Università degli Studi di Pisa
Dipartimento di Statistica e Matematica
Applicata all'Economia

Report n. 287

Nonparametric M-quantile Regression using Penalized Splines

Monica Pratesi, M. Giovanna Ranalli, Nicola Salvati

Pisa, Novembre 2006
- Stampato in Proprio –

Via Cosimo Ridolfi, 10 - 56124 PISA - Tel. Segr. Amm. 050 2216231 Segr. Stud. 050 2216317 Fax 050 2216375
Cod. Fisc. 80003670504 - P. IVA 00286820501 - Web http://statmat.ec.unipi.it/ - E-mail: dipstat@ec.unipi.it

Nonparametric M-quantile Regression using Penalized Splines

Monica Pratesi* M. Giovanna Ranalli † Nicola Salvati‡

November 30, 2006

Abstract

Quantile regression investigates the conditional quantile functions of a response variable in terms of a set of covariates. M-quantile regression extends this idea by a "quantile-like" generalization of regression based on influence functions. In this work we extend it to nonparametric regression, in the sense that the M-quantile regression functions do not have to be assumed to have a certain parametric form, but can be left undefined and estimated from the data. Penalized splines are employed to estimate them. This choice makes it easy to move to bivariate smoothing and additive modeling. The asymptotic properties of model estimates are sketched and an algorithm based on iteratively reweighted penalized least squares to actually fit the model is also proposed. Simulation studies are presented that show the finite sample properties of the proposed estimation technique. The method is then applied to small area estimation for the prediction of the mean Acid Neutralizing Capacity for each 8-digit Hydrologic Unit Codes in the Northeastern states of the US.

Keywords: Small area estimation; Robust regression; Natural Resources Survey; Iteratively Reweighted Least Squares.

1 Introduction

Regression analysis is a standard tool for modeling the relationship between a response variable y and some covariates $\boldsymbol{x}$. It summaries the average behavior of y given $\boldsymbol{x}$ and has been one of the most important statistical methods for applied research for many decades. However, in some circumstances the mean does not give a complete picture of a distribution.

*Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa, m.pratesi@ec.unipi.it

†Dipartimento di Economia, Finanza e Statistica, Università degli Studi di Perugia, giovanna@stat.unipg.it

‡Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa, salvati@ec.unipi.it

It does not consider, for example, the extreme behavior of y conditional on $\boldsymbol{x}$. For this reason, a method that allows for direct modeling of the relationship between the dependent variable and the explanatory variables for these extreme values is needed. In other words, it may be useful to investigate the conditional quantile functions. Such a modeling exercise is referred to as *quantile regression* (Koenker & Bassett, 1978; Koenker & D'Orey, 1987). Quantile regression has been widely used in a broad range of application settings: from financial economics (see Hendricks & Koenker, 1991; Manning et al., 1995) to labor markets studies, which benefit from quantile regression for analyzing wage and income data (Buchinsky, 1994). Other applications concern ecology (Pandey & Nguyen, 1999), event history analysis (Koenker & Geling, 2001) and medicine (Cole & Green, 1992).

M-quantile regression extends this idea by a "quantile-like" generalization of regression based on influence functions (Breckling & Chambers, 1988). For a specified quantile q, in a linear M-quantile regression model the quantile is a linear function of the predictors, i.e. $Q_q(\boldsymbol{x}, \psi) = \boldsymbol{x}\boldsymbol{\beta}_\psi(q)$, where ψ denotes the influence function associated with the qth M-quantile. Practical advantages of M-quantile regression over quantile regression are (a) the guaranteed convergence of the algorithm used to a single solution and (b) the enhanced flexibility associated with the chosen influence function. While nonparametric smoothing has been usefully applied to quantile regression (see e.g. He, 1997; Takeuchi, Le, Sears & Smola, 2005), little or no work has been done on extending M-quantile regression with nonparametric modeling. Our proposal is to extend it by using *penalized splines* (Eilers & Marx, 1996; Ruppert et al., 2003).

The outline of the paper is the following. A short review on M-quantile regression is in Section 2. In Section 3 nonparametric M-quantile regression based on penalized splines is introduced and its properties studied. In Section 4 we report on the results from some simulation studies. The attention is on the performance of the proposed method when a single covariate model expresses the true underlying relationship between y and x, and especially when a bivariate model is considered. This second case is meanly relevant to test the empirical properties of the method when the study variable has a clear spatial pattern as a function of its position in space represented by its geographical coordinates. In Section 5, we extend the M-quantile small area estimation approach of Chambers & Tzavidis (2006) to the setting

in which the functional form of the relationship between the variable of interest and the covariates is left unspecified. A nonparametric model could have significant advantages when the functional form of the relationship between the variable of interest and the covariates is not linear and an erroneous specification of the model can lead to biased estimates. The method is then applied to the estimation of the mean Acid Neutralizing Capacity for each 8-digit Hydrologic Unit Codes (HUCs) in the Northeastern states of the US. Here a survey of 334 lakes in a population of 21,026 has been conducted between the years 1991 and 1996. Finally, in Section 6 we present and discuss our main findings.

2 M-quantile regression

Quantile regression is a generalization of median regression and has been developed by Koenker & Bassett (1978). In the linear case, quantile regression leads to a family of hyper-planes indexed by the value of the corresponding quantile coefficient $q \in (0,1)$. Given a set of covariates $\boldsymbol{x}$ and a response variable y, for each value of q the corresponding model $Q_q(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}(q)$ explains how the q^{th} quantile of the conditional distribution of y given $\boldsymbol{x}$ varies with $\boldsymbol{x}$. The set of regression quantiles parameter estimates satisfies the criterion of minimum sum of absolute asymmetrically weighted residuals: given a sample of n observations, the vector $\boldsymbol{\beta}(q)$ is estimated by minimizing

$$\sum_{i=1}^{n} |r_i[\boldsymbol{\beta}(q)]| \{(1-q) I(r_i[\boldsymbol{\beta}(q)] \leq 0) + q I(r_i[\boldsymbol{\beta}(q)] > 0)\},$$

where $r_i[\boldsymbol{\beta}(q)] = y_i - \boldsymbol{x}_i\boldsymbol{\beta}(q)$, with respect to $\boldsymbol{\beta}(q)$ by using linear programming methods (Koenker & D'Orey, 1987).

Note that regression quantile hyper-planes are not comparable with the regression ones based on ordinary least-squares that describe how the mean of y changes with $\boldsymbol{x}$ (Breckling & Chambers, 1988). In fact, the former are based on an absolute deviations criterion, while the latter on a least-squares one. A first generalization of expectation was suggested by Newey & Powell (1987) through the use of *expectile* lines. M-quantile regression further extends this idea by a "quantile-like" generalization of regression based on influence functions (Breckling & Chambers, 1988). In particular, $Q_q(\boldsymbol{x}, \psi) = \boldsymbol{x}\boldsymbol{\beta}_\psi(q)$ and the general M-estimator of $\boldsymbol{\beta}_\psi(q)$

can be obtained by solving the set of estimating equations

$$\sum_{i=1}^{n} \psi_q(y_i - \boldsymbol{x}_i\boldsymbol{\beta}_\psi(q))\boldsymbol{x}_i^T = \mathbf{0} \tag{1}$$

with respect to $\boldsymbol{\beta}_\psi(q)$, assuming that

$$\psi_q(t) = 2\psi\{s^{-1}(t)\}\{(1-q)I(t \leq 0) + qI(t > 0)\}$$

where s is a robust estimate of scale. Robust regression models can be fitted using an Iterative Reweighted Least Squares algorithm (IRLS) that guarantees the convergence to a unique solution (Kokic et al., 1997).

The advantages of M-quantile regression models are (a) the simplicity of the algorithm used to fit the model and (b) the great flexibility in modeling by using a wide range of influence functions (i.e. Huber function, Huber proposal 2, Hampel function). A drawback for all quantile-type fitted regression plans is the phenomenon of quantile crossing and it is due to model misspecification, collinearity or huge outlying values. He (1997) proposes a restricted version of regression quantiles that avoids the occurrence of crossing while maintaining sufficient modeling flexibility. Another method to overcome this problem is described in Koenker (1984) by forcing proper ordering of the percentile curves. The author considers parallel quantile planes for linear models, but they do not cater to the needs of heteroscedastic models.

In the following section, M-quantiles are extended through nonparametric regression to allow for unknown and maybe complicated relationships between the covariates and the response. We will estimate them through penalized splines.

3 Penalized Splines M-quantile Regression

3.1 The method

Nonparametric regression is a popular technique that extends linear regression by relaxing the assumption of a pre-specified functional relationship between the mean value of y and the covariates $\boldsymbol{x}$. Such relationship does not have to be assumed linear or polynomial, but only an unknown smooth function. Techniques like kernels, local polynomials or smoothing splines can then be used to learn this function from the data (see e.g. Hastie et al., 2001,

for a review of techniques). Smoothing has been usefully applied to quantile regression (see e.g. He, 1997; Takeuchi, Le, Sears & Smola, 2005), but little or no work has been done on extending M-quantile regression with nonparametric modeling. Here we will do so by using penalized splines. Penalized splines are now often referred to as p-splines and have been recently brought up to attention by Eilers & Marx (1996). P-splines provide an attractive smoothing method for their simplicity of implementation, being a relatively straightforward extension of linear regression, and flexibility to be incorporated in a wide range of modeling contexts. Ruppert, Wand & Carroll (2003) provide a thorough treatment of p-splines and their applications.

Let us first consider only smoothing with one covariate x_1; we will then move to bivariate smoothing and semiparametric modeling. Given an influence function ψ, a nonparametric model for the q^{th} quantile can be written as $Q_q(x_1, \psi) = \tilde{m}_{\psi,q}(x_1)$, where the function $\tilde{m}_{\psi,q}(\cdot)$ is unknown and, in the smoothing context, usually assumed to be continuous and differentiable. Here, we will assume that it can be approximated sufficiently well by the following function

$$m_{\psi,q}[x_1; \boldsymbol{\beta}_\psi(q), \boldsymbol{\gamma}_\psi(q)] = \beta_{0\psi}(q) + \beta_{1\psi}(q)x_1 + \ldots + \beta_{p\psi}(q)x_1^p + \sum_{k=1}^{K} \gamma_{k\psi}(q)(x_1 - \kappa_k)_+^p, \quad (2)$$

where p is the degree of the spline, $(t)_+^p = t^p$ if $t > 0$ and 0 otherwise, κ_k for $k = 1, \ldots, K$ is a set of fixed knots, $\boldsymbol{\beta}_\psi(q) = (\beta_{0\psi}(q), \beta_{1\psi}(q), \ldots, \beta_{p\psi}(q))^T$ is the coefficient vector of the parametric portion of the model and $\boldsymbol{\gamma}_\psi(q) = (\gamma_{1\psi}(q), \ldots, \gamma_{K\psi}(q))^T$ is the coefficient vector for the spline one. The latter portion of the model allows for handling nonlinearities in the structure of the relationship. If the number of knots K is sufficiently large, the class of functions in (2) is very large and can approximate most smooth functions. In particular, in the p-splines context, a knot is placed every 4 or 5 observations at uniformly spread quantiles of the unique values of x_1. For large datasets, this rule-of-thumb can lead to an excessive number of knots (and therefore parameters), so that a maximum number of allowable knots, say 40, may be recommended. Note that, on the contrary, the degree of the spline does not have to be particularly large: it is usually taken to be between 1 and 3. The spline model (2) uses a truncated polynomial spline basis to approximate the function $\tilde{m}_{\psi,q}(\cdot)$. Other bases can be used; in particular we will later use radial basis functions to handle bivariate

smoothing. More details on bases and knots choice can be found in Ruppert et al. (2003, Chapters 3 and 5).

Given the large number of knots, model (2) can be over-parametrized and the resulting approximation would look too wriggly. The influence of the knots is limited by putting a constraint on the size of the spline coefficients: typically $\sum_{k=1}^{K} \gamma_{k\psi}^2(q)$ is bounded by some constant, while the parametric coefficients $\boldsymbol{\beta}_\psi(q)$ are left unconstrained. Therefore, estimation can be accommodated by mimicking penalization of an objective function and solving the following set of estimating equations

$$\sum_{i=1}^{n} \psi_q(y_i - \boldsymbol{x}_i\boldsymbol{\beta}_\psi(q) - \boldsymbol{z}_i\boldsymbol{\gamma}_\psi(q))(\boldsymbol{x}_i, \boldsymbol{z}_i)^T + \lambda \begin{bmatrix} \mathbf{0}_{(1+p)} \\ \boldsymbol{\gamma}_\psi(q) \end{bmatrix} = \mathbf{0}_{(1+p+K)}, \tag{3}$$

where $\boldsymbol{x}_i$ here is the i-th row of the $n \times (1+p)$ matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{11}^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{1n}^p \end{bmatrix},$$

while $\boldsymbol{z}_i$ is the i-th row of the $n \times K$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} (x_{11} - \kappa_1)_+^p & \cdots & (x_{11} - \kappa_K)_+^p \\ \vdots & \ddots & \vdots \\ (x_{1n} - \kappa_1)_+^p & \cdots & (x_{1n} - \kappa_K)_+^p \end{bmatrix},$$

and λ is a Lagrange multiplier that controls the level of smoothness of the resulting fit. Note that the set of estimating equations (3) resembles that employed in the linear case in (1) excluding the penalization bit of the spline portion of the model.

The following section explores asymptotic properties of $\hat{\boldsymbol{\beta}}_\psi(q)$ and $\hat{\boldsymbol{\gamma}}_\psi(q)$, while Section 3.3 provides an algorithm to effectively compute them. Once those estimates are obtained, $\hat{m}_{\psi,q}[x_1] = m_{\psi,q}[x_1; \hat{\boldsymbol{\beta}}_\psi(q), \hat{\boldsymbol{\gamma}}_\psi(q)]$ can be computed as an estimate for $Q_q(x_1, \psi)$. The approximation ability of this final estimate will heavily depend on the value of the smoothing parameter λ. Generalized Cross Validation (GCV) has been usefully applied in the context of smoothing splines (Craven & Wahba, 1979) and will be used here too. Details on the criterion are given in Section 3.3.

As we have just dealt with flexible smoothing of M-quantiles in scatterplots, we can now handle the way in which two continuous variables affect the M-quantiles of the response

without any structural assumptions: $Q_q(x_1, x_2, \psi) = \tilde{m}_{\psi,q}(x_1, x_2)$, i.e. we can deal with *bivariate* smoothing. It is of central interest in a number of application areas as environment and public health. It has particular relevance when geographically referenced responses need to be converted to maps. As seen earlier, p-splines rely on a set of basis functions to handle nonlinear structures in the data. Bivariate smoothing requires bivariate basis functions; Ruppert et al. (2003, Chapter 13) advocate the use of radial basis functions to derive *Low-rank thin plate splines*. In particular, we will assume the following model at quantile q for unit i:

$$m_{\psi,q}[x_{1i}, x_{2i}; \boldsymbol{\beta}_\psi(q), \boldsymbol{\gamma}_\psi(q)] = \beta_{0\psi}(q) + \beta_{1\psi}(q)x_{1i} + \beta_{2\psi}(q)x_{2i} + \boldsymbol{z}_i\boldsymbol{\gamma}_\psi(q). \tag{4}$$

Here $\boldsymbol{z}_i$ is the i-th row of the following $n \times K$ matrix

$$\boldsymbol{Z} = [C(\tilde{\boldsymbol{x}}_i - \boldsymbol{\kappa}_k)]_{\substack{1\leqslant i\leqslant n \\ 1\leqslant k\leqslant K}} [C(\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'})]^{-1/2}_{1\leqslant k\leqslant K}, \tag{5}$$

where $C(\boldsymbol{t}) = ||\boldsymbol{t}||^2 \log ||\boldsymbol{t}||$, $\tilde{\boldsymbol{x}}_i = (x_{1i}, x_{2i})$ and $\boldsymbol{\kappa}_k$, $k = 1, \ldots, K$ are knots. The derivation of the $\boldsymbol{Z}$ matrix as in (5) from a set of radial basis functions is lengthy and goes beyond the scope of this paper; Ruppert et al. (2003, Chapter 13), Kammann & Wand (2003) and French, Kammann & Wand (2001) give a thorough treatment of it. Here, it is enough to notice that the $C(\cdot)$ function is applied so that in the full rank case – i.e. when knots correspond to all the observations – the model for classical bivariate smoothing leads to *Thin plate splines* (see e.g. Green & Silverman, 1994). In addition, the second part of the right hand expression in (5) is a transformation used so that the estimation procedure simplifies; in particular, it can again be written as in (3), with $\boldsymbol{x}_i = (1, \tilde{\boldsymbol{x}}_i)$.

The choice of knots in two dimensions is more challenging than in one. One approach could be that of laying down a rectangular lattice of knots, but this has a tendency to waste a lot of knots when the domain defined by x_1 and x_2 has an irregular shape. In one dimension a solution to this issue is that of using quantiles. However, the extension of the notion of quantiles to more than one dimension is not straightforward. Two solutions suggested in literature that provide a subset of observations nicely scattered to cover the domain are *space filling designs* (Nychka & Saltzman, 1998) and the *clara* algorithm (Kaufman & Rousseeuw, 1990, Chapter 3). The first one is based on the maximal separation principle of K points among the unique $\tilde{\boldsymbol{x}}_i$ and is implemented in the `fields` package of the `R` language

(R Development Core Team, 2005). The second one is based on clustering and selects K representative objects out of n; it is implemented in the package `cluster` of R.

It should be noted, then, that the estimating equations in (3) can be used to handle univariate smoothing and bivariate smoothing by suitably changing the parametric and the spline part of the model, i.e. once the $\boldsymbol{X}$ and the $\boldsymbol{Z}$ matrices are set up. Finally, other continuous or categorical variables can be easily inserted parametrically in the model by adding columns to the $\boldsymbol{X}$ matrix. This allows for semiparametric modeling, as intended in Ruppert et al. (2003), to be inherited and applied to M-quantile regression.

3.2 Asymptotic properties

The asymptotic properties of $\hat{\boldsymbol{\beta}}_\psi(q)$ and $\hat{\gamma}_\psi(q)$ will be briefly discussed by using the results in Breckling & Chambers (1988) on M-quantiles of a univariate distribution and those in Huber (1981, Chapter 7) on robust regression. In general, the asymptotic properties of the parameter estimates will be related to those obtained for the M-median, which in turn depend on the shape of the influence curve given by ψ (see the discussion on the choice of ψ in Serfling, 1980, Chapter 7). Here, complications arise as of bias (i) in the M-quantile context, from the fact that for $q \neq 0.5$ the influence function is skewed asymmetric and (ii) in the p-splines context, from the presence of the penalization λ in the estimation procedure.

Let F denote the distribution of y given $\boldsymbol{x}$ underlying the data and $\zeta_F(Q) = \int \psi_q(y - Q)\mathrm{d}F(y|\boldsymbol{x})$. Then the M-quantile of order q is defined as the solution $Q_q(\boldsymbol{x}, \psi)$ of $\zeta_F(Q) = 0$. For simplicity here, let us consider only the univariate case for which the M-quantile takes the spline form in (2) and let $\boldsymbol{\eta}_\psi(q) = (\boldsymbol{\beta}_\psi(q)^T, \boldsymbol{\gamma}_\psi(q)^T)^T$. For a sample $y_1, \ldots, y_n$ from F, the penalized M-estimate is the solution $\hat{\boldsymbol{\eta}}_\psi(q)$ to equation (3) in which the penalty λ, and the number and position of knots are considered fixed. The gross error sensitivity depends on $\mu = \max(q, 1 - q)$ and so does the asymptotic bias of $\hat{\boldsymbol{\eta}}_\psi(q)$. The asymptotic bias also depends on the penalty λ. In fact, for a given λ, $\hat{\boldsymbol{\eta}}_\psi(q)$ resembles a ridge regression estimate, to which it reverts when ψ is the identity function.

Now, $\hat{\boldsymbol{\eta}}_\psi(q)$ is a consistent estimator of $\boldsymbol{\eta}_\psi(q)$ and the asymptotic distribution of $\hat{\boldsymbol{\eta}}_\psi(q)$ can be written as

$$n^{1/2}(\hat{\boldsymbol{\eta}}_\psi(q) - \boldsymbol{\eta}_\psi(q)) \underline{d} N(0, V(\hat{\boldsymbol{\eta}}_\psi(q)),$$

with

$$V(\hat{\boldsymbol{\eta}}_\psi(q)) = \frac{\int \psi_q^2(y - Q_q(\boldsymbol{x}, \psi))\mathrm{d}F(y|\boldsymbol{x})}{[\zeta_F'(Q_q(\boldsymbol{x}, \psi))]^2} \left([\boldsymbol{X}\ \boldsymbol{Z}]^T[\boldsymbol{X}\ \boldsymbol{Z}]\right)^{-1},$$

provided $\zeta_F'(Q_q(\boldsymbol{x}, \psi)) \neq 0$. For this result to hold, $1+p+K$ remains fixed and the following conditions should be satisfied (Yohai & Maronna, 1979; Huber, 1981, Chapter 7):

- the design matrix $[\boldsymbol{X}\ \boldsymbol{Z}]$ has full rank $(1 + p + K)$ and the smallest eigenvalue of $[\boldsymbol{X}\ \boldsymbol{Z}]^T[\boldsymbol{X}\ \boldsymbol{Z}]$ tends towards infinity;
- $\psi(x)$ is nondecreasing and bounded;
- the errors are i.i.d. and such that $E_F(\psi_q^2(y_i - Q_i)) = \nu < \infty$ and $E_F(\psi_q(y_i - Q_i)) = 0$.

In addition to the previous classical assumptions for robust regression estimates, assumptions on the behavior of λ as n grows should be made. In particular, we will require that $\lim_{n\to\infty} n^{-1}\lambda_n = \lambda_*$. In other words, consistency of the estimates holds if λ remains constant or goes to zero as n grows, or if λ grows with n but at a slower rate. Note that both the asymptotic bias and the variance of the asymptotic distribution will depend on q. In particular (e.g. Breckling & Chambers, 1988, Section 2), for the Huber's Proposal 2 type estimator, both the asymptotic bias and variance would increase without bound as q tends towards zero or one.

3.3 The algorithm

Let us rewrite the set of estimating equations in (3) as follows

$$\sum_{i=1}^{n} \psi_q(y_i - \boldsymbol{u}_i\boldsymbol{\eta}_\psi(q))\boldsymbol{u}_i^T + \lambda \boldsymbol{G}\boldsymbol{\eta}_\psi(q) = \boldsymbol{0}_{(1+p+K)}, \tag{6}$$

where $\boldsymbol{u}_i = (\boldsymbol{x}_i, \boldsymbol{z}_i)$, $\boldsymbol{\eta}_\psi(q) = (\boldsymbol{\beta}_\psi(q)^T, \boldsymbol{\gamma}_\psi(q)^T)^T$ and $\boldsymbol{G} = \mathrm{diag}\left\{\boldsymbol{0}_{(1+p)}, \boldsymbol{1}_K\right\}$. If we define the weight function $w(e) = \psi(e)/e$ and let $w_i = w(e_i)$, then (6) can be written as

$$\sum_{i=1}^{n} w_i(y_i - \boldsymbol{u}_i\boldsymbol{\eta}_\psi(q))\boldsymbol{u}_i^T + \lambda \boldsymbol{G}\boldsymbol{\eta}_\psi(q) = \boldsymbol{0}_{(1+p+K)}.$$

Solving this set of estimating equations is a penalized weighted least squares problem in which weights, residuals and coefficients depend one upon another. Further, the value of the smoothing parameter λ has to be chosen. The GCV criterion to be optimized (minimized)

to this end is the following

$$GCV(\lambda) = \frac{\sum_{i=1}^{n} \{(\boldsymbol{I} - \boldsymbol{S}_\lambda)\boldsymbol{y}\}_i}{(1 - n^{-1}\theta \operatorname{tr}(\boldsymbol{S}_\lambda))^2},$$

where $\boldsymbol{S}_\lambda$ is the smoother matrix associated with $\hat{m}_{\psi,q}[\boldsymbol{u}_i]$, i.e. $\hat{m}_{\psi,q}[\boldsymbol{u}_i] = \{\boldsymbol{S}_\lambda\}_i \boldsymbol{y}$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, and θ is a constant that penalizes additional degrees of freedom given by the trace of the smoother matrix.

An iterative solution to this problem through *Iteratively Reweighted Penalized Least Squares* is here proposed. In what follows we will consider the influence function ψ and the quantile of interest q fixed; we will then drop suffixes and indexes for ease of notation when this will not lead to ambiguity. The algorithm is the following:

1. Select initial estimates $\boldsymbol{\eta}^0$.
2. At each iteration t, calculate residuals $e_i^{(t-1)} = y_i - \boldsymbol{u}_i \boldsymbol{\eta}^{(t-1)}$ and associated weights $w_i^{(t-1)}$ from the previous iteration.
3. Optimize the $GCV(\lambda)$ criterion over a grid of λ values and obtain λ^*.
4. Calculate the new weighted penalized least squares estimates as

$$\boldsymbol{\eta}^t = \left[\boldsymbol{U}^T \boldsymbol{W}^{(t-1)} \boldsymbol{U} + \lambda^* \boldsymbol{G}\right]^{-1} \boldsymbol{U}^T \boldsymbol{W}^{(t-1)} \boldsymbol{y},$$

   where $\boldsymbol{U} = \{\boldsymbol{u}_i\}_{i=1,\ldots,n}$ and $\boldsymbol{W}^{(t-1)} = \operatorname{diag}\left\{w_i^{(t-1)}\right\}$ is the current weight matrix.

Iterate steps 2, 3 and 4 until convergence. R code that implements this algorithm is available from the authors.

4 Simulation studies

In this section we report on some Monte Carlo simulation studies carried out to investigate the performance of the p-splines M-quantile regression – PSPL – as compared to standard linear M-quantile – LIN. We first report on simulations with a single covariate and then move to the bivariate case.

4.1 A single covariate

The following four models are used to generate the true underlying relationship between the covariate x and the response variable y:

Linear. $m(x) = 1 + 2(x - 0.5)$;

Exponential. $m(x) = \exp(6x)/400$;

Cycle. $m(x) = 2\sin(2\pi x)$;

Jump. $m(x) = 1 + 2(x - 0.5)I(x \leqslant 0.5) + 0.5I(x > 0.5)$.

The first case represents a situation in which LIN is a good representation of the true model and PSPL may be too complex and overparametrized. The second and the third model define an increasingly more complicated structure of the relationship between y and x, while the last one is a discontinuous function for which both LIN and PSPL are misspecified. More in detail, $n = 200$ x values are generated from a Uniform distribution in $[0, 1]$; y values are generated at each replicate by adding errors to the signals defined above. Two different settings are considered: Gaussian errors with mean 0 and standard deviation 0.4 and Cauchy errors with location parameter 0 and scale parameter 0.05. The first setting is considered as a situation of "regularly" noisy data with a signal-to-noise ratio of about 2 for all signals, but the Exponential function for which it is only about 0.4. The second one, on the contrary, defines a situation of more noisy data with the likely presence of extreme and outlying observations. This provides a 4×2 design of simulations.

For each simulation and each of the $R = 1000$ replicates, LIN and PSPL parameter estimates and response estimates at observed x points are calculated at the deciles using the Huber 2 influence function. In addition, for PSPL a truncated linear bases is used, i.e. $p = 1$, with $K = 39$ knots set at x quantiles; the smoothing parameter λ has been chosen via GCV with $\theta = 2$; this means that each additional degree of freedom used to approximate the underlying signal is penalized twice. It is common to use a value of θ between 1 and 3.

For each technique the following quantities are computed at each quantile to compare performances:

MCEV. Monte Carlo Expected Value, defined for each i and each q as

$$R^{-1} \sum_{r=1}^{R} \hat{m}^r_{\psi,q}[x_i];$$

MASE. Mean Average Squared Errors, defined for each q as

$$(Rn)^{-1} \sum_{i=1}^{n} \sum_{r=1}^{R} (\hat{m}^r_{\psi,q}[x_i] - m_{\psi,q}[x_i])^2;$$

MADE. Mean Absolute Deviation Error, defined for each q as

$$(Rn)^{-1} \sum_{i=1}^{n} \sum_{r=1}^{R} |\hat{m}^r_{\psi,q}[x_i] - m_{\psi,q}[x_i]|.$$

Figure 1 shows the MCEV for both LIN and PSPL for all simulations, together with the true value of the signal for the nine deciles investigated. LIN works well in the linear case for all quantiles; PSPL, on the other hand, seems to work well with more complicated structures and is able to capture even the Cycle signal with a Cauchy error. These findings were somehow expected and are supported by Table 1. It reports the values of the ratios of LIN MASEs to the PSPL ones. Large gains in efficiency of PSPL over LIN are shown for the more complicated structures as expected. In the Linear case, the performance of the two methods is similar in the Gaussian case, while for Cauchy errors PSPL looses in efficiency.

[Figure 1 about here.]

[Table 1 about here.]

Figure 2 reports boxplots of the values of MADE taken by PSPL and LIN for the eight simulations. These plots provide an insight on the variability of the performance of the techniques over the quantiles considered and, therefore, an overall measure of precision. PSPL gives the best performance in all cases but the Linear Cauchy one. This latter behavior can be explained by the fact that the M estimators are not in general robust against the effect of leverage points, as it is the case with Cauchy type errors; a nonparametric approximation of the data may be less robust in these cases than a linear one.

[Figure 2 about here.]

4.2 Bivariate case

In this section simulation studies conducted using two covariates are presented. In particular, x_1 and x_2 take uniformly spread values in the interval $[-1, 1]$ to form a grid of $n = 256$ points. Two model surfaces have been considered:

Plane. $m(x_1, x_2) = 0.5x_1 + 0.2x_2$;

Mountain. $m(x_1, x_2) = \cos\sqrt{(1.2\pi x_1)^2 + (1.2\pi x_2)^2}$.

Figure 3 shows the perspective plots of these two models. Response values are generated at each simulation replicate by adding errors to the surfaces introduced. As in the previous section, two settings are considered: Gaussian errors with mean 0 and standard deviation 1 and Cauchy errors with location parameter 0 and scale parameter 1: Signal to noise ratios for the Gaussian settings take values of 0.11 for the Plane surface and 0.26 for the Mountain one; these represent less good-quality datasets compared to the univariate case, but it was important to test them in view of the application in the following section. This becomes especially true for the Cauchy errors distribution case. A 2×2 design of simulations is therefore set up. For each of the $R = 1000$ replicates LIN and PSPL parameters and surface estimates have been computed; in particular, PSPL uses the radial basis mentioned in Section 3.1 with $K = 50$ knots laid down on a regular grid. The performance quantities computed for the two techniques are the same as those explored for the univariate case.

[Figure 3 about here.]

Plots of MCEV for all cases would be too space consuming and are not reported here, although available from the authors. Here we report only those for the Plane and Mountain with Gaussian errors simulations and a subset of quantiles. Figures 4 and 5 are arranged with quantiles on rows and, respectively, the true surface, LIN and PSPL MCEVs on columns. Biases look negligible in all cases except for the LIN approximation of the Mountain surface as expected.

[Figure 4 about here.]

[Figure 5 about here.]

Table 2 reports MASE ratios for all quantiles and the four simulations. Gains in efficiency for PSPL are shown as expected for the Mountain response surface. Such gains are more remarkable for the Gaussian errors distribution. Losses in efficiency are shown for the Plane surface and central quantiles. In Figure 6 MADE boxplots are reported for all four simulations and show that PSPL may again be useful and reliable when an overall precision tool is required.

[Table 2 about here.]

[Figure 6 about here.]

5 Application to small area estimation

5.1 The methodology

In many surveys it is common to compute estimates for portions (small areas, small domains) of the population of interest such as a mean, a total or a proportion of a variable y. It may happen that sample sizes for such portions tend to be too small, sometimes non-existent, to provide reliable direct – design-based – estimates. Consequently, small area estimation techniques have been developed to satisfy the need for small area statistics without further burdening the already constrained budget for the survey. In a model based perspective to small area estimation, methods based on M-quantile regression focus on the quantiles of the distribution of the study variables (Chambers & Tzavidis, 2006) . When the functional form of the relationship between the q^{th} quantile and the covariates is not linear, a PSPL model may have significant advantages compared to the LIN model. In fact, an erroneous specification of the M-quantile model can lead to biased estimators of the small area parameters.

PSPL is applied to the estimation of a small area mean as follows. The first step is to estimate the M-quantile coefficients q_i for each unit i in the probabilistic sample s of size n without reference to the m small areas of interest. This is done defining a fine grid of values on the interval $(0, 1)$ and using the sample data to fit the PSPL functions at each value q on this grid. If a data point lies exactly on the qth fitted curve, then the coefficient of the corresponding sample unit is equal to q. Otherwise, to obtain q_i, a linear interpolation over the grid is used.

If a hierarchical structure does explain part of the variability in the population data, we expect units within clusters defined by this hierarchy to have similar M-quantile coefficients. Therefore, an estimate of the mean quantile for area j, $\bar{q}_j$, is obtained by taking the corresponding average value of the sample M-quantile coefficient of each unit in area j, $\hat{\bar{q}}_j = \sum_{i=1}^{n_j} q_i$. In case of out of sample areas, $\hat{\bar{q}}_j$ can be set to 0.5. The small area estimator

of the mean $\bar{y}_j$ is then

$$\hat{\bar{y}}_j = \frac{1}{N_j}\Big\{\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j} \hat{y}_{ij}\Big\} \tag{7}$$

where s_j and r_j denote the sampled and non sampled units in area j, respectively, with $U_j = s_j \cup r_j$, and N_j is the known population size of area j. Note that the unobserved value for population unit $i \in r_j$ is predicted using $\hat{y}_{ij} = \boldsymbol{x}_{ij}\hat{\boldsymbol{\beta}}_\psi(\hat{\bar{q}}_j) + \boldsymbol{z}_{ij}\hat{\boldsymbol{\gamma}}_\psi(\hat{\bar{q}}_j)$ where $\hat{\boldsymbol{\beta}}_\psi(\hat{\bar{q}}_j)$ and $\hat{\boldsymbol{\gamma}}_\psi(\hat{\bar{q}}_j)$ are the coefficient vectors of the parametric and spline portion, respectively, of the fitted PSPL function at $\hat{\bar{q}}_j$.

The estimator of the small area mean can be biased for small areas containing outliers. This has already been noted in Tzavidis & Chambers (2006) for the estimator under the LIN model. They propose an adjustment for bias based on the Chambers & Dunstan (1986) estimator (denoted by a subscript CD) of the small area distribution function. The adjusted small area distribution function is

$$\hat{F}_{\mathrm{CD},j}(t) = \frac{1}{N_j}\Big\{\sum_{i\in s_j} I(y_{ij} \le t) + \frac{1}{n_j}\sum_{i\in r_j}\sum_{k\in s_j} I\left(\{\hat{y}_{ij} + [y_{kj} - \hat{y}_{kj}]\} \le t\right)\Big\}, \tag{8}$$

where $\hat{y}_{ij}$ and $\hat{y}_{kj}$ are the predicted values for the population units in r_j and s_j, respectively. The corresponding bias-adjusted estimator for the mean is then

$$\hat{\bar{y}}_j = \frac{1}{N_j}\Big\{\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j} \hat{y}_{ij} + \frac{N_j - n_j}{n_j}\sum_{i\in s_j}(y_{ij} - \hat{y}_{ij})\Big\}. \tag{9}$$

Other quantiles of the distribution function can be obtained by appropriately integrating the CD estimator of the distribution function.

Following the approach described in Chandra & Chambers (2005) and Chambers & Tzavidis (2006), for fixed q and λ, the $\hat{\bar{y}}_j$ can be written as the following linear combination of the observed y_i,

$$\hat{\bar{y}}_j = \frac{1}{N_j}\sum_{i\in s} w_{ij} y_i, \tag{10}$$

with weights $\boldsymbol{w}_j = (w_{1j}, \ldots, w_{nj})^T$ given by

$$\boldsymbol{w}_j = \frac{N_j}{n_j}\mathbf{1}_{s_j} + \boldsymbol{W}(\hat{\bar{q}}_j)[\boldsymbol{X}\ \boldsymbol{Z}]\Big([\boldsymbol{X}\ \boldsymbol{Z}]^T\boldsymbol{W}(\hat{\bar{q}}_j)[\boldsymbol{X}\ \boldsymbol{Z}] + \lambda\boldsymbol{G}\Big)^{-1}\left(\boldsymbol{T}_{r_j} - \frac{N_j - n_j}{n_j}\boldsymbol{T}_{s_j}\right) \tag{11}$$

with $\mathbf{1}_{s_j}$ the n-vector with i^{th} component equal to one whenever the corresponding sample unit is in area j and to zero otherwise, $\boldsymbol{W}(\hat{\bar{q}}_j)$ a diagonal matrix that contains the final set

of weights produced by the iteratively reweighted penalized least squares algorithm used to estimate the regression coefficients (see Section 3.3), and with $\boldsymbol{T}_{r_j}$ and $\boldsymbol{T}_{s_j}$ the totals of the covariates for the non-sampled and the sampled units in area j, respectively.

The weights derived from (11) are treated as fixed and a "plug in" estimator of the mean squared error of estimator (10)

$$MSE(\hat{\bar{y}}_j) = var(\hat{\bar{y}}_j - \bar{y}_j) + [bias(\hat{\bar{y}}_j)]^2 \tag{12}$$

can be proposed by using the standard methods for robust estimation of the variance of unbiased weighted linear estimators (Royall & Cumberland, 1978) and by following the results due to Tzavidis & Chambers (2006). The prediction variance of (10) can be approximated by

$$var(\hat{\bar{y}}_j - \bar{y}_j) \approx \frac{1}{N_j^2}\Big(\sum_{i\in s_j}\Big\{d_{ij}^2 + \frac{N_j - n_j}{n_j - 1}\Big\}var(y_{ij}) + \sum_{i\in s\setminus s_j} d_{ij}^2 var(y_{ij})\Big) \tag{13}$$

with $d_{ij} = w_{ij} - 1$ if $i \in s_j$ and $d_{ij} = w_{ij}$ otherwise, and $s\setminus s_j$ the set of sampled units out of area j. The bias can be written as

$$bias(\hat{\bar{y}}_j) \approx \frac{1}{N_j}\Big(\sum_{k=1}^{m}\sum_{i\in s_k} w_{ij}\tilde{y}_{ik} - \sum_{i\in U_j}\tilde{y}_{ij}\Big) \tag{14}$$

where $\tilde{y}_{ik} = \boldsymbol{x}_{ik}\beta_\psi(\hat{\tilde{q}}_k) + \boldsymbol{z}_{ik}\gamma_\psi(\hat{\tilde{q}}_k)$ are the study variable values under the PSPL model. Following the area level residual approach (Tzavidis & Chambers, 2006), we can interpret $var(y_{ij})$ conditionally to the specific area j from which y_i is drawn and hence replace $var(y_{ij})$ in (13) by $(y_{ij} - \hat{y}_{ij})^2$. An estimate of the bias is obtained replacing $\tilde{y}_{ik}$ by $\hat{y}_{ik}$ in (14). A robust estimator of the mean squared error of (10) is given by the sum of the estimator of the variance

$$\widehat{var}(\hat{\bar{y}}_j) = \frac{1}{N_j^2}\left[\sum_{i\in s_j}\left\{d_{ij}^2 + \frac{N_j - n_j}{n_j - 1}\right\}(y_{ij} - \hat{y}_{ij})^2 + \sum_{i\in s\setminus s_j} d_{ij}^2(y_{ij} - \hat{y}_{ij})^2\right] \tag{15}$$

and the squared estimate of the bias

$$\hat{b}^2(\hat{\bar{y}}_j) = \frac{1}{N_j^2}\left(\sum_{k=1}^{m}\sum_{i\in s_k} w_{ij}\hat{y}_{ik} - \sum_{i\in U_j}\hat{y}_{ij}\right)^2 \tag{16}$$

Since the bias-adjusted nonparametric M-quantile estimator is an approximately unbiased estimator of the small area mean, the squared bias term will not impact significantly the mean squared error estimator. The main limitation of the MSE estimator is that it does not account for the variability introduced in estimating the area specific q's and λ. Thus it may underestimate the true MSE. We note also that we can obtain an estimate only for areas where there are at least two sampled units. For all these reasons, we are currently investigating the use of bootstrap as an alternative approach for estimating the MSE.

5.2 The estimation of Acid Neutralizing Capacity at HUC level in NE lakes

Between 1991 and 1996, the Environmental Monitoring and Assessment Program (EMAP) of the U.S. Environmental Protection Agency conducted a survey of lakes in the Northeastern states of the U.S. The survey is based on a population of 21,026 lakes from which 334 lakes were surveyed, some of which were visited several times during the study period. The total number of measurements is 551.

This data set, developed by EMAP, was supplied to us by Space-Time Aquatic Resources Modeling and Analysis Program (STARMAP) at Colorado State University. Figure 7*a* displays the region of interest and the locations of the sampled lakes. The small areas are defined by 8-digit Hydrologic Unit Codes (HUC) within the region of interest. Note that 27 of 113 HUCs are out of sample areas. The target parameter is the mean Acid Neutralizing Capacity (ANC) for each of 113 small areas. ANC is often used as an indicator of the acidification risk of water bodies in water resource surveys. Figure 7*b* shows the distribution of the observed ANC values. The distribution is skewed and there are outlying observations.

[Figure 7 about here.]

For this data set Opsomer et al. (2005) suggest that the dependence of ANC from the geographical position of the lake as represented by the geographical coordinates (UTM coordinate system) of its centroid is more complex than a plane in the space. It can be approximated using a set of radial basis functions in the p-splines context. The method described in Section 5.1 has been applied to the data using bivariate splines with geographical coordinates

of the centroid of each lake as covariates and an extra parametric term for elevation. Units belonging to the same HUC have similar M-quantile coefficients. The mean ANC at HUC level has been estimated by expression (9).

Figure 8 shows the map of the design based direct estimates of the average ANC computed only for the HUCs in the sample (8*a*) and the estimated means for the sampled and not sampled HUCs under the PSPL model (8*b*). Compared to the map in Figure 8*a*, the small area estimation map in 8*b* is made robust against outlying data values. This characteristic can be noted for outlying positive values that lead to higher direct estimates of the mean ANC with respect to those obtained by the PSPL model. The method provides a useful tool to detect those HUCs in which lakes are at risk of acidification (red spots, values of ANC smaller than about 200) accounting for the skewed distribution of the response. This approach can also be usefully applied to the estimation of many area specific parameters including the quantiles (e.g. medians and percentiles) of the distribution of ANC in the different HUCs.

[Figure 8 about here.]

In order to appreciate the results obtained with the introduction of spatial information, we computed the estimates of mean ANC also through a LIN model that uses the same covariates. Figure 9 reports the estimated means for each HUC with the PSPL model (x-axis) and the LIN one (y-axis). There are clear differences for estimates obtained in out of sample HUCs: the shrinkage effect of the prediction by LIN model is avoided by the use of the p-splines. It seems that the spatial spline term allows for improving the model predictions for these areas by "borrowing strength" from related observed units belonging to neighboring HUCs. This behavior is also shown by the results obtained in Opsomer et al. (2005) using mixed effects models.

[Figure 9 about here.]

Let us now look at the estimated precision of the estimates. The estimator under the PSPL model is less variable in each small area than under the LIN one. Table 3 shows the percentage joint distribution of the estimated means (classified as in Figure 9) and

their estimated Coefficient of Variation $\widehat{CV} = \sqrt{mse(\hat{\bar{y}})}/\hat{\bar{y}}$, computed by expressions (15) and (16) for both the PSPL and the LIN approach (LIN in parentheses). In both cases the estimates refer to sampled areas. We can note that the average $\widehat{CV}$ of the estimator under the PSPL model is lower than in LIN case (44.0% vs 112.7%) and the median value is 24.2% in the PSPL case, while it is 50.9% in the LIN case.

[Table 3 about here.]

6 Conclusions

In this paper we propose an extension to M-quantile regression when the functional form of the relationship between the variable of interest and the covariates is not linear or some other pre-specified parametric form. The nonparametric modeling via penalized splines, beyond having all the properties of M-quantile models, allows for dealing with undefined functional forms that can be estimated from the data. To fit the model we propose an algorithm based on penalized iteratively reweighted least squares. Asymptotic properties of the parameter estimators of the p-spline model are discussed. Relative performances of the nonparametric M-quantile regression (PSPL) are evaluated through Monte Carlo experiments. Results from the simulation studies indicate that this approach works well and competes with the conventional M-quantile regression models when the underlying structure of the relationship between the response and the covariates is more complicated than linear.

The PSPL can be widely used in many important application areas, such as financial and economic statistics and environmental and public health modeling. In this work the PSPL models are used for small area estimation. Also in this case they appear to be an useful tool when the functional form of the relationship between the variable of interest and the covariates is left unspecified and the data are characterized by complex patterns of spatial dependence.

Acknowledgements

The work reported here has been developed under the support of the project PRIN *Metodologie di stima e problemi non campionari nelle indagini in campo agricolo-ambientale* awarded

by the Italian Government to the Universities of Cassino, Florence, Perugia, Pisa and Trieste. The authors would like to thank Ray Chambers for his help and support and they are grateful to Space-Time Aquatic Resources Modeling and Analysis Program (STARMAP) for data availability. The views expressed here are solely those of the authors.

References

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

BUCHINSKY, M. (1994). The demand for alcohol: the differential respnse to price. *Econometrica* **62**, 405–458.

CHAMBERS, R. & DUNSTAN, P. (1986). Estimating distribution function from survey data. *Biometrika* **73**, 597–604.

CHAMBERS, R. & TZAVIDIS, N. (2006). M-quantile models for small area estimation. *Applied Statistics* **36**, 383–393.

CHANDRA, H. & CHAMBERS, R. (2005). Comparing eblup and c-eblup for small area estimation. *Statistics in Transition* **7**, 637–648.

COLE, T. & GREEN, P. (1992). Smoothing reference centile curves: the lms method and penalized likelihood. *Statist. Med.* **11**, 1305–1319.

CRAVEN, P. & WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377–403.

EILERS, P. H. C. & MARX, B. D. (1996). Reply to comments on "Flexible smoothing with *B*-splines and penalties". *Statistical Science* **11**, 115–121.

FRENCH, J., KAMMANN, E. & WAND, M. (2001). Comment on paper by ke and wang. *Journal of the American Statistical Association* **96**, 1285–1288.

GREEN, P. J. & SILVERMAN, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman & Hall Ltd.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: with 200 Full-color Illustrations.* Springer-Verlag Inc.

HE, X. (1997). Quantile curves without crossing. *The American Statistician* **51**, 186–192.

HENDRICKS, W. & KOENKER, R. (1991). Hierarchical spline models for conditional quan-

tiles and the demand for electricity. *Journal of the American Statistical Association* **87**, 58–68.

HUBER, P. J. (1981). *Robust Statistics.* John Wiley & Sons.

KAMMANN, E. E. & WAND, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society - Series C,* **52**, 1–18.

KAUFMAN, L. & ROUSSEEUW, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York.

KOENKER, R. (1984). A note on l-estimators for linear models. *Statistics and Probability Letters* **2**, 323–325.

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

KOENKER, R. & D'OREY, V. (1987). Computing regression quantiles. *Biometrika* **93**, 255–268.

KOENKER, R. & GELING, R. (2001). Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association* **96**, 458–468.

KOKIC, P., CHAMBERS, R., BRECKLING, J. & BEARE, S. (1997). A measure of production performance. *Journal of Business and Economic Statistics* **10**, 419–435.

MANNING, W., BLUMBERG, L. & MOULTON, L. (1995). The demand for alcohol: the differential respnse to price. *Journal of Health Economics* **14**, 123–148.

NEWEY, W. & POWELL, J. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55**, 819–847.

NYCHKA, D. & SALTZMAN, N. (1998). Design of air quality monitoring networks. In *Nychka, Douglas, Piegorsch, Walter W. and Cox, Lawrence H. (eds), Case studies in environmental statistics.*

OPSOMER, J., CLAESKENS, G., RANALLI, M., KAUERMANN, G. & BREIDT, F. (2005). Nonparametric small area estimation using penalized spline regression. In *Preprint Series,* I. S. U. Department of Statistics, ed. Iowa.

PANDEY, G. & NGUYEN, V. (1999). A comparative study of regression based methods in regional flood frequency analysis. *J. Hydrol* **225**, 92–101.

R DEVELOPMENT CORE TEAM (2005). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-

0.

ROYALL, R. & CUMBERLAND, W. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association* **73**, 351–358.

RUPPERT, D., WAND, M. P. & CARROLL, R. (2003). *Semiparametric Regression.* Cambridge University Press, Cambridge, New York.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons.

TAKEUCHI, I., LE, V., SEARS, T. & SMOLA, A. (2005). Nonparametric quantile regression. *Journal of Machine Learning Research* **7**, 1001–1032.

TZAVIDIS, N. & CHAMBERS, R. (2006). Bias adjusted distribution estimation for small areas with outlying values. In *S3RI Methodology Working Papers*, ed. Southampton Statistical Sciences Research Institute, ed. Southampton.

YOHAI, V. J. & MARONNA, R. A. (1979). Asymptotic behavior of M-estimators for the linear model. *The Annals of Statistics* **7**, 258–268.

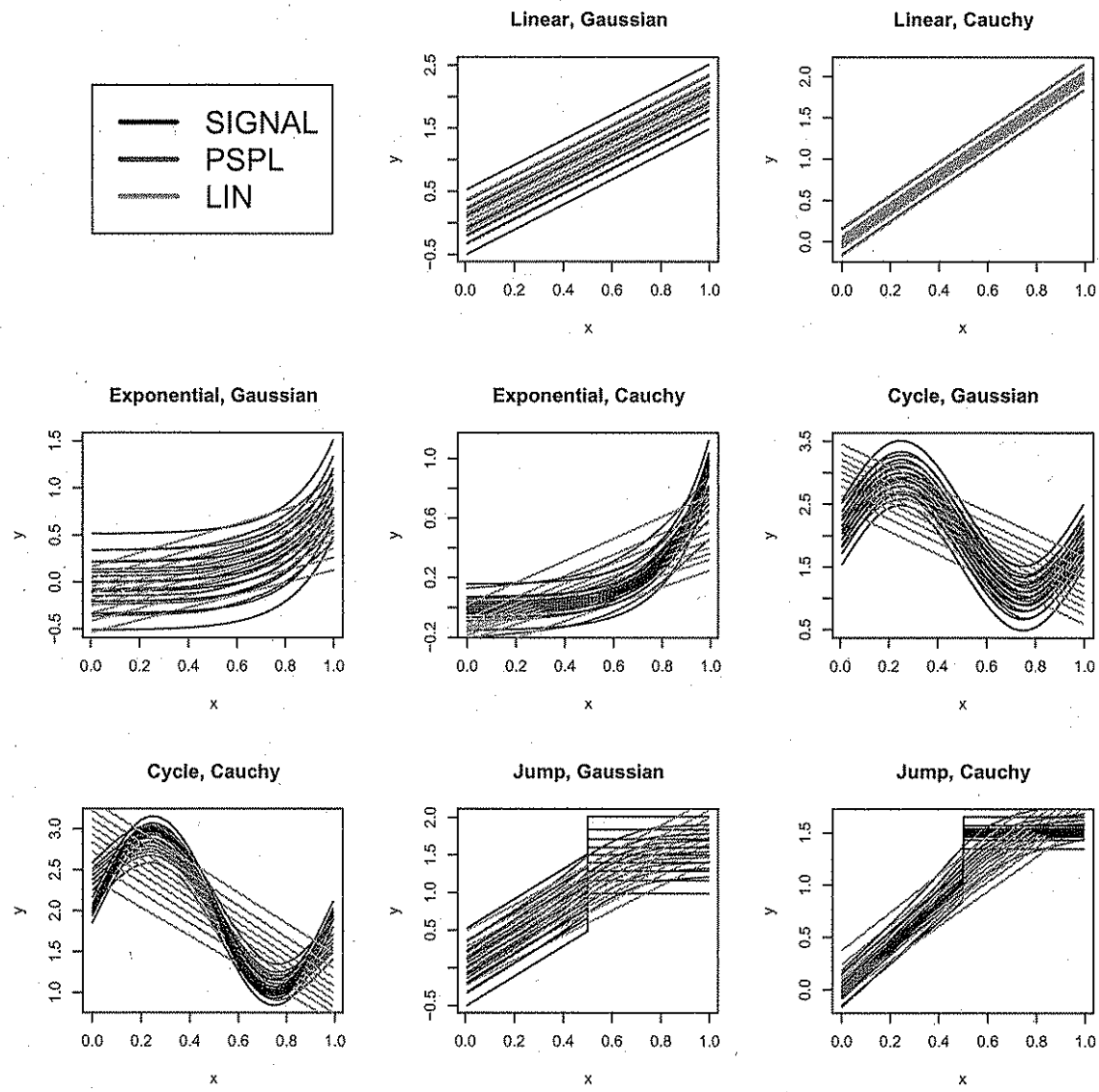


Figure 1: MCEV for LIN and PSPL together with the true quantile functions for all univariate simulation studies.

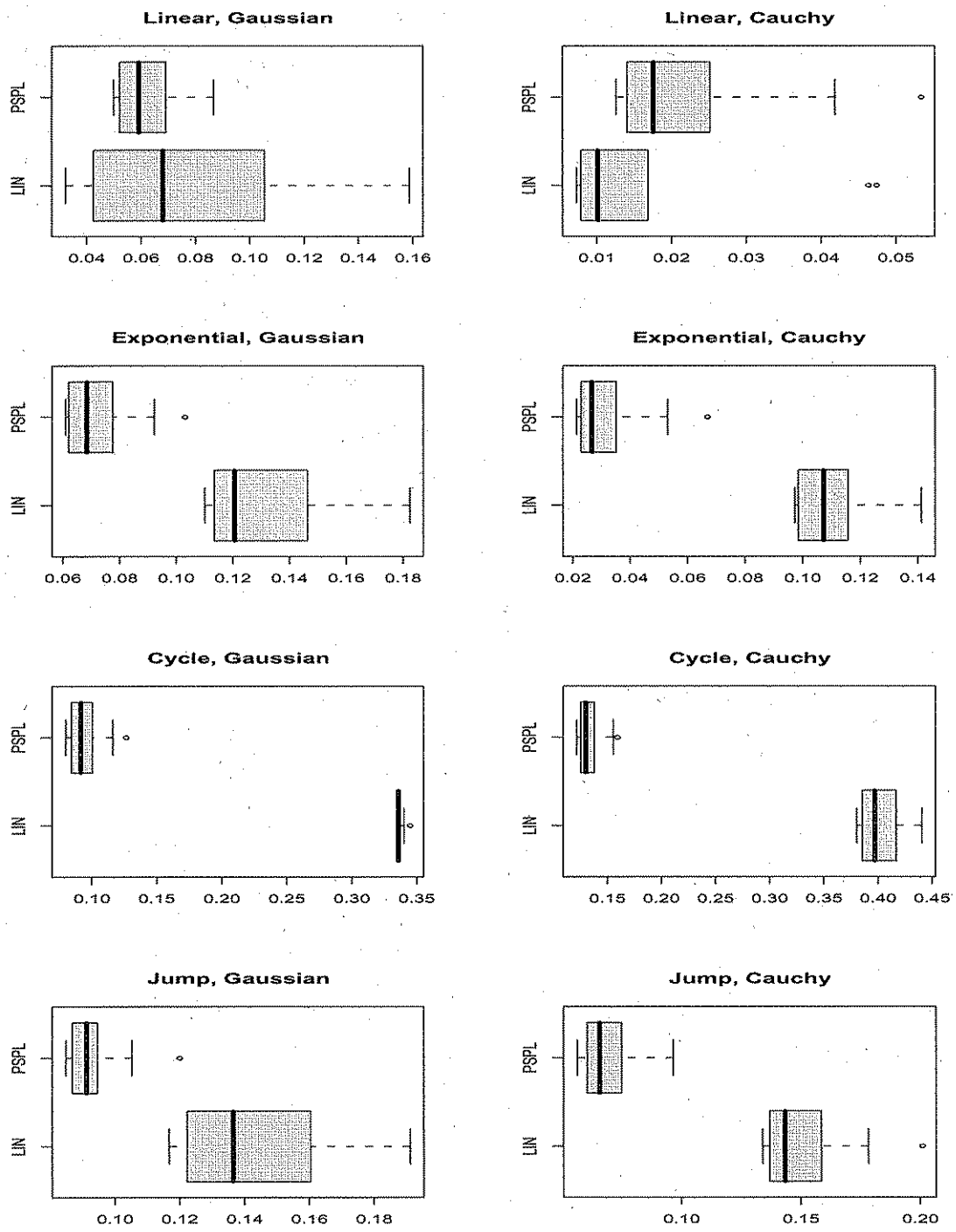


Figure 2: Boxplots of MADE for LIN and PSPL for all univariate simulation studies.

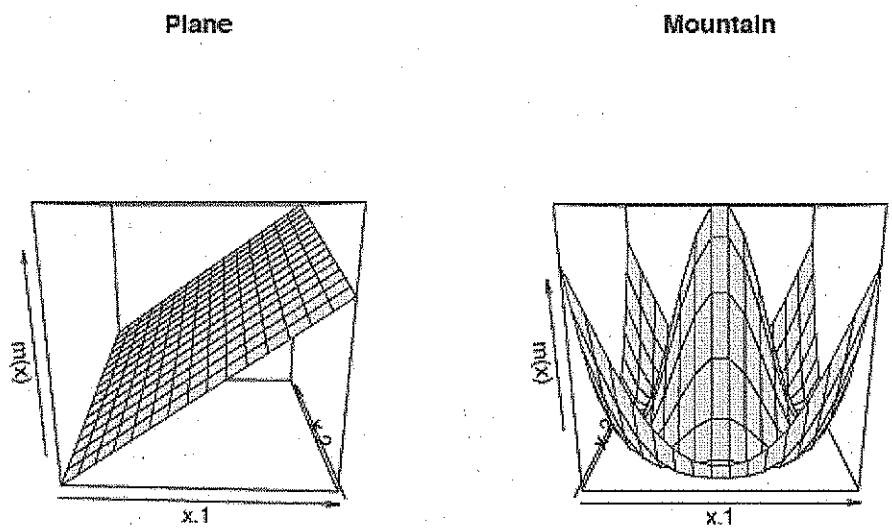


Figure 3: Perspective plots of the two models used in the bivariate simulation studies.

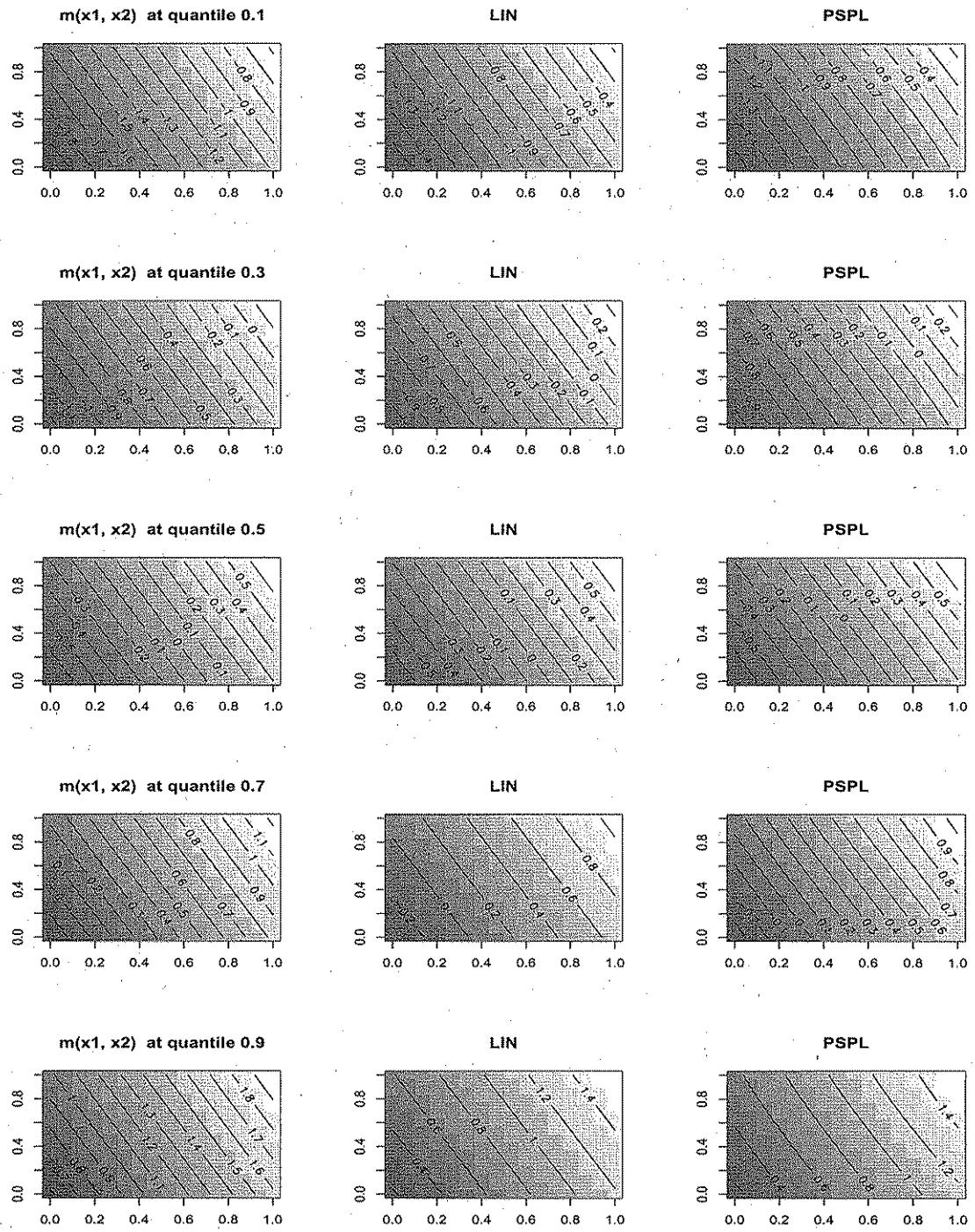


Figure 4: Images of the true quantile function and MCEVs for LIN and PSPL at five quantiles for the Plane with Gaussian errors simulation.

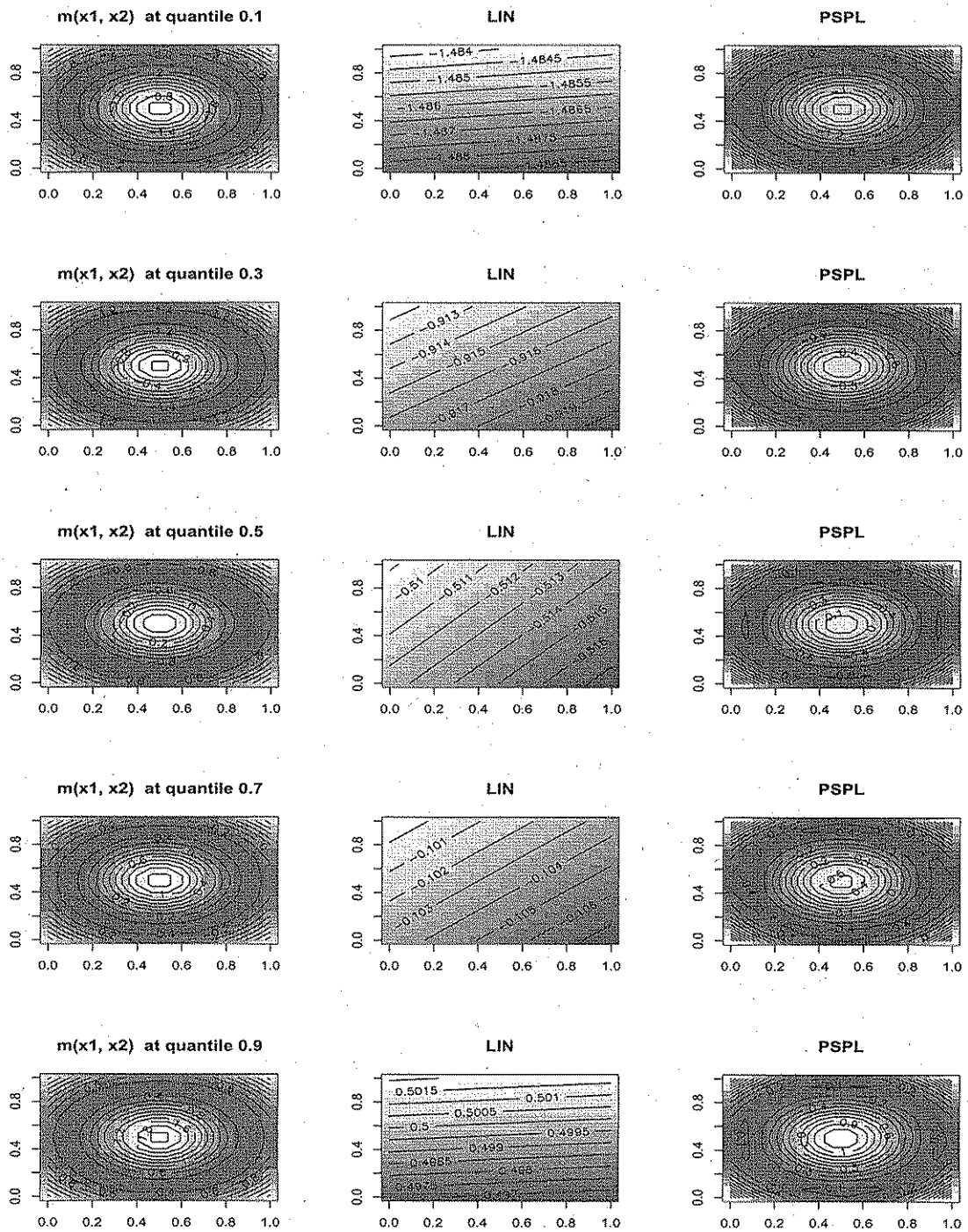


Figure 5: Images of the true quantile function and MCEVs for LIN and PSPL at five quantiles for the Mountain with Gaussian errors simulation.

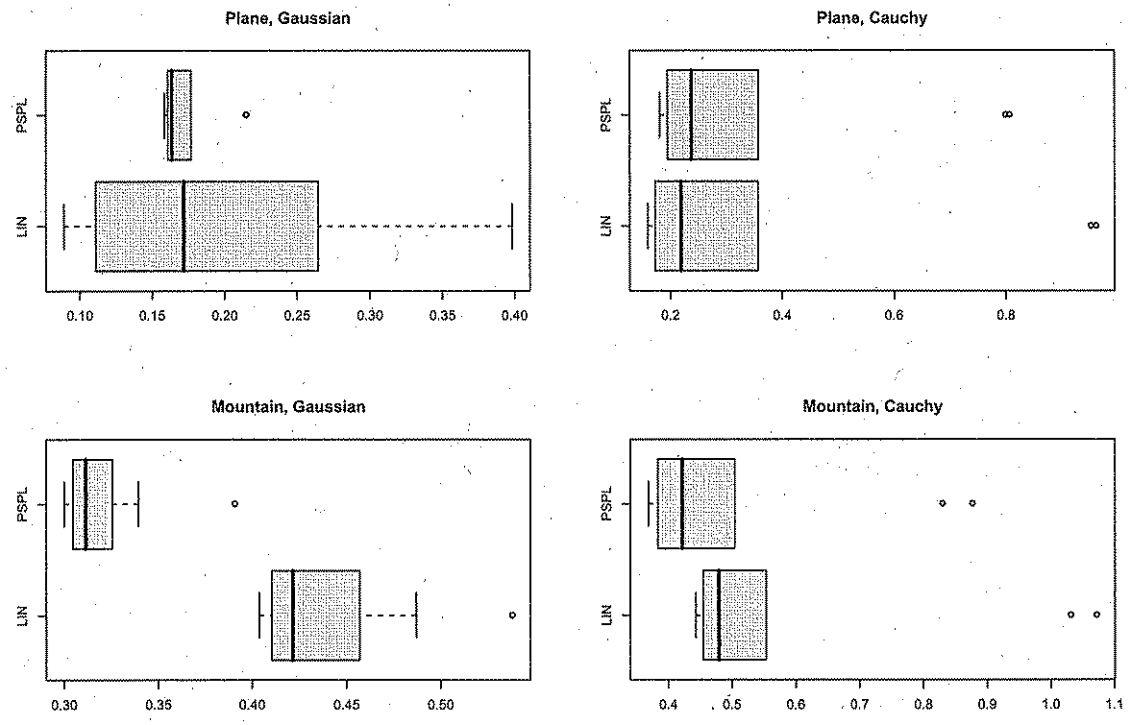
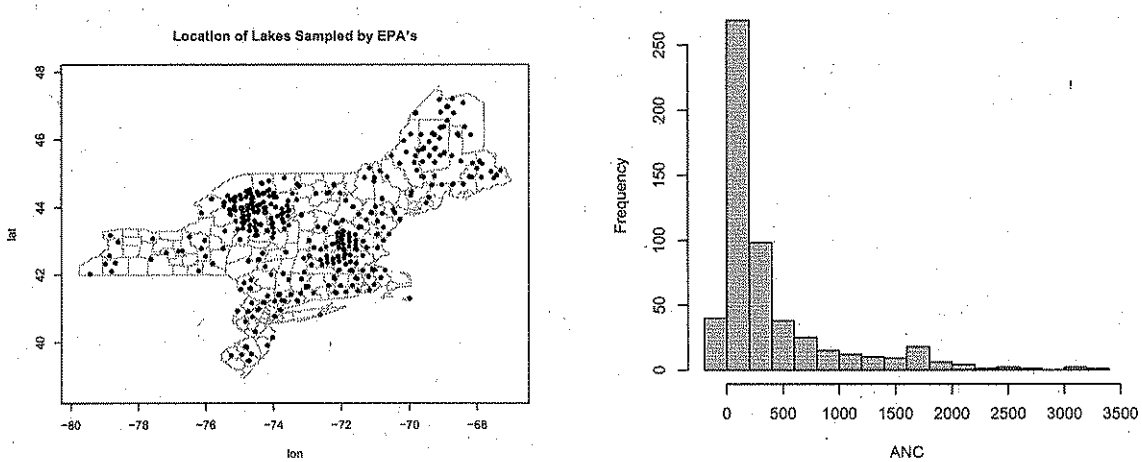
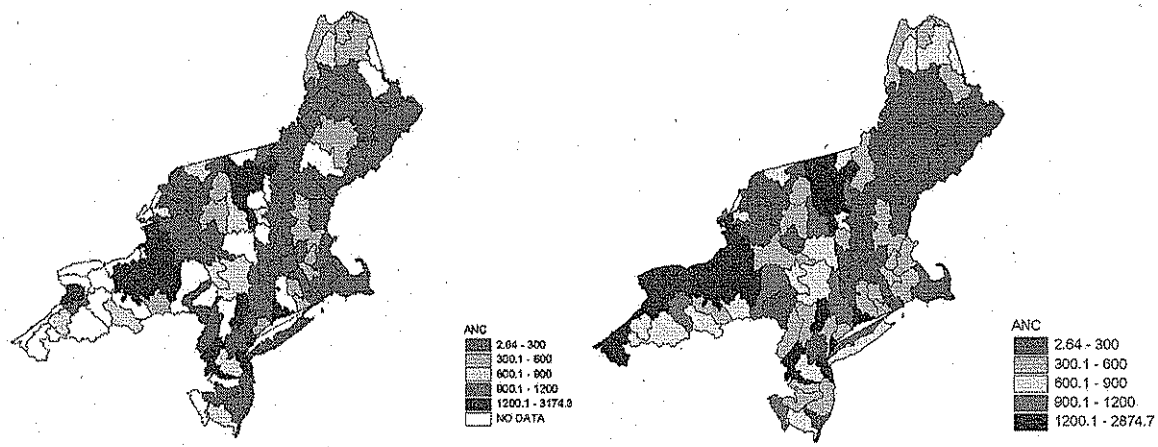


Figure 6: Boxplots of MADE for LIN and PSPL for all bivariate simulation studies.



(a) (b)

Figure 7: (*a*) Location of Northeastern Lakes sampled by EPA's EMAP in the years 1991-1996 (some revisited). (*b*) Distribution of ANC over the 551 visits.



(a) (b)

Figure 8: (a) Map of design based direct estimates of ANC means for each HUC. (b) Map of model predicted ANC means for each HUC under PSPL model.

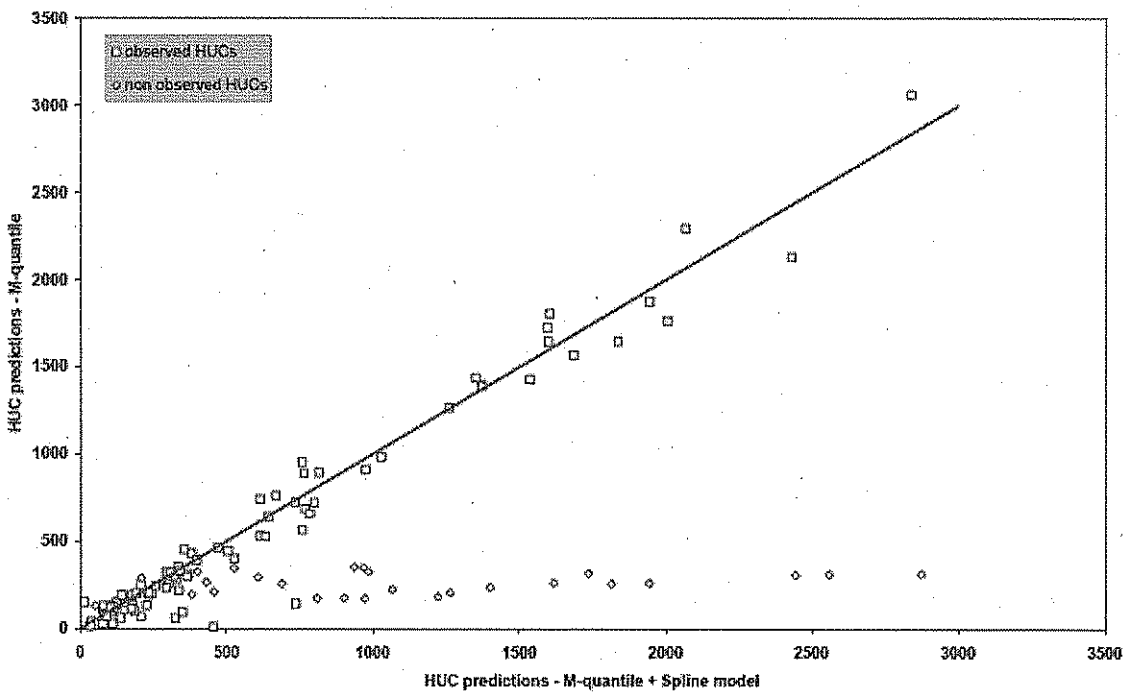


Figure 9: Comparison of HUC predictions under PSPL model and LIN model (solid line is 45-degree line).

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Linear, Gaussian	1.0	1.0	0.9	0.9	0.8	0.8	0.9	0.9	1.0
Linear, Cauchy	0.9	0.7	0.6	0.5	0.6	0.6	0.7	0.9	1.1
Exponential, Gaussian	1.3	1.8	2.5	3.5	4.0	3.4	2.5	1.8	1.3
Exponential, Cauchy	2.3	4.7	6.8	8.2	9.1	9.9	9.8	7.2	3.0
Cycle, Gaussian	5.5	10.4	17.5	26.1	30.7	26.2	17.6	10.5	5.6
Cycle, Cauchy	5.0	5.4	5.2	5.0	4.9	4.9	5.1	5.2	4.7
Jump, Gaussian	1.2	1.6	1.9	2.2	2.4	2.2	1.9	1.6	1.2
Jump, Cauchy	2.1	2.8	3.0	3.1	3.2	3.3	3.3	3.2	2.6

Table 1: MASE values for LIN for each decile and univariate simulation study; MASE for PSPL = 1.

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Plane, Gaussian	0.9	0.9	0.8	0.7	0.6	0.7	0.8	0.9	0.9
Plane, Cauchy	1.0	0.8	0.7	0.7	0.7	0.7	0.8	0.9	1.0
Mountain, Gaussian	1.3	1.9	2.5	2.9	3.1	2.9	2.5	1.9	1.3
Mountain, Cauchy	1.0	0.9	1.0	1.1	1.1	1.1	1.0	1.0	1.0

Table 2: MASE values for LIN for each decile and bivariate simulation study; MASE for PSPL = 1.

	$\widehat{CV}$ (%)					
ANC	0-10	10-20	20-30	30-50	>50	Tot.
2.64–300	0.0 (0.0)	2.7 (0.0)	17.9 (5.4)	12.3 (11.0)	11.0 (37.0)	43.9 (53.4)
300.1–600	1.4 (0.0)	15.1 (1.3)	4.1 (6.8)	2.7 (8.2)	0.0 (2.8)	23.3 (19.1)
600.1–900	2.7 (0.0)	2.7 (0.0)	4.1 (0.0)	5.5 (4.1)	0.0 (5.4)	15.0 (9.5)
900.1–1200	0.0 (0.0)	1.4 (0.0)	1.3 (1.4)	0.0 (1.4)	0.0 (0.0)	2.7 (2.8)
1200.1–2874.7	4.1 (0.0)	6.9 (0.0)	0.0 (1.4)	4.1 (6.9)	0.0 (6.9)	15.1 (15.2)
Tot.	8.2 (0.0)	28.8 (1.3)	27.4 (15.0)	24.6 (31.6)	11.0 (52.1)	100 (100)

Table 3: Joint class distribution of the level of ANC and the estimates of the Coefficient of Variation of small area estimates under the PSPL model (in parentheses those under the LIN model).