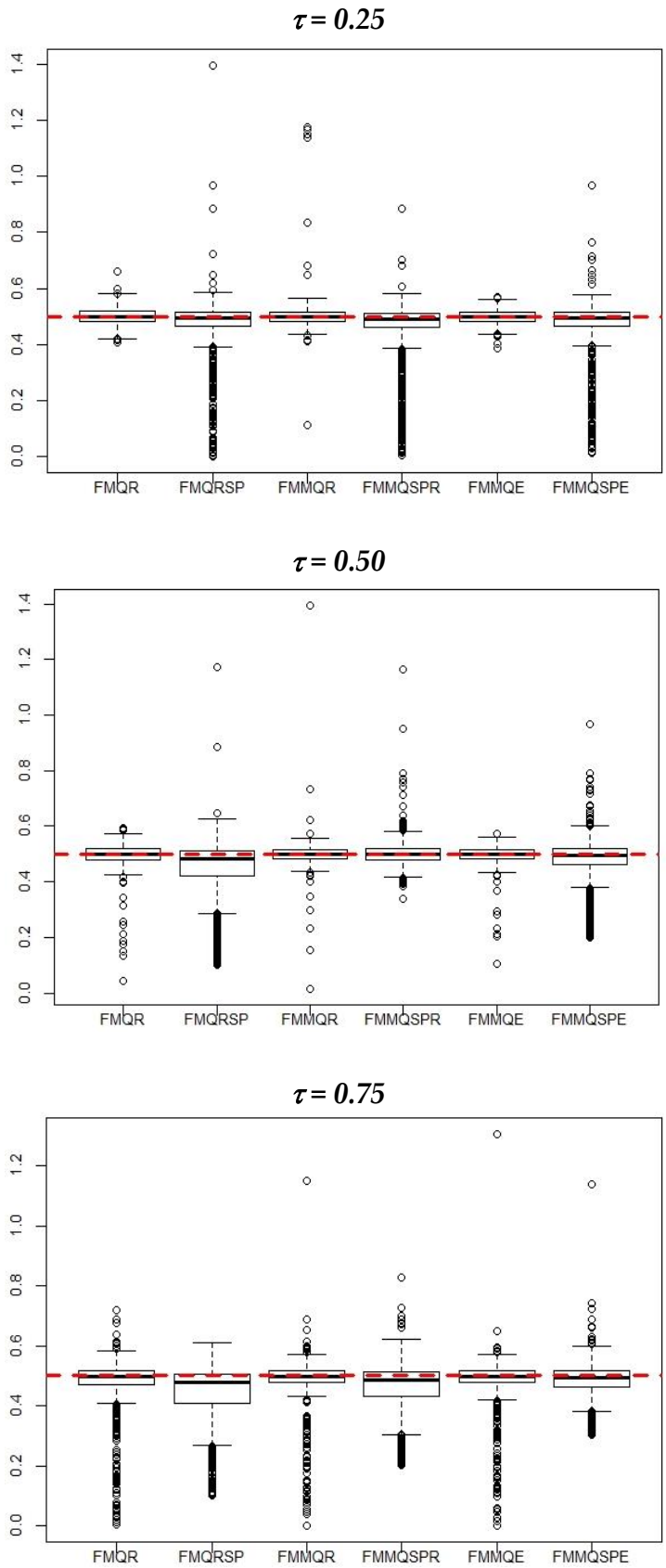

Figure A.5 Boxplot for distribution of β_{24} estimates (scenario 1)

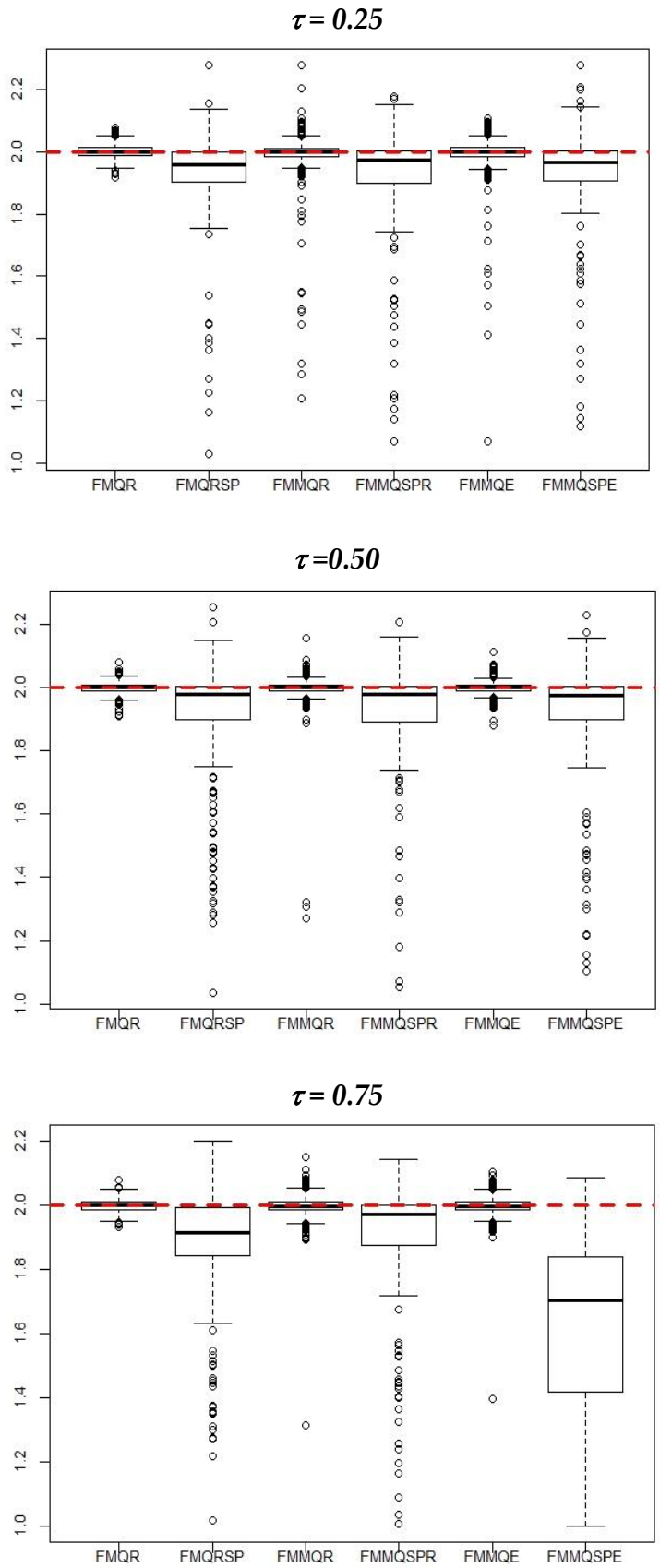

Figure A.6 Boxplot for the didtribution of the fixed parameter estimates (scenario 2)

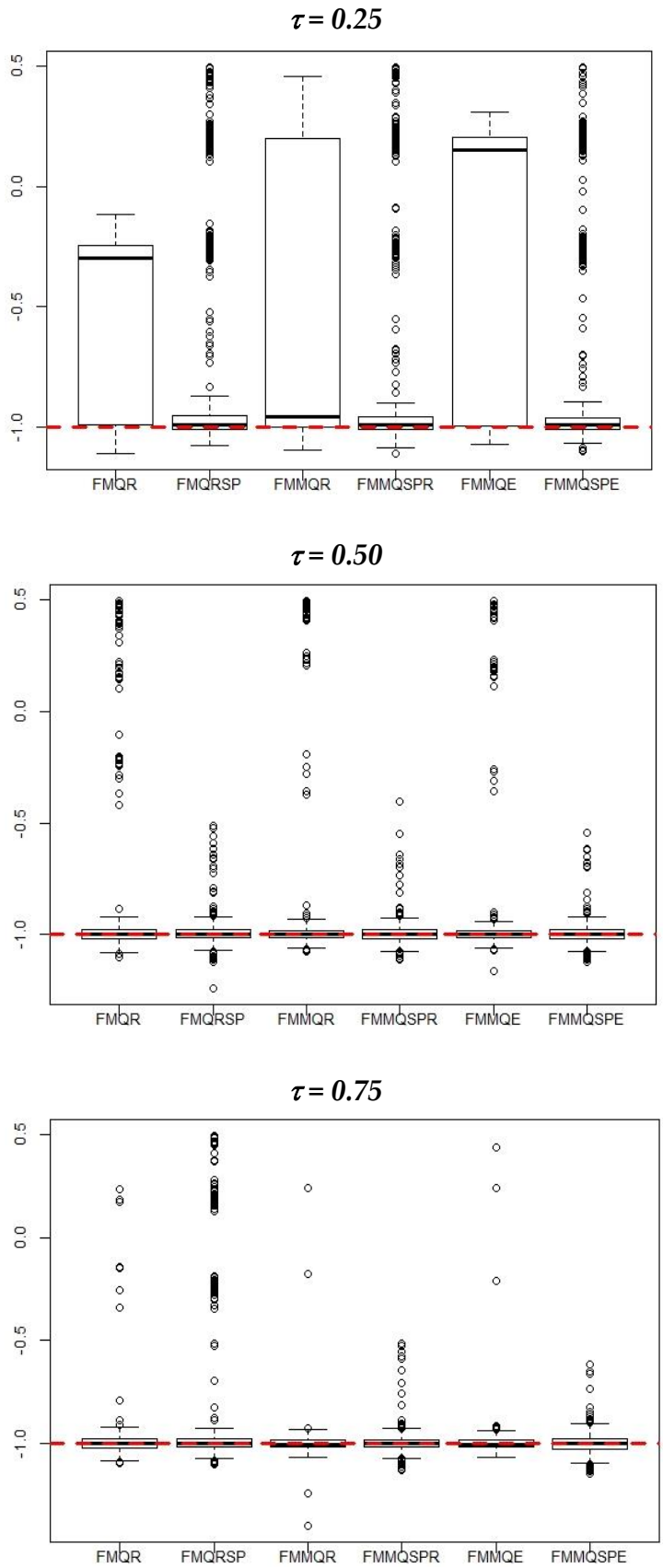

Figure A.7 Boxplot for distribution of β_{21} estimates (scenario 2)

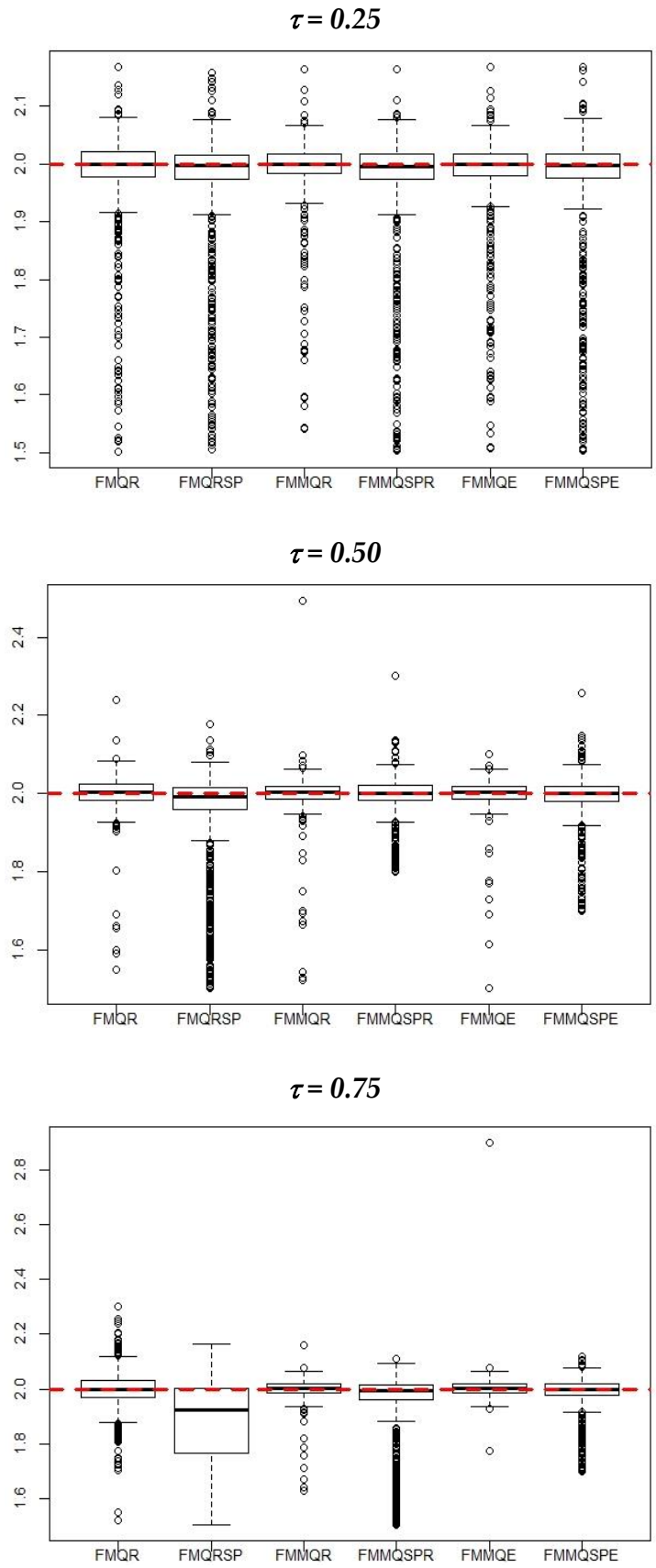

Figure A.8 Boxplot for distribution of β_{22} estimates (scenario 2)

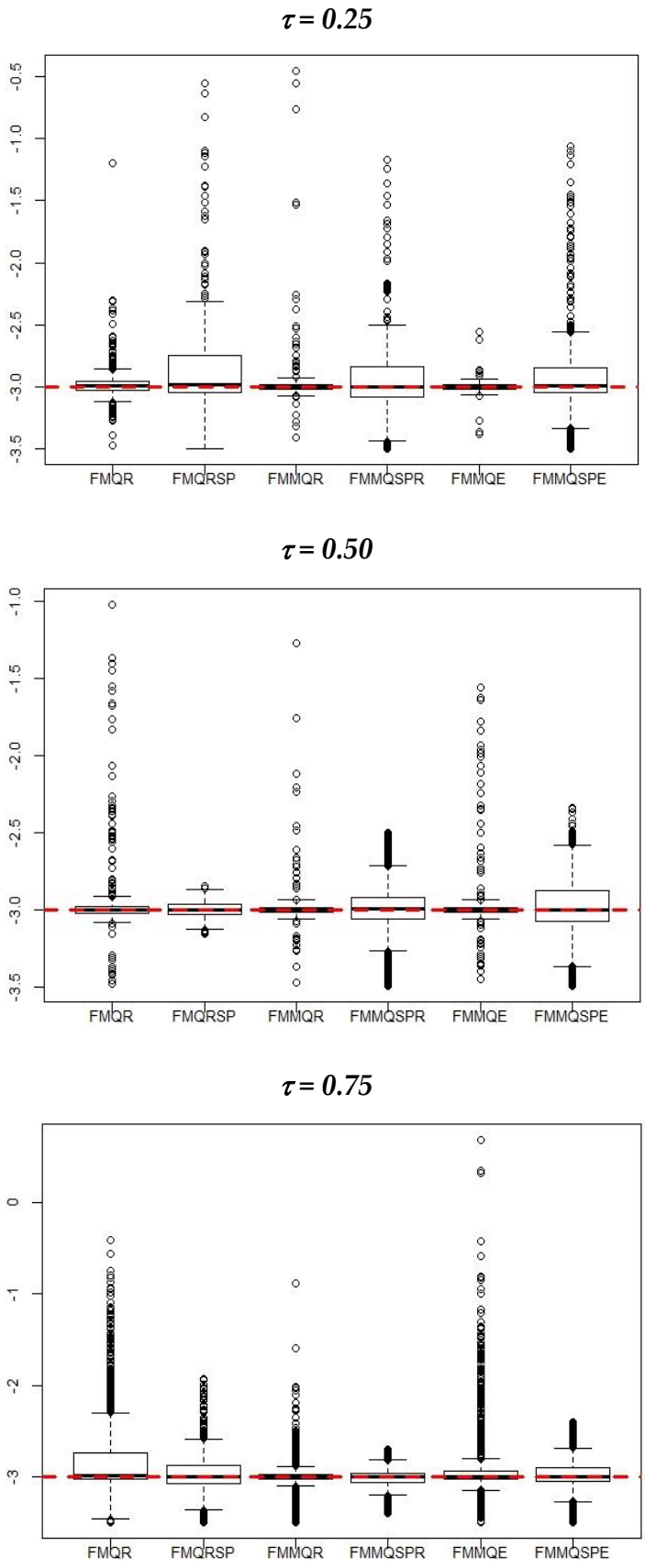

Figure A.9 Boxplot for distribution of β_{23} estimates (scenario 2)

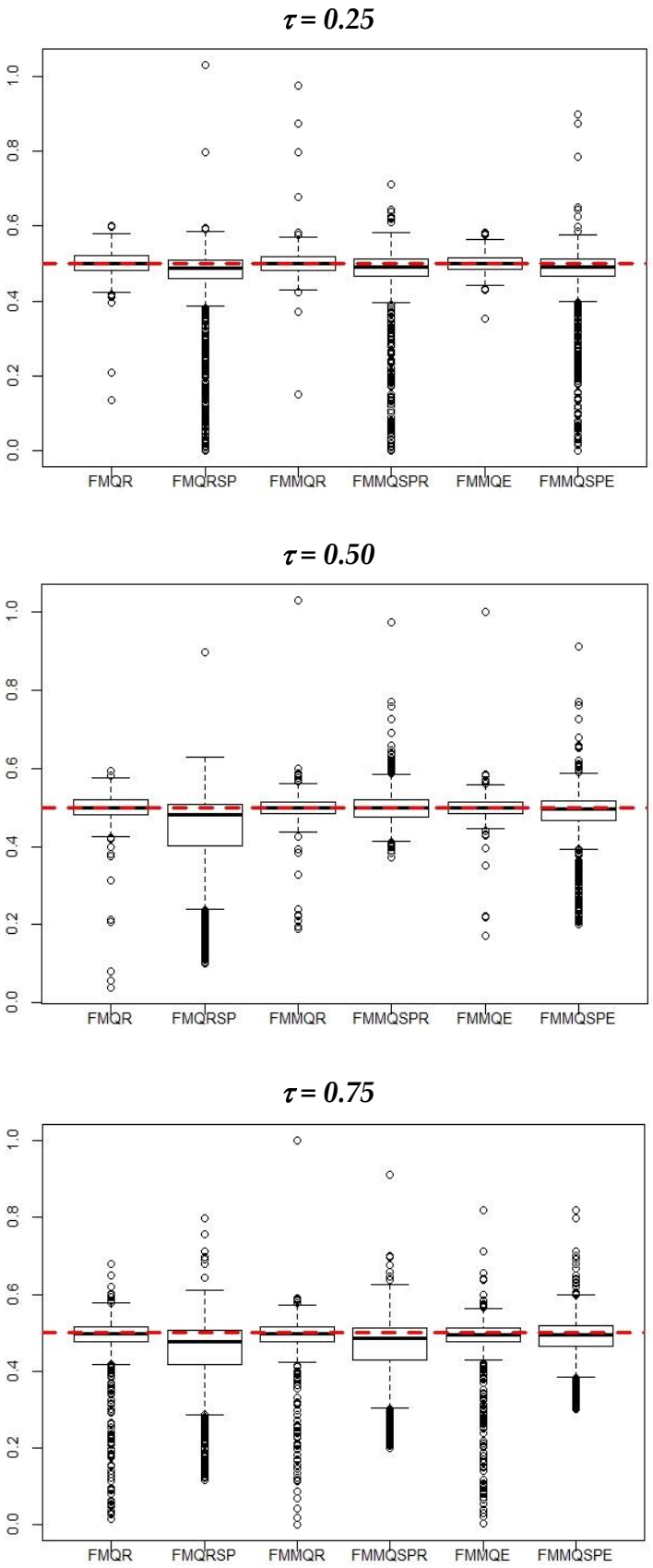

Figure A.10 Boxplot for distribution of β_{24} estimates (scenario 2)

References

AITKIN, M. (1996). A General Maximum Likelihood analysis of Overdispersion in Generalized Linear Models. Statistics and Computing, 6:127-130.

AITKIN, M. (1999). A General Maximum Likelihood analysis of Variance Components in Generalized Linear Models. *Biometrics*, 55:117-128.

AITKIN, M., FRANCIS, B. AND HINDE, J. (2005). Statistical Modelling in GLIM. *Oxford University Press* 2nd edition.

ALFÒ, M. AND TROVATO, G. (2004). Semiparametric mixture models for multivariate count data, with application. *Econometrics Journal*, 7:426-454.

ALFÒ, M., NIEDDU, L. AND VICARI, D. (2008) A finite mixture model for image segmentation. *Stat Comput*. vol 18:137-150

ALFÒ, M., NIEDDU, L. AND VICARI, D. (2009) Finite Mixture Models for Mapping Spatially Dependent Disease Counts. *Statistical Methods in Epidemiology*. vol 51:84-97

ALFÒ, M., SALVATI, N. AND RANALLLI, M.G. (2017). Finite mixtures of quantile and M-quantile regression models. *Stat Comput. vol 27: 547-570*

ANDERSON, R. (1973). Consumer Dissatisfaction: The Effect of Disconfirmed Expectancy on Perceived Product Performance. *Journal of Marketing Research*, 10: 38-44

ABRAMOWITZ, M. AND STEGUN, I. (1964). Handbook of Mathematical Functions. *National Bureau of standards*, Washington DC.

BARRODALE, I. AND ROBERTS, F.D.K. (1974). Algorithm 478: Solution of an Overdetermined System of Equations in the *L*1 Norm. *Communications of the ACM*, 17(6):319-326.

BESAG, J., YORK, J. AND MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* 43, 1–59.

BESAG, J. (1975). Statistical analysis of non-lattice data. *The statistician*, 24 :192–236.

BIANCHI, A., FABRIZI, E., SALVATI, N. AND TZAVIDIS, N. (2015). M-quantile regression: diagnostics and parametric representation of the model. Working paper, available at http://www.sp.unipg.it/surwey/dowload/publications/24-mq-diagn.html.

BIANCHI, A. AND SALVATI, N. (2015). Asymptotic properties and variance estimators of the M-quantile regression coefficients estimators. *Communications in Statistics – Theory and Methods*, 44:2416-2429.

BÖHNING, D. (1982). Convergence of simar's algorithm for finding the maximum likelihood estimate of a compound poisson process. *Annals of Statistics*10, 1006-1008

BOOTH, J. G. AND HOBERT, J. P. (1999). Maximizing generalized linear models with an automated Monte Carlo EM algorithm, *Journal of the Royal Statistical Society* B, 61, 625-85.

BORGONI R., BIANCO P.D., SALVATI N., SCHMID T. AND TZAVIDIS N. (2018) Modelling the distribution of health-related quality of life of advanced melanoma patients in a longitudinal multi-centre clinical trial using m-quantile random effects regression. *Stat Methods Med Res* 27:549–563

BORGONI, R., CARCAGNÍ, A., SALVATI, N. AND SCHMID, T. (2019). Analysing radon accumulation in the home by flexible M-quantile mixed effect regression. *Stochastic Environmental Research and Risk Assessment*, 33(2), 375-394.

BRECKLING, J. AND CHAMBERS, R. (1988). M-quantiles. *Biometrika* 75, 761-771.

CANAY, I. (2011). A Simple Approach to Quantile Regression for Panel Data, *The Econometrics Journal*, vol 14, 3:368-386

DESARBO, W. AND CRON, W. (1988). A maximum likelihood methodology for cluster wise regression. *Journal of Classification*, 5:249-282.

DEL SARTO, S., MARINO, M. F., RANALLI, M. G. AND SALVATI N. (2019).Using finite mixtures of M-quantile regression models to handle unobserved heterogeneity in assessing the effect of meteorology and traffic on air quality. *Stochastic Environmental Research and Risk Assessment*, vol. 33, 7:1345–1359.

FARCOMENI, A. (2012). Quantile Regression for longitudinal data based on latent Markov subject-specific parameters. *Statistics and Computing*, 22, 141-152

FRIEDL, H. AND KAUERMANN, G. (2000). Standard errors for EM estimates in generalized linear models with random effects. Biometrics, 56:761-767.

GEMAN, S. AND GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6(6)*, 721–741.

GERACI, M. AND BOTTAI, M. (2007) Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, 8, 140–154.

GERACI, M. AND BOTTAI, M. (2014) Linear quantile mixed models. *Statist. Comput.*, 24, 461–479.

GEYER, C. AND THOMPSON, E. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society B*, 54:657-699.

GOLDSTEIN, H. (2003). *Multilevel Statistical Models*. Wiley, NewYork.

GREEN P.J. AND RICHARDSON S. (2002). Hidden Markov Models and Disease Mapping. *Journal of the American Statistical Association*. 97:1055-1070

HAINING, R. (1990). Spatial data analysis in the social and environmental sciences. Cambridge University Press.

HAMMERSLEY, J. M. AND CLIFFORD P. (1971). Markov Fields on Finite Graphs and Lattices. *Unpublished*.

HARVILLE A.D., (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. Journal of the American Statistical Association, 72, 320-338.

HENNIG, C. (2000). Identifiability of models for cluster wise linear regression. *Journal of Classification*, 17:273-296.

HOFFMAN, S. AND DUNCAN, G. (1988). Multinomial and conditional logit discrete-choice models in demography, *Demography*, vol. 25,3: 415-427

HUBER, P. J. (1964) Robust Estimation of a Location Parameter. Ann. Math. Statist. 35, no. 1, 73—101

HUBER, P. J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 221–233. Berkeley: University of California Press.

HUBER, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Ann. Statist. 1*, no. 5, 799--821.

HUBER, P. J. (1981) *Robust Statistics*. New York: Wiley.

HUBERT, L. AND ARABIE, P. (1985). Comparing Partitions, *Journal of the Classification*, 2:193-218.

JONES, M. C. (1994), Expectiles and *M*-Quantiles Are Quantiles, *Statistics and Probability Letters*, 20, 149–153.

KINDERMANN, R. AND SNELL, J.L. (1999). *Markov Random Fields and Their Applications*, vol. 1. American Mathematical Society, Providence

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

KOENKER, R. (2005). *Quantile Regression*. New York: Cambridge University Press.

KOENKER, R. AND BASSETT, G. (1978). Regression quantiles. *Econometrica*46, 33–50.

KOENKER, R. AND MACHADO, J. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* 94, 1296–309.

KOENKER, R. AND D'OREY, V. (1987). Algorithm AS 229: Computing Regression Quantiles. *Applied Statistics*, 36(3):383-393.

KOKIC, P., CHAMBERS, R., BRECKLING, J. AND BEARE, S. (1997). A measure of production performance. *Journal of Business and Economic Statistics*, 15, 445-451.

LAMARCHE, C. (2010). Robust Penalized Quantile Regression Estimation for Panel Data, *Journal of Econometrics*, vol 157, 2:396-408

LAIRD, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73:805-811.

LEE, D. AND NEOCLEOUS, T. (2010). Bayesian quantile regression for count data with application to environmental epidemiology. *Journal of the Royal Statistical Society Series* C 59, 905–920.

LINDSAY, B. G. (1983a). The Geometry of Mixture Likelihoods: A General Theory. *Ann. Statist.* 11, 1:86-94.

LINDSAY, B.G. (1983b). The geometry of mixture likelihoods, Part I: A general theory. Annals of Statistics 11, 783-792

LIU, Q. AND PIERCE, D. (1994). A note on Gaussian-Hermite quadrature. *Biometrika*, 81:624-629.

LIU, Y. AND BOTTAI, M. (2009). Mixed-effects models for conditional quantiles with longitudinal data. Internation Journal of Biostatistics, 5:1-22.

LIPSITZ, S., FITZMAURICE, G., MOLENBERGHS, G. AND ZHAO, L. (1997). Quantile regressionmethods for longitudinal data with drop-outs: application to cd4 cell counts of patients infected with the human immunodeficiency virus. *Journal of the Royal Statistical society*, series C, 46:463-476

LOUIS, T. A. (1982). Finding the Observed Information Matrix When Using the EM Algorithm, *Journal of the Royal Statistical Society*. Series B: Methodological 44(2):226-233

MENG, X. AND VAN DYK, D. (1998). Fast EM-Type Implementations for Mixed Effects Models. *Journal of the Royal Statistical Society*. Series B, Vol. 60, 3:559-578

MCCULLOCH, C. (1994). Maximum likelihood estimation of variance components for binary data. *Journal of the American Statistical Association*, 89:330-335.

MCCULLOCH, C. E. AND NEUHAUS, J.M. (2005) *Generalized Linear Mixed Models.*

MCCULLOCH, C. E. AND SEARLE, S. R. (2000) *Generalized, Linear, and Mixed Models*, Wiley Series in Probability and Statistics

MCFADDEN, D. (1973) Conditional Logit Analysis of Qualitative Choice Be. In: Zarembka, P., Ed., *Frontiers in Econometrics*, Academic Press, New York, 105-142

MUNDLAK, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica*, vol. 46, 1:69-85.

MUNKIN, M. K. AND TRIVEDI, P. K. (1999). Simulated maximum likelihood estimation of multivariate mixed-Poisson regression models, with application. *The Econometrics Journal*, 2:29-48.

NEWEY,W. K. AND POWELL,J. L. (1987) "Asymmetric Least Squares Estimation and Testing," *Econometrica*, 55, 819–847

NG, C. T. AND JOE, H. (2014) Model comparison with composite likelihood information criteria. *Bernoulli*. vol. 20, 4: 1738--1764.

OAKES, D. (1999). Direct Calculation of the Information Matrix via the EM Algorithm. *Journal of the Royal Statistical Society*, series B, 61:479-482.

OPSOMER J., CLAESKENS G., RANALLI M., KAUERMANN G. AND BREIDT F. (2008) Nonparametric small area estimation using penalized spline regression. *J R Stat Soc Ser B* 70(1):265–283

PINHEIRO, J. AND BATES, D. (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics*, 4:12-35.

PRATESI M., RANALLI M. AND SALVATI N. (2009) Nonparametric m-quantile regression using penalized splines. *J Nonparametr Stat* 21:287–304.

PRESS, W., TEUKOLSKY, S., VETTERLING, W. AND FLANNERY, B. (2007). Numerical recipes: the art of scientific computing. New York: Cambridge University Press.

CANTONI, E. AND RONCHETTI, E. (2001). Robust Inference for Generalized Linear Models. *Journal of the American Statistical Association*. Vol.96, 455:1022-1030.

REICH, B. J., BONDELL, H. D. AND WANG, H. J. (2010). Flexible Bayesian quantile regression for independent and clustered data, *Biostatistics*, 11, 337–352.

RICHARDSON, A. AND WELSH, A. (1995). Robust Restricted Maximum Likelihood in Mixed Linear Models. *Biometrics.* vol. 51, 4:1429-1439.

RIZOPOULOS, D., VERBEKE, G. AND LESAFFRE, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society*, Series B. vol 71:637–654.

RUPPERT D., WAND M. AND CARROLL R. (2003) Semiparametric regression. *Cambridge University Press*, Cambridge.

SEARLE, S.R., CASELLA, G. AND MCCULLOCH, C.E., (1992). *Variance components*. New York: Wiley.

SIMAR, L. (1976). Maximum likelihood estimation of a compound poisson process. *Annals of Statistics*4:1200-1209

SINHA, S.K. AND RAO, J.N.K. (2009). Robust small area estimation. *Canad. J. Stat.*, 37, 381-399.

STRAUSS, D.J. (1977). Clustering on coloured lattices. *J. Appl. Probab.*14(1), 135–143

STRAUSS, D. AND IKEDA**,** M.(1990). Pseudolikelihood estimation for social networks, *Journal of the American statistical association*, 85: 204-212.

STREET, J. O., CARROLL, R. J. AND RUPPERT, D.(1988). A note on computing robust regression estimates via iteratively reweighted least squares, *The American Statistician*, 42, 152–154.

Tobler W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region, *Economic Geography*, 46: 234-240

TZAVIDIS, N., MARCHETTI, S. AND CHAMBERS, R. (2010). Robust prediction of small area means and distributions. *Aust. & NZ J. Stat.*, 52, 167-186.

TZAVIDIS, N., SALVATI, N., SCHMID, T., FLOURI, E. AND MIDOUHAS, E.(2016) Longitudinal analysis of the strengths and difficulties questionnaire scores of the Millennium Cohort Study children in England using M-quantile random-effects regression. *J R Stat Soc Ser A Stat Soc*. 179(2):427-452.

VAN DUIJN, M. A. J., GILE, K. AND HANDCOCK, M. S. (2007). Comparison of Maximum Pseudo Likelihood and Maximum Likelihood Estimation of Exponential Family Random Graph Models. *Working Paper, Center for Statistics and the Social Sciences University of Washington*, 74.

VON EYE, A. AND SCHUSTER, C. (1998). Regression Analysis for Social Sciences. *Elsevier*.

VERBEKE, G. AND MOLENBERGHS, G. (2000).Linear Mixed Models for Longitudinal Data. Springer.

YU, K. AND MOYEED, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4): 437 - 447.

YU, K. AND ZHANG, J. (2005). A three-parameter asymmetric Laplace distribution and its extension, *Communications in Statistics - Theory and Methods*, 34, 1867–1879.

YUAN, Y. AND YIN G. (2010). Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics*. vol.66, 1:105-14

WANG, P., PUTERMAN, M., COCKBURN, I. AND LE, N. (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics*, 52:381-400.

WANG, Y., LIN, X., ZHU, M. AND BAI, Z. (2007). Robust estimation using the Huber function with a data-dependent tuning constant. *Journal of Computational and Graphical Statistics*, 16 (2):468-481.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*. Vol. 48, 4:817-838