Predicting Gas Emissions in a Cement Kiln Plant using Hard and Soft Modeling Strategies

Dulce Gabriel\textsuperscript{1,2}, Tiago Matias\textsuperscript{1}, J. Costa Pereira\textsuperscript{2} and Rui Araújo\textsuperscript{1}
\textsuperscript{1}Institute of Systems and Robotics (ISR-UC), and Department of Electrical and Computer Engineering (DEEC-UC), University of Coimbra, Pólo II, PT-3030-290 Coimbra, Portugal.
\textsuperscript{2}Department of Chemistry (DQ-UC), University of Coimbra, PT-3004-535 Coimbra, Portugal.
\texttt{dracag@isr.uc.pt}, \texttt{tmatias@isr.uc.pt}, \texttt{jcpereira@qui.uc.pt}, \texttt{rui@isr.uc.pt}

Abstract

In this work, two alternative methodologies for modeling and predicting gas emissions of NO, NO\textsubscript{2} and SO\textsubscript{2} are presented. The first method involves hard modeling strategies with Parsimonious Multivariate Least Squares (PMLS) assuming simple polynomial functions as base model. The second is a soft modeling approach using Extreme Learning Machine (ELM). In this work we found that both methods have similar capabilities for data description, providing an in depth analysis of the system under study. Results also reveal further insights in predicting gas emissions and enlighten on which of the factors can be useful for prediction, and consequently for system characterization and emission abatement.

1 Introduction

The production of cement is a complex process composed of several procedures, which require considerable technological support, and that implies several environmental impacts [11]. The first approach to overcome environmental problems was based in the development of protecting mechanisms in order to minimize its prejudicial effects. Now, a greater effort is being made in understanding the industrial process in order to improve its efficiency and identify causes on which it may be possible to actuate in order to minimize the formation of these environmental harmful agents.

During the cement production process, there are large amounts of emissions of particulate matter, nitrogen oxides (NO\textsubscript{x}), mainly in NO and NO\textsubscript{2} forms, sulfur dioxide (SO\textsubscript{2}), carbon dioxide and carbon monoxide. Other trace quantities of volatile organic compounds, acid gases, trace metals, and organic micro pollutants may also be emitted [6]. Due to the volume and environmental effects, the focus is given to NO\textsubscript{x} and SO\textsubscript{2} emissions.

NO\textsubscript{x} represents highly reactive compounds containing nitrogen and oxygen in varying amounts. In atmospheric chemistry these oxides refer specifically to NO (nitric oxide) and NO\textsubscript{2} (nitrogen dioxide) since they are the most abundant. Industrial NO\textsubscript{x} formation depends on (i) the amount of nitrogen present in the fuel, (ii) the combustion temperature, and (iii) the stoichiometric conditions for the combustion reaction [3]. However, many parameters that affect the conditions of combustion and temperature have to be considered. Similarly, most representative species present in the emission of SO\textsubscript{x} are SO\textsubscript{2} and SO\textsubscript{3}, due to their relative stability.

For economical purposes and also for waste disposal and reuse, industrial kilns are progressively using hazardous wastes and thus, subject to greater degrees of regulatory scrutiny and more rigorous emission control requirements. Reduction of environmental manufacturing footprints by restrictions on pollutant emissions is one of the measures proposed by the European Union (EU) commission. Europe focuses on steadily reducing allowed emissions, increasing the penalty risks and mitigation expenses for the companies.

The aim of this article is to identify the most relevant variables that affect the NO, NO\textsubscript{2} and SO\textsubscript{2} gas emissions in a cement kiln plant and in this way to be able to describe emissions levels in order to predict and control the amounts of this gaseous major pollutants emissions. Emission predictions are useful variables used in the relevant process control systems. This assures the control of the pollutants output on the stack within the legal limits by optimally and automatically selecting, monitoring, and controlling the most relevant parameters of the cement production process, always taking in consideration the non-violation of regulations. Emission estimations enable to make adjustments to the process more frequently, and facilitate to stabilize the plant, optimize the production, and manage and correct process disruptions - all to ensure optimum plant performance.

The paper is organized as follow. In Section 2 the selection of the relevant variables is described. Section 3 presents the strategy used for data set manipulation. The Case Study is described in Section 4. Section 5 presents and discusses the results. Section 6 presents concluding remarks.
2 Selection of the Relevant Variables

The selection of the relevant process variables is crucial for the reduction of the NO\textsubscript{x} and SO\textsubscript{2} emissions. However, as the cement production process is a sequential and sometimes simultaneous complex combination of events, physical operations such as martial feed rate, fuel and air flow, kiln speed as well as physiochemical reactions, several problems have to be solved in order to determine the most relevant variables [11].

The first major problem in dealing with this highly multivariate system is the large number of variables available in the process, most of them having nonlinear and positive/negative correlations with the emission variables.

The second problem is the time delay between the sequential steps of the process: to model the gas emissions at the end of the process there is a variable time gap to each previous step that has to be accounted and estimated to maximize the ability to describe the output response. Another difficulty is that the industrial process usually does not have constant feeding and out coming rates, but these rates generally oscillate between extreme values.

The third is the process buffering ability – although the process is industrially confined, each physiochemical event evolves in time and space dimensions, subjected to interference and diffusional dispersion, which are responsible for natural smoothing and residual memory-effect latency (sharp time events tend to become suppressed by bulky events). This is a huge problem when dealing with modeling gas emissions at the end of a complex industrial process.

Time series analysis may be used to estimate these delays [1] but since this analysis uses one-to-one correlations, this may not be a good option since the combination of several hill correlated process details may also have a drastic impact in the result.

In our point of view, in order to avoid some of these problems when dealing with a continuous process, two extreme approaches may be used. The first one is based on mutual maximal information [12] and on the determination of appropriate time gaps to compensate “out-of-phase” situations. The second suggestion is to mimic the system with moving window averaging methods in order to include “memory” into the variables and somehow compensate delayed responses and system buffering.

3 Strategy

In order to predict gas emissions in a cement kiln plant, after a standard preprocessing step, two approaches were used to retrieve system information and predict emissions based on hard and soft modeling procedures.

3.1 Pre-processing

Recorded variables usually present different unities, scales and ranges. This is an undesirable situation for modeling purposes. To circumvent these heterogeneities an initial transformation of sample \( j \) of variable \( i \), \( \chi_{ji} \), is performed for all \( i, j \), using the transform

\[
x_{ji} = \frac{2\chi_{ji} - \chi_{\max} - \chi_{\min}}{\chi_{\max} - \chi_{\min}},
\]

where \( x_{ji} \) is the value of \( \chi_{ji} \) scaled to the [-1,1] interval, and \( \chi_{\max} \) and \( \chi_{\min} \) are maximum and minimum values of the \( i \)-th variable, respectively.

After these variables scaling, process variables were classified into \( n \) causes (independent variables) and \( l \) responses (dependent variables) and separated into two subspaces: the predictors sub-space, \( \chi_{(n \times m)} \), and response, \( y_{(n \times 1)} \). Each response is treated separately and similarly.

3.2 Modeling

An industrial system may be viewed as a “black-box” multivariate system where there is a huge set of input variables (controlled and uncontrolled process conditions) and output variables (responses or dependent system outcomes related to some inputs).

Retrieving information from such multivariate systems is a computational challenge since they are a) intrinsically multivariate, b) have non-linear relations, c) present simultaneous interactions (in-phase correlated variables), and d) are subject to propagation (delays in time and dispersion in space), and e) with subsequent further interactions (out-of-phase correlations).

Response modeling is usually the best way to obtain functional information and knowledge of the process. Excluding the possibility to develop a fully physicochemical-based model able to accurately describe all the multivariate system, this modeling usually involves an adequate functional development process in order to able to describe experimental values and predict results for newer situations.

In this work, two different multivariate approaches were used: the first is based on hard modeling, and the second is related to soft modeling.

3.2.1 Hard Modeling Approach

Hard modeling approach involves the previous stipulation of a explicit model and relies on the use of standard least squares (LS) method to estimate parameters that allows an adequate model data fit of dependent variables [5].

A polynomial approach [13] is usually a good solution when dealing with unknown black-box complex systems since a) the estimation equation is linear in the parameters (parameters may be directly estimated with no need of optimization algorithms), b) it is statistically supported (error estimates may also be accurately estimated), allowing a c) flexible adaptive approach (further additional terms may be included) with d) any desired complexity - polynomial dependence may account for a specific variable influence (first order predictor or other degree dependence) and mutual interaction (variables internal product).

Assuming an \( m \) dimensional predictor space, \( \chi_{(n \times m)} \), the most simple polynomial model may only be a first degree dependence with \( p = m + 1 \) parameters like

\[
y_j = \sum_{i=1}^{m} b_i x_{ji},
\]

where \( y_j \) is the value predicted by the model for the \( j \)-th exemplar, \( b_0 \) stands for a constant and \( b_i \) represent the sensitivity coefficients for the predictors (input variables), \( x_{ji} \), reflecting the parameter dependency.
This simpler situation may be further developed in order to overcome some of modeling difficulties like correlated variables (cooperative/destructive interferences) in a $p = m(m + 1)/2$ parameter model

$$y_i = b_0 + \sum_{t=1}^{m} b_i x_{j1} + \sum_{t=1}^{m-1} \sum_{k=m+1}^{m} b_{ik} x_{j1} x_{jk}, \quad (3)$$

where $b_{ik}$ accounts for the interaction between variables $x_{j1}$ and $x_{jk}$.

Rewriting (2) into matrix form, and considering modeling error, yields

$$X b = y^d + e, \quad (4)$$

where $X_{(n \times p)}$, $y^d_{(n \times 1)}$, and $y_{(n \times 1)}$ are the predictors matrix, response vector, and predicted response vector, respectively. $e = y^d - y$, and $b_{(p \times 1)}$ represent model lack-of-fit (error), and estimated parameter vector, respectively. In order to minimize the fitting error $e$ in the LS sense, and considering that all columns of $X$ are linearly independent. $b$ can be estimated using the Moore-Penrose pseudoinverse $X^+$. So

$$\hat{b} = X^+ y^d = (X^T X)^{-1} X^T y^d = C X^T . y^d, \quad (5)$$

where $X^T_{(p \times n)}$ is the predictors transposed matrix, and $C = C_{(p \times p)}$.

In the LS standard approach (unweighted LS), $C_{(p \times p)}$ in conjunction with the model residual standard error, may be used to estimate the parameter standard errors (parameter covariance matrix) and thus evaluate the respective uncertainty [13]. This allows the use of a statistical approach to assess the most significant variables, assuming a linear first degree dependency.

In eq. (3), there is an infinity of parameter choices with polynomial modeling, and over parameterization may be frequently a problem - in general, an excellent fit is not necessarily the best way to approach the problem since it usually conduces to over complicated models with unpredictable behaviors. Several approaches were developed in order to circumvent this difficulty. A good choice is to conceive an unsupervised and statistically based process.

Parsimonious Multivariate Least Squares (PMLS) is an unsupervised method statistically based on a standard least squares fitting strategy. The general principle of parsimonious data modeling states that if two models in some way adequately model a given set of data, the one that is described by a fewer number of parameters will have better predictive ability given new data. As first approach, data are previously over-fitted with an adequate complete model using excessive $p$ parameters. This model is successively simplified, stripping one parameter at a time (the less significant parameter), leading to a $(p-1)$ parameter model. This iterative process may be interrupted based on some criteria (lack-of-fit, residual error level, etc) or allowed to stop at the last iterative cycle with $p = 1$. There is a need to evaluate the impact of parameter removal from the initial previous model comparing the fitting error evolution in comparison with a pure error estimate [2].

This iterative method starts, at each implemented modeling situation, with $p$ initial parameters. After parameter estimation, the least significant parameter is removed from the initial model and a model with $(p-1)$ parameters is defined for the next iteration step. The process is allowed to iteratively run until a last parameter is retained. Compiling the iterative fitting information it is possible to evidence the most relevant models with a parsimony test [2]. These relevant models allow to access the most relevant variables used to describe a specific response and thus to predict new situations.

### 3.2.2 Soft Modeling Approach

In the soft modeling approach the selection of the most relevant variables will be seen as an input variable selection procedure in a neural networks (NN) model. In a NN model that tries to describe a target response with a large number of input variables, there may exist some redundant or irrelevant variables that slow down the learning process and cause the over-fitting of the model [14].

For each separately considered target output variable (NO, NO$_2$, and SO$_2$), a sequential backward search (SBS) approach [10] will be used for the selection of the input variables. However, instead of using error backpropagation, an extreme learning machine (ELM) will be employed for learning a single hidden-layer feedforward NN model with sigmoidal and linear activation functions in the hidden-layer and output layer, respectively. The ELM learning algorithm was proposed in [7] and was inspired in the concept of the random weights NN [8]. In ELM the weights and bias of the connections of the first layer are randomly assigned and the output weights are obtained by least squares. In [7], it was shown that the learning of the NN by ELM is much faster than conventional learning algorithms and that a good generalization performance can be obtained.

In the SBS methodology, the objective is the successive removal of the redundant/irrelevant variables, retaining as model inputs the variables which are more relevant to the desired output. In the first step, the performance of the NN model with the whole set of input variables is determined. Then, the performances of each input variable are computed and the input variable with lowest saliency is removed. Then, the performance of the NN model without this input variable is determined, and the loop starts again until a certain criterion is satisfied. The concept of saliency is a concept that expresses the importance of an input variable. An input variable with a large saliency is more important than an input variable with a small saliency.

The performance of the NN model is measured using the mean of the estimation error in a 10-fold cross-validation procedure: the data set is divided into 10 subsets and 10 iterations are performed: in each iteration the ELM is trained using nine data subsets (training subsets), and the mean square error (MSE) between the estimated output and the desired output in the remaining data subset (validation subset) is obtained. The performance will be the mean of the MSE in the ten different validation subsets. The saliency of each variable is measured using the performance of the NN model when such input variable is removed.

However, as previously mentioned, in large industrial processes, input variables do not influence the output instantly, but after some time delay that depends on the in-
put variable. So a delay selection procedure is needed in order to determine the time that each input takes to influence the target output variable. The procedure used in this approach is the maximization of the mutual information between each input variable and the target variable. The delay of each input variable $i$ is given by:

$$d_i = \arg \max_{k \in D} I(x_{(j-k)}, y^d), \quad i = 1, \ldots, m, \quad (6)$$

where $x_{(j-k)}$ is the $i$-th normalized input variable delayed by $k \in D \subset \mathbb{N}$ data samples, $y^d$ is the target/desired output variable, and $I$ is the mutual information given by [11]:

$$I(x_{(j-k)}, y^d) = H(x_{(j-k)}) + H(y^d) - H(x_{(j-k)}, y^d), \quad (7)$$

where $H(x_{(j-k)})$ and $H(y^d)$ are the entropy of variables $x_{(j-k)}$ and $y^d$, and $H(x_{(j-k)}, y^d)$ is the joint entropy of $x_{(j-k)}$ and $y^d$.

### 3.3 Performance Indicators

In order to evaluate the ability of model description and response prediction some performance were used. Goodness-of-fit was evaluated using the Pearson’s multivariate correlation coefficient, $CC$, and model error estimates were evaluated via a parametric ($\% RMSE$) and a non-parametric estimator ($\% AD$).

The multivariate Pearson’s correlation coefficient, $CC(y, y^d)$, estimated the interrelation between the real response, $y^d$, and its model estimate, $y$, according to:

$$CC(y, y^d) = \frac{\sigma_{y,y^d}^2}{\sqrt{\sigma_y^2 \times \sigma_{y^d}^2}}, \quad (8)$$

where $\sigma_{y,y^d}^2$ represents the covariance between the variable and its estimate, and $\sigma_y^2$ and $\sigma_{y^d}^2$ represent the respective variances.

Model bias parametric estimate was evaluated in terms of normalized root mean squared error:

$$\% RMSE = 100 \times \frac{RMSE}{\Delta}, \quad (9)$$

where $\Delta$ is the response amplitude and $RMSE$ represents the residual mean squared error,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y^d_i)^2}{n - p}}, \quad (10)$$

evaluated over $n$ data points using a parameter model estimates, $y$.

Non-parametric model bias was estimated in terms of percentiles of relative absolute deviation, $\% AD$:

$$\% AD\alpha = 100 \times \frac{P\alpha(|y - y^d|)}{\Delta}, \quad (11)$$

where $P\alpha$ represents the $\alpha$ percentile of the module of the estimation error. Percentiles 2.5, 50.0 and 97.5 were used in order to define central estimate (median, $P_{50}$) and central confidence interval ($P_{2.5}$ to $P_{97.5}$).

Two strategies were developed and used to organize the tests. In the first strategy, all the available data was fitted simultaneously. In the second strategy, to evaluate the model predicting ability, the data set was divided into 4 quartiles and the model was fitted, at each time, to a specific quartile and used to estimate the values in the other sections. In both strategies, the same $CC$, $\% RMSE$, and $\% AD$ performance indicators were used to evaluate the model prediction performance.

### 4 Case Study

In this work, a data set\(^1\) from a cement kiln plant was used. The database has multiple process measurements in the system, regarding temperatures, pressures, concentrations, manual and laboratorial entries, and several other variables of the cement production process, from the preheater (cyclone) tower until the chimney and cement mill [9].

The experimental data refers to operation of an industrial kiln process of a cement plant. The data refers to 194 monitored variables, recorded with a sampling interval of $T = 1$ [min], during over a month period leading to 43469 entries. This huge data set contains most process relevant information related with the actual laboring imposed / controlled conditions (system inputs or independent variables, here named as predictors, and processing results (process outcomes or responses)). In the procedure for the selection of the variables delays, the considered maximum sample delay is 360 samples (6 hours), so that the set of possible delays is $D = \{ k : k = 0, 1, \ldots, 360 \}$.

A first step was to check among these 194 accessed variables which contain pertinent process information, and to separate predictors from responses. The separation was performed based on knowledge about the process. Thus, it was possible to downsize the initial system from 194 to 108 variables, where 79 are predictors ($x_i$), and 28 are responses including the NO, NO$_2$ and SO$_2$ gas emissions that are studied in this paper.

As the data was captured from an industrial environment, failure in measurement, in hardware, or in transmission can cause the presence of outliers in the data. For the outliers identification process it was considered that an outlier is a sample that was not in the confidence interval of $P_{01}$ to $P_{99}$, where $P_\alpha$ represents the percentile $\alpha$. Each identified outlier was replaced by the average of the 10 most recent non-outlier values of the corresponding variable.

### 5 Results

#### 5.1 Preliminary Correlation Diagnosis

As a first approach, a multivariate correlation matrix was computed in order to evidence the most relevant “in-phase” (zero delay) variable dependencies upon the measured NO, NO$_2$ and SO$_2$ gas emissions. Such most relevant variables are (Table 1): temperature of the coal ($x_{26}$), pressure at the end of the coal cyclone ($x_{32}$), temperature in the coal cyclone ($x_{35}$), injection of O$_2$ at the kiln entrance ($x_{39}$), pressure of the burnt air ($x_{43}$), kiln motor intensity ($x_{46}$), temperature of the burnt of alternative fuels

\(^1\)Provided by “Acontrol - Automação e Controle Industrial, Lda”, Coimbra, Portugal.
Table 1: Highest direct “in-phase” Pearson’s correlation coefficients (CC) values for NO, NO$_2$ and SO$_2$ responses concerning to the 79 most relevant monitored predicting variables ($|CC| \geq 0.20$).

<table>
<thead>
<tr>
<th>CC</th>
<th>$x_{36}$</th>
<th>$x_{32}$</th>
<th>$x_{35}$</th>
<th>$x_{39}$</th>
<th>$x_{43}$</th>
<th>$x_{46}$</th>
<th>$x_{52}$</th>
<th>$x_{54}$</th>
<th>$x_{62}$</th>
<th>$x_{63}$</th>
<th>$x_{77}$</th>
<th>$x_{78}$</th>
<th>$x_{79}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>-0.230</td>
<td>0.529</td>
<td>0.266</td>
<td>0.244</td>
<td>-0.230</td>
<td>-0.238</td>
<td>-0.238</td>
<td>-0.230</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO$_2$</td>
<td>-0.301</td>
<td>0.511</td>
<td>0.219</td>
<td>0.274</td>
<td>0.379</td>
<td>-0.341</td>
<td>-0.271</td>
<td>-0.338</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SO$_2$</td>
<td>-0.269</td>
<td>0.258</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

($x_{52}$), motor intensity ($x_{54}$), cooler motor intensity ($x_{62}$), pressure in the filter ($x_{64}$), grill velocity ($x_{77}$), ventilator velocity ($x_{78}$), and $x_{79}$. Table 1 evidences a very poor relation between “in-phase” variables which can be explained as a result of “out-of-phase” (delayed responses) and also due to the gas dispersion along the cement kiln industrial process path. Consequently, a PCA technique would be unable to perform a drastic reduction in dimensionality of this multivariate problem.

A further effort was attempted in order to retrieve “out-of-phase” correlations using the classical time series approach [4]. Focusing on the evidenced predictors (Table 1), and running until a 24 [h] delay (excessive time gap). No significantly higher correlations were found with this effort, indicating that a) it is difficult to define specific delays for each predictor-response relation, and b) these responses are intrinsically complex and do not meet a direct 1-to-1 relation, evaluated with Pearson’s correlation coefficients.

5.2 Selection of Input Variables

In this section it is shown the variables most related to the emission variables using the two modeling approaches - PMLS (hard modeling) and ELM (soft modeling).

5.2.1 PMLS

Considering exclusively the most simple polynomial regressor (2), where first degree relations and a constant term are used, a model with $p = 80$ parameters is obtained. In this case, the description of responses was quite good taking into account the acquired correlation coefficients of the obtained parsimonious models: 0.812, 0.705, and 0.770 of the models for NO, SO$_2$ and NO$_2$, respectively. However, after variable selection, these models require a large number of parameters ($p = 70$, 57, and 65) to implement the respective responses. Comparing to the results of Sec. 5.1 ($p = 1$), these results reinforce that the responses jointly depend upon several simultaneously considered predictors ($p \gg 1$).

Better fitting results could be obtained if the interaction terms (cross products, $x_{ij}x_{kj}$) were included into the polynomial model (3). However, the number of initial parameters will increase from $(1+79)$ to $1+(79+79^2)/2 = 3161$ which, taking into account data matrix sizes, results in a useless computational challenge. So this approach was discarded.

In order to maintain a simple model and also to circumvent delayed and dispersed responses an integral approach was adopted in conjunction with the first order regressor (2). Selecting an appropriate constant window size, $h$, and averaging past results over a moving window in order to compute the actual values, it is possible to include into variables the respective past information. The goal for using this “memory-effect” is to compensate for unsynchronized and dispersed events.

Table 2 presents the effect of the window size, $h$, on the number of parameters, $p$, correlation coefficient, CC, and relative residual error ($\%RMSE$, $\%AD_{50}$ and $\%AD$) of the parsimonious model. It is notorious that the increase in the window size, conduces to a general reduction in the number of required parameters, and in the relative model error ($\%RMSE$, $\%AD_{50}$ and $\%AD$), and to an improvement on the correlation coefficient (CC).

In this study it is possible to see that some variables are consistently used to describe the responses. There are some core variables which are selected in all tested situations with different $h$ or in more than 80% cases, while other variables are only present in rare situations or are inconsistently present into models (diffuse variables). In our point of view, these “core variables” reveal the real dependency between response and process conditions while those less relevant are essentially used to ensure a better fitting purpose.

For NO emissions there are a total of 19 core variables - 5 are always present (variables representative of pressures on the cyclone tower, temperature on the coal cyclone, NO$_2$ at the kiln input and fan flow) and 14 present over 80% cases. These core and most selected variables are related to gas and fuel admission, input and output gas pressure, kiln temperature and other process relevant conditions in cement kiln production. Considering SO$_2$ emissions there are a total of 23 core variables - 10 are always present (feed to the cyclone tower, pressure, temperature and CO on the cyclone tower, temperature from the coal, pressure, temperature and O$_2$ of the coal cyclone, pressure in the kiln input and pressure from the alternative fuels) and 13 present over 80% of the cases. With respect to NO$_2$ emissions there are a total of 20 core variables - 7 always present (similar to NO variables except on O$_2$ on the kiln input and pressure from the kiln to the cyclone) and 13 present over 80% of the cases.

Comparing all emissions, there are 4 variables in common, which are related to the gas pressure of the process at different points ($x_{32}$, $x_{30}$ and $x_{31}$) and to the temperature of admission charcoal ($x_{34}$). As expected, there are also 9 variables in common between NO and NO$_2$ ($x_{65}$, $x_{38}$, $x_{39}$, $x_{40}$, $x_{47}$, $x_{61}$, $x_{65}$, $x_{72}$, $x_{75}$) which are also related to gas pressure, oxygen admission, temperature and cooling, revealing an interdependency of these two forms while for SO$_2$ these variables are not evidenced.

As already described, a wider window size, $h$, reduces the number of required parameters to describe the responses. To discuss the effect of the window size, $h$, 3
Table 2: Effect of the averaging filter window size (h) on the number of required parameters of the parsimonious polynomial model (p), on the ability of the parsimonious model obtained by PMLS to describe the experimental results (correlation coefficient, CC, and relative residual error - residual mean squared error, %RMSE - relative absolute deviation with AD50, %AD50, and relative absolute deviation confidence interval (AD2.5 to AD97.5), [%AD]).

<table>
<thead>
<tr>
<th>h</th>
<th>p</th>
<th>NO</th>
<th>%RMSE</th>
<th>%AD50</th>
<th>[%AD]</th>
<th>NO2</th>
<th>%RMSE</th>
<th>%AD50</th>
<th>[%AD]</th>
<th>SO2</th>
<th>%RMSE</th>
<th>%AD50</th>
<th>[%AD]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>70</td>
<td>0.811</td>
<td>20.5</td>
<td>7.0</td>
<td>0.3-30.9</td>
<td>65</td>
<td>0.770</td>
<td>22.4</td>
<td>6.7</td>
<td>0.3-30.4</td>
<td>57</td>
<td>0.704</td>
<td>24.9</td>
</tr>
<tr>
<td>3</td>
<td>66</td>
<td>0.834</td>
<td>19.4</td>
<td>6.7</td>
<td>0.3-29.1</td>
<td>61</td>
<td>0.800</td>
<td>21.1</td>
<td>6.2</td>
<td>0.3-28.6</td>
<td>55</td>
<td>0.726</td>
<td>24.1</td>
</tr>
<tr>
<td>5</td>
<td>55</td>
<td>0.845</td>
<td>18.8</td>
<td>6.6</td>
<td>0.2-28.6</td>
<td>56</td>
<td>0.816</td>
<td>20.3</td>
<td>5.9</td>
<td>0.2-28.0</td>
<td>52</td>
<td>0.741</td>
<td>23.6</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>0.861</td>
<td>17.9</td>
<td>6.4</td>
<td>0.2-27.4</td>
<td>51</td>
<td>0.836</td>
<td>19.3</td>
<td>5.7</td>
<td>0.2-26.8</td>
<td>48</td>
<td>0.767</td>
<td>22.6</td>
</tr>
<tr>
<td>15</td>
<td>50</td>
<td>0.873</td>
<td>17.2</td>
<td>6.3</td>
<td>0.3-25.9</td>
<td>50</td>
<td>0.849</td>
<td>18.6</td>
<td>5.6</td>
<td>0.3-25.3</td>
<td>49</td>
<td>0.787</td>
<td>21.8</td>
</tr>
<tr>
<td>20</td>
<td>49</td>
<td>0.884</td>
<td>16.6</td>
<td>6.0</td>
<td>0.3-25.1</td>
<td>45</td>
<td>0.858</td>
<td>18.2</td>
<td>5.5</td>
<td>0.2-25.2</td>
<td>50</td>
<td>0.801</td>
<td>21.2</td>
</tr>
<tr>
<td>30</td>
<td>50</td>
<td>0.909</td>
<td>15.3</td>
<td>5.5</td>
<td>0.4-21.9</td>
<td>49</td>
<td>0.872</td>
<td>17.4</td>
<td>5.4</td>
<td>0.4-22.0</td>
<td>45</td>
<td>0.818</td>
<td>20.4</td>
</tr>
<tr>
<td>45</td>
<td>48</td>
<td>0.915</td>
<td>14.5</td>
<td>5.1</td>
<td>0.4-20.4</td>
<td>41</td>
<td>0.894</td>
<td>16.2</td>
<td>4.8</td>
<td>0.4-21.4</td>
<td>39</td>
<td>0.848</td>
<td>19.1</td>
</tr>
<tr>
<td>60</td>
<td>38</td>
<td>0.927</td>
<td>13.8</td>
<td>5.0</td>
<td>0.2-18.5</td>
<td>32</td>
<td>0.905</td>
<td>15.4</td>
<td>5.0</td>
<td>0.2-20.1</td>
<td>38</td>
<td>0.873</td>
<td>17.8</td>
</tr>
<tr>
<td>90</td>
<td>39</td>
<td>0.937</td>
<td>13.0</td>
<td>4.8</td>
<td>0.2-18.9</td>
<td>38</td>
<td>0.922</td>
<td>14.3</td>
<td>4.9</td>
<td>0.2-18.5</td>
<td>31</td>
<td>0.872</td>
<td>17.9</td>
</tr>
<tr>
<td>120</td>
<td>26</td>
<td>0.955</td>
<td>10.9</td>
<td>3.9</td>
<td>0.3-14.1</td>
<td>28</td>
<td>0.945</td>
<td>12.2</td>
<td>4.0</td>
<td>0.2-16.1</td>
<td>34</td>
<td>0.919</td>
<td>14.9</td>
</tr>
<tr>
<td>180</td>
<td>24</td>
<td>0.964</td>
<td>10.2</td>
<td>3.4</td>
<td>0.2-11.7</td>
<td>32</td>
<td>0.954</td>
<td>11.6</td>
<td>4.4</td>
<td>0.3-13.1</td>
<td>35</td>
<td>0.929</td>
<td>14.4</td>
</tr>
<tr>
<td>300</td>
<td>31</td>
<td>0.975</td>
<td>8.8</td>
<td>3.7</td>
<td>0.2-10.6</td>
<td>32</td>
<td>0.970</td>
<td>9.8</td>
<td>3.5</td>
<td>0.2-11.9</td>
<td>35</td>
<td>0.949</td>
<td>12.8</td>
</tr>
</tbody>
</table>

Cases are considered - narrow filter bandwidth (h ≤ 3 min, case A), intermediate (10 ≤ h ≤ 30 min, case B) and wide filter bandwidth (h ≥ 90 min, case C). Situation A (narrow window size) is the most conservative case - variables almost preserve their original values with little “memory effect” added into the signal. This situation is unable to compensate experimental delays and dispersion effects. Situation C (wide window size) is the least conservative case - smoothed variables are essentially representative of the selected window in detriment to the actual value. This situation may be able to circumvent delays and dispersion but introduces a severe data smoothing that may be prejudicial when trying to estimate events. In our perspective the best solution is the intermediate - situation B (intermediate window size) - were the actual value is assembled with recently past values in a ratio of 1:10 to 1:30. In order to preserve the maximal recent information a bandwidth of 10 will be selected.

5.2.2 ELM

As previously described in Section 3.2.2, in the soft modeling approach a SBS methodology was used in order to select the most relevant input variables for each target variable. Variables are sorted increasing order of relevance according to the order by which the variables are removed from the NN model (the first removed variable is the least relevant and the last removed input variable is the most relevant). The best set of variables is defined as the one that exactly includes the variables remaining in the NN model when the error is minimal. However, as this is a long process, before the selection of the most relevant input variables it is necessary to determine the delay of each variable using equation (6).

The results of the selection of the input variables with the respective delays obtained by the soft modeling approach are presented in the Table 3. It can be observed that variable $x_{39}$ was considered relevant for all the target variables, however with different delays. This variable represents the measure of NO$_2$ at the input of the cement kiln. There are also some variables that are common only with NO and NO$_2$ estimation: $x_{25}$ and $x_{47}$. These variables represent, respectively, the SO$_2$ at the input of the kiln and the temperature of the clinker at the output of the kiln. It is known that there is a strong relation between the NO and NO$_2$, and therefore the selection of these three common variables can indicate that they are very relevant for the NO and NO$_2$ emissions. Furthermore, the selection of these three variables shows that despite of gas dispersion from the kiln to the chimney, the chemical reactions inside the kiln are the most important factors to the generation of these pollutants gases.

With respect to the SO$_2$ emission, the most important variables beyond the ones already mentioned, are, by decreasing order of importance, the flow of the gases returned from the kiln to the cyclones ($x_{34}$), the temperatures of the coal ($x_{20}$ and $x_{27}$), the debit of the cooler fan ($x_{7}$) and the quantity of oxygen at the output of the coal mill ($x_{32}$).

Several tests were also made by removing the pollutant variables measured at the input of the kiln from the data set of available variables. It was verified that satisfactory responses description can be obtained by ELM in such case. However, due to space limitations these results are not presented.

Table 3: Selected input variables by the ELM soft modeling approach sorted by decreasing order of relevance for the three target outputs (NO, NO$_2$ and SO$_2$ emissions).

<table>
<thead>
<tr>
<th>NO emission</th>
<th>NO$_2$ emission</th>
<th>SO$_2$ emission</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{39}(k-3)$, $x_{7}(k)$, $x_{35}(k-2)$, $x_{27}(k)$, $x_{34}(k-18)$</td>
<td>$x_{54}(k-323)$, $x_{37}(k-46)$, $x_{37}(k-355)$, $x_{34}(k)$, $x_{46}(k-3)$</td>
<td>$x_{54}(k-1)$, $x_{20}(k)$, $x_{7}(k-127)$, $x_{39}(k-9)$, $x_{27}(k-241)$, $x_{32}(k-19)$</td>
</tr>
</tbody>
</table>
Figure 1: Response description ability of PMLS (a, c and e) and ELM (b, d and f) of variables NO, NO\textsubscript{2} and SO\textsubscript{2} using the selected prediction models.

5.3 Model Fitting and Prediction Ability

Figure 1 presents the overlay of responses predicted by the PMLS and ELM models over the actual responses. As previously mentioned, in the PMLS the data was filtered with an intermediate moving average filter with a window size of 10 samples.

In order to verify whether the selected input variables are or not in fact related with emissions, the prediction capabilities of the models selected by each modeling methodology were estimated. The data set was divided into 4 fractions (quartiles) and each quartile was used, each at a time, to obtain a model to predict the other 3 quartiles. The fitting performance is measured for each trained quartile, and for the corresponding remaining 3-quartile data the prediction performance was measured. Average results are compiled together in Table 4. From the analysis of the table, two conclusions can be obtained. First, both methods (PLMS and ELM) conduct to quite satisfactory data description. The overall fitting correlation coefficient ($CC$) varies from 0.59 to 0.82, and average relative error varies in the [4, 13]\% range, while the most probable value is about [2, 8]\% bias, and the highest fitting error estimate is below 33\%. These results seem to be very satisfactory in terms of data description. It should be noted that the data used in the soft approach (ELM) is not filtered. A similar situation is verified with the prediction ability of PMLS and ELM - correlation coefficients are generally above 0.50 and may achieve 0.71 values; prediction error is slightly higher, [16, 21]\%, with major incidence in the [7, 15]\% range; the highest prediction error may overcome [36, 56]\% as comprehensibly expected when comparing to fitting error.

Secondly, PMLS retains more input variables for NO, NO\textsubscript{2} and SO\textsubscript{2} emission estimation (50, 48 and 51, respectively) than ELM models (10, 6 and 3) with the beneficial of increased correlations and lower model errors. In the obtained PMLS models the parameter $b_0$ was one of the first parameters to be removed and thus the number of variables was equal to the number of parameters.

6 Conclusion

The presented PMLS and ELM methodologies have enabled the data-based understanding and modeling of the
Table 4: Global results of the fitting and prediction abilities of the PMLS and ELM methods with respect to NO, NO$_2$ and SO$_2$ emissions (responses).

<table>
<thead>
<tr>
<th>Emission</th>
<th>Algorithm</th>
<th>Fitting ability</th>
<th>Prediction ability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CC</td>
<td>%RMSE</td>
<td>%AD</td>
</tr>
<tr>
<td>NO</td>
<td>PMLS</td>
<td>0.803</td>
<td>12.9</td>
</tr>
<tr>
<td></td>
<td>ELM</td>
<td>0.779</td>
<td>15.2</td>
</tr>
<tr>
<td>NO$_2$</td>
<td>PMLS</td>
<td>0.759</td>
<td>12.1</td>
</tr>
<tr>
<td></td>
<td>ELM</td>
<td>0.759</td>
<td>12.1</td>
</tr>
<tr>
<td>SO$_2$</td>
<td>PMLS</td>
<td>0.803</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>ELM</td>
<td>0.590</td>
<td>11.3</td>
</tr>
</tbody>
</table>

NO, NO$_2$, and SO$_2$ emissions on a cement kiln plant. With both approaches, it is possible to describe and predict the emission values. The methods select the most important input variables for the characterization of the emission variables and for the corresponding prediction settings.

In-phase correlations evidenced low direct interdependencies between target responses with the pool of predictor variables. Highest values were obtained with $x_{33}^1$-NO$_2$ $[0.529]$ and $x_{38}^1$-NO$_2$ $[0.511]$ and $x_{55}^1$-SO$_2$ $[0.379]$ and $x_{54}^1$-SO$_2$ $[-0.341]$ (Table 1). These results are indicative that the system is not directly linear and responses are subject to several anomalies such as delays, interferences and dispersion.

Simple polynomial multivariate parsimonious hard modeling (PMLS) in conjunction with an intermediate averaging filter (with a 10 min delay) was able to effectively describe and predict responses, Fig. 1 and Table 4, maybe due to a large number of parameters for NO, SO$_2$ and NO$_2$ respectively ($p = 50, 48$ and $51$). With this methodology, it was useful to evidence the existence of “core variables” in each response - 19 for NO, 23 for SO$_2$, and 20 for NO$_2$. These results overcome the achievements with “in-phase” direct evaluation.

Soft modeling evidences that the measure of NO$_x$ at the input of the kiln was considered relevant for all the target variables, however with different delays. There are also some variables that are common only with NO and NO$_2$ emission estimation: the NO$_x$ at the input of the kiln, and the temperature of the clinker at the output of the kiln. These variables are possibly the most important variables in the NO$_x$ emissions. This shows that a chemical reaction inside the kiln plays a crucial role in the pollutant emissions. In respect to the SO$_2$ emission, was obtained by soft modeling that the most important variables, by decreasing order of importance, are the flow of the gases returned from the kiln to the cyclones, the temperatures of the coal, the debit of the cooler fan, the measure of NO$_x$ at the input of the kiln, and the quantity of oxygen at the output of the coal mill.

Acknowledgment

This work was supported by Project FAir-Control “Factory Air Pollution Control” (reference: E/E498), supported by the Eurostars Programme of the EUREKA network, financed by “Fundação para a Ciência e a Tecnologia” (FCT), do Ministério da Educação e Ciência, “Agência de Inovação” (AdI), and the Seventh Framework Programme for Research and Technological Development (FP7) of the European Union.

References