Predicting trace gas concentrations using quantile regression models

M. Conde-Amboage, W. González-Manteiga and C. Sánchez-Sellero

Version: Author's Accepted Manuscript
This is a post-peer-review, pre-copyedit version of an article published in Stoch Environ Res Risk Assess. The final authenticated version is available online at: https://doi.org/10.1007/s00477-016-1252-4

HOW TO CITE

Conde-Amboage, M., González-Manteiga, W. and Sánchez-Sellero, S. (2017) Predicting trace gas concentrations using quantile regression models. Stochastic Environmental Research and Risk Assessment, 31, 1359-1370.

FUNDING

Research has been funded by project INNPAR2D (reference code MTM2016-76969-P), from the Ministry of Economy and Competitiveness and the European Regional Development Fund (ERDF).

Noname manuscript No.
(will be inserted by the editor)

Predicting trace gas concentrations using quantile regression models

Mercedes Conde-Amboage · Wenceslao González-Manteiga · César Sánchez-Sellero

Received: date / Accepted: date

Abstract Quantile regression methods are evaluated for computing predictions and prediction intervals of NO_x concentrations measured in the vicinity of the power plant in As Pontes (Spain). For these data, smaller prediction errors were obtained using methods based on median regression compared with mean regression. A new method to construct prediction intervals involving median regression and bootstrapping the prediction error is proposed. This new method provides better coverage for NO_x data compared with classical and bootstrap prediction intervals based on mean regression, as well as simpler prediction intervals based on quantile regression. A simulation study illustrates the features of this proposed method that lead to a better performance for obtaining prediction intervals for these particular NO_x concentration data, as well as for any other environmental dataset that do not meet assumptions of homoscedasticity and normality of the error distribution.

Keywords quantile regression; NO_x concentration; prediction errors; prediction intervals; bootstrapping; median regression.

1 Introduction

The power plant at As Pontes (A Coruña, Spain) is an important facility of Endesa Generación S.A. Figure 1 shows a picture of the power plant, and its geographical location within Europe. The plant comprises a thermal power station and a combined cycle power station. Its activity releases NO_x in quantities that need to

M Conde-Amboage
Department of Statistics and O.R., University of Santiago de Compostela, Spain
E-mail: mercedes.amboage@usc.es

W González-Manteiga
Department of Statistics and O.R., University of Santiago de Compostela, Spain
E-mail: wenceslao.gonzalez@usc.es

C Sánchez-Sellero
Department of Statistics and O.R., University of Santiago de Compostela, Spain
E-mail: cesar.sanchez@usc.es

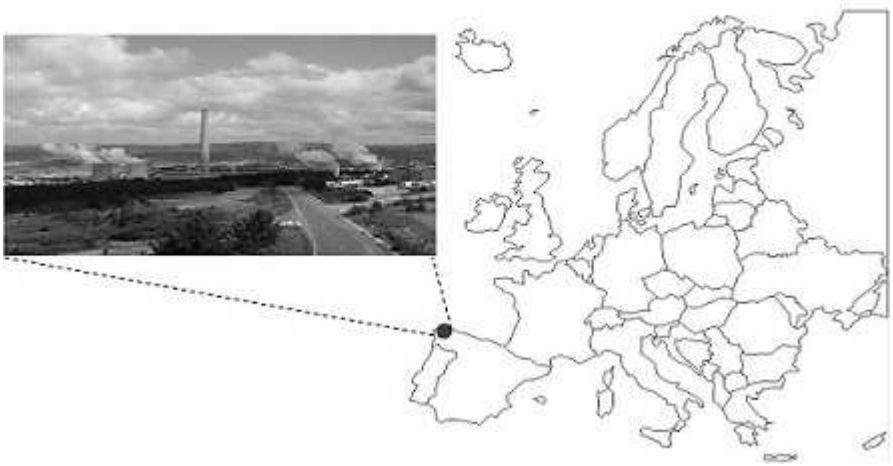

Fig. 1 Picture of the As Pontes power plant and its geographical location within Europe.

be monitored for both legal and ecological reasons. European legislation imposes threshold levels on ambient NO_x concentrations to protect human and environmental health. In addition, the location of the power plant near to natural enclaves of high ecological value requires special care to be exercised to mitigate pollution of the local environment. As a consequence, the power plant possesses several systems of pollution control. In particular, it has a 'Network of Vigilance of Atmospheric Quality', comprising seven automatic analyzers for sulfur dioxide (SO_2), oxides of nitrogen (NO_x), particles in suspension, temperature, and oxygen, located in several positions around the power plant. A meteorological station also provides information to help assess and predict contamination. Predictions of 30 minute in advance are necessary, because it takes about 30 minute for countermeasures to be implemented at the power plant, and to arrange for other contributors to the national power grid to compensate these effects on energy production.

Throughout the years, several methods have been proposed to predict future pollution in the surrounding areas of the power plant at As Pontes. García-Jurado et al. (1995) proposed a semiparametric prediction system for a time series that generalizes the Box–Jenkins model. Prada-Sánchez and Febrero-Bande (1997) introduced the concept of a historical matrix, which summarizes the information on past pollution events in a semiparametric model. Prada-Sánchez et al. (2000) considered partially linear models within an environmental context, which allowed the user to introduce additional information as meteorological variables. Fernández-Castro et al. (2003) used neural network models to predict the evolution of certain pollutant elements. Fernández-Castro et al. (2005) and Fernández-Castro and González-Manteiga (2008) employed several functional techniques for predicting sulfur dioxide levels. Roca-Pardiñas et al. (2004) and Roca-Pardiñas et al. (2005) used a generalized additive model with an unknown link function to predict the binary time series defined using a SO_2 concentration threshold. Along similar lines,

a study of correlations between various contaminants around four coal-fired power plants in Greece was provided by Nanos et al. (2015) and regression modeling of atmospheric NO_x concentration in urban London can be found in Shi and Harrison (1997).

Clearly, most works rely on least squares methods and prediction of mean pollutant levels. In contrast, the purpose of our work is to provide prediction methods for NO_x concentration using quantile regression models. Quantile regression models were introduced by Koenker and Bassett (1978), with the purpose of estimating certain quantiles of a response variable conditional to values of its predictors. In this way, a more complete description of the conditional distribution can be given, where the central and best known quantile is the median, but lower or upper quantiles are also taken into account. Thus, these models describe the effects of the predictors not only on the central values of the response variable, but also on its lower or upper range of values. Moreover, quantile regression is estimated in a more robust manner than common mean regression models, and does not require stringent assumptions to be satisfied, such as homoscedasticity and normality of the error distribution. A thorough description of quantile regression methods can be found in Koenker (2005).

Quantile regression was successfully applied to environmental data by several authors in recent years. Sousa et al. (2009) made use of quantile regression to predict ozone concentrations in Oporto, Northern Portugal. Salama (2005) showed that median regression analysis is more useful for detecting relationships between environmental performance and corporate financial performance than ordinary least squares regression. A hierarchical Bayesian spatial quantile regression model was proposed by Fontanella et al. (2015) to analyze indoor radon concentrations. Cade and Noon (2003) provide a nice review of applications of quantile regression. Quantile regression has also proven to be very useful for obtaining prediction intervals. Meinshausen (2006) and Mayr et al. (2012) made use of estimated quantiles to define the endpoints of prediction intervals, while Zhou and Portnoy (1996) proposed a relatively simple correction to prediction intervals to improve their coverage.

In this paper, quantile regression is shown to be more accurate than similar ones based on least squares methods for estimating prediction errors for NO_x concentrations. Thus, a new method for computing prediction intervals is proposed here, based on quantile regression estimation and bootstrap approximation of the prediction error. Its performance is evaluated using real data. In addition, simulations are provided to illustrate features of this model that make it suitable for other environmental datasets.

2 Data and Methods

Most of the previous work on pollution around As Pontes power plant was focused on SO_2 levels, because this was the main pollutant from the power plant during its first years of operation, when combustion of local coal was the main power source. Lately, local coal has been replaced by imported coal to reduce SO_2 emissions. This change in source material, together with a new combined cycle generator, have resulted in NO_x pollution becoming more relevant. For this reason, we focus our attention on NO_x levels in this study.

The concentration of NO_x is measured every minute, and recorded by an automatic monitoring system. Simultaneously, the local meteorological station records temperature, wind speed and wind direction every minute. Our purpose is to predict the concentration of NO_x at a time $(t+30)$ based on available information at time t, where t and $(t+30)$ are measured in minutes. Thus, a regression model of the following type was considered:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t,grad} + \beta_3 Z_{1t} + \beta_4 Z_{2t} + \beta_5 Z_{3t} + \varepsilon_t \tag{1}$$

where X_t is the NO_x concentration at time t; $X_{t,grad} = X_t - X_{t-5}$ represents the gradient of NO_x concentration over the last 5- minute interval; Z_{1t}, Z_{2t} and Z_{3t} are the mean values of temperature, wind speed and wind direction for the interval covering the last 6 minutes (from $(t-5)$ to t); $Y_t = X_{t+30}$ is the NO_x concentration at time $(t+30)$, taken as the response variable; and ε_t represents the error. In this way, measurements for the latest 6-minute interval are used to predict $Y_t = X_{t+30}$. Wind direction is treated as a scalar variable because we measure the absolute value of the deviation angle from true north.

We included all five predictors in our model because they are usually considered to affect local pollution around this power plant (see Prada-Sánchez et al. (2000)). In particular, the NO_x concentration at time t, X_t, is expected to have a positive effect on the same concentration at time $t+30$, Y_t. Then, a linear effect with expected positive coefficient seems to be adequate for this predictor us, a linear effect, with a positive coefficient was selected for this predictor. The remaining four predictors, having a smaller effect on the response variable, but can also be modeled with linear terms; more complex effects are not expected to play a role. In particular, given that low ambient temperature facilitates the dispersion of pollutants, then higher temperatures would promote higher pollutant concentrations local to the power plant. Similarly, high wind speed is associated with pollutant dispersion, resulting in lower pollutant concentrations near to the power plant. Pollutant gradients and wind direction had least association with the response variable, but are included for the purposes of comparison with the literature. In Section 3, a test of linearity is applied to check the validity of our linear model.

Our model (1) is fitted using observations covering a period of 10 days; these are defined as the training sample. The subsequent 10-day period is used as the evaluation sample to assess the performance of our predictions and prediction intervals. Given that our model includes observations from the last 6 minutes of measurements as predictors, while the response value is scheduled for 30 minutes later, then these data are divided into blocks of 36 observations, with no predictors or response values for two of these blocks. This circumvents any possible autocorrelation issues. In Section 3, autocorrelation tests are applied to validate our model. Thus, after removing some missing data from the training sample, we had a sample size of 338 blocks, with corresponding observations $(X_{t-5}, \ldots, X_t, Z_{1t}, Z_{2t}, Z_{3t}, Y_t)$. Likewise, the evaluation sample comprises another 338 blocks.

2.1 Least squares versus quantile methods

The regression model given in (1) can be interpreted as a mean regression model, if we assume that the error has an expectation of zero, that is, $E(\varepsilon_t) = 0$. In this

case, the model can be estimated by minimizing the sum of squared residuals

$$\widehat{\beta}_{LS} = \arg\min_{\beta} \sum_{t} \left(Y_t - \beta' P_t\right)^2$$

where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$ is the vector of coefficients to be estimated, $P_t = (1, X_t, X_{t,grad}, Z_{1t}, Z_{2t}, Z_{3t})'$ is the vector of predictors, and $\widehat{\beta}_{LS}$ is the least squares estimator.

For the same expression (1), instead of assuming $E(\varepsilon_t) = 0$, one can surmise that the error has a τ-quantile equal to zero, i.e., $P(\varepsilon_t \leq 0) = \tau$ with $\tau \in (0, 1)$. This implies that the proportion of negative (non-positive) errors is expected to be τ, which is equivalent to the proportion of observations below the regression function equal to τ. In this way, the regression function is no longer the conditional expectation, but rather the conditional τ-quantile of the response variable given the predictors.

To estimate the coefficients $\beta(\tau)$ for a certain τ, we observed that while mean regression minimizes the sum of squared residuals, the τ-quantile minimizes the sum of weighted absolute values of the residuals. Thus, the estimator $\widehat{\beta}(\tau)$ is given by

$$\widehat{\beta}(\tau) = \arg\min_{\beta} \sum_{t} \rho_\tau \left(Y_t - \beta' P_t\right)$$

for each $\tau \in (0, 1)$, where ρ_τ is the quantile loss function:

$$\rho_\tau(z) = \begin{cases} (1-\tau) \cdot |z| & \text{if } z \leq 0 \\ \tau \cdot |z| & \text{if } z > 0 \end{cases}$$

Here, ρ_τ produces a weighting effect on the residuals. Positive residuals, which correspond to observations above the regression function, are weighted by the factor τ. Negative residuals receive the weighting factor $(1 - \tau)$.

Because different regression models will be compared, validation of these models is critical. We applied well-known procedures to check whether the residuals satisfy the assumptions of homoscedasticity and normality. More details are given in Section 3, where our model is applied to the example atmospheric data. However, the validation of quantile regression models is not well addressed in the literature. One reason is that quantile regression does not require any stringent assumptions about the error distribution. The linearity of the quantile model, which is the most critical assumption, can be checked using the lack-of-test developed by Conde-Amboage et al. (2015). This test evaluates the fit of a parametric quantile regression model with many predictors (in our case, a linear quantile model with five predictors) versus any other possible model, i.e., a nonparametric alternative. The test is based on the cumulative sum of residuals with respect to unidimensional linear projections of the covariates. Moreover, a wild bootstrap mechanism is used to approximate the critical values of the test (Conde-Amboage et al. (2015)). In Section 3, p-values for this test are provided to assess the linear quantile regression model for different values of τ.

2.2 Assessment of prediction methods

The pointwise prediction for a value of Y_{t_0} at a future time t_0 using $P_{t_0} = (1, X_{t_0}, X_{t_0,grad}, Z_{1t_0}, Z_{2t_0}, Z_{3t_0})'$ as predictors, can be obtained from the mean regression model by means of $\widehat{Y}_{t_0,LS} = \widehat{\beta}'_{LS} P_{t_0}$. An alternative prediction can be obtained from the median regression model in a similar way, where $\widehat{Y}_{t_0,\tau=0.5} = \widehat{\beta}(\tau = 0.5)' P_{t_0}$.

To compare the performance of these two prediction methods, we use two criteria: the mean absolute error (MAE) and the mean squared error (MSE), given by

$$MAE = n_0^{-1} \sum_{t_0} \left| Y_{t_0} - \widehat{Y}_{t_0,m} \right| \qquad MSE = n_0^{-1} \sum_{t_0} \left(Y_{t_0} - \widehat{Y}_{t_0,m} \right)^2$$

where m represents the prediction method, either a least squares regression (LS) or median regression ($\tau = 0.5$); the times t_0 in the summation are those of the evaluation sample; and n_0 is the evaluation sample size. Note that the estimations $\widehat{\beta}_{LS}$ and $\widehat{\beta}(\tau = 0.5)$ are calculated from the training sample.

2.3 Prediction intervals: conditional and unconditional coverage

A prediction interval for a value Y_{t_0} is an interval that is expected to contain the true value Y_{t_0} with a (presumably) high probability $(1 - \alpha)$, usually called the confidence level. Let us denote a prediction interval as (L_{t_0}, U_{t_0}), where the endpoints L_{t_0} and U_{t_0} are obtained as functions of the training sample, and the values of the predictors P_{t_0} at time t_0. It would be expected that

$$P\left(Y_{t_0} \in (L_{t_0}, U_{t_0})\right) = 1 - \alpha.$$

In this expression, the probability is defined for all possible training samples and new observations. We call this *unconditional coverage*. However, because the value of the predictors for new observation, P_{t_0}, is known, it is reasonable to define the above probability as conditional to these predictors, that is,

$$P\left(Y_{t_0} \in (L_{t_0}, U_{t_0}) | P_{t_0} = p_{t_0}\right).$$

We call this probability the *conditional coverage*. The unconditional coverage can be obtained as an average of the conditional coverage, with respect to the predictors distributions. A sample analogue for the unconditional coverage would be the proportion of prediction intervals that contain the new observation in the entire evaluation sample, while the conditional coverage is the same proportion, but with evaluation samples taken at a certain value of the predictor P_{t_0}. Mayr et al. (2012) provides further explanation of these concepts.

The immediate consequence of these definitions is that: if the conditional coverage respects the nominal level $(1 - \alpha)$, then the unconditional coverage will also respect it. The reverse is not necessarily true. Retaining a conditional coverage at the nominal level $(1 - \alpha)$ is therefore a more stringent condition, requiring more detailed use of the information gathered by the predictor P_{t_0}.

Below, we outline a number of known methods for obtaining prediction intervals, together with our new proposed method. Each method is valid for a certain set of restrictive assumptions on the error distribution or on the conditional variability. They fail to provide unconditional or conditional coverage, when these assumptions are not satisfied. In particular, misspecification of the error distribution affects the unconditional coverage, while misspecification of the conditional variability affects the conditional coverage. The goal of our proposed method is to provide a prediction interval with appropriate conditional and unconditional coverage based on a quantile regression estimation and a bootstrapping procedure. We found that quantile methods were particularly useful because of their robustness and flexibility under more general conditions.

2.4 Methods for obtaining prediction intervals

Here, we consider four published methods for obtaining prediction intervals: two based on mean regression, and two based on quantile regression. In addition, a new method is proposed here, based on quantile regression estimation and a bootstrapping method.

M1 A prediction interval for Y_{t_0} with level $(1-\alpha)$ is traditionally obtained from mean regression by

$$\left(\widehat{Y}_{t_0,LS} - t_{n-6,\alpha/2}\,\widehat{\sigma}\,\sqrt{1+P'_{t_0}(X'X)^{-1}P_{t_0}},\right.$$
$$\left.\widehat{Y}_{t_0,LS} + t_{n-6,\alpha/2}\,\widehat{\sigma}\,\sqrt{1+P'_{t_0}(X'X)^{-1}P_{t_0}}\right),$$

where $t_{n-6,\alpha/2}$ is the $(1-\alpha/2)$ quantile of the Student's t-distribution with $(n-6)$ degrees of freedom; $\widehat{\sigma}^2 = (n-6)^{-1}\sum_t (Y_t - \widehat{\beta}'_{LS}P_t)^2$ is the error variance estimate based on the training sample; and X is the design matrix of the training sample. This type of interval was used from the very beginning for estimating prediction intervals (see Seber (1977)) and it is still the most common method used to obtain prediction intervals using linear regression models (see Fahrmeir et al. (2013)). The main drawback of this method is that it heavily depends on the assumptions of homoscedasticity and error normality.

M2 Stine (1985) proposed a bootstrapping method to circumvent the error normality condition involved in constructing a prediction interval using mean regression. Homoscedasticity is still be required for this method.

Q1 A prediction interval for Y_{t_0} of level $(1-\alpha)$ can be obtained from quantile regression models as

$$\left(\widehat{Y}_{t_0,\tau=\alpha/2}, \widehat{Y}_{t_0,\tau=1-\alpha/2}\right),$$

where the endpoints are estimations of the $\alpha/2$ and $(1-\alpha/2)$ quantiles of Y_{t_0} conditional to the values of the predictors P_{t_0}. Intervals of this kind were used by several authors, including Meinshausen (2006) and Mayr et al. (2012), that outline the construction of these prediction intervals and their main advantages. Such intervals do not require homoscedasticity and adapt to any error distribution. Their drawback is that a parametric (commonly linear) model is assumed at extreme quantiles, which effects estimation, leading to an empirical coverage that is smaller than the nominal one.

Q2 This method is similar to the previous one, but has a small correction to account for its effects on estimation. This prediction interval is defined as

$$\left(\widehat{Y}_{t_0,\tau=\alpha/2-\delta}, \widehat{Y}_{t_0,\tau=1-\alpha/2+\delta}\right),$$

where $\delta = 0.5(z_{1-\alpha/2}/n)$, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution, and n is the training sample size. The factor 0.5 comes from the expression $\sqrt{\tau(1-\tau)}$, where τ is the quantile to be estimated (here $\tau = 0.5$). This modification was proposed by Zhou and Portnoy (1996). There are other more elaborate procedures to improve the empirical coverage of this kind of prediction intervals, involving extremal quantile regression methods, such as those of Chernozhukov (2005).

Q3 This is a new method, proposed here, where the prediction interval is computed as

$$\left(\widehat{Y}_{t_0,\tau=0.5} + G^{\star-1}(\alpha/2), \widehat{Y}_{t_0,\tau=0.5} + G^{\star-1}(1-\alpha/2)\right),$$

where $G^{\star-1}$ represents the quantile function associated with the bootstrap distribution of the prediction error, denoted by $G^{\star}$. Here, $G^{\star-1}(\alpha/2)$ and $G^{\star-1}(1-\alpha/2)$ denote $\alpha/2$ and $(1-\alpha/2)$ quantiles of the bootstrap distribution $G^{\star}$, respectively. These quantities are obtained as follows:

Step 1 Bootstrap replicates of the training sample and the new observation are determined from

$$\begin{aligned} Y_t^{\star} &= \widehat{\beta}(\tau=0.5)'P_t + \varepsilon_t^{\star} \qquad t \in \{1,\ldots,n\} \\ Y_{t_0}^{\star} &= \widehat{\beta}(\tau=0.5)'P_{t_0} + \varepsilon_{t_0}^{\star} \end{aligned}$$

$\widehat{w}here\beta(\tau=0.5)$ is an estimate of the coefficients in the median regression equation obtained from the training sample. The bootstrap errors are $\varepsilon_t^{\star} = w_t|r_t|$, where $|\cdot|$ denotes the absolute value and $r_t = Y_t - \widehat{\beta}(\tau=0.5)'P_t$ gives the residuals in the original training sample. The weights w_t are drawn from the discrete distribution at values of 1 and -1, with equal probability 0.5. This distribution was proposed by Feng et al. (2011) in the context of wild bootstrap for quantile regression.

The bootstrap error for the new observation is given by $\varepsilon_{t_0}^{\star} = w_{t_0}|r_{t_0}|$, where w_{t_0} follows the same two-point distribution as w_t, while the residual r_{t_0} is drawn from the following estimate of the conditional distribution of the error for the value of the predictor P_{t_0}:

$$\widehat{F}(r|P_{t_0}) = \sum_{t=1}^{n} I(r_t \le r) W_{t,P_{t_0}}$$

where $I(r_t \le r)$ is the indicator function with value 1, if the condition $r_t \le r$ is satisfied, else value 0; and

$$W_{t,P_{t_0}} = \frac{K((\widehat{\beta}(\tau=0.5)'P_t - \widehat{\beta}(\tau=0.5)'P_{t_0})/h)}{\sum_{s=1}^{n} K((\widehat{\beta}(\tau=0.5)'P_s - \widehat{\beta}(\tau=0.5)'P_{t_0})/h)}$$

are nonparametric smoothing weights. The smoothing parameter was chosen as $h = cn^{-1/5}$, where c is a constant that depends on several unknown

quantities, and $n^{-1/5}$ is the conventional rate for this type of Nadaraya–Watson non parametric estimator. See Hall et al. (1999) for more detail on this type of estimator and an outline of the bootstrapping method used to select the value of h. Here, we propose simpler rules to those given by Li and Racine (2007), where a rule-of-thumb is used, taking the constant c to be the standard deviation of the covariate, i.e., the variable $\widehat{\beta}(\tau = 0.5)' P_t$ in our framework. In our empirical evaluation, an even simpler rule, taking $c = 1$, was used with satisfactory results.

Step 2 Based on the bootstrap training sample, a bootstrap replicate of the coefficient estimate can be obtained. Let us denote it as $\widehat{\beta}^\star(\tau = 0.5)$. Thus, bootstrap prediction errors are computed as

$$D^\star = Y^\star_{t_0} - \widehat{\beta}^\star(\tau = 0.5)' P_{t_0}.$$

Step 3 Steps 1 and 2 are repeated B times to compute a sample of differences $D^\star_1, \ldots, D^\star_B$. The empirical distribution of this sample is a Monte Carlo approximation of the distribution function $G^\star$, from which the quantiles $G^{\star -1}(\alpha/2)$ and $G^{\star -1}(1 - \alpha/2)$ are determined.

2.5 Theoretical discussion

Here, we discuss the expected properties of these various prediction methods, with particular emphasis on the newly proposed method Q3 as it compares to published methods. The expected properties are determined based on empirical outcomes of predicting NO_x concentrations using real or simulated atmospheric data, as outlined in Sections 3 and 4. Convergence results for the new method are also provided.

Methods M1, M2, and the new method Q3 have in common that they are based on the estimation of a central quantity of the conditional distribution, i.e., the conditional mean in methods M1 and M2, and the conditional median in method Q3, as well as estimation of the prediction error distribution. Estimating a mean regression, as in methods M1 and M2, is very efficient under normality, but is inefficient and lacks robustness for more general error distributions. This is one of the main reasons why we proposed a median regression estimation for our method Q3. To estimate the prediction error distribution, the classical method M1 applies a simple rule based on stringent assumptions of homoscedasticity and error normality. This is the best method under these assumptions, but it yields a poor coverage approximation, when these assumptions are not satisfied. Method M2 makes use of a bootstrapping method to estimate the error distribution, but still assumes homoscedasticity. The proposed method Q3 applies a bootstrapping method adapted for quantile regression and a heteroscedastic setup. Hence, this new method is applicable under very general conditions, and overcomes the limitations of the stringent assumptions in methods M1 and M2.

Methods Q1 and Q2 are not based on estimating any central quantity of the conditional distribution, but directly obtain the lower and upper endpoints of the prediction interval through quantile estimation. Method Q1 does not address the problem of prediction error, while method Q2 applies a simple correction for this problem. The main virtue of these two methods is that their quantile procedures of estimation are very flexible with respect to the error distribution type; it is not

required to be Gaussian or similar. However, estimating non-central quantities, especially relatively extreme quantiles, has two main drawbacks: there may be few data points available for estimating these extreme quantiles, causing what is known as the problem of sparsity described in the literature dealing with quantile regression; and estimation will usually require a model assumption (most commonly, linearity) that restricts its real world applications. Heteroscedasticity can be considered for methods Q1 and Q2, but only under a specific model for the conditional variability. In other words, methods Q1 and Q2 will work well under linear heteroscedasticity, but will fail under a more general heteroscedastic pattern. This means that estimating a complex model for extreme quantiles is often unfeasible in practice. Because of these restrictions, we opted to estimate a central quantile in Q3, i.e., the conditional median, and to use a bootstrap approximation of the prediction error to account for general heteroscedasticity and general error distributions.

The convergence properties of the proposed method Q3 are derived using similar arguments to those given in Stine (1985). Thus, the bootstrap prediction errors can be expressed as:

$$D^\star = \varepsilon^\star_{t_0} + \left(\widehat{\beta}(\tau = 0.5) - \widehat{\beta}^\star(\tau = 0.5)\right)' P_{t_0}.$$

Given that the summands on the right are generated independently, the bootstrap distribution of the prediction error is the convolution of two distributions:

$$G^\star = \hat{F}_{t_0} * Z^\star$$

where $\hat{F}_{t_0}$ is the distribution of $\varepsilon^\star_{t_0}$ and $Z^\star$ is the distribution of the second summand, i.e., the bootstrap approximation of the parameter estimation error multiplied by the predictors. Feng et al. (2011) obtained the consistency of $Z^\star$ under the bootstrapping mechanism proposed here. Hall and Yao (2005) provided the consistency of the estimator $\widehat{F}(r|P_{t_0})$, where smoothing is applied to projected predictors, as performed here. Since $\hat{F}_{t_0}$ is constructed from $\widehat{F}(r|P_{t_0})$ by including the bootstrap multipliers given by Feng et al. (2011), bootstrap validity depends on the consistency of $\widehat{F}(r|P_{t_0})$. Although Stine (1985) makes use of an empirical distribution function of the residuals, a locally smoothed version of the residual distribution is used here. Thus, the asymptotic coverage is attained using a smoothed version of Theorem 2 in Stine (1985).

3 Results

Model (1) is adjusted for mean regression and for regression with different quantiles, using the training sample described in Section 2. The results for the mean and the median regression are similar, in the sense that the most significant predictors are: X_t (the current concentration of NO_x); and Z_{2t} (the mean value of wind speed). Higher values of the current concentration produce higher predictions for the 30 minute future concentration, as would be expected. Higher wind speed is associated with lower future concentrations. This is consistent with the premise that wind carries NO_x away from the power plant surroundings.

Quantile regression provides a more detailed interpretation of the predictors' effects on each quantile of the future concentration. Figure 2 shows the estimated

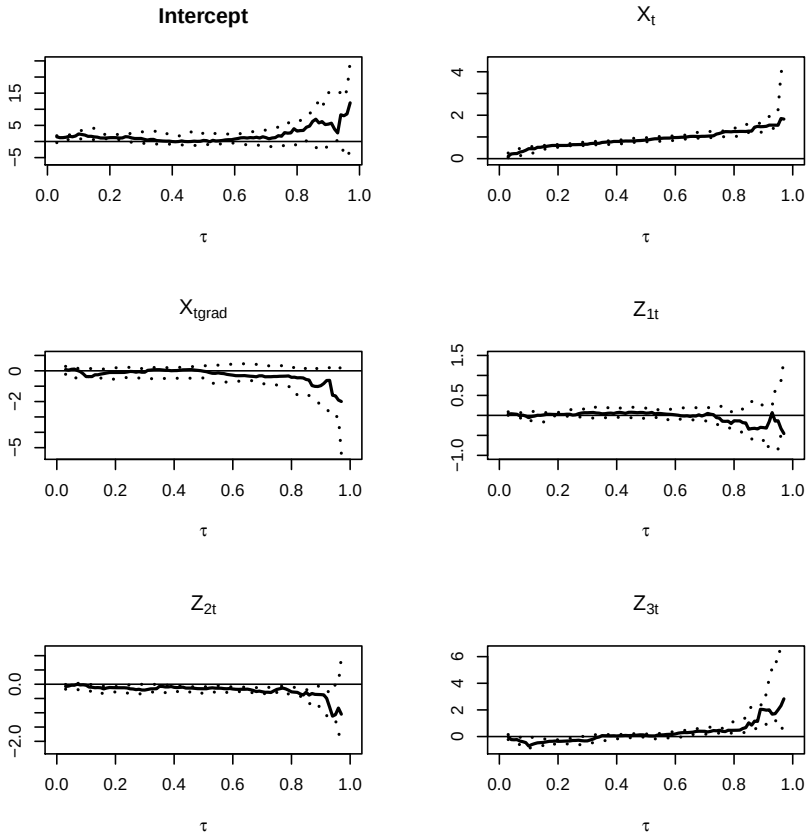

Fig. 2 Coefficients associated with each predictor as a function of the order τ of the quantile. The solid line represents the coefficients, while the dotted lines represent the endpoints of confidence intervals for the coefficients.

coefficients as a function of the quantile order, τ, together with their confidence intervals. The confidence intervals for the estimated parameters have been calculated by inverting a rank test, as described in Koenker (1994). Clearly, the effect of each predictor is different at each order τ. The most significant predictor, the current NO_x concentration (X_t), has a coefficient that is positive for all quantiles, but is larger for larger τ (representing the upper range of the future NO_x concentration). The negative effect of wind speed is less dependent on a particular quantile. All coefficients show larger confidence intervals for upper quantiles, related to the higher variability in the high-end range of NO_x concentrations.

The validity of the mean regression model was determined to explore whether its assumptions were satisfied for our case study. First, a scatter plot of the residuals versus the fitted values from the model was produced (3). This plot shows atypical observations. Clearly, more variability is found for higher-fitted values of the response, having a heteroscedastic pattern. Second, a QQ-Plot was constructed

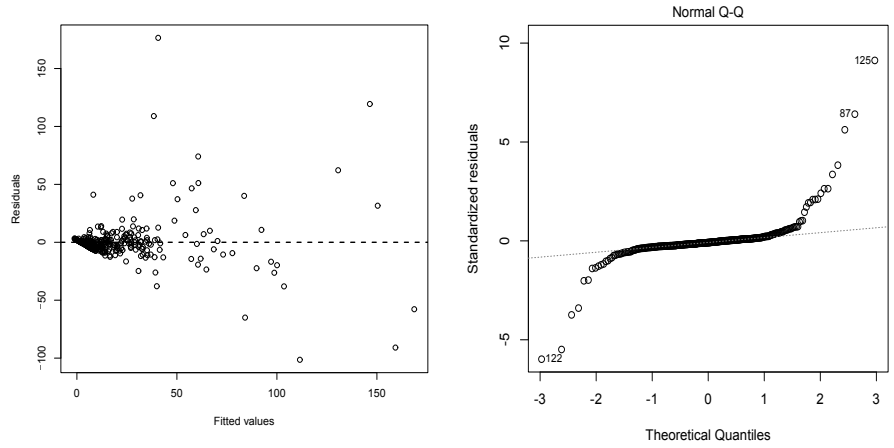

Fig. 3 Validation of the mean regression model for our case study. Left side: Scatterplot of residuals versus fitted values. Right side: Normal QQ-Plot of standardized residuals.

Table 1 p-values for the linearity test carried out for different quantiles.

τ	0.025	0.05	0.10	0.5	0.9	0.95	0.975
p-value	0.6702	0.2672	0.1454	0.1332	0.3596	0.3400	0.8322

(3) to detect deviations from normality. Deviations linked to extreme values, much larger than expected from the normal distribution, are visible. A Shapiro–Wilk test of normality showed a highly significant deviation from normality (p-value smaller than 2.2×10^{-6}). Because the data in the training sample are obtained as a time series, autocorrelation may occur. Hence, we applied a Durbin–Watson test of one-lag autocorrelation, and a Ljung–Box test of two-lag autocorrelation. No significant autocorrelation was found in either of the tests, with a p-value=0.1656 for the Durbin–Watson test, and a p-value=0.2702 for the Ljung–Box test.

The quantile regression model is flexible to more general conditions, with less constraints on the error distribution and conditional variability. Thus, it is not necessary to check homoscedasticity or error normality. The only assumption to be tested is that of linearity, i.e., the assumption that the predictors effects can be explained by the linear function given by (1). We evaluated this assumption using the test proposed by Conde-Amboage et al. (2015). Table 1 contains the p-values associated with this test, carried out for different quantiles. A linear model was acceptable for all quantiles evaluated.

To evaluate different methods, we compared the mean regression models with our median regression model by means of the prediction errors obtained for the evaluation sample. Table 2 shows the MAE and the MSE for the median and the mean regression models for each of the 10 days in the evaluation sample. The last row of the table gives the average value. We observed that the median regression model had smaller prediction errors, both in terms of MAE and MSE.

We computed prediction intervals using the five methods described in Section 2.4. Table 3 shows the empirical coverage of these prediction intervals, computed as

Table 2 Mean absolute error (MAE) and mean squared error (MSE) associated with predictions obtained using median and mean regression models.

	Mean Absolute Error		Mean Squared Error	
	Median	Mean	Median	Mean
Day 1	4.1008	5.0523	35.925	41.570
Day 2	3.1304	4.1902	19.488	33.362
Day 3	8.3023	8.7387	246.563	248.473
Day 4	8.4403	8.8304	151.346	151.759
Day 5	5.0827	6.2487	122.750	126.733
Day 6	2.3943	3.7248	8.982	19.069
Day 7	8.2810	8.5096	487.657	478.338
Day 8	2.9789	4.0173	27.353	31.836
Day 9	2.3820	4.8768	11.525	30.377
Day 10	3.3201	5.1076	42.249	64.301
Average	4.8413	5.9296	115.384	122.582

the percentage of times that the real value of NO_x concentration in the evaluation sample belonged to the prediction interval. This was done for two nominal levels: 90% and 95%. The nominal levels are compared with the actual conditional and unconditional coverages. Our validation of these regression models indicated that we are working in a heteroscedastic context. Thus, it is reasonable to expect that the variability of the response variable (NO_x concentration at time t) is not the same for high and low concentrations. In this case, heteroscedasticity is related to the value of the response. Therefore, conditional coverages were computed for five intervals, I1 to I5, each with the same number of observations, defined by evenly splitting the ordered Y-values. Because the evaluation sample size is 338, we considered these intervals to have a reasonable number of elements (around 67) in each interval. In addition, the variability of the response variable within each interval is not large. The unconditional coverage is the average of the conditional coverages in the five intervals.

Clearly, both conditional and unconditional coverages shown in Table 3 are much larger than the nominal level for methods M1, M2, Q1 and Q2, while the proposed method Q3 provides coverages quite close to the nominal level for each interval and the overall average. The fact that assumptions of normality and homoscedasticity are not satisfied, likely affected the behavior of methods M1, M2, Q1 and Q2. These effects are discussed in more detail in Section 4.

4 Simulation study

We carried out a simulation study to show how deviations, such as those present in our data, from the common assumptions of the classical linear models of mean regression, lead to inadequate predictions and prediction intervals. In such situations, quantile regression is clearly a better option for prediction, while the proposed method Q3 provides a good alternative for computing prediction intervals.

Table 3 Coverage (in percentage) of prediction intervals obtained using the five methods described in Section 2.4, for nominal levels 90% and 95%. Intervals I1 to I5 are defined by splitting the ordered Y-values by their empirical quantiles. M1 and M2 are based on mean regression models, Q1 and Q2 are based on quantile regression, and Q3 is the median regression model proposed herein.

	Level = 90%					Level = 95%				
	M1	M2	Q1	Q2	Q3	M1	M2	Q1	Q2	Q3
I1	100.00	97.05	88.23	88.23	91.17	100.00	100.00	94.11	97.05	97.05
I2	100.00	97.05	100.00	100.00	91.17	100.00	100.00	100.00	100.00	94.11
I3	100.00	97.05	100.00	100.00	94.11	100.00	100.00	100.00	100.00	94.11
I4	100.00	97.05	97.05	97.05	88.23	100.00	100.00	100.00	100.00	97.05
I5	100.00	88.23	94.11	94.11	85.29	100.00	100.00	97.05	100.00	91.17
Average	100.00	95.29	95.88	95.88	90.00	100.00	100.00	98.23	99.41	94.70

Our simulated model is a linear model, with five explanatory variables, as in our case study,

$$Y = 1 + X_1 + X_2 + X_3 + X_4 + X_5 + \sigma(X_1, \ldots, X_5)\varepsilon,$$

where $X_1, \ldots, X_5$ are independent and have an uniform distribution on the unit interval $(0, 1)$; $\sigma(X_1, \ldots, X_5)$ represents the effect of the predictors on the standard deviation of the response variable; and ε is an random error variable, independent of these predictors.

Three types of conditional standard deviations are considered:

Ho A homoscedastic model, where $\sigma(X_1, \ldots, X_5) = 1$.

He1 A heteroscedastic model, where $\sigma(X_1, \ldots, X_5) = (1 + X_1 + X_2 + X_3 + X_4 + X_5)/2$. Note that in this model, the conditional standard deviation is a linear function of the predictors.

He2 A heteroscedastic model, where $\sigma(X_1, \ldots, X_5) = 1 + (X_1 + X_2 + X_3 + X_4 + X_5)^4/100$. Note that this conditional standard deviation is a non-linear function of the predictors.

These three models explore the conditions that should be verified in all datasets to achieve a good performance in estimating prediction intervals for each computation method. Given that methods M1 and M2 are designed under the assumption of homoscedasticity, we would expect a relatively poor performance for heteroscedastic models, He1 and He2. Importantly, He1 and He2 differ in whether their conditional standard deviation is a linear or non-linear function of the predictors, respectively. This creates a big difference in methods Q1 and Q2, because they are based on linear estimations of lower and upper quantiles. Such methods work well under linear heteroscedasticity, like that considered in model He1, but may be misleading under a non-linearly heteroscedastic model like He2. Finally, we would expect a good performance of method Q3 for all considered models, Ho1, He1 and He2.

From all of these models, training samples of independent observations were drawn of size n (different values will be considered for n) to provide estimates for both quantile and mean regression models. For each training sample, an evaluation sample was drawn of the same size to compute the empirical coverage of

Table 4 Coverage (in percentage) of prediction intervals obtained using the five methods described in Section 2.4, with homoscedastic model Ho. Values are for a nominal level of 90%, as well as for different error distributions and sample sizes. M1 and M2 are based on mean regression models, Q1 and Q2 are based on quantile regression, and Q3 is the median regression model proposed herein.

	$n = 100$					$n = 1000$				
ε	M1	M2	Q1	Q2	Q3	M1	M2	Q1	Q2	Q3
N(0,1)	90.01	88.41	84.47	85.76	88.50	90.02	89.49	89.47	89.64	89.50
Uniform(-1,1)	93.13	88.17	84.26	85.65	88.76	94.90	89.26	89.44	89.61	89.65
χ^2_2	92.55	90.40	84.24	85.66	88.53	92.90	90.38	89.47	89.63	89.55
Cauchy(0,1)	95.95	92.08	83.56	85.09	88.18	98.76	92.98	89.43	89.60	89.46

the prediction intervals. One thousand training samples and their corresponding evaluation samples were used to compute mean values of the prediction errors and coverage errors. Moreover, five hundred bootstrap replicates were considered. For reasons of brevity, prediction errors are omitted, and only coverages of prediction intervals are presented and discussed below.

Table 4 contains the empirical coverages obtained for the homoscedastic model (Ho), with a nominal level of 90%, and for different error distributions and sample sizes. The error distributions investigated were standard normal, uniform on the interval $(-1, 1)$, chi-square with two degrees of freedom, and a standard Cauchy distribution. The classical prediction intervals based on linear mean regression show that method M1 provides very accurate results under the standard normal error distribution, while the other three distributions provide empirical coverage that is higher than the nominal level. Method M2 based on linear mean regression with a bootstrap approximation of the prediction error provides accurate coverage for all distributions (with better accuracy for larger sample size), with the only exception being for the Cauchy distribution. The Cauchy distribution does not have a mean, which makes the classical estimator of the linear mean regression inconsistent. The two quantile-based methods, Q1 and Q2, show a coverage below the nominal level for all distributions, although this under-estimation goes to zero, with increasing sample size. Method Q2 performs marginally better than Q1. The proposed method Q3 exhibits an accurate coverage for all four distributions, even for small sample sizes.

Table 5 shows the empirical coverages under the first heteroscedastic model (He1), for a nominal level of 90%, and two sample sizes. The error distribution was a standard normal one. In this way, we could analyze the specific effect of heteroscedasticity, without incorporating deviation from normality. Intervals I1 to I5 are defined by means of the quantiles of equal probability of the distribution of the linear function $X_1 + X_2 + X_3 + X_4 + X_5$, because this is the underlying cause of heteroscedasticity. Coverage for each interval is, in a sense, an indicator of conditional coverage. Simulations show that methods M1 and M2 provide inaccurate conditional coverage, where some intervals (I1 and I2) have under-coverage, and others (I4 and I5) have over-coverage. This effect does not diminish with increasing sample size. Observations for the first intervals have a smaller conditional standard deviation (and are more likely contained within the prediction intervals), while the

Table 5 Coverage (in percentage) of prediction intervals obtained by the five methods described in Section 2.4, with heteroscedastic model He1. Values are for a nominal level of 90% and two sample sizes. I1 to I5 represent five intervals of ordered expected values of the response variable. M1 and M2 are based on mean regression models, Q1 and Q2 are based on quantile regression, and Q3 is the median regression model proposed herein.

	$n = 100$					$n = 1000$				
	M1	M2	Q1	Q2	Q3	M1	M2	Q1	Q2	Q3
I1	97.13	96.17	82.82	84.26	92.62	97.27	96.95	89.30	89.43	91.28
I2	93.43	91.95	84.29	85.76	90.10	93.64	93.12	89.44	89.63	90.15
I3	90.42	88.61	84.80	86.05	88.54	90.48	89.88	89.53	89.68	89.54
I4	87.03	85.15	84.89	86.38	87.06	87.11	86.50	89.47	89.63	89.02
I5	81.79	79.52	84.86	86.43	84.55	81.75	81.01	89.51	89.67	88.25
Average	89.96	88.28	84.33	85.78	88.57	90.05	89.49	89.45	89.61	89.65

last intervals have larger ones. This is a natural consequence of heteroscedasticity, since methods M1 and M2 assume homoscedasticity. Coverage derived from methods Q1 and Q2 is somewhat smaller than the nominal level for small sample size, but converges to the nominal level as sample size increases. This reflects the fact that these two methods are based on estimating extreme quantiles using a linear model, which is valid under the linear heteroscedasticity of model He1. Method Q3 provides reasonably accurate coverage at any interval, with better performance observed for larger sample sizes.

Table 6 presents empirical coverages for the second heteroscedastic model (He2), with a nominal level of 90% and two sample sizes. The error distribution is a standard normal one, and the five intervals I1 to I5 are constructed, as described for Table 5. The four methods M1, M2, Q1 and Q2 are unable to provide accurate coverage in this case. Methods M1 and M2 show the same over-coverage for the first intervals and under-coverage for the last ones, as observed under the He1 model. Methods Q1 and Q2 show an overall under-coverage up to a sample size of $n = 1000$; for larger sample sizes, the non-linearity of the heteroscedasticity produces over-coverage in some intervals, I2 to I4, and under-coverage in others, I1 and I5. Meanwhile, method Q3 is robust to heteroscedasticity of any form, and even under this non-linear model, it provides accurate coverage, with better results for larger sample sizes.

5 Conclusions

Quantile regression methods are evaluated as an alternative to mean regression for prediction and calculation of prediction intervals of NO_x concentrations around the power plant in As Pontes, Spain. We show that for these data, median regression provides smaller prediction errors than mean regression. Heteroscedasticity and a non-normal error distribution were found to characterize these data, which deviate from the assumptions for classical mean regression models and likely explain the better performance of quantile methods for these data.

Although two known methods based on quantile regression were explored for obtaining prediction intervals, because of the special features of our atmospheric

Table 6 Coverage (in percentage) of prediction intervals obtained using the five methods described in Section 2.4, with heteroscedastic model He2. Values are for a nominal level of 90% and two sample sizes. I1 to I5 represent five intervals of ordered expected values of the response variable. M1 and M2 are based on mean regression models, Q1 and Q2 are based on quantile regression, and Q3 is the median regression model proposed herein.

	$n = 100$					$n = 1000$				
	M1	M2	Q1	Q2	Q3	M1	M2	Q1	Q2	Q3
I1	98.40	97.38	78.02	79.53	92.16	98.68	98.02	85.12	85.34	90.51
I2	96.75	95.20	86.95	88.08	91.44	97.22	96.26	92.38	92.52	90.73
I3	94.22	92.04	88.18	89.39	90.49	94.49	93.20	92.84	92.96	90.28
I4	88.92	86.27	87.35	88.76	87.67	89.24	87.47	91.60	91.75	89.34
I5	74.68	71.39	81.08	82.65	81.12	74.61	72.42	85.34	85.54	86.96
Average	90.59	88.46	84.32	85.68	88.58	90.85	89.47	89.46	89.62	89.56

data, we also proposed an additional method based on quantile regression estimation and bootstrap approximation of the prediction error. Our new method gave a markedly better performance than other methods evaluated here. In a simulation study, we showed how deviations from the assumptions of homoscedasticity and normality affected other methods for computing prediction intervals. The coverage accuracy of our new method was shown for both real and simulated scenarios.

Acknowledgements This study was supported by Project MTM2013-41383P from the Spanish Ministry of Economy and Competitiveness, as well as the European Regional Development Fund (ERDF). Support from the IAP network StUDyS from the Belgian Science Policy is also acknowledged. M. Conde-Amboage was supported by FPU grant AP2012-5047 from the Spanish Ministry of Education. Comments and suggestions from two referees are gratefully acknowledged.

References

Cade BS, Noon BR (2003) A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment* **1**: 412-420.

Chernozhukov V (2005) Extremal quantile regression. *The Annals of Statistics* **33**: 806-839.

Conde-Amboage M, Sánchez-Sellero C, González-Manteiga W (2015) A lack-of-fit test for quantile regression models with high-dimensional covariates. *Computational Statistics and Data Analysis* **88**: 128-138.

Fahrmeir L, Kneib T, Lang S and Marx B. 2013. *Regression: Models, Methods and Applications*. Springer.

Feng X, He X and Hu J. 2011. Wild bootstrap for quantile regression. *Biometrika* **98**: 995-999.

Fernández-Castro BM, Prada-Sánchez JM, González-Manteiga W, Febrero-Bande M, Bermúdez-Cela JL and Hernández-Fernádez JJ. 2003. Prediction of SO_2 levels using neural networks. *Journal of the Air and Waste Management Association* **53**: 532-538.

Fernández-Castro BM, Guillas S and González-Manteiga W. 2005. Functional samples and bootstrap for predicting sulfur dioxide levels. *Technometrics* **47**: 212-222.

Fernández-Castro BM and González-Manteiga W. 2008. Boosting for real and functional samples: an application to an environmental problem. *Stochastic Environmental Research and Risk Assessment* **22**: 27-37.

Fontanella L, Ippoliti L, Sarra A, Valentini P and Palermi S. 2015. Hierarchical generalised latent spatial quantile regression models with applications to indoor radon concentration. *Stochastic Environmental Research and Risk Assessment* **29**: 357-367.

García-Jurado I, Gonzalez-Manteiga W, Prada-Sánchez JM, Febrero-Bande M and Cao R. 1995. Predicting using Box-Jenkins, nonparametric and bootstrap techniques. *Technometrics* **37**: 303-310.

Hall P and Yao Q. 2005. Approximating conditional distribution functions using dimension reduction. *The Annals of Statistics* **33**: 1404-1421.

Hall P, Wolff R.C.L. and Yao Q. 1999. Methods for estimating a conditional distribution fuction. *Journal of the American Statistical Association* **94**: 154-163.

Koenker, R. 1994. Confidence intervals for regression quantiles. *Proceedings of the 5th Prague Symposium on Asymptotic Statistics, New York: Springer-Verlag*: 349-359.

Koenker R. 2005. *Quantile Regression*. Cambridge: Cambridge University Press.

Koenker R and Bassett G. 1978. Regression quantiles. *Econometrica* **46**: 33-50.

Li Q and Racine JS. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Mayr A, Hothorn T and Fenske N. 2012. Prediction intervals for future BMI values of individual children - a non-parametric approach by quantile boosting. *BMC Medical Research Methodology* **12**: 6.

Meinshausen N. 2006. Quantile Regression Forests. *Journal of Machine Learning Research* **7**: 983-999.

Nanos N, Grigoratos T, Rodríguez-Martín JA and Samara C. 2015. Scale-dependent correlations between soil heavy metals and As around four coal-fired power plants of northern Greece. *Stochastic Environmental Research and Risk Assessment* **29**: 1531-1543.

Prada-Sánchez JM and Febrero-Bande M. 1997. Parametric, Non-Parametric and Mixed approaches to prediction of sparsely distributed pollution incidents: a case study. *Journal of Chemometrics* **11**: 13-32.

Prada-Sánchez JM, Febrero-Bande M, Cotos-Yáñez T, González-Manteiga W, Bermúdez-Cela JL and Lucas-Domínguez T. 2000. Prediction of SO_2 pollution incidents near a power station using partially linear models and a historical matrix of predictor-response vectors. *Environmetrics* **11**: 209-225.

Roca-Pardiñas J, González-Manteiga W, Febrero-Bande M, Prada-Sánchez JM and Cadarso-Suárez C. 2004. Predicting binary time series of SO_2 using generalized additive models with unknown link function. *Environmetrics* **15**: 729-742.

Roca-Pardiñas J, Cadarso-Suárez C and González-Manteiga W. 2005. Testing for interactions in generalized additive models. Application to SO_2 pollution data. *Statistics and Computing* **15**: 289-299.

Salama A. 2005. A note on the impact of environmental performance on financial performance. *Structural Change and Economic Dynamics* **16**: 413-421.

Seber GAF. 1977. *Linear Regression Analysis*. Wiley.

Shi JP and Harrison RM. 1997. Regression modelling of hourly NO_x and NO_2 concentrations in urban air in London. *Atmospheric Environment* **31**: 4081-4094.

Sousa SIV, Pires JCM, Martins FG, Pereira MC and Alvim-Ferraz MCM. 2009. Potentialities of quantile regression to predict ozone concentrations. *Environmetrics* **20**: 147-158.

Stine RA. 1985. Bootstrap prediction intervals for regression. *Journal of the American Statistical Association* **80**: 1026-1031.

Zhou KQ and Portnoy SL. 1996. Direct use of regression quantiles to construct confidence sets in linear models. *The Annals of Statistics* **24**: 287-306.