Environmental Over-Threshold Event Forecasting 
using NARX Models *

Enrico Pisoni ∗ Marcello Farina ∗∗ Giorgio Pagani ∗∗ Luigi Piroddi ∗∗

Abstract: Over-threshold event forecasting is of paramount importance in the monitoring of environmental variables, such as those related to air pollution. This paper explores the use of nonlinear polynomial NARX models for the prediction of ozone concentration data, using specific cost functions and identification algorithms devised to enhance the prediction accuracy at the peak values of the signal, in order to improve the over-threshold event detection. Some preliminary results of the experimental data analysis carried out on observed time histories are illustrated to show the effectiveness of the presented methodology.

Keywords: System identification, Nonlinear systems, Data processing, Prediction methods, NARX models, Environmental modeling.

1. INTRODUCTION

Ozone ($O_3$) gas forms in the lower atmosphere as a result of photochemical reactions that occur when precursor pollutants such as nitrogen oxides (NOx) and volatile organic compounds (VOC) are exposed to solar radiation (Carnevale et al., 2008). It is one of the most harmful ground level atmospheric pollutants (Salazar-Ruiz et al., 2008), and for this reason current regulations, now framed in the 2008/50 EU Directive, set ozone concentration thresholds. In particular, for human health protection the threshold is defined with respect to the “maximum daily eight-hour average”, with a target value of 120 $\mu g/m^3$ not to be exceeded on more than 25 days per calendar year. As every threshold exceeding implies a damage to human health, it is important to reduce the ozone exposition of all population, children and elders in particular. For this reason, accurate forecasting of ozone concentrations is of extreme importance.

Direct modeling of ozone levels and of the associated physical/chemical phenomena (Saarikoski et al., 2007) involves using partial differential equations, and requires many and precise input data, related to emissions, meteorology, land cover, etc. The resulting models are extremely detailed, but complex and expensive to develop and maintain (Coman et al., 2008). A cheaper alternative is to use statistical models, i.e. models derived directly from the processing of observed data (Schlink and Volta, 2000). A survey of statistical techniques applied to the ozone forecasting problem is reported in (Schlink et al., 2006). In particular, nonlinear statistical models have been applied with some success, employing also meteorological variables as inputs, since these are known to be related to ozone concentration through complex nonlinear phenomena. Neural network-oriented models have often been used for this task, see e.g. (Osowski and Garanty, 2007), (Niska et al., 2004), (Gomez-Sanchis et al., 2006), (Brunelli et al., 2007), (Ibarra-Berastegi et al., 2008), (Solaiman et al., 2008), although as argued by Cobourn (2007) alternative nonlinear techniques such as Generalized Additive Models (GAM), fuzzy systems and NonLinear Regressions (NLR) should provide equivalent results. Neural networks have the drawback that their structure is hidden in the input-output functional dependence, so that their processing and interpretation a posteriori is complex. Another disadvantage is that, in the presence of data with highly varying amplitude, such as environmental data with sharp peaks interleaved with inactive periods, most models tend to average the prediction accuracy between low and high signal samples. This, which is mostly due to the statistical calibration method employed, results in low accuracy in the peak prediction task, which, unfortunately, is what matters mostly in these applications.

Regarding the former issue, it is here suggested to employ polynomial nonlinear autoregressive models with exogenous variables (NARX) (Leontaritis and Billings, 1985), which are a quite general model class widely used for nonlinear model identification and prediction purposes. Polynomial expansions guarantee sufficient model flexibility (they are universal approximators) and the resulting model structure is linear-in-the-parameters, so that standard Least Squares type algorithms can be used for parameter estimation. Moreover, some kind of structural analysis of the model is possible by inspection of the quadratic or cubic or bilinear dependencies in the given variables. Due to the so called curse of dimensionality of polynomial expansions, it is important to perform a selection of the most appropriate terms to include in the model, using one of several available methods (Korenberg et al., 1988), (Piroddi and Spinelli, 2003), (Li et al., 2005). A comparison with neural network-based structures has been conducted in (Pisoni et al., * This work has been partially supported by the Italian Ministry of University and Research project “Identification and adaptive control of industrial systems”.

Copyright by the International Federation of Automatic Control (IFAC)
2009) on the ozone forecasting problem, showing that indeed NARX models are a viable solution.

In the same work it was suggested that a suitable weighting on the model identification cost function can be used to improve peak estimation accuracy, by enhancing algorithm action with respect to peak values of ozone concentration. This idea is here developed in two directions that both result in an enhanced model reliability. First, motivated by the fact that excessive pollution situations are detected in practice by simply setting appropriate thresholds for the ozone concentration amplitude, a different form of the weighting function is introduced, designed to enhance algorithm action only in an interval of amplitude values centered on the attention threshold. Furthermore, an iterative estimation method is developed to reduce false peak estimation, by enhancing algorithm action not only in the occurrence of high amplitude measured samples, but also in correspondence of incorrectly estimated peaks.

Finally, it is shown that using hourly data for model identification purposes can yield much more accurate daily predictions as opposed to models developed directly on daily data, (Pisoni et al., 2009). Notice that the usual practice in statistical modeling is to build models based on daily data for daily thresholds, and to employ hourly models only in case of thresholds defined on an hourly basis, but apparently the hourly dynamics cannot be overlooked for accurate peak estimation in any case.

Some preliminary analysis results are provided using data collected from a measuring station in the Milan urban area. These data are particularly challenging since, as far as the analysis performed in (Pisoni et al., 2009) went, the use of weighting did not convey appreciable improvement in that case. It is here shown that, on the contrary, significant improvements can be obtained in the peak prediction accuracy using hourly data and the ideas mentioned before.

The paper is structured as follows. Section 2 sets up the ozone peak forecasting problem, with particular emphasis on the metrics and time scale issues. The identification techniques for enhanced peak prediction based on polynomial NARX models are explained in Section 3. Some preliminary results of the experimental data analysis are illustrated and commented in Section 4, followed by some conclusions in Section 5.

2. METRICS AND TIME SCALE FOR PEAK PREDICTION

Ozone ground level concentrations are formed through highly nonlinear reactions, driven by emissions of nitrogen oxides and volatile organic compounds, and by particular meteorological conditions (as, e.g., high solar radiation, stagnant wind conditions, etc.). Due to these complex nonlinear chemical and physical phenomena it is very hard to correctly model the ozone dynamics in the atmosphere. This is especially true when the objective is to model ozone concentration peaks, that can be generated both by local emissions or transported and accumulated from surrounding areas.

2.1 The metrics

Current regulations are concerned with ozone concentration peaks exceeding a threshold of 120 μg/m², computed over an 8-hour average of the concentration. If the general behavior of a model can be described by “classical” statistics (as mean value, variance, etc.), in the case of peak forecasting it is also important to focus on the so-called “categorical” indexes. Denote as ME and PE the number of measured and predicted peak events, respectively, and as DE the number of correctly predicted ones (i.e., excluding false alarms). Then, the Detection Rate (DR) is defined as the fraction of the observed peak events actually detected by the model:

\[ DR = \frac{DE}{ME}, \]

and the Reliability Rate (RR) is evaluated in terms of the fraction of actual peak events detected with respect to the total number of predicted ones:

\[ RR = \frac{DE}{PE}. \]

Both objectives are important to assess the performance of different models, and a reasonable compromise between them should be obtained. An aggregate quadratic index \( AI = RR^2 + DR^2 \) will also be used to compare different solutions.

2.2 Time scale

The main objective of the models developed in this work is to predict the dynamic behaviour of the maximum daily 8-hour average ozone concentration, with special emphasis on the peak values. While it is common practice to identify models which perform one-day ahead predictions of such variable based on the values of the previous days, the approach taken in this work is different.

More precisely, the models are here identified based on ozone concentration hourly data, and carry out 24 hours ahead predictions of \( \text{O}_3 \) concentration values. The predicted values obtained in this way are then used to compute the one-day ahead prediction of the maximum daily 8-hour average concentration of ozone which, as previously discussed, is the real variable of interest. In the following we briefly describe how we perform this computation, in line with the EU Directive 2008/50/EC.

Let \( x(d,h) \) denote the (predicted) average ozone concentration in the time interval \([h : 00, h + 1 : 00]\) of day \( d \) (hourly data), \( h \) denoting the hour of the day. The 8-hour average concentration \( \bar{x}_{8h}(d,h) \) is defined as:

\[ \bar{x}_{8h}(d,h) = \frac{1}{8} \sum_{i=h}^{h+7} x(d,i) \]  \[ (3) \]

where, if \( i < 0 \), we substitute \( (d - 1,i + 24) \) for \( (d,i) \). Finally, the maximum daily 8-hour average concentration is computed as

\[ \bar{x}_{\text{max8h}}(d) = \max_{h=0, \ldots, 23} \bar{x}_{8h}(d,h) \]

In other words, the maximum daily 8-hour average concentration \( \bar{x}_{\text{max8h}}(d) \) is evaluated extracting the maximum of the 8-hour running averages \( \bar{x}_{8h}(d,h) \), with \( h = 0, \ldots, 23 \). Notice that the \( \bar{x}_{8h}(d,h) \) values are computed from the hourly data \( x(d,h) \) and updated every hour.

For example, the first sample \( \bar{x}_{8h}(d,0) \) of day \( d \) is computed, according to (3), on the basis of concentrations \( x(d - 1,i) \), with \( i = 17, \ldots, 23 \) and of \( x(d,0) \) while the last sample \( \bar{x}_{8h}(d,23) \) of day \( d \) is computed on the basis of concentrations \( x(d,i) \), with \( i = 16, \ldots, 23 \).
3. IDENTIFICATION TECHNIQUES FOR ENHANCED PEAK PREDICTION

3.1 Polynomial NARX models

NARX models (Leontaritis and Billings, 1985) are nonlinear black-box models represented by a discrete time input-output recursive expression of the form:

\[ y(t) = f(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u)) + \xi(t), \]

where \( u(\cdot) \), \( y(\cdot) \) are the model input and output, \( n_y, n_u \) are the respective maximum lags, and \( \xi(\cdot) \) is a (white) gaussian noise term.

If function \( f(\cdot) \) is a polynomial expansion, equation 5 is referred to as a polynomial NARX model. In that case, \( y(t) \) is obtained as a linear combination of monomials in the arguments \( y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u) \) plus noise, which can be reformulated as a linear regression:

\[ y(t) = \varphi(t)^T \theta + \xi(t), \]

where the regressor vector \( \varphi(t) \) is a column vector of the monomials and \( \theta \) is the vector of unknown coefficients. In vector form, Eq. (6) becomes:

\[ Y = \Phi \theta + \Xi, \]

where \( Y = [y(1) \ldots y(N)]^T \), \( \Phi = [\varphi(1) \ldots \varphi(N)]^T \) and \( \Xi = [\xi(1) \ldots \xi(N)]^T \), \( N \) being the size of the available data-set.

Conventional model performance is measured in terms of the Mean Squared Error (MSE):

\[ J = \frac{1}{N} \sum_{i=1}^{N} \epsilon(t)^2, \]

where \( \epsilon(t) = y(t) - \hat{y}(t) \) and \( \hat{y}(t) = \varphi(t)^T \hat{\theta} \) is the model prediction. The parameter estimation task amounts to solving an ordinary Least Squares (LS) problem, obtaining:

\[ \hat{\theta} = (\Phi^T \Phi)^{-1} \Phi^T Y, \]

provided the information matrix \( \Phi^T \Phi \) is full rank.

Full polynomial expansions are seldom used due to the curse of dimensionality and model structure selection must be performed to obtain models of manageable size and sufficient robustness. Forward regression methods are often adopted for this purpose (see e.g. (Korenberg et al., 1988), (Billings et al., 1989), (Haber and Unbehauen, 1990)), which iteratively augment the model structure and rely on orthogonalization techniques to decouple the estimation of different regressors. These methods tend to find over-parameterized models unless model reduction is also performed. The following iterative model building scheme is adopted in this work, which combines model augmentation and model reduction phases to yield a parsimonious model structure, (Piroddi and Spinelli, 2003). Let \( \Pi_{in} \) and \( \Pi_{out} \) denote the sets of regressors included in and excluded from the model, respectively, and \( R \) the complete set of regressors.

1. Set \( \Pi_{in} = \emptyset \) and \( \Pi_{out} = R \).
2. For each not yet selected regressor \( r_{out} \in \Pi_{out} \), the current model is augmented with this regressor, estimated and evaluated. If the best augmented model is better than the current one (by a given threshold), the corresponding regressor is included in the model (\( \Pi_{in}^{mp} = \Pi_{in} \cup \{ r_{out} \} \)), otherwise the algorithm ends.
3. For each regressor in the model, \( r_{in} \in \Pi_{in}^{mp} \), the sub-model obtained after its elimination (corresponding to the regressor set \( \Pi_{in}^{mp} \setminus \{ r_{in} \} \)) is considered, its parameters estimated and the corresponding performance index computed. If the best sub-model is still better than the model obtained at the end of the previous iteration (corresponding to \( \Pi_{in} \)), the sub-model is retained (\( \Pi_{in}^{mp} = \Pi_{in}^{mp} \setminus \{ r_{in} \}, \Pi_{out} = \Pi_{out} \cup \{ r_{in} \} \)) and step (3) is repeated, otherwise \( \Pi_{in} = \Pi_{in}^{mp} \) and step (2) is executed.

The explained method does not employ regressor orthogonalization, since the latter does not easily adapt to model reduction. In fact, if one of the orthogonalized regressors is removed, then not all the parameters are recovered.

The algorithm rationale is to add one regressor at each iteration, selecting the one that best improves the output explained variance. Then, all included regressors are tested for possible elimination, provided that a performance improvement is still obtained overall at each iteration. Notice that the method is intrinsically suboptimal, in that the local improvement of the model is actually optimized as opposed to the global performance. Also, the solution depends on the initialization of the regressor set, although the model reduction robustifies the results.

3.2 Weighted prediction criteria

As already observed, standard identification methods based on quadratic criteria are not very efficient in combination with non-stationary signals that display abrupt peaks interlaced with inactive periods. In fact, those criteria perform a uniform weighting of the error, independently of the output amplitude, which results in an accuracy trade-off between peak regions and inactive ones. Since accuracy in the peak points is mainly of interest in the current application, a non uniform weighting of the error is here used to enhance error reduction at peak points.

This results in a weighted MSE cost function (WMSE) defined as follows:

\[ J = \frac{1}{N} \sum_{i=1}^{N} w(y(t)) \epsilon(t)^2, \]

where two different weighting functions have been employed to emphasize the errors corresponding to high values of \( y(t) \):

- a sigmoidal function, which is designed to give emphasis to values of \( y(t) \) above the threshold value,
  \[ w_1(y(t)) = w_{min} + (1 - w_{min}) \frac{y(t)^H}{y(t)^H + y(t)} \]
  \[ \text{and} \quad (11a) \]
- a double sigmoidal function, which is designed to give emphasis only to the values of \( y(t) \) which are in the neighborhood of the threshold value,
  \[ w_2(y(t)) = \min\{w_{min} + (1 - w_{min}) \frac{y(t)^H}{y(t)^H + y(t)}, w_{min} + (1 - w_{min}) \frac{(2 y_{thres} - y(t))^H}{y(t)^H + (2 y_{thres} - y(t))^H} \} \]
  \[ \text{and} \quad (11b) \]

Notice that \( w_1(\cdot) \) is sigmoidal and \( w_2(\cdot) \) is a double-sigmoid, provided that \( H > 1 \). These functions assign different weights to error terms depending on the value of \( y(t) \), with a more or less gradual slope depending on \( H \) (see Figure 1). Conventionally, the maximum weighting applied is 1, while the lowest value is set by \( w_{min} \in [0, 1] \). In our case, \( y_{thres} = 120 \) is the ozone concentration threshold value. For both weights, parameter \( y \) estimated.
Fig. 1. Sigmoidal (top) and double sigmoidal (bottom) weighting functions (11).

identifies the slope change of the curve, and should be related to the ozone threshold.

The use of the weighted cost function requires minor modifications of the explained identification routines. Briefly, the following Weighted LS (WLS) solution must be used instead of Eq. (9):

$$\hat{\theta} = (\Phi^T W \Phi)^{-1} \Phi^T W Y,$$

where $W = \text{diag}(w_1(y(1)), \ldots, w_i(y(N))), i = 1, 2.$

3.3 Estimation iterations to reduce false peak detection

The weighting function $w_1(\cdot)$ (respectively $w_2(\cdot)$) has been introduced in the previous section, to improve model precision in correspondence to over-threshold data (respectively, data values in the neighborhood of the threshold). In other words, the use of the weighting functions tends to increase the detection rate $DR$, i.e., by increasing the probability that the model detects the peak event when $y(t)$ reaches the threshold.

However, experience shows that increasing the $DR$ of these prediction models may, in many cases, hamper the reliability of the model. In fact, the probability of false peak detections may increase, resulting in a dramatic decrease of the reliability rate $RR$. To overcome this drawback, the identification procedure should emphasize not only the data samples where $y(t)$ is high, but also the samples when the predicted value $\hat{y}(t)$ reaches large values (which is generally associated to large values of some of the regressors), thus preventing the loss of relevant information due to weighting. This is achieved using an iterative procedure, summarized as follows:

1) at iteration $k = 1$ compute (12) with $W = \text{diag}(w_1(y(1)), \ldots, w_i(y(N))), i \in \{1, 2\}$, and obtain predictions $\hat{y}^{(1)}(t)$;
2) set $k = k + 1$;
3) at iteration $k$ compute (12) with $W = \text{diag}(w_1(\text{max}(y(1), \hat{y}^{(k-1)}(1))), \ldots, w_i(\text{max}(y(N), \hat{y}^{(k-1)}(N))), i \in \{1, 2\}$, and obtain predictions $\hat{y}^{(k)}(t)$;
4) go to step 2) unless $k = \bar{k}$.

For simplicity, the stopping condition has been defined as the attainment of a fixed maximum number of iterations (specifically, the condition $\bar{k} = 5$ has been used in the documented simulations). More complex termination conditions can be introduced, in order to further increase the effectiveness of the algorithm.

4. EXPERIMENTAL DATA ANALYSIS

4.1 Experimental data-set and application domain

A data-set of hourly ozone concentration and temperature values from 2000 to 2006, measured in the urban area of the city of Milan, in northern Italy, has been used in this study. The considered area is characterized by high urban and industrial emissions, and is known to frequently display high ozone level episodes due to critical anthropogenic emissions, stagnating meteorological conditions and Mediterranean solar radiation (Gabusi and Volta, 2005). Attention has been focused on a monitoring station in which the use of weighting functions proved ineffectual, (Pisoni et al., 2009).

Since some data points are missing due to occasional failures of the monitoring station, for any given model structure the linear regression (6) has been restricted to time samples such that all monomials in $\Phi(t)$ are available. In the preliminary analysis documented here, since model selection is a time-consuming process, only the data from year 2000 have been used for identification purposes, while validation has been performed over year 2004, that is considered an “average” meteorological year. While not conclusive, the simulation results can be directly compared to those obtained by Pisoni et al. (2009) on daily data, where the same year had been chosen for validation.

Several NARX models were identified using different weighting functions, with the temperature as input signal ($u(t)$) and the ozone concentration as output ($y(t)$). The measured temperature has been used, on the grounds that reliable temperature prediction can be obtained using deterministic models, as in Borge et al. (2008). This allows to focus specifically on the ozone predictor modeling, without introducing further uncertainty due to the temperature prediction issue. In view of this, the obtained results are to be interpreted as an upper bound of the achievable performance. Future work will address the validation of the presented method using predicted temperature values. The ozone concentration is predicted 24-steps ahead (corresponding to 24 hours) using past ozone values (up to 24 hours before) and temperature (up to the previous hour) in the model. More specifically, the set of candidate regressors used in the simulations is defined as $R = \{y(t - 24), \ldots, y(t - 60), y(t - 28), y(t - 48), \ldots, u(t - 50), u(t - 72), u(t - 24), \ldots, u(t - 28), u(t - 48), \ldots, u(t - 50), u(t - 72)\}$, i.e., it emphasizes the dependence on both daily and hourly dynamics.

4.2 Numerical results

Previous weighted NARX model identification tests conducted on daily data (Pisoni et al., 2009) resulted in the performance represented in Fig. 2. In this and the following figures each circle in the graph represents the peak prediction performance on validation data of a NARX model identified with a given weighting function, evaluated in terms of the $RR$ and $DR$ indices. The unweighted predictor model is at the crossing of the horizontal and vertical dashed lines. Models whose performance dominates that of the unweighted predictor with respect to both indices are in the highlighted rectangular region. Circular lines represent isolines of the aggregate quadratic index $AI$. The highlighted annular region indicates improvement of $AI$ with respect to the unweighted predictor.

10562
Apparently, from inspection of Fig. 2, the unweighted predictor was one of the best models in that case, only one weighting combination resulting in a dominant solution. In absolute terms, the unweighted predictor scores $RR = 0.61$ and $DR = 0.59$ only slightly improved by the dominant solution ($H = 10$, $w_{\min} = 0$, $\bar{y} = 60$), which displays $RR = 0.62$ and $DR = 0.61$. Although many solutions could be argued to be more convenient than the unweighted one, e.g. in terms of $AI$, it can be also acknowledged that most weighting combinations determined an improvement of $DR$ at the cost of a reduced $RR$. This shows that, in this case, the introduction of weighting in the cost function increases the number of both true and false estimated peaks, ultimately reducing the predictor model reliability (that is well described by $RR$).

Much more promising results are obtained using hourly data for model identification (see Fig. 3). First of all, the unweighted model already improves significantly both indices ($RR = 0.82$ and $DR = 0.62$), and particularly the robustness index, that indicates that more than 80% of the estimated over-threshold events are actually true. This improved reliability feature is generally shared by the whole cloud of solutions (compare Figs. 2-4), showing that indeed the hourly dynamics cannot be overlooked for accurate peak event estimation. What is more, the use of weighting in the cost function appears now to yield a more decisive benefit with a significant number of dominant solutions (see the highlighted rectangular region in Fig. 3). The $DR$ index can be enhanced by a 10% without decreasing $RR$, and, vice versa, 2.5% can be gained on $RR$ without $DR$ loss. One of the best compromise solutions scores $RR = 0.85$ and $DR = 0.65$.

In these and the following results, various configurations of the weighting functions have been studied with parameters ranging in the intervals $w_{\min} \in \{0,0.25,0.5,0.75\}$, $\bar{y} \in \{60,80,100,120\}$, and $H \in \{1,2,3,5,20\}$. The preliminary analysis discussed here focuses on some significant parameter combinations in the mentioned ranges only. The threshold parameter $T_{\text{thres}}$ was set to 120 for all simulations. In terms of these parameters, most of the dominant solutions were obtained using intermediate values of $w_{\min}$ (0.25 or 0.5), low $\bar{y}$ values (between 60 and 80), and low $H$ values (up to 3.5), i.e. with smoothly degrading weighting functions. A more detailed parameter sensitivity analysis will be the object of future research.

In the tests synthetically represented in Figure 3 both the single and double sigmoidal weighting functions were used. The tests were conducted both with no estimation repetitions and with 5 iterations. Although the analysis regarding the parameters of the weighting function is by no means exhaustive, some general facts have been observed:

- On average, the double sigmoidal weighting function improves the $RR$ index (at the cost of reducing $DR$);
- As expected, the use of estimation iterations based on the estimated peaks has an analogous effect on the two indices;
- On average, the iterative approach improves also the aggregate index $AI$.

A comparison with Fig. 4 were the over-threshold events are accounted for on an hourly basis discovers another important fact. Apparently, the use of weighting in that context is almost always beneficial (most solutions are in the highlighted annular region), and when it is not the cost to pay is comparably small. However, this benefit does not always translate to an equivalent improvement when the indices are calculated on a daily basis, as required by current regulations.

5. CONCLUSION

In this work a novel approach to the forecasting of peak (daily maximum of 8-hour averages) secondary pollution concentrations has been presented. The challenges of this problem are related to the nonlinearity of the phenomena that bring to secondary pollution formation and accumulation and to the non-
Fig. 4. NARX model peak prediction performance on validation data in terms of indices $RR$ (x axis) and $DR$ (y axis), varying weighting functions and parameters: hourly indices evaluated using models estimated on hourly data.

Linearity of the over-threshold index to be reproduced. The presented approach is based on NARX models, using different cost functions specifically tailored to represent the peak behavior, and applying iterative estimation methods to reduce false peak estimations. Furthermore the approach has been applied both using hourly and daily data, to catch the underlying dynamics of the available time series. The results of the presented case study (ozone concentrations at Milan location, in Italy) show that such an approach can be beneficial in improving ozone peak forecasting.

REFERENCES


