Atmospheric Environment 95 (2014) 151–157

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Predicting SO_2 pollution incidents by means of additive models with optimum variable selection

Marta Sestelo [a, b], Javier Roca-Pardiñas [b], Celestino Ordóñez [c, *]

[a] Department of Mathematics, Autonomous University of Barcelona, 08193 Cerdanyola del Vallès, Spain
[b] Department of Statistics and Operations Research, University of Vigo, 36208 Vigo, Spain
[c] Department of Mining Exploitation and Prospecting, University of Oviedo, 33600 Oviedo, Spain

HIGHLIGHTS

- A mathematical for the detection of SO_2 emission episodes was developed.
- A generalized additive model and an algorithm for variable selection were used.
- SO_2 concentrations and meteorological variables were considered.
- The best prediction is reached with only two terms of the time series.
- Meteorological variables were found not to be significant covariates.

ARTICLE INFO

Article history:
Received 5 November 2013
Received in revised form
9 May 2014
Accepted 11 June 2014
Available online 12 June 2014

Keywords:
Variable selection
Bootstrap
Additive models
Nonparametric regression
Pollution incident

ABSTRACT

The aim of this paper is to predict time series of SO_2 concentrations emitted by coal-fired power stations in order to estimate in advance emission episodes and analyze the influence of some meteorological variables in the prediction. An emission episode is said to occur when the series of bi-hourly means of SO_2 is greater than a specific level. For coal-fired power stations it is essential to predict emission episodes sufficiently in advance so appropriate preventive measures can be taken. We proposed a methodology to predict SO_2 emission episodes based on using an additive model and an algorithm for variable selection. The methodology was applied to the estimation of SO_2 emissions registered in sampling locations near a coal-fired power station located in Northern Spain. The results obtained indicate a good performance of the model considering only two terms of the time series and that the inclusion of the meteorological variables in the model is not significant.

1. Introduction

The combustion of fuel oil or coal releases sulphur dioxide into the atmosphere with negative consequences for the environment. Current Spanish legislation governing the control of environmental pollution in the vicinity of potential point sources of pollution such as coal-fired power stations places a limit on the mean of 24 successive determinations of SO_2 concentration taken at 5 min intervals. An emission episode is said to occur when the series of bi-hourly means of SO_2 is greater than a specific level v. In this framework, it will be in the best interest for the plant, both economically and well as environmentally, to be able to predict when the legal limit will be exceeded so there will be sufficient time for effective countermeasures to be taken.

In previous works (García-Jurado et al., 1995; Prada-Sánchez and Febrero-Bande, 1997; Prada-Sánchez et al., 2000; Roca-Pardiñas et al., 2004), semiparametric, partially linear models and generalized additive models with unknown link function were applied to the prediction of atmospheric SO_2 pollution incidents in the vicinity of a coal/oil-fired power station. Neural networks were used to solve the same problem (Fernández de Castro et al., 2003) and the results compared with those obtained with a semiparametric model. Sabah and Saleh (2008) also used neural networks to predict sulphur dioxide concentrations from a refinery in Oman including five meteorological parameters that were expected to affect SO_2 concentrations. An adaptative Linear Neural Network and a Generalized Regression Neural Network were applied to predict the SO_2 concentrations in the city of Salamanca (México); results showed that the linear regression neural network improves the estimation (Cortina et al., 2008). Nunnari et al. (2004) compared the estimation of SO_2 critical episodes obtained using neural

* Corresponding author.

http://dx.doi.org/10.1016/j.atmosenv.2014.06.025
1352-2310/

networks, fuzzy logic and generalized additive models and linear time-series models; the results showed that none of these models produce a better prediction compared with others in terms of the performed indices that were considered.

In this paper we propose a new approach to this problem, where we also try to predict a new emission episode, but focussing our attention on the importance of finding the best combinations of variables — both emission and meteorological variables — to obtain the best prediction. The developed methodology includes the combination of two topics. On the one hand, the selection of the best combinations of q variables by stepwise-based selection procedure, and on the other hand, the determination of the number of covariates to be included in the model based on bootstrap resampling techniques. The combination of these two topics leads to obtain the "best" model, the model with the best prediction capacity.

The paper is structured as follows: Section 2 describes the possible variables for the prediction purpose and the proposed methodology for their selection. Section 3 shows the results obtained in the prediction of a real pollution incident, and finally, Section 4 presents the conclusions.

2. Materials and methods

As we said in the Introduction, the propose of this paper is to predict SO_2 pollution incidents by means of additive models. Because of the precise nature of the legal limits on pollution in Spain, the time series of interest were series of "two-hour mean" SO_2 values. Denoting X_t as the value obtained by this series in the instant t (5-min temporal instants) and being k the prediction horizon, the interest is to predict $Y_t = X_{t+k}$ from a vector of predictive covariates $\mathbf{X}_t$ whose information will be recorded before the instant t. For instance, our prediction might be $\widehat{Y}_t = \widehat{m}(\mathbf{X}_t)$, where $\widehat{m}$ is an estimate of the regression function $m(\mathbf{X}_t) = E[Y_t|\mathbf{X}_t]$.

2.1. Initial selection of explicative variables

In some common situations, $\mathbf{X}_t$ can be obtained from some transformed vector of X_t, X_{t-1}, X_{t-2}, etc. For instance, in Prada-Sánchez et al. (2000) the use of the covariates $\mathbf{X}_t = (X_t, X_{t-3})$ was proposed. Another option is based on adding to the model a vector of exogenous meteorological explicative variables. In this context, following the ideas of Prada-Sánchez et al. (2000), a vector of $p = 14$ covariates was considered

$$\mathbf{X}_t = \left(X_t, X_{t-3}, T_t^{80-10}, \ldots, E_{t-12}\right)$$

The explanation of these $p = 14$ covariates is as follows:

1) X_t: SO_2 emission in the actual instant t.
2) X_{t-3}: SO_2 emission in the $t - 3$ instant (15 min before).
3) T_t^{8010}: Temperature gradient between 80 and 10 m of altitude in t.
4) T_{t-12}^{8010}: Temperature gradient between 80 and 10 m of altitude in the $t - 12$ instant (60 min before).
5) T_t^{10}: Temperature at 10 m of altitude in t.
6) T_{t-12}^{10}: Temperature at 10 m of altitude in $t - 12$.
7) V_t^{80}: Wind speed at 80 m of altitude in t.
8) V_{t-12}^{80}: Wind speed at 80 m of altitude in $t - 12$.
9) R_t: Solar Radiation in t.
10) R_{t-12}: Solar Radiation in $t - 12$.
11) H_t: Humidity in the actual instant t.
12) H_{t-12}: Humidity in the $t - 12$.
13) E_t: SO_2 emission of the power station chimney in t.
14) E_{t-12}: SO_2 emission of the power station chimney in $t - 12$.

The data used here were obtained from 9 control stations sited at distances from the coal station between 5 and 20 km. All the above variables were measured in the control stations except E which was measured in the power station.

2.2. Sample

The nature of the SO_2 time series makes prediction especially challenging. Most of the time X_l displays values close to zero. Sometimes this situation is interrupted at random intervals which can range from a few days to several weeks. When this happens the values of X_l rise to high levels and then fall back to zero. In order for predictions to be based on data representing a reasonably large number of incidents, one could take as its samples $\{(\mathbf{X}_i,Y_i)\}_t$, the 2000 rows of a historical matrix $\mathbf{M}_t$ that was constructed and updated as follows: First, we determined the range of 2-hourly means observed over the course of the previous 2 years. We then divided the non-zero region of this range into 20 strata containing an approximately equal number of observations, randomly selected for 100 values $Y_i = X_{i+k}$ for each stratum, and each associated with the corresponding predictor values of the $p = 14$ covariates in $\mathbf{X}_i$. The 2000 entries so formed made up the 2000 rows of $\mathbf{M}_0$, the seed of the historical matrix. Thereafter, during on-line processing, the historical matrix was updated by identifying the stratum to which X_t belonged and then substituting $(\mathbf{X}_{t-k}, X_t)$ for the oldest row of M_{t-1}. In this way, given the present instant t, there are 2000 available entries in the historical matrix $\mathbf{M}_t = \{(\mathbf{X}_i, Y_i)\}_{i=1}^{2000}$. In this study, due to problem arisen in the sampling, the used matrix had 1600 entries.

2.3. Predictive models

Let Y be the response variable, and $\mathbf{X} = X_1,\ldots,X_p$ be the p-vector of associated covariates and denoting $m(\mathbf{X})=E[Y|\mathbf{X}]$, the well known multivariate linear regression model (LM) takes the form:

$$m(\mathbf{X}) = \alpha_0 + \alpha_1 \cdot X_1 + \cdots + \alpha_p \cdot X_p$$

where $(\alpha_0,\alpha_1,\ldots,\alpha_p)$ is a vector of coefficients. In some instances, LMs can be very restrictive, since they assume linearity in the covariates. Consequently, if the parametric model fails then the conclusions will be erroneous. This constraint can be avoided by replacing the linear index $\alpha_0 + \alpha_1 \cdot X_1 + \cdots + \alpha_p \cdot X_p$ with a nonparametric structure. Accordingly, here we will concentrate on the additive model (AM, Hastie and Tibshirani, 1990), which is a generalization of the LM, by introducing one-dimensional, nonparametric functions instead of linear components. Specially, AMs express the conditional mean as

$$m(\mathbf{X}) = \alpha + m_1(X_1) + m_2(X_2) + \cdots + m_p(X_p) \quad (1)$$

where a is a constant and m_j is the unknown smooth partial function or effect curve associated to each continuous covariate X_j. Note that identification is guaranteed by introducing a constant α into the model and requiring a zero mean for the partial functions. The AM is widely used as an extension of the traditional LMs specially when continuous covariates are present. The AM is more flexible than the LM, since the researcher does not assume a parametric form for the effects of the continuous covariates, but only assumes that these effects may be represented by arbitrary unknown smooth functions. The AMs are easy to interpret, because the additive components simply describe the influence of each