



**HAL**  
open science

## Conditional quantiles with functional covariates: an application to ozone pollution forecasting

Hervé Cardot, Christophe Crambes, Pascal Sarda

► **To cite this version:**

Hervé Cardot, Christophe Crambes, Pascal Sarda. Conditional quantiles with functional covariates: an application to ozone pollution forecasting. INCONNU, 2004, /, France. pp.769-776. hal-00134251

**HAL Id: hal-00134251**

**<https://hal.science/hal-00134251>**

Submitted on 1 Mar 2007

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# CONDITIONAL QUANTILES WITH FUNCTIONAL COVARIATES : AN APPLICATION TO OZONE POLLUTION FORECASTING

Hervé Cardot<sup>a</sup>, Christophe Crambes<sup>b</sup> and Pascal Sarda<sup>b,c</sup>

*Key words:* functional data analysis, conditional quantiles, robustness,  $B$ -spline functions, weighted least squares, backfitting algorithm, Ozone forecasting.

*COMPSTAT 2004 section:* Functional Data Analysis.

**Abstract:** We are interested in estimating conditional quantiles when we have functional covariates, aiming to forecast Ozone pollution. We modelize conditional quantiles as a continuous linear functional of the covariates and we propose a spline estimator of the coefficient which minimizes a  $L^1$ -type penalized criterion. Then we give some insights on the asymptotic behaviour of this estimator. At last, we apply this approach to pollution data in the area of Toulouse.

## 1 Presentation of the pollution data

Pollution forecasting is nowadays of primary importance, particularly for prevention. In the city of Toulouse (France), ORAMIP<sup>1</sup> deals with measures stations of specific pollutants. These measures, as well as meteorological measures, are made each hour. So we have functional variables (see Ramsay and Silverman, 1997 and 2002) known in some discretisation points.

More precisely, the data consist in hourly measurements during the period going from the 15<sup>th</sup> May to the 15<sup>th</sup> September for the years 1997, 1998, 1999 and 2000, for the following variables : Nitrogen Monoxide, Nitrogen Dioxide, Dusts, Ozone, Wind Direction, Wind Speed, Temperature, Humidity and Sun Radiance. These variables are observed in six different stations in Toulouse. There are some missing data, principally because of breakdowns or missing measurement apparatus.

A PCA of these data has shown that the behaviour of these variables does not vary much from a station to another. That is why we considered the mean over the stations for each variable. This will also have the advantage to remove the problem of missing data. However, for some variables, the missing data were too numerous, so, in the following, we will just consider the five following variables : Nitrogen Monoxide (NO), Nitrogen Dioxide (N<sub>2</sub>), Ozone (O<sub>3</sub>), Wind Direction (WD) and Wind Speed (WS).

The aim of the study is to predict the maximum of Ozone for a day knowing the (functional) values of the above five variables the day before. We will achieve this goal in section 3 by means of estimating the conditional median of the Ozone. Before this, we study in section 2 the general problem of estimating a conditional quantile for a (or several) functional predictor(s).

<sup>1</sup> "Observatoire Régional de l'Air en Midi-Pyrénées"

## 2 Quantile regression for functional covariates

### 2.1 Conditional quantiles for functional covariates

Let us consider a sample  $(X_i, Y_i)_{i=1, \dots, n}$  of pairs of random variables, independent and identically distributed, with the same distribution as  $(X, Y)$ , with  $X$  belonging to the functional space  $L^2([0, 1])$  of square integrable functions defined on  $[0, 1]$ , and  $Y$  belonging to  $\mathbb{R}$ . Without loss of generality, we suppose that  $X$  is a centered variable, that is to say  $\mathbb{E}(X) = 0$ . Assume that  $H$ , the range of  $X$ , is a closed subspace of  $L^2([0, 1])$ . Let  $\alpha$  be a real number in  $]0, 1[$  and  $x$  a function in  $H$ . The *conditional  $\alpha$ -quantile* of  $Y$  given  $[X = x]$  is defined as the scalar  $g_\alpha(x)$  such that

$$P(Y \leq g_\alpha(x) | X = x) = \alpha, \quad (1)$$

where  $P(\cdot | X = x)$  is the conditional probability given  $[X = x]$ .

Provided that  $\mathbb{E}|Y| < \infty$ ,  $g_\alpha(x)$  can be defined in an equivalent way as the solution of the minimization problem

$$\min_{a \in \mathbb{R}} \mathbb{E}(l_\alpha(Y - a) | X = x), \quad (2)$$

where  $l_\alpha$  is the function defined by  $l_\alpha(u) = |u| + (2\alpha - 1)u$  (see Koenker and Bassett, 1978).

We assume now that  $g_\alpha$  is a linear and continuous functional. This condition can be seen as the direct generalization of the model introduced by Koenker and Bassett, the difference being that here, the covariates are functions. Then, there exists a unique function  $\Psi_\alpha \in L^2([0, 1])$  such that

$$g_\alpha(X) = \langle \Psi_\alpha, X \rangle = c + \int_0^1 \Psi_\alpha(t) X(t) dt, \quad (3)$$

where the notation  $\langle \cdot, \cdot \rangle$  refers to the usual inner product of  $L^2([0, 1])$ . The norm of  $L^2([0, 1])$  induced by this inner product will be noted  $\|\cdot\|$ .

### 2.2 Spline estimator of $\Psi_\alpha$

Our goal is now to introduce an estimator of the function  $\Psi_\alpha$ . In the case where the covariate  $X$  is real, Koenker and Bassett (1978) have proposed an estimator based on the minimization of the empirical version of (2); for nonparametric modelling, estimators have already been proposed: see for example Bhattacharya and Gangopadhyay (1990), Fan, Hu and Truong (1994) or Lejeune and Sarda (1988). He and Shi (1994) proposed a spline estimator and although our setting is quite different, the estimator of  $\Psi_\alpha$  defined below is of the same type as the one introduced by He and Shi (based on regression splines). However in our (functional) case there is a need to introduce a penalization in the criterion to be minimized.

We consider a space of spline functions : for this we choose a degree (fixed) and knots in the interval  $[0, 1]$ . For given integers  $q$  and  $k = k_n$ , with  $k \neq 0$ , we consider splines of degree  $q$  and  $k - 1$  equispaced knots on  $[0, 1]$  (see de Boor, 1978). This space of spline functions is a vectorial space of dimension  $k + q$ . A basis of this vectorial space is the set of the so-called  $B$ -spline functions, that we note  $\mathbf{B}_{\mathbf{k}, \mathbf{q}} = {}^t(B_1, \dots, B_{k+q})$ .

We estimate  $\Psi_\alpha$  by a linear combination of the  $B_l$  functions. This leads us to find a vector  $\boldsymbol{\theta} = {}^t(\hat{\theta}_1, \dots, \hat{\theta}_{k+q})$  in  $\mathbb{R}^{k+q}$  such that

$$\widehat{\Psi}_\alpha = \sum_{l=1}^{k+q} \hat{\theta}_l B_l = {}^t \mathbf{B}_{\mathbf{k}, \mathbf{q}} \widehat{\boldsymbol{\theta}}, \quad (4)$$

with  $\widehat{\boldsymbol{\theta}}$  solution of the following minimization problem, which is the penalized empirical version of (2),

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{k+q}} \left\{ \frac{1}{n} \sum_{i=1}^n l_\alpha(Y_i - c - \langle {}^t \mathbf{B}_{\mathbf{k}, \mathbf{q}} \boldsymbol{\theta}, X_i \rangle) + \rho \| ({}^t \mathbf{B}_{\mathbf{k}, \mathbf{q}} \boldsymbol{\theta})^{(m)} \|^2 \right\}, \quad (5)$$

where  $({}^t \mathbf{B}_{\mathbf{k}, \mathbf{q}} \boldsymbol{\theta})^{(m)}$  is the  $m$ -th derivative of the spline function  ${}^t \mathbf{B}_{\mathbf{k}, \mathbf{q}} \boldsymbol{\theta}$  and  $\rho$  is a penalization parameter which role is to control the smoothness of the estimator. This criterion is similar to the one introduced by Cardot *et. al.* (2003) for the estimation of the conditional mean in the functional linear model, the quadratic function being here replaced by the loss function  $l_\alpha$ . In this case, we have to deal with an optimization problem that does not have an explicit solution, contrary to the case of the functional linear model. That is why we adopted the strategy proposed by Lejeune and Sarda (1988). At first, let us note that the minimization problem (5) is equivalent to

$$\min_{c \in \mathbb{R}, \boldsymbol{\theta} \in \mathbb{R}^{k+q}} \left\{ \frac{1}{n} \sum_{i=1}^n \delta_i(\alpha) | Y_i - c - \langle {}^t \mathbf{B}_{\mathbf{k}, \mathbf{q}} \boldsymbol{\theta}, X_i \rangle | + \rho \| ({}^t \mathbf{B}_{\mathbf{k}, \mathbf{q}} \boldsymbol{\theta})^{(m)} \|^2 \right\}. \quad (6)$$

where the function  $\delta_i$  is defined by  $\delta_i(\alpha) = 2\alpha \mathbb{1}_{\{Y_i - c - \langle {}^t \mathbf{B}_{\mathbf{k}, \mathbf{q}} \boldsymbol{\theta}, X_i \rangle \geq 0\}} + 2(1 - \alpha) \mathbb{1}_{\{Y_i - c - \langle {}^t \mathbf{B}_{\mathbf{k}, \mathbf{q}} \boldsymbol{\theta}, X_i \rangle < 0\}}$ . The algorithm for solving (6) described below consists in performing iterative reweighted least squares (see Ruppert and Carroll, 1988) : at step  $j + 1$ ,  $\delta_i(\alpha)$  is replaced by the value  $\delta_i^j(\alpha)$  evaluated at step  $j$  and the absolute value is replaced by the ratio of square residuals (at step  $j + 1$ ) on the root of residuals computed from step  $j$ .

– **Initialization**

We determine  $\boldsymbol{\beta}^1 = {}^t(c^1, \boldsymbol{\theta}^1)$  solution of the minimization problem

$$\min_{c \in \mathbb{R}, \boldsymbol{\theta} \in \mathbb{R}^{k+q}} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - c - \langle {}^t \mathbf{B}_{\mathbf{k}, \mathbf{q}} \boldsymbol{\theta}, X_i \rangle)^2 + \rho \| ({}^t \mathbf{B}_{\mathbf{k}, \mathbf{q}} \boldsymbol{\theta})^{(m)} \|^2 \right\},$$

which solution  $\beta^1$  is given by  $\beta^1 = \frac{1}{n}(\frac{1}{n} {}^t\mathbf{D}\mathbf{D} + \rho\mathbf{K})^{-1} {}^t\mathbf{D}\mathbf{Y}$ , with

$$\mathbf{D} = \begin{pmatrix} 1 & \langle B_1, X_1 \rangle & \dots & \langle B_{k+q}, X_1 \rangle \\ \vdots & \vdots & & \vdots \\ 1 & \langle B_1, X_n \rangle & \dots & \langle B_{k+q}, X_n \rangle \end{pmatrix} \quad \text{and} \quad \mathbf{K} = \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{G} \end{pmatrix},$$

where  $\mathbf{G}$  is the  $(k+q) \times (k+q)$  matrix such that  $\mathbf{G}_{jl} = \langle B_j^{(m)}, B_l^{(m)} \rangle$ .

– **Step j+1**

Knowing  $\beta^j = {}^t(c^j, \theta^j)$ , we determine  $\beta^{j+1} = {}^t(c^{j+1}, \theta^{j+1})$  solution of the minimization problem

$$\min_{c \in \mathbb{R}, \theta \in \mathbb{R}^{k+q}} \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\delta_i^j(\alpha)(Y_i - c - \langle {}^t\mathbf{B}_{k,q}\theta, X_i \rangle)^2}{[(Y_i - c - \langle {}^t\mathbf{B}_{k,q}\theta, X_i \rangle)^2 + \eta^2]^{1/2}} + \rho \|({}^t\mathbf{B}_{k,q}\theta)^{(m)}\|^2 \right\},$$

where  $\eta$  is a strictly positive constant that allows us to avoid a denominator equal to zero. Let us define the  $n \times n$  diagonal matrix  $\mathbf{W}_j$  with diagonal elements given by

$$[\mathbf{W}_j]_{ii} = \frac{\delta_i^j(\alpha)}{n[(Y_i - c - \langle {}^t\mathbf{B}_{k,q}\theta, X_i \rangle)^2 + \eta^2]^{1/2}}.$$

Then,  $\beta^{j+1} = ({}^t\mathbf{D}\mathbf{W}_j\mathbf{D} + \rho\mathbf{K})^{-1} {}^t\mathbf{D}\mathbf{W}_j\mathbf{Y}$ .

Let us notice that, at each step of the algorithm, the estimator depends on many parameters : the number  $k$  of knots, the degree  $q$  of the spline functions, the order  $m$  of derivation, and the parameter  $\rho$  of the penalization. In our experience, we found that the penalization parameter  $\rho$  is of primary importance at least when the number of knots is large enough (see also Marks and Eilers, 1996, Besse *et. al.*, 1997). For this reason, we choose in our study (see section 3) to fix  $q = 3$  and  $m = 2$ , which are values commonly used to reach a sufficient degree of regularity. After several attempts, we fix  $k$  to be 8 *i.e.* a number of knots which is not too small. Now, the parameter  $\rho$  is chosen at each step of the algorithm by minimizing the generalized cross validation criterion (see Wahba, 1990)

$$GCV(\rho) = \frac{\frac{1}{n} \sum_{l=1}^n (Y_l - \widehat{Y}_l)^2}{\left(1 - \frac{1}{n} \text{Tr}(\mathbf{H}(\rho))\right)^2}, \quad (7)$$

where  $\widehat{\mathbf{Y}} = \mathbf{H}(\rho)\mathbf{Y}$  and  $\mathbf{H}(\rho) = \frac{1}{n}\mathbf{D}(\frac{1}{n} {}^t\mathbf{D}\mathbf{W}_j\mathbf{D} + \rho\mathbf{K})^{-1} {}^t\mathbf{D}\mathbf{W}_j$ .

### 2.3 Multiple conditional quantiles

Let us notice that this estimation procedure can be easily extended to the case where there is more than one covariate. Considering now  $v$  functional covariates  $X^1, \dots, X^v$ , we have the following model

$$P(Y_i \leq c + g_\alpha^1(X_i^1) + \dots + g_\alpha^v(X_i^v) / X_i^1 = x_i^1, \dots, X_i^v = x_i^v) = \alpha. \quad (8)$$

If we assume as before that  $g_\alpha^1, \dots, g_\alpha^v$  are linear and continuous functionals from a closed subspace of  $L^2([0, 1])$ , we can write  $g_\alpha^r(X_i^r) = \langle \Psi_\alpha^r, X_i^r \rangle$  for  $r = 1, \dots, v$  with  $\Psi_\alpha^1, \dots, \Psi_\alpha^v$  in  $L^2([0, 1])$ . The estimation of each functional  $\Psi_\alpha^r$  is obtained using the iterative backfitting algorithm (see Hastie and Tibshirani, 1990) combined with the algorithm described previously.

## 2.4 An asymptotic result

In this section we give some insights on the asymptotic behavior of  $\widehat{\Psi}_\alpha$ . As a matter of fact, we give an upper bound for the  $L^2$  convergence : the proof of this result can be found in Cardot *et al.* (2004). Let us now introduce some notations that we use in the following. If we assume that  $\mathbb{E}\|X\|^2 < \infty$ , the covariance operator of  $X$ , denoted by  $\Gamma_X$ , is defined for all  $u$  in  $L^2([0, 1])$  by  $\Gamma_X u = \mathbb{E}(\langle X, u \rangle X)$ . This operator is non-negative, so we can associate it a semi-norm noted  $\|\cdot\|_2$  and defined by  $\|u\|_2^2 = \langle \Gamma_X u, u \rangle$ . Using notations from Cardot *et al.* (2003), we consider the  $(k + q) \times (k + q)$  matrix  $\widehat{\mathbf{C}}$  with general term  $\langle \Gamma_n(B_j), B_l \rangle$  where  $\Gamma_n$  is the empirical version of  $\Gamma$ . Moreover, we set  $\widehat{\mathbf{C}}_\rho = \widehat{\mathbf{C}} + \rho \mathbf{G}$  where the matrix  $\mathbf{G}$  is defined in section 2.2. It is possible to find a sequence  $(\eta_n)_{n \in \mathbb{N}}$  of non negative reals such that  $\Omega_n = \{\omega / \lambda_{\min}(\widehat{\mathbf{C}}_\rho) > c\eta_n\}$ , where  $\lambda_{\min}(\widehat{\mathbf{C}}_\rho)$  is the smallest eigenvalue of  $\widehat{\mathbf{C}}_\rho$ , has probability going to 1 when  $n$  goes to infinity (see Cardot *et al.*, 2003)

To prove the convergence result of the estimator  $\widehat{\Psi}_\alpha$ , we assume that the following hypotheses are satisfied.

(H.1)  $\|X_i\| \leq C_0 < +\infty$ , as.

(H.2) The function  $\Psi_\alpha$  is supposed to have a  $p'$ -th derivative  $\Psi_\alpha^{(p')}$  lipschitz continuous of order  $\nu \in [0, 1]$ .

In the following, we set  $p = p' + \nu$  and we suppose that  $q \geq p \geq m$ .

(H.3) The eigenvalues of  $\Gamma_X$  are strictly positive.

(H.4) For  $x \in H$ ,  $Y$  has a conditional density function  $f_Y^x$  given  $[X = x]$  continuous and strictly positive at the  $\alpha$ -quantile.

Then, we have the following result.

**Theorem 1** *Under hypotheses (H.1) – (H.4), if we also suppose there exists  $\beta, \gamma$  in  $]0, 1[$  such that  $k_n \sim n^\beta$ ,  $\rho \sim n^{-\gamma}$  and  $\eta_n \sim n^{-\beta - (1-\delta)/2}$ , then*

(i)  $\widehat{\Psi}_\alpha$  exists except on a set whose probability goes to zero as  $n$  goes to infinity,

(ii)  $\mathbb{E}(\|\widehat{\Psi}_\alpha - \Psi_\alpha\|_2^2 | X_1, \dots, X_n) = O_P\left(\frac{1}{k_n^{2p}} + \frac{1}{n\eta_n} + \frac{\rho^2}{k_n\eta_n} + \rho k_n^{2(m-p)}\right)$ .

### 3 Application to Ozone prediction

We want to predict the maximum of the variable Ozone one day  $i$ , noted  $Y_i$ , using the functional covariates observed the day before until 5:00 pm. We consider covariates with length of 24 hours. We can assume that, beyond 24 hours, the effects of the covariate are negligible knowing the last 24 hours, so each curve  $X_i$  begins at 6:00 pm the day  $i - 2$ .

We randomly splitted the initial sample  $(X_i, Y_i)_{i=1, \dots, n}$  into two sub-samples :

- a learning sample  $(X_{a_i}, Y_{a_i})_{i=1, \dots, n_l}$  with  $n_l = 332$ , used to determine the estimators  $\hat{c}$  and  $\hat{\Psi}_\alpha$ ,
- a test sample  $(X_{t_i}, Y_{t_i})_{i=1, \dots, n_t}$  with  $n_t = 142$ , used to evaluate the quality of the models and to make a comparison.

We use the conditional median to predict the value of  $Y_i$ , *i.e.*  $\alpha = 0.5$ . To judge the quality of the models, we give a prediction of the maximum of Ozone for each element of the test sample,

$$\widehat{Y}_{t_i} = \hat{c} + \int_D \widehat{\Psi}_\alpha(t) X_{t_i}(t) dt.$$

Then, we consider three criteria given by

$$C_1 = \frac{\frac{1}{n_t} \sum_{i=1}^{n_t} (Y_{t_i} - \widehat{Y}_{t_i})^2}{\frac{1}{n_t} \sum_{i=1}^{n_t} (Y_{t_i} - \bar{Y}_l)^2},$$

$$C_2 = \frac{1}{n_t} \sum_{i=1}^{n_t} |Y_{t_i} - \widehat{Y}_{t_i}|,$$

$$C_3 = \frac{\frac{1}{n_t} \sum_{i=1}^{n_t} l_\alpha(Y_{t_i} - \widehat{Y}_{t_i})}{\frac{1}{n_t} \sum_{i=1}^{n_t} l_\alpha(Y_{t_i} - q_\alpha(Y_l))},$$

where  $\bar{Y}_l$  is the empirical mean of the learning sample  $(Y_{a_i})_{i=1, \dots, n_l}$  and  $q_\alpha(Y_l)$  is the empirical  $\alpha$ -quantile of the learning sample  $(Y_{a_i})_{i=1, \dots, n_l}$ . This last criterion  $C_3$  is similar to the one proposed by Koenker and Machado (1999). We remark that, the more these criteria take low values (close to 0), the better is the prediction. After testing all the possible models with one to five covariates, we finally kept the model using the four covariates Ozone, Nitrogen Monoxide, Nitrogen Dioxide and Wind Speed (we have put some of the results obtained in table 1). For this model, figure 1 represents predicted Ozone vs measured Ozone. Except for one outlier, the prediction seems rather good. The most efficient covariate to estimate the maximum of Ozone is the Ozone curve the day before; however, we noticed an improvement adding other covariates. We can also think that the results we obtained can be improved by other covariates that were not available, like the curve of temperature for example. Finally, let us note that we can similarly estimate conditional quantiles of  $Y_i$  to derive some kind of predictive intervals.

Models	Variables	$C_1$	$C_2$	$C_3$
1 covariate	N2	0,814	16,916	0,906
	<b>O3</b>	<b>0,414</b>	<b>12,246</b>	<b>0,656</b>
	WS	0,802	16,836	0,902
2 covariates	O3, NO	0,413	11,997	0,643
	O3, N2	0,413	11,880	0,637
	O3, WS	0,414	12,004	0,635
3 covariates	O3, NO, N2	0,412	12,127	0,644
	O3, N2, WD	0,409	12,004	0,645
	O3, N2, WS	0,410	11,997	0,642
4 covariates	<b>O3, NO, N2, WS</b>	<b>0,400</b>	<b>11,718</b>	<b>0,634</b>
5 covariates	O3, NO, N2, WD, WS	0,401	11,750	0,639

TABLE 1 – Forecast quality for some models of median regression.

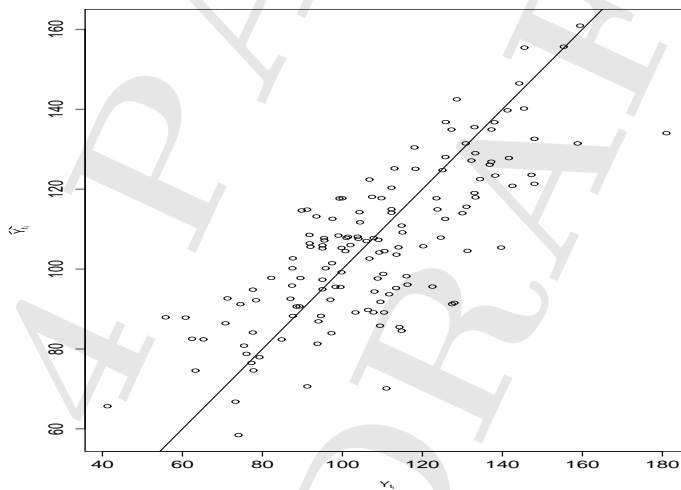


FIG. 1 – Predicted Ozone vs measured Ozone (prediction with the variables O3, NO, N2 and WS).

### Références

- [1] Besse, P.C., Cardot, H. and Ferraty, F. (1997). Simultaneous nonparametric regression of unbalanced longitudinal data. *Comput. Statist. and Data Anal.*, **24**, 255-270.
- [2] Bhattacharya, P.K. and Gangopadhyay, A.K. (1990). Kernel and Nearest-Neighbor Estimation of a Conditional Quantile. *Ann. Statist.*, **18**, 1400-1415.



- [3] Cardot, H., Crambes, C. and Sarda, P. (2004). Spline Estimators of Conditional Quantiles for Functional Covariates, *Preprint*.
- [4] Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline Estimators for the Functional Linear Model. *Statistica Sinica*, **13**, 571-591.
- [5] de Boor, C. (1978). *A Practical Guide to Splines*. Springer, New-York.
- [6] Fan, J., Hu, T.C. and Truong, Y.K. (1994). Robust Nonparametric Function Estimation. *Scand. J. Statist*, **21**, 433-446.
- [7] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman and Hall, New-York.
- [8] He, X. and Shi, P. (1994). Convergence Rate of  $B$ -Spline Estimators of Nonparametric Conditional Quantile Functions. *Nonparametric Statistics*, **3**, 299-308.
- [9] Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica*, **46**, 33-50.
- [10] Koenker, R. and Machado, J. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association*, **94**, 1296-1310.
- [11] Lejeune, M. and Sarda, P. (1988). Quantile Regression : A Nonparametric Approach. *Computational Statistics and Data Analysis*, **6**, 229-239.
- [12] Marx, B.D. and Eilers P.H. (1999). Generalized Linear Regression on Sampled Signals and Curves : A  $P$ -Spline Approach. *Technometrics*, **41**, 1-13.
- [13] Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis*. Springer-Verlag.
- [14] Ramsay, J.O. and Silverman, B.W. (2002). *Applied Functional Data Analysis*. Springer-Verlag.
- [15] Ruppert, D. and Carroll, J. (1988). *Transformation and Weighting in Regression*. Chapman and Hall.
- [16] Wahba, G. (1990). *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia.

*Acknowledgement:*

We would like to thank ORAMIP that provided the pollution data

*Address:*

<sup>a</sup>INRA Toulouse, Biométrie et Intelligence Artificielle, Chemin de Borde-Rouge, BP 27, 31326 Castanet-Tolosan Cedex, France.

<sup>b</sup>Université Paul Sabatier, Laboratoire de Statistique et Probabilités, UMR C5583, 118 route de Narbonne, 31062 Toulouse Cedex, France.

<sup>c</sup>Université Toulouse-le-Mirail, GRIMM, EA 2254, 5 allées Antonio Machado, 31058 Toulouse Cedex 9, France.

*E-mail:*

cardot@toulouse.inra.fr

crambes@cict.fr

Pascal.Sarda@math.ups-tlse.fr