Journal of Mathematics in Industry
a SpringerOpen Journal

RESEARCH Open Access

Semiparametric prediction models for variables related with energy production

Wenceslao González-Manteiga[1,2†], Manuel Febrero-Bande[1,2*†] and María Piñeiro-Lamas[3†]

*Correspondence:
manuel.febrero@usc.es
[1] MODESTYA group, Technological Institute for Industrial Mathematics (ITMATI), Santiago de Compostela, Spain
[2] Dept. of Statistics, Mathematical Analysis and Optimization, Fac. of Mathematics, Universidade de Santiago de Compostela, Santiago de Compostela, Spain
Full list of author information is available at the end of the article
† Equal contributors

Abstract

In this paper a review of semiparametric models developed throughout the years thanks to an extensive collaboration between the Department of Statistics and Operations Research of the University of Santiago de Compostela and a power station located in As Pontes (A Coruña, Spain) property of Endesa Generation, SA, is shown. In particular these models were used to predict the levels of sulphur dioxide in the environment of this power station with half an hour in advance. In this paper also a new multidimensional semiparametric model is considered. This model is a generalization of the previous models and takes into account the correlation structure of errors. Its behaviour is illustrated in a simulation study and with the prediction of the levels of two important pollution indicators in the environment of the power station: sulphur dioxide and nitrogen oxides.

Keywords: Semiparametric prediction models; Pollution indicators; Cointegration

1 Introduction: an environmental problem

The coal-fired power station in As Pontes is one of the production centers owned by Endesa Generation SA in the Iberian Peninsula. It is located in the town of As Pontes de García Rodríguez, northeast of A Coruña province.

This power station was designed and built to make use of lignite from the mine located in its vicinity. This solid fuel was characterized by its high moisture and sulphur contents and its low calorific value. Throughout the years the plant has undergone several transformation processes in their facilities with the aim of reducing emissions of sulphur dioxide (SO_2). The power station completed its last adaptation in 2008 to consume, as primary fuel, imported subbituminous coal, characterized by its low sulphur and ash contents.

The location of the power plant close to natural sites of high ecological value, such as the Natural Park *As Fragas do Eume* and existing legislation, mean that it has existed since the beginning a great concern for its impact on the environment. Therefore the station has a *Supplementary Control System of Air Quality* that allows it to make changes in operating conditions in order to reduce emissions when the weather conditions are adverse to the spread of the emitted smoke plume, specifically containing SO_2, and there are significant episodes of impaired air quality. Spanish law, by rules and regulations, sets maximum concentrations that can be achieved for these gases in a given period of time. In particular, for this plant the only limit that might be exceeded at any time, is one that is established on the hourly mean (continuously computed) from the concentration of SO_2 in the soil, the value of 350 μg/m^3.

The problem is to be able to predict, using the information received continuously at sampling stations and the past information, the future values for SO_2 levels. Statistical forecast models are the key to get these predictions and suggest a course of action to the plant operators.

In recent years, new statistical models have been designed to obtain the simultaneous prediction of two pollution indicators in the environment due to the changes in the environmental legislation, in the power station itself, and the construction of a new natural gas combined cycle station in the vicinity. The fuels that are going to be used make that the main interest lies in predicting the values of the nitrogen oxides (NO_x) which is emitted by both facilities simultaneously with the values of SO_2 which is only emitted by the power station.

All these changes have created a new problem: predicting hourly mean concentrations of sulphur dioxide and nitrogen oxides, measured in the environment of the two facilities. Faced with this new approach, the statistical forecast models are again an effective tool. Thus, a multidimensional prediction general model is designed (see Sect. 3).

2 Methods: one-dimensional predictive models

2.1 Models designed to solve the environmental problem

Resulting from the collaboration over the past years between the Department of Statistics and Operations Research at the University of Santiago de Compostela and the Environment Section of the power station, the *Integrated System of Statistical Prediction of the Immision* (SIPEI, in Spanish) have been created employing statistical models to provide predictions for the levels of SO_2 with a half an hour horizon.

Due to data availability with minutal frequency in real-time and current legislation, the hourly mean is considered from both of the values of SO_2 and NO_x, for predictions of future values of both pollutants. Thus, two time series are constructed, $X_{1,t}$ and $X_{2,t}$, for which the subscript t represents a minutal instant, and each value will be an average of the actual values for the last hour:

$$X_{1,t} = \frac{1}{60}\sum_{i=0}^{59} \mathrm{SO}_2(t-i) \quad \text{and} \quad X_{2,t} = \frac{1}{60}\sum_{i=0}^{59} \mathrm{NO}_x(t-i),$$

where $SO_2(t)$ and $NO_x(t)$ represent the concentration of SO_2 and NO_x, respectively, at time t, measured in $\mu g/m^3$.

The series of hourly SO_2 means has a characteristic behaviour, highly influenced by weather conditions and local topography. It takes values close to zero for long periods of time, and it can suddenly and sharply increase (episodes) in bad meteorological conditions for the dispersion of the smoke plume. Nowadays, the series of hourly NO_x means has a similar behaviour to that of SO_2, but on a smaller scale (see Fig. 1). The main objective of the developed statistical models is to predict the episodes, so our interest is centred on the values that occur less frequently along the time series.

Because of this, a kind of memory called *Historical Matrix* was designed (Prada-Sánchez and Febrero-Bande [14]), which will be essential to the behaviour of all developed models so far. This matrix is composed of a large number of vectors based on $(X_{t-l}, \ldots, X_t, X_{t+k})$:

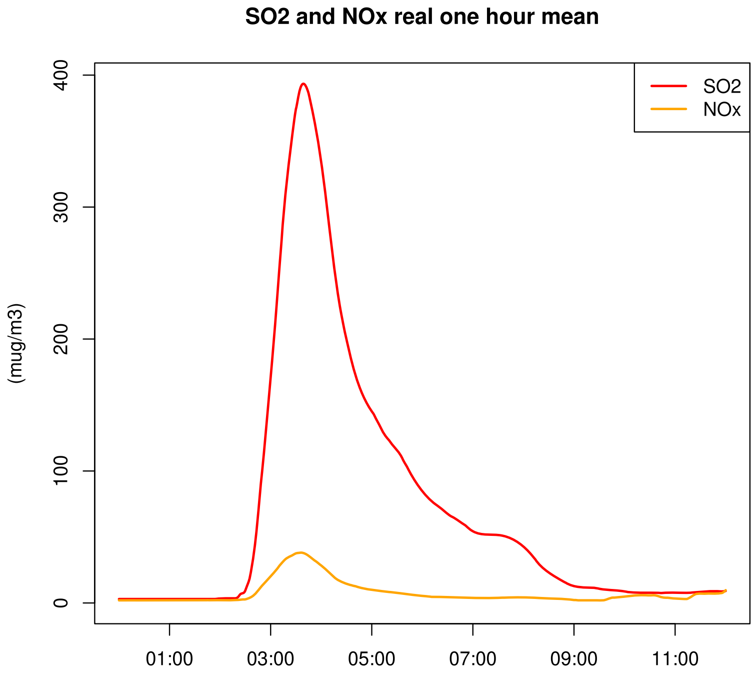

Figure 1 Episode depicted in one of sampling stations. The one hour mean of SO_2 and NO_x are, respectively, drawn in red and orange

real data of bihourly SO_2 or NO_x means, chosen so as to cover the full range of variable in question and make the role of historical memory. To ensure that cover the entire range of the variable, the matrix is divided into blocks according to the level of the response variable, X_{t+k}. To update the memory, in every instant, when a new observation is received, the historical matrix is renewed in the following way: the class to which the new observation belongs is found and then the oldest datum in such class leaves the matrix and the new observation enters it. With a sample built this way, makes sure that always have updated information on the full variation range of the interest variable, and over the years this concept has been adapted to the different statistical techniques used.

2.1.1 The first semiparametric model

In the early years of development, the data transmission frequency to SIPEI was pentaminutal, and also, the legislation in force at that time established the limit values for the two hour mean of the SO_2. For this reason, the prediction models for SO_2 levels initially worked with series of bihourly means. The objective was to obtain the prediction, with a half an hour horizon, for this time series. Therefore, each time it receives a new observation, X_t, it has to predict the value at six times ahead, X_{t+6}.

A *semiparametric approach* was considered (García-Jurado et al. [8]) which generalizes the traditional Box–Jenkins models as follows:

$$X_{t+\kappa} = \varphi_\kappa(X_t, X_{t-l}) + Z_{t+\kappa}, \qquad \kappa, l \in \mathbb{Z}^+,$$

where Z_t has an ARIMA structure of mean zero independent of X_t (Box et al. [1]).

In particular at each time t, the regression function $\varphi_6(X_t, X_{t-1}) = \mathbb{E}(X_{t+6}/X_t, X_{t-1})$ is estimated with the well-known Nadaraya–Watson kernel type estimator (see Nadaraya [13] and Watson [19]) using the information provided by the historical matrix. The second step is to calculate the residual time series $\hat{Z}_{t-64}, \ldots, \hat{Z}_t$ relative to the last six hours, where $\hat{Z}_i = X_i - \hat{\mathbb{E}}(X_i/X_{i-6}, X_{i-7})$ for each i and fits an appropriate ARIMA model for it. Finally we get the Box–Jenkins prediction of $\hat{Z}_{t+6}$. The final point prediction proposed is given by: $\hat{\mathbb{E}}(X_{t+6}/X_t, X_{t-1}) + \hat{Z}_{t+6}$.

2.1.2 Partially linear model

The information used by the previous semiparametric models to obtain the predictions is the past of the time series; however it might be useful to introduce additional information in order to improve these predictions. Specifically, meteorological and emission variables have been used with, the so-called *partially linear models* (Prada-Sánchez et al. [15]) to estimate bihourly mean values of SO_2 with one hour in advance.

Data in the form of (V_t, Z_t, Y_t) is considered, where V_t is a vector of exogenous variables, $Z_t = (X_t, X_{t-l})$ and $Y_t = X_{t+12}$ being X_t the series of bihourly SO_2 means; and it is assumed that this series conform to the following partially linear model: $Y_t = V_t^t \beta + \varphi(Z_t) + \epsilon_t$, where ϵ_t is an error term of mean equals to zero.

This model can easily estimated following Speckman [18] and allow us to extend the horizon to one hour maintaining the same level of accuracy as the semiparametric model for half an hour horizon. In any case, the incorporation of external information slightly improves the prediction because the measure point for the meteorological variables is located at 80 m over ground level which is relatively far away (and so, uncorrelated) respect to the typical height of the emitted smoke plume (above 800 m over ground level). Emission information is also of little interest because these signals are almost constant specially when the facility is working not describing at all the reasons that make the smoke plume falls to the ground. By these reasons, meteorological or emission information was not considered in the following models.

2.1.3 Neural networks

The change in the interest series established by the European Council Directive 1999/30/CE, from bihourly means to hourly means, causes the time series to be less smooth. At the beginning, the previous semiparametric model was adapted to work on the new series of hourly means. The results showed a considerable increase in terms of the variability of the given predictions, regarding the results usually obtained for the series of two hour means.

In an attempt to improve the response given by the SIPEI, and in particular, its point predictions with half an hour horizon, new predictors based on *neural networks models* were developed (Fernández de Castro et al. [6]).

A neural network model has been designed to provide predictions of one hour mean values of SO_2 with half an hour in advance. It consists of an input layer, one hidden layer and an output layer. The number of nodes in the output layer is determined by the size of the response to be obtained from the network; in this case interested in a prediction for X_{t+6}. As input to the network it has been taken the bidimensional vector (X_{t-3}, X_t) and the nodes in the hidden layer have been taken as the activation function of a logistic function, and in the output layer, the identity function.

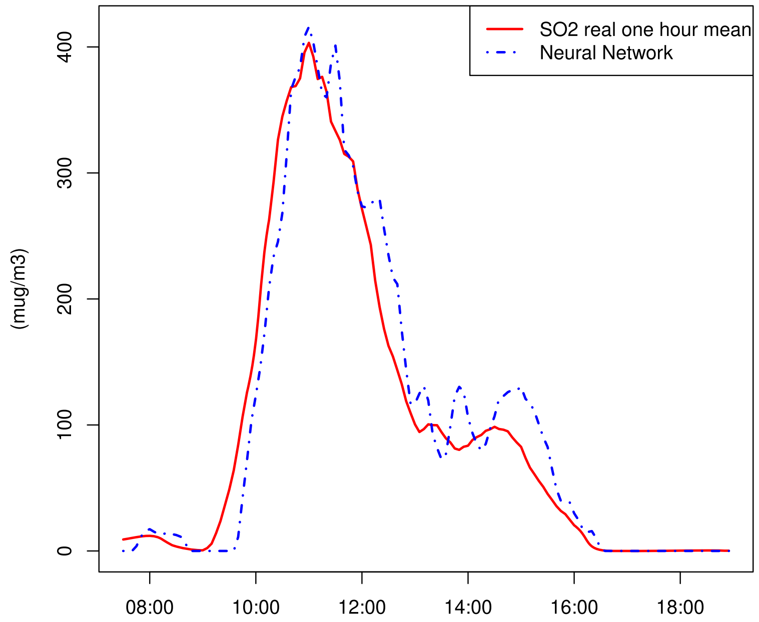

Figure 2 Episode of SO$_2$ depicted in one of sampling stations (red) jointly with the prediction provided by the neural network (blue) (Fernández de Castro et al. [6])

The predictor given by the neural network has the following expression:

$$\hat{X}_{t+6} = o_1 = \sum_{j=1}^{L} \omega_{1j}^{o} f_j^{h}\left(\theta_j^{h} + \omega_{j1}^{h} X_{t-3} + \omega_{j2}^{h} X_t\right)$$

with $f_j^h(z) = \frac{1}{1+e^{-z}}$.

The weights $\{\omega_{j1}^h, \omega_{j2}^h, \omega_{1j}^o; j = 1, \ldots, L\}$ and the trends $\{\theta_j^h; j = 1, \ldots, L\}$ are determined during the training process, as well as the final L number of hidden layer nodes, that is chosen like the value which neural network provides better results, after having trained networks with identical architecture and different values of L. To design the training set of the neural network it have been considered historical matrices, formerly introduced, suitably adapted.

Figure 2 shows the forecasts given half an hour before by the neural network with 50 nodes in its hidden layer for an episode depicted in one of the measuring stations. The good behaviour of the forecast (dotted line) can easily be seen. The procedures based on neural networks accurately predict the real one hour mean SO$_2$ air quality values (solid line). These models were optimized later with boosting learning techniques (Fernández de Castro and González-Manteiga [4]).

2.1.4 Functional data model

The one hour mean values of SO$_2$ can be treated as observations of a stochastic process in continuous time. The interest is, as it was discussed above, to predict a half-hour horizon, so that each of the curves is an interpolated data on half an hour. In this case curves were obtained by considering six pentaminutal consecutive observations, with sampling points for each functional data. Therefore, we use random variables with values in Hilbert space $H = L^2([0,6])$ with the form $X_t(u) = x(6t + u)$.

The following statistical model is considered $X_t = \rho(X_{t-1}) + \epsilon_t$, where ϵ_t is a Hilbertian strong white noise and $\rho : H \to H$ is the operator to estimate. For the estimation of ρ, a *functional kernel estimator* has been used in the *autoregressive Hilbertian of order-one framework*. Furthermore, it has been conveniently adapted the concept of historical matrix to the case where the data are curves (Fernández de Castro et al. [5]).

2.1.5 Other approaches designed to predict probabilities

The models described, so far, provide point predictions of SO_2, but other techniques have also been developed in order to predict probabilities. The aim of these alternative models is to estimate the probability that the series of bihourly SO_2 measures exceeds a certain level r with an hour anticipation, namely in our case, we predict $\mathbb{P}(Z_t) = \mathbb{P}(X_{t+12} > r|Z_t)$ being $Z_t = (X_t, X_t - X_{t-3})$. To do it *additive models with an unknown link function* (Roca-Pardiñas et al. [17]) have been used.

It has also been considered more complex *generalized additive models* (GAM) with second-order interaction terms (Roca-Pardiñas et al. [16]). They have shown that the GAM with interactions detects the onset of episodes earlier than it does GAM on its own.

2.2 Alternative one-dimensional models: additive models

In the statistical literature there is a wide range of one-dimensional models which can be used to predict the levels of SO_2. We will focus on the techniques we will use in the next section to construct our multidimensional model: additive models for continuous response.

There have been a number of proposals for fitting the additive models. Friedman and Stuetzle [7] introduced a backfitting algorithm and Buja et al. [2] studied its properties. Mammen et al. [12] proposed the so called *smooth backfitting* by employing projection arguments. Let $\{(Y_t, Z_t)\}_{t=1}^{T}$ be a random sample of a strictly stationary time series, with Y_t one-dimensional and Z_t q-dimensional following the model:

$$Y_t = m(Z_t) + \epsilon_t, \quad t \in \mathbb{Z}, \tag{1}$$

where $\{\epsilon_t\}$ is a white noise process and $\mathbb{E}[\epsilon_t|Z_t] = 0$.

Typically, it is assumed that the function m is additive with component functions m_j, for $j = 0, \dots, q$, thus

$$Y_t = m_0 + m_1(Z_{1,t}) + \cdots + m_q(Z_{q,t}) + \epsilon_t. \tag{2}$$

A generalized kernel nonparametric estimation can be given using *smooth backfitting* for the functions $m_1, \dots, m_q$ (see again the above mentioned papers).

In all the models described above it is usually necessary the selection of a regularization parameter (bandwidth with kernel smoothing, number of neurons in the hidden layer for neural networks, ...). The calibration of this parameter was developed using cross-validation techniques with the information of the updated Historical Matrix.

3 Methods: multidimensional semiparametric prediction

The new goal is to incorporate the prediction of NO_x with half an hour in advance, as well as to continue getting the predictions of SO_2, as has already been commented. The idea is to generalize the one-dimensional semiparametric approach proposed by García-Jurado et al. [8] taking into account the structure of correlation between the vectorial series that is intended to predict.

3.1 The model

Be $(Y,Z) = (Y_l, Z_l)$, $l = 0, \pm 1, \pm 2, \ldots$ a vectorial strictly stationary time series, where Y_l is a r-dimensional response series and Z_l is a q-dimensional covariables series and, let $\{(Y_t, Z_t)\}_{t=1}^{T}$ be a random sample of (Y,Z). The following model is considered

$$Y_t = \varphi(Z_t) + \mathcal{E}_t, \tag{3}$$

where $Y_t = (Y_{1,t}, \ldots, Y_{r,t})^t$, $Z_t = (Z_{1,t}, \ldots, Z_{q,t})^t$ and $\mathcal{E}_t = (\mathcal{E}_{1,t}, \ldots, \mathcal{E}_{r,t})^t$. Let us consider two possible structures for the multidimensional residuals series:

P1. Each $\mathcal{E}_{k,t}$ is a stationary AR(p_k) process of the form

$$\mathcal{E}_{k,t} = \sum_{i=1}^{p_k} \phi_k^i \mathcal{E}_{k,t-i} + \xi_{k,t} \quad \text{for all } t \in \mathbb{Z}, k = 1, \ldots, r$$

independent of Z_t, where $\xi_{k,t}$ is a white noise process with variance σ_k^2, for $k = 1, \ldots, r$.

P2. $\mathcal{E}_t$ has a VAR(p) structure of the form

$$\mathcal{E}_t = \sum_{i=1}^{p} \Phi_i \mathcal{E}_{t-i} + \xi_t \quad \text{for all } t \in \mathbb{Z},$$

independent of Z_t, where the Φ_i are fixed $(r \times r)$ coefficients matrices and ξ_t is a r-dimensional white noise process, i.e. $\mathbb{E}(\xi_t) = 0$, $\mathbb{E}(\xi_t \xi_t') = \Sigma_\xi$ and $\mathbb{E}(\xi_t \xi_s') = 0$ for $t \neq s$.

Our main objective is to predict Y_t using a sample of size T, κ instants ahead. The prediction of $Y_{t+\kappa}$ is then defined by

$$\dot{Y}_{t+\kappa} = \hat{\varphi}_\kappa(Z_t) + \dot{\mathcal{E}}_{t+\kappa}, \tag{4}$$

where $\hat{\varphi}_\kappa(Z_t)$ is a nonparametric estimate of $\varphi_\kappa(Z_t) = \mathbb{E}[Y_{t+\kappa}/Z_t]$ and $\dot{\mathcal{E}}_{t+\kappa}$ the prediction given, κ instants ahead, for the residual series constructed as $\hat{\mathcal{E}}_{t+\kappa} = Y_{t+\kappa} - \hat{\varphi}_\kappa(Z_t)$.

3.2 Estimations

We suppose that the model (3) is verified. The first step is to make a nonparametric estimation of φ independently for each of the r components of Y_t: $\varphi(Z_t) = (\varphi_1(Z_t), \ldots, \varphi_r(Z_t))$. Furthermore, we assume that the functions φ_k are additive with component functions φ_k^j, for $k = 1, \ldots, r$ and $j = 0, \ldots, q$, thus

$$\varphi_k(Z_t) = \varphi_k^0 + \varphi_k^1(Z_{1,t}) + \cdots + \varphi_k^q(Z_{q,t}), \quad k = 1, \ldots, r. \tag{5}$$

Therefore, r additive models with q covariates are estimated using the smooth backfitting technique. We have to take into account that the process $\mathcal{E}_t$ is not observable since the function φ is not known. Thus, we have to replace $\mathcal{E}_t$ by the residuals

$$\hat{\mathcal{E}}_t = Y_t - \hat{\varphi}(Z_t)$$

and use these approximations to $\mathcal{E}_t$ in the maximum likelihood estimations later defined.

To estimate the parametric part of the model, we must consider the two possible error structures proposed above:

P1. The parameters $\phi_k = (\phi_k^1, \ldots, \phi_k^{p_k})$ of the error process $\{\mathcal{E}_{k,t}\}$ are estimated by standard maximum likelihood methods. In particular, we use a conditional maximum likelihood estimator for every component of the form

$$\hat{\phi}_k = \arg\max_{\phi_k \in \Phi} \hat{l}(\phi_k),$$

where Φ is a compact parameter space and $\hat{l}$ is the conditional log-likelihood given by

$$\hat{l}(\phi_k, \sigma_k^2) = -\frac{T}{2}\log(2\pi) + \frac{1}{2}\log(\sigma_k^{-2}) - \frac{1}{2}\sum_{t=p_k+1}^{T} \bigl((\hat{\mathcal{E}}_{k,t} - \hat{\mathcal{E}}_{k,t}(\phi_k))/\sigma_k\bigr)^2$$

with $\hat{\mathcal{E}}_{k,t}(\phi_k) = \sum_{i=1}^{p_k} \phi_k^i \hat{\mathcal{E}}_{k,t-i}$.

P2. The coefficients matrices $(\Phi_1, \ldots, \Phi_p)$ of the r-dimensional error process $\{\mathcal{E}_t\}$ are also estimated by generalized maximum likelihood methods (Hamilton [10]). First, we need to establish the following notation: $\Phi^t = [\Phi_1\ \Phi_2\ \ldots\ \Phi_p]$ denote the $(r \times rp)$ coefficients matrix, let X_t be a $(rp \times 1)$ vector containing p lags of each of the elements of $\mathcal{E}_t$: $X_t^t = [\mathcal{E}_{t-1}^t\ \mathcal{E}_{t-2}^t\ \ldots\ \mathcal{E}_{t-p}^t]$.

The theoretical conditional log-likelihood function to be optimized has the following expression:

$$l(\Phi, \Sigma_\xi) = -\frac{rT}{2}\log(2\pi) + \frac{r}{2}\log\bigl|\Sigma_\xi^{-1}\bigr| - \frac{1}{2}\sum_{t=1}^{T}\bigl[\bigl(\mathcal{E}_t - \Phi^t X_t\bigr)^t\ \Sigma_\xi^{-1}\bigl(\mathcal{E}_t - \Phi^t X_t\bigr)\bigr].$$

Thus the conditional log-likelihood is:

$$\hat{l}(\hat{\Phi}, \hat{\Sigma}_\xi) = -\frac{rT}{2}\log(2\pi) + \frac{r}{2}\log\bigl|\hat{\Sigma}_\xi^{-1}\bigr| - \frac{1}{2}\sum_{t=1}^{T}\bigl[\bigl(\hat{\mathcal{E}}_t - \hat{\Phi}^t \hat{X}_t\bigr)^t\ \hat{\Sigma}_\xi^{-1}\bigl(\hat{\mathcal{E}}_t - \hat{\Phi}^t \hat{X}_t\bigr)\bigr].$$

3.3 Other considerations: the phenomenon of cointegration

Sometimes the vectorial processes can be cointegrated, so one has to take into account the *structure of correlation* between the series. The notion of cointegration has been one of the most important concepts in time series since Granger [9] and Engle and Granger [3] that formally developed it. The issue has broad applications in the analysis of economic data as well as several publications in the economic literature.

Let $Y_t = (Y_{1,t}, \ldots, Y_{r,t})^t$ be a vector of r time series integrated of order 1 ($I(1)$). Y_t is said to be *cointegrated* if a linear combination of them exists that it is stationary ($I(0)$), i.e., if there exists a vector $\beta = (\beta_1, \ldots, \beta_r)^t$ such as

$$\beta^t Y_t = \beta_1 Y_{1,t} + \cdots + \beta_r Y_{r,t} \sim I(0).$$

The vector β is called the *cointegration vector*. This vector is not unique since for any scalar c the linear combination $c\beta^t Y_t = \beta^{*t} Y_t \sim I(0)$. Therefore, normalization is often assumed to identify an unique β. A typical normalization is $\beta = (1, -\beta_2, \ldots, -\beta_r)^t$.

Johansen [11] addresses the issue of the cointegration within an error correction model in the framework of vector autoregressive models (VAR). Consider then a general model VAR(p) for the vector of r series Y_t

$$Y_t = \Phi_0 D_t + \Phi_1 Y_{t-1} + \cdots + \Phi_p Y_{t-p} + \xi_t, \quad t = 1, \ldots, T,$$

where D_t contains deterministic terms (constant, trend, ...).

Suppose Y_t is $I(1)$ and possibly cointegrated. Then, the VAR representation is not the most suitable representation for analysis because the cointegrating relationships are not explicitly apparent. The cointegrating relationships become apparent if the VAR model is transformed to a *vector error correction model* of order p (VECM(p))

$$\Delta Y_t = \Phi_0 D_t + \Pi Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + \cdots + \Gamma_{p-1} \Delta Y_{t-p+1} + \xi_t,$$

where $\Pi = \Phi_1 + \cdots + \Phi_p - I_r$, $\Gamma_k = -\sum_{j=k+1}^{p} \Phi_j$, $k = 1, \ldots, p-1$ and $\Delta Y_t = Y_t - Y_{t-1}$. The matrix Π is called the *long-run impact matrix* and Γ_k are the *short-run impact matrices*. Moreover, the rank of the singular matrix Π provides information on the number of cointegration relations that exist, i.e., *the rank of cointegration*. Johansen proposes a sequential procedure of likelihood ratio tests to estimate this range.

3.4 Prediction scheme

We present now the prediction scheme step by step:

1. Every instant t, $\varphi_\kappa(Z_t)$ is estimated with the smooth backfitting technique independently for each of r components using the data $(Y_l, Z_{l-\kappa})$, $l = \kappa + 1, \ldots, T$.
2. The residuals series $\hat{\mathcal{E}}_{t+\kappa}$ is computed by

$$\hat{\mathcal{E}}_{t+\kappa} = Y_{t+\kappa} - \hat{\varphi}_\kappa(Z_t), \quad t = 1, \ldots, T - \kappa.$$

3. The following step is to make an appropriate adjustment on the model error structure (VECM) and to obtain the prediction κ instants ahead: $\dot{\mathcal{E}}_{T+\kappa}$.
4. The proposed final prediction is given by (4).

This scheme is a natural generalization of the one-dimensional prediction models described in Sect. 2.1.1. In the next two sections simulation examples and real data analysis are considered.

4 Results and discussion

4.1 A simulation study

To analyze the behavior of the proposed prediction procedure, a simulation study has been performed generating samples from artificial series and making a prediction study to k lags using, in all cases, $Z_t = Y_{t-1}$.

The following models are considered:

Series 1. Two independent AR(3) with constant trend:

$$Y_t = \varphi + \begin{pmatrix} \mathcal{E}_{1,t} \\ \mathcal{E}_{2,t} \end{pmatrix},$$

being

$$\mathcal{E}_{1,t} = 0.50\mathcal{E}_{1,t-1} - 0.525\mathcal{E}_{1,t-2} + 0.75\mathcal{E}_{1,t-3} + \eta_{1,t},$$
$$\mathcal{E}_{2,t} = 0.1875\mathcal{E}_{2,t-1} - 0.50\mathcal{E}_{2,t-2} + 0.05\mathcal{E}_{2,t-3} + \eta_{2,t},$$

where $\eta_{1,t} \sim N(0, 0.25^2)$, $\eta_{2,t} \sim N(0, 0.10^2)$ and $\varphi = (25, 10)^t$.

Series 2. VAR(3) with constant trend:

$$Y_t = \varphi + \Pi_1(Y_{t-1} - \varphi) + \Pi_2(Y_{t-2} - \varphi) + \Pi_3(Y_{t-3} - \varphi) + \eta_t,$$

being $\Pi_1 = \begin{pmatrix} 0.50 & 0.3150 \\ 0.75 & 0.1875 \end{pmatrix}$, $\Pi_2 = \begin{pmatrix} -0.525 & 0 \\ 0 & -0.50 \end{pmatrix}$, $\Pi_3 = \begin{pmatrix} 0.75 & 0.375 \\ -0.50 & 0.050 \end{pmatrix}$, $\eta_t \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.25^2 & 0 \\ 0 & 0.10^2 \end{pmatrix}\right)$ and φ as in Series 1.

Series 3. NPVAR(1) with independent VAR(3) noise:

$$Y_t = \varphi(Y_{t-1}) + \mathcal{E}_t,$$

being $\varphi(y) = \begin{pmatrix} \varphi_1(y) \\ \varphi_2(y) \end{pmatrix} = \begin{pmatrix} 5\cos(|y_1|) \\ 5\cos(|y_2|) \end{pmatrix}$ and $\mathcal{E}_t = \Pi_1\mathcal{E}_{t-1} + \Pi_2\mathcal{E}_{t-2} + \Pi_3\mathcal{E}_{t-3} + \eta_t$, where Π_1, Π_2, Π_3 and η_t are similar to those of the previous series.

Series 4. VECM with constant trend:

$$Y_t = \varphi + \begin{pmatrix} Y_{1,t} \\ Y_{2,t} \end{pmatrix},$$

being $Y_{1,t} = Y_{1,t-1} + v_t$ and $Y_{2,t} = -Y_{1,t} + u_t$, where $v_t \sim N(0, 0.5^2)$, $u_t = 0.75u_{t-1} + \eta_t$, $\eta_t \sim N(0, 0.5^2)$ and φ is similar to that of the first series.

In each case, 1000 bidimensional series of length 500 were generated from the models given above ($Y_1^i, \ldots, Y_{500}^i$ with $1 \leq i \leq 1000$). These values correspond to the generation after an initial period of stabilization (starting at zero and neglecting the first 500 values drawn). For every sample, $M = 500$ possible continuations of the series were obtained for k periods ahead ($Y_{500+k}^{i1}, \ldots, Y_{500+k}^{i500}$), which were compared with the prediction that was made from the sample $Y_1^i, \ldots, Y_{500}^i$.

For each of these series, three predictors are compared:

(a) The nonparametric predictor using additive models with the estimation of each component independently (NPM).

(b) The semiparametric predictor for additive models to estimate the trend of each bidimensional series component independently with model P1 for the residuals (SPM).

(c) The semiparametric predictor for additive models to estimate the trend of each bidimensional series component independently with VAR modelling for the vector residuals proposed in the previous section as P2 (SPBM).

Thus, as noted above, by calling $Y_1^i, \ldots, Y_{500}^i$, $i = 1, \ldots, N = 1000$, each of the simulated series and, considering $\hat{Y}_{500+k}^{i(a)}$, $\hat{Y}_{500+k}^{i(b)}$ and $\hat{Y}_{500+k}^{i(c)}$, $k = 1, \ldots, 30$ as each of the predictors according to the methods (a), (b) and (c) respectively, methods are compared using Mean Square Prediction Errors:

$$\text{MSPE}(l) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}\left(Y_{500+k}^{ij} - \hat{Y}_{500+k}^{i(l)}\right)^2, \tag{6}$$

where Y^{ij}_{500+k} represents the observed value of the jth prolongation of the ith series, $j = 1,\ldots,M = 500$, $l = a, b$ or c and $k = 1,\ldots,30$.

The results are summarized in Tables 1 to 4. "MSPE(a,b,c)" means the mean square prediction error (6) for the methods (a), (b) and (c), respectively. It can be seen that the proposed semiparametric method improves the behavior the other two, specially in the first lags and as the lags grow the differences between the three methods become smaller. This is illustrated in Fig. 3 which compares the distribution of the prediction errors obtained with three predictors for the second model.

Table 1 AR independent model (Series 1)

Series 1	MSPE(a)		MSPE(b)		MSPE(c)	
Lags	Var. 1	Var. 2	Var. 1	Var. 2	Var. 1	Var. 2
1	0.1637	0.0134	0.0646	0.0103	0.0640	0.0102
2	0.1704	0.0135	0.0819	0.0107	0.0804	0.0105
3	0.1735	0.0134	0.0856	0.0129	0.0845	0.0127
10	0.1701	0.0134	0.1595	0.0137	0.1480	0.0135
20	0.1690	0.0134	0.2140	0.0141	0.1695	0.0135
30	0.1689	0.0134	0.2696	0.0147	0.1733	0.0135

Table 2 VAR(3) model (Series 2)

Series 2	MSPE(a)		MSPE(b)		MSPE(c)	
Lags	Var. 1	Var. 2	Var. 1	Var. 2	Var. 1	Var. 2
1	0.1689	0.2299	0.0709	0.0513	0.0644	0.0103
2	0.1965	0.2782	0.0950	0.0773	0.0842	0.0494
3	0.1963	0.2650	0.0979	0.1295	0.0852	0.0673
10	0.2066	0.2803	0.1898	0.2455	0.1794	0.2313
20	0.2101	0.2824	0.2060	0.2856	0.2058	0.2732
30	0.2114	0.2828	0.2103	0.2875	0.2106	0.2813

Table 3 Model NPAR(1) with VAR(3) noise (Series 3)

Series 3	MSPE(a)		MSPE(b)		MSPE(c)	
Lags	Var. 1	Var. 2	Var. 1	Var. 2	Var. 1	Var. 2
1	1.3530	1.0066	1.2277	0.9937	1.2281	0.9895
2	6.4228	6.8469	6.3825	6.8125	6.3962	6.8373
3	12.4008	14.9944	12.5475	15.0671	12.5544	15.0604
10	19.9809	19.8974	20.0418	20.0296	20.0456	20.0320
20	20.5862	20.0169	20.6387	20.1749	20.6455	20.1776
30	20.8909	19.6885	20.9619	19.8530	20.9679	19.8559

Table 4 Pure VECM model (Series 4)

Series 4	MSPE(a)		MSPE(b)		MSPE(c)	
Lags	Var. 1	Var. 2	Var. 1	Var. 2	Var. 1	Var. 2
1	6.9370	8.8101	0.3426	3.0539	0.2699	2.6542
2	16.8852	18.7127	2.9435	5.8296	2.8263	5.1088
3	25.5642	27.4790	7.1160	10.2645	6.9875	9.3126
10	47.1114	49.5927	23.1711	28.1485	21.3344	24.0975
20	51.5780	54.1854	33.4512	41.7827	25.4007	28.2645
30	54.3195	56.9187	46.7582	59.7599	28.0551	30.8976

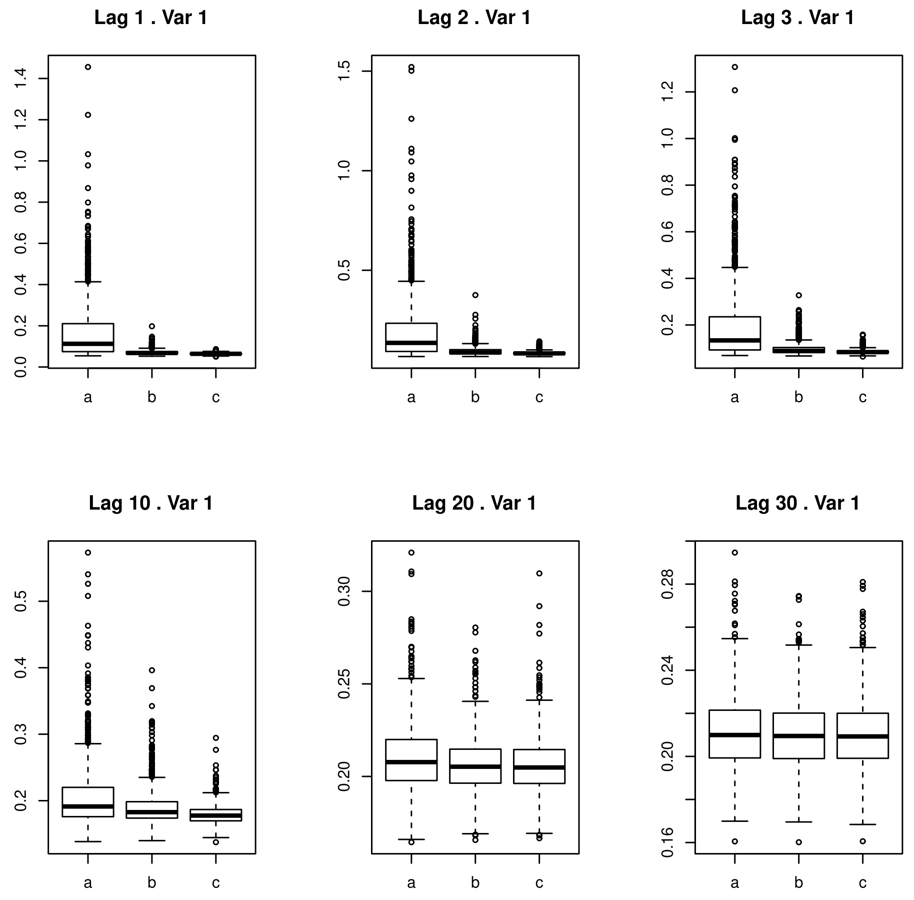

Figure 3 Boxplots of the prediction errors obtained with three predictors for the pure VAR model. (**a**) nonparametric predictor for each component independently, (**b**) semiparametric predictor for each component independently and (**c**) semiparametric predictor with VAR modelling of the residuals

4.2 Real data application

The general model proposed in Sect. 3.1 was implemented for the particular case of the prediction of levels of SO_2 and NO_x in the vicinity of power station and combined cycle.

Let X_t be the bidimensional series formed by the one hour mean series of SO_2 and NO_x at each minute t. In terms of equation (3), we consider $Y_t = X_{t+\kappa}$ and $Z_t = (X_t, X_t - X_{t-5})$. If $\hat{X}_i$ denotes the observed values for past instants ($i \leq t$) and the best prediction for future instants ($i > t$), the aim is to predict X_{t+30} following the next algorithm:

- Every instant t, $\varphi(Z_t)$ is estimated with additive models and the information provided by the historical matrix, independently for each component. The estimate of φ is done at 30 instants ahead: $\dot{Y}_t = \dot{X}_{t+30} = \hat{\varphi}_{30}(Z_t) + \dot{e}_{t+30}$.
- The residuals series $\hat{e}_t$ is computed by $\hat{e}_t = Y_t - \hat{\varphi}_{30}(Z_t)$ and a test of model adequacy is performed (for instance, the Ljung–Box test) for each component of the series concerning the last four hours (240 observations).
- If any of the components of the residuals series is not white noise, a test is performed to explore if the vectorial residual series is cointegrated. If this is the case, an adequate VECM is adjusted. If the series is not cointegrated, a VAR model is fitted.
- Thus $\dot{e}_{t+30}$ is obtained.

- The proposed final prediction given by the *Semiparametric Bidimensional Model* with the nonparametric part estimated at 30 instants (SPBM) is:

$$\dot{X}_{t+30} = \hat{\varphi}_{30}(Z_t) + \dot{e}_{t+30}.$$

To observe the behaviour of the prediction model, we have evaluated its performance on two episodes of air quality alteration, whose information has not been included in the historical matrix.

Figure 4 shows the forecasts given half an hour before by the proposed models for an episode depicted in one of sampling stations. The good behaviour of the forecasts can easily be seen. The proposals estimate quite well the real one hour mean of SO_2 and NO_x values. This is confirmed in Table 5. This table contains three measures of accuracy for the pure nonparametric predictor (NPM) and the proposed semiparametric predictor (SPBM), based on the following criteria:

(a) Squared error: SE = $\sum_t (y_t - \hat{y}_t)^2$.

(b) Absolute error: AE = $|y_t - \hat{y}_t|$.

(c) Relative absolute error (%): RAE = $100|\frac{y_t - \hat{y}_t}{y_t}|$.

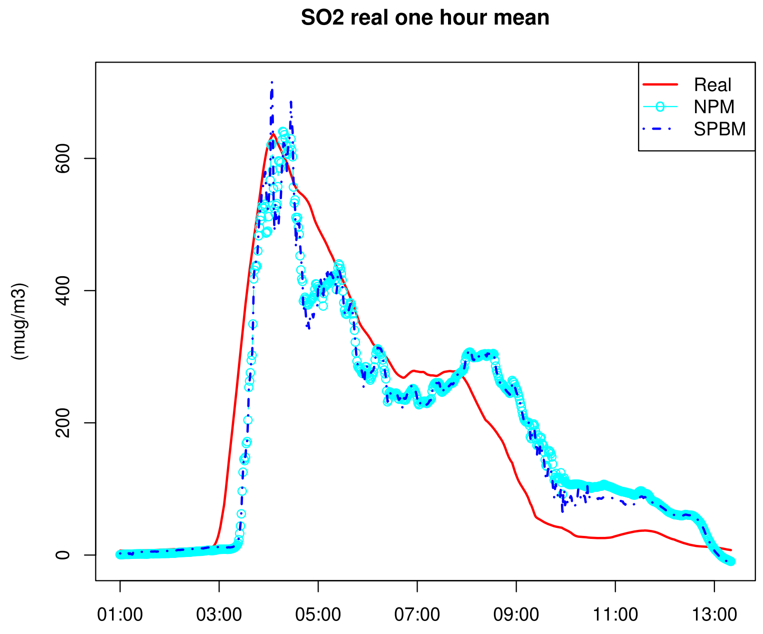

Figure 4 Episode depicted in one of the sampling stations. Predictions given by the bidimensional semiparametric models (NPM and SPBM) for the one hour SO_2 mean

Table 5 SO_2 and NO_x Forecast errors

Model	SO_2						NO_x					
	SE		AE		RAE		SE		AE		RAE	
	M	*Md*	*M*	*Md*	*M*	*Md*	*M*	*Md*	*M*	*Md*	*M*	*Md*
SPBM	1265.28	635.65	27.91	25.21	27.15	18.33	15.83	8.87	3.25	2.97	21.35	9.28
NPM	1043.23	372.02	24.20	19.29	24.18	15.99	30.35	18.28	4.42	4.28	29.77	12.17

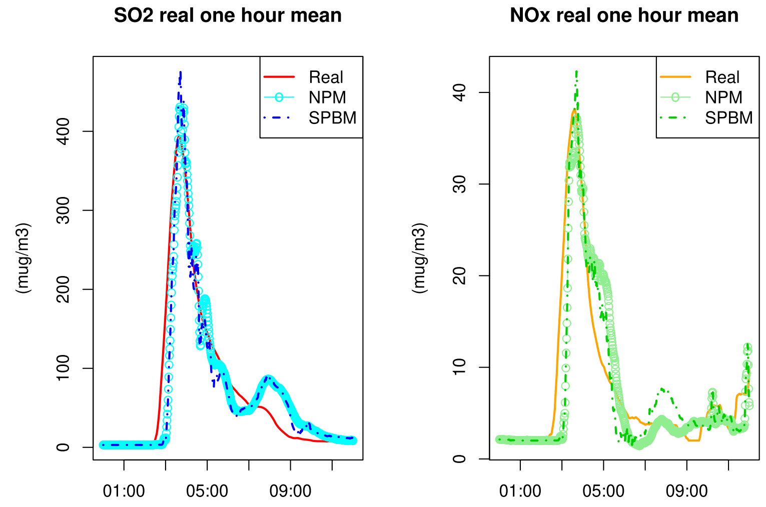

Figure 5 Episode depicted in one of sampling stations. Predictions given by the bidimensional semiparametric models for the one hour SO_2 (left) and NO_x (right) means

Table 6 SO_2 Forecast errors

Model	SO_2					
	SE		AE		RAE	
	M	*Md*	*M*	*Md*	*M*	*Md*
SPBM	5782.18	2566.27	62.40	50.66	38.48	16.31
NPM	5833.85	3055.29	64.17	55.27	43.92	16.39

The mean (M) and the median (Md) of these three measures have been computed for the period covering the pollution incident proper (02.00 to 10.00 hours). The SO_2 nonparametric prediction with the historical matrix captures very well the behaviour of the real series (RAE: 24.18%) while the semiparametric prediction is not able to overcome (RAE: 27.15%). However, the NO_x prediction given by SPBM (RAE: 21.35%) notably improves one obtained by the NPM (RAE: 29.77%). Furthermore, the residuals series was detected as cointegrated 123 times (8.37%), mainly when the episode higher values occur.

In another SO_2 episode depicted in one of the sampling stations (see Fig. 5) the behaviour of the predictors is somewhat different. The SO_2 prediction given by NPM (RAE: 43.92%) does not entirely capture the behaviour of the real series and so, the semiparametric prediction (RAE: 38.48%) can improve that results as shown in Table 6. In this episode, the NO_x values are very low (practically zero) and therefore there are no cointegration relationships.

5 Conclusions

This paper reviews several prediction models that have been implemented along the years for the prediction of SO_2 in the vicinity of a power station. This evolution reflects the adaptation of the statistical models to the change of improved environmental rules and the availability of new technological resources that allows the estimation in more complex situations.

The last part of the paper is devoted to a new proposal that, having in mind the same philosophy applied to the previous univariate models, extends the semiparametric model to the multivariate framework. In particular, the paper deals with the joint prediction of SO_2 and NO_x levels using natural extensions of the model in the univariate framework. These models, originally developed for financial applications, are successfully adapted to the environmental problem showing good results in the simulation studies and in the real data application. The semiparametric joint predictor (SPBM) obtains similar results as the nonparametric (NPM) and the semiparametric independent predictor (SPM) in those scenarios where the components of the response are not related (see Table 1). Recall that predictors NPM and SPM are constructed under this assumption. In the scenarios with dependence among components, predictor SPBM clearly beats its competitors (see Tables 2–4) showing also good results in the real data application.

Acknowledgements
The work by Wenceslao González-Manteiga and Manuel Febrero-Bande was partially supported by projects MTM2013-41383-P and MTM2016-76969-P from the Spanish Ministry of Science and Innovation and European Regional Development Fund and IAP network StUDyS from Belgian Science Policy.

Funding
Endesa Generation S.A. has signed along the last 20 years several contracts with academic departments for the development of an integral program for contamination prevention around the facility. The three authors have been involved in some of those contracts under different roles.

Abbreviations
SIPEI, Integrated System of Statistical Prediction of the Inmision; ARIMA, Autoregressive Integrated Moving Average; VECM, Vector Error Correction Model; NPM, Nonparametric Model; SPM, Semiparametric Model; SPBM, Semiparametric Bidimensional Model; MSPE, Mean Square Prediction Error.

Availability of data and materials
Please contact authors for data requests.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
The three authors are equally contributors to this paper. All authors read and approved the final manuscript.

Author details
[1]MODESTYA group, Technological Institute for Industrial Mathematics (ITMATI), Santiago de Compostela, Spain. [2]Dept. of Statistics, Mathematical Analysis and Optimization, Fac. of Mathematics, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [3]CIBER Epidemiología y Salud Pública, Complexo Hospitalario da Universidade de Santiago, Santiago de Compostela, Spain.

Publisher's Note

Received: 28 March 2018 Accepted: 2 August 2018 Published online: 23 August 2018

References

1. Box G, Jenkins M, Reinsel C. Time series analysis: forecasting and control. New York: Wiley; 2008.
2. Buja A, Hastie T, Tibshirani R. Linear smoothers and additive models. Ann Stat. 1989;17:453–510.
3. Engle RF, Granger CWJ. Co-integration and error correction: representation, estimation and testing. Econometrica. 1987;57:251–76.
4. Fernández de Castro B, González-Manteiga W. Boosting for real and functional samples: an application to an environmental problem. Stoch Environ Res Risk Assess. 2008;22(1):27–37.
5. Fernández de Castro B, Guillas S, González-Manteiga W. Functional samples and bootstrap for predicting sulfur dioxide levels. Technometrics. 2005;47(2):212–22.
6. Fernández de Castro B, Prada-Sánchez J, González-Manteiga W, Febrero-Bande M, Bermúdez Cela J, Hernández Fernández J. Prediction of SO2 levels using neural networks. J Air Waste Manage Assoc. 2003;53(5):532–9.
7. Friedman J, Stuetzle W. Projection pursuit regression. J Am Stat Assoc. 1981;76(376):817–23.
8. García-Jurado I, González-Manteiga W, Prada-Sánchez J, Febrero-Bande M, Cao R. Predicting using Box–Jenkins, nonparametric, and bootstrap techniques. Technometrics. 1995;37(3):303–10.
9. Granger C. Co-integrated variables and error-correcting models. PhD thesis, Discussion Paper 83-13. Department of Economics, University of California at San Diego; 1983.
10. Hamilton JD. Time series analysis. vol. 2. Princeton: Princeton University Press; 1994.
11. Johansen S. Statistical analysis of cointegration vectors. J Econ Dyn Control. 1988;12(2):231–54.

12. Mammen E, Linton O, Nielsen J. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. Ann Stat. 1999;27(5):1443–90.
13. Nadaraya EA. On estimating regression. Theory Probab Appl. 1964;9(1):141–2.
14. Prada-Sánchez J, Febrero-Bande M. Parametric, non-parametric and mixed approaches to prediction of sparsely distributed pollution incidents: a case study. J Chemom. 1997;11(1):13–32.
15. Prada-Sánchez J, Febrero-Bande M, Cotos-Yáñez T, González-Manteiga W, Bermúdez-Cela J, Lucas-Domínguez T. Prediction of SO2 pollution incidents near a power station using partially linear models and an historical matrix of predictor-response vectors. Environmetrics. 2000;11(2):209–25.
16. Roca-Pardiñas J, Cadarso-Suárez C, González-Manteiga W. Testing for interactions in generalized additive models: application to SO2 pollution data. Stat Comput. 2005;15(4):289–99.
17. Roca-Pardiñas J, González-Manteiga W, Febrero-Bande M, Prada-Sánchez J, Cadarso-Suárez C. Predicting binary time series of SO2 using generalized additive models with unknown link function. Environmetrics. 2004;15(7):729–42.
18. Speckman P. Kernel smoothing in partial linear models. J R Stat Soc, Ser B, Stat Methodol. 1988;50:413–36.
19. Watson GS. Smooth regression analysis. Sankhya, Ser A. 1964;26:359–72.