Forecasting peak air pollution levels using NARX models

Enrico Pisoni a,*, Marcello Farina b, Claudio Carnevale a, Luigi Piroddi b

a Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza Leonardo da Vinci, 32, I-20133 Milan, Italy
b Dipartimento di Elettronica per l’Automazione, Università degli Studi di Brescia, Via Branze 38, I-25123 Brescia, Italy

Abstract

Air pollution has a negative impact on human health. For this reason, it is important to correctly forecast over-threshold events to give timely warnings to the population. Nonlinear models of the nonlinear autoregressive with exogenous variable (NARX) class have been extensively used to forecast air pollution time series, mainly using artificial neural networks (NNs) to model the nonlinearities. This work discusses the possible advantages of using polynomial NARX instead, in combination with suitable model structure selection methods. Furthermore, a suitably weighted mean square error (MSE) (one-step-ahead prediction) cost function is used in the identification/learning process to enhance the model performance in peak estimation, which is the final purpose of this application. The proposed approach is applied to ground-level ozone concentration time series. An extended simulation analysis is provided to compare the two classes of models on a selected case study (Milan metropolitan area) and to investigate the effect of different weighting functions in the identification performance index. Results show that polynomial NARX are able to correctly reconstruct ozone concentrations, with performances similar to NN-based NARX models, but providing additional information, as, e.g., the best set of regressors to describe the studied phenomena. The simulation analysis also demonstrates the potential benefits of using the weighted cost function, especially in increasing the reliability in peak estimation.

Keywords:
Polynomial NARX models
Artificial neural networks
Prediction-error based identification
Identification cost function
Air pollution forecasting
Tropospheric ozone

1. Introduction

The forecasting of extreme air pollution levels is an important issue, since it allows to provide timely warnings to the population to start preventive actions, given the negative impact of air pollution on human health. Ozone is one of the main pollution problems in Europe, due to its harmful effect on vegetation and population (Salazar-Ruiz et al., 2008). It is a secondary pollutant, formed in a nonlinear way from precursor emissions (that is to say volatile organic compounds, VOC, and nitrogen oxides, NOx) and solar radiation (Carnevale et al., 2008). Accurate forecasting of ozone concentration levels is a difficult task, because of the complex mechanisms that determine ozone formation and accumulation.

Two families of models are employed for air quality forecasting, i.e., deterministic (Saarikoski et al., 2007) and statistical (Slíni et al., 2006) models. Deterministic models use partial differential equations to forecast air pollution concentrations, and require numerous and accurate input data, as emissions, meteorology, land cover, etc., to be run. These models give a very detailed description of atmospheric variables, but at the same time are very expensive to develop and maintain (Coman et al., 2008). On the contrary, statistical models are easier to implement, and describe the input/output dynamics using simpler mathematical functions (Coman et al., 2008). Deterministic and statistical nonlinear models are proved to be particularly effective in describing the complex mechanisms of formation and accumulation of certain pollutants, e.g., ozone.

Various statistical techniques and methodologies are available in the literature, mostly based on artificial neural networks (NNs). A detailed review of prediction techniques can be found in Schlink et al. (2003, 2006), in which 15 different statistical techniques for ozone forecasting have been applied to different data sets throughout Europe. The results of the study show that in general the best statistical models to predict ground-level ozone are NNs and generalized additive models (GAMs), as these can correctly manage nonlinear phenomena. The work of Cobourn (2007) argues that nonlinear techniques as NNs, GAMs, fuzzy systems and nonlinear regressions (NLR) are expected to exhibit equivalent performances at forecasting ozone, if starting from the same set of regressor variables and applying proper model training and optimization methods. Different NN implementations are proposed in the literature, for wavelet coefficient identification (Osowski and Garanty, 2007), or taking advantage of genetic algorithms for input selection (Niska et al., 2004). Some examples of applications of NNs on different geographical areas are presented in Gomez-Sanchis et al. (2006) and Ibarra-Berastegi et al. (2008).
One of the most convenient model forms for prediction purposes is the nonlinear autoregressive model with exogenous variables (NARX) (Leontaritis and Billings, 1985a, b), which is a quite general formulation where the current output value is made dependent on the past values of the input and output signals through a suitable nonlinear static function. The model is dynamic in view of the lagged inputs and outputs that feed the nonlinear function, but the model identification problem actually consists of finding the static mapping between these signals and the current output. The nonlinear function that provides such mapping can be represented with various functional expansions depending on the identification method of choice (Sjöberg et al., 1995). For example, it can be described as a feed-forward neural network, or as a polynomial expansion, or it can be formulated using radial basis functions, wavelets, support vector machines (SVM), etc. From a flexibility viewpoint all these functional expansions are equivalent, in that they are universal approximators.

In this study we focus in particular on polynomial NARX models. Using truncated polynomial expansions to approximate the nonlinear mapping has several advantages. The model turns out to be linear-in-the-parameters, which implies that simple algorithms of the least squares family can be employed for parameter estimation. Also, its dynamic relationships can be easily interpreted, since a polynomial model directly shows the presence of quadratic or cubic or bilinear dependencies in the given variables. The main difficulty in the identification of polynomial NARX models is the selection of the model structure, since the number of terms in a (full) polynomial expansion grows rapidly with the number of arguments and the nonlinearity degree. The selection of the most appropriate terms of the expansion is therefore crucial for model accuracy and robustness. Fortunately, many efficient model selection techniques are available in the literature (see, e.g., Korenberg et al., 1988; Piroldi and Spinelli, 2003a; Li et al., 2005). Frequency analysis tools in the nonlinear domain have also been developed (Billings and Tsang, 1989). Notwithstanding these important advances, the class of polynomial NARX models has not been employed before, to the authors’ knowledge, for ground-level ozone forecasting. More precisely, nonlinear regression models have been presented in Cobourn (2007) and Lin and Cobourn (2007) in the Kentucky metropolitan area, using as input a nonlinear function, but this approach did not include the autoregressive part of the NARX models. Also, in the cited study there was no automatic selection of the best nonlinear terms to reconstruct ozone concentrations. Other functional expansions, such as NNs and SVMs may be more convenient than polynomial expansions in the presence of nonlinearities requiring high order polynomials to be accurately described. For comparison purposes the NN functional expansion has also been studied here.

In the present work polynomial NARX models are presented and compared with NN-based NARX models for a case study in the Milan urban area. A further analysis is aimed at assessing the potential benefits of using a weighted mean square error (WMSE) cost function to enhance the peak estimation quality of the identified model. Actually, the plain MSE criterion averages the (squared) error irrespective of the signal amplitude, so that typically peak concentrations are mis-estimated due to the influence of the remaining portion of the data in the identification process. This behavior is typical of environmental data with high amplitude variability, that are characterized by abrupt high amplitude peaks alternated with low amplitude signal portions, a behavior which is difficult to capture with a single model. The weighting in the performance index is precisely aimed at emphasizing the modeling errors associated with high amplitude output values in the averaging performed by the MSE.

The paper is structured as follows. In Section 2 the proposed statistical approaches are formalized. The case study setup presented in this paper is described in Section 3. Finally, Sections 4 and 5 present results and conclusions.

2. Identification of NARX models

2.1. The NARX model class

NARX models (Leontaritis and Billings, 1985a) are the nonlinear generalization of the well-known ARX models, which constitute a standard tool in linear black-box model identification (Ljung, 1999). These models can represent a wide variety of nonlinear dynamic behaviors and have been extensively used in various applications. A NARX model is formulated as a discrete time input–output recursive equation:

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

where \( y(t) \) and \( u(t) \) are the model input and output, \( n_y \) and \( n_u \) are the respective maximum lags, and \( \xi(t) \) is a noise term, generally assumed gaussian and white. The optimal predictor form of this model is

\[ \hat{y}(t) = f(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u)), \]

where \( \hat{y}(t) \) denotes the one-step ahead prediction of \( y(t) \). Depending on how function \( f(\cdot) \) is represented and parameterized, different NARX model structures and, consequently, identification algorithms are derived. In the sequel, both polynomial expansions and neural networks will be used for this purpose.

2.2. Polynomial NARX models

Representing \( f(\cdot) \) in (1) as a polynomial expansion has various practical advantages, since it yields a linear in the parameters model structure, for which simple least squares type algorithms can be devised. Also, a direct interpretation of polynomial terms is possible in several applications. Another advantage of the polynomial NARX model class is the availability of nonlinear frequency analysis tools based on high order frequency response functions (Billings and Tsang, 1989; Peyton Jones and Billings, 1989). On the downside, to avoid the curse of dimensionality, only polynomial expansions truncated to a very low degree of nonlinearity \( f \) are generally considered, at the cost of a reduced model accuracy. Even so, a careful model structure selection must be performed to obtain models of manageable size and sufficient robustness. The recent literature has witnessed many contributions on polynomial NARX models, regarding both methodological aspects and applications (see, e.g., Mao and Billings, 1997; Aguirre and Jácome, 1998; Aguirre et al., 2000; Piroldi and Palumbo, 2000a, b; Chiras et al., 2001; Boaghe et al., 2002; Leva and Piroldi, 2002; Piroldi and Spinelli, 2003a; Wei et al., 2004; Spinelli et al., 2005).

The general form of the polynomial NARX model is as follows:

\[ y(t) = \hat{y}(t) = a_0 + \sum_{i=1}^{n_y} \sum_{j=1}^{n_u} \sum_{k=1}^{n_x} a_{i,j,k} x_{i,j,k} t + \xi(t), \]

where \( x = [y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u)] \) is the \( n = n_y + n_u + n_x \) dimensional vector of the base regressors. In other words, \( y(t) \) is obtained from a linear regression of terms (regressors) obtained as products of powers of past values of \( u \) and \( y \).

Full polynomial expansions are seldom employed, to avoid having an excessive number of regressors for parameter estimation, thus exceeding acceptable model complexity, not to mention the numerical ill-conditioning problems that may arise due to
over-parameterization (Ljung, 1999; Aguirre and Billings, 1995). Structure selection is thus a critical issue in NARX identification, and several approaches have been proposed in the literature. Typically, the model structure is constructed iteratively, adding regressors one by one on the basis of a convenient accuracy criterion (generally the MSE), exploiting regression and orthogonalization techniques (Haber and Unbehauen, 1990). Given the linearity of model (3) with respect to the parameters, simple algorithms of the least squares type are conventionally employed for parameter estimation, and orthogonalization is used to decouple the estimation of different regressors. An example of such methods is the forward-regression orthogonal estimator (FROE) (Korenberg et al., 1988; Billings et al., 1989). Other works suggest the use of pruning techniques to reduce model complexity and of simulation error minimization criteria to improve model robustness (Piroddi and Spinelli, 2003a,b). The interested reader is also directed to Mendes and Billings (2001), Li et al. (2005), Kukreja et al. (2006) and Peng et al. (2006) for some recently proposed alternative approaches.

In the sequel, the following forward regression procedure with pruning is applied, which is derived from Piroddi and Spinelli (2003a), albeit using only prediction-oriented performance indexes:

1. Initially the model structure is empty, i.e., \( P_{in} = \emptyset \) and \( P_{out} = R \), where \( P_{in} \) and \( P_{out} \) denote the sets of regressors included in and excluded from the model, respectively, and \( R \) is the complete set of regressors.

2. For each not yet selected regressor \( r_{out} \in P_{out} \), the current model is augmented with this regressor, estimated and evaluated in terms of its MSE. If the best augmented model is better than the current one (by a given threshold), the regressor is included in the model \( (P_{in}^{tmp} = P_{in} \cup \{r_{out}\}, P_{out} = P_{out} \setminus \{r_{out}\}) \), otherwise the algorithm ends.

3. For each regressor in the model, \( r_{in} \in P_{in}^{tmp} \), the sub-model obtained after its elimination (corresponding to the regressor set \( P_{in}^{tmp} \setminus \{r_{in}\} \)) is considered, its parameters re-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 \( P_{in} \)), the sub-model is retained \( (P_{in}^{tmp} = P_{in}^{tmp} \setminus \{r_{in}\}, P_{out} = P_{out} \cup \{r_{in}\}) \) and step (3) is repeated, otherwise \( P_{in} = P_{in}^{tmp} \) and step (2) is executed.

The first step initializes the algorithm with an empty model structure, the second one increments the model with the regressor that improves best the prediction accuracy of the model, and the third one iteratively prunes the model. Concerning the pruning step, the least significant regressor of the current model structure is first found as the one whose elimination minimally increases the model’s MSE. If its elimination yields a model which is still better than the one obtained at the previous iteration, the regressor is actually eliminated, and a new check for eliminable regressors is effected. When no further regressor can be eliminated, the pruning phase ends. A complete iteration of the algorithm either adds a new regressor to the current model or substitutes one or more of its terms with it, providing that this exchange of terms improves the model accuracy. Moreover, the performance index is always decreasing between iterations. The pruning mechanism is particularly effective when a correct regressor is inserted: typically, a number of spurious terms is deleted from the actual regressor set. Notice that the regressor set selected at each iteration is a local minimum for the loss function, and the iteration solution is dependent on the initial regressor set. The method is thus intrinsically suboptimal.

The flow diagram of the NARX model selection algorithm is shown in Fig. 1.

2.3. Neural network-based NARX models

The function \( f(.) \) in Eq. (1) for NARX models (1) can also be modeled by means of artificial neural networks. A huge literature is available on NN model theory and applications (see, e.g., Gardner and Dorling, 1998). Here, only a brief description of NNs is presented. Neural networks are composed by simple connected elements (denoted as neurons) operating in parallel. Each neuron’s output is obtained by filtering a weighted sum of its inputs through a usually nonlinear function, i.e., the so-called activation function. The weights associated with the network connections are tuned during the training (learning) phase in order to reduce a given cost function, such as MSE (or a modified cost function, as proposed in this paper). The neural network structure used in this study is a standard feed-forward neural network (Fig. 2). This kind of network computes a vector function \( f_{NN}: \mathbb{R}^Q \rightarrow \mathbb{R}^L \) where \( Q \) and \( L \) are the dimensions of the input and output vectors of the net, respectively, the \( n \)-th element of the vector function \( f_{NN} \) for the \( n \)-th pattern \( (n^e \in \mathbb{R}^Q) \) is defined as \( (M \) is the number of the neurons in the hidden layer)\n
\[
f_{NN}(n^e) = a_{f_2} \left( \sum_{m=1}^{M} (OW_{lm} \cdot a_m) + g_l \right).
\]

where

\[
a_m = a_{f_1} \left( \sum_{q=1}^{Q} (IW_{mq} \cdot v_q) + b_m \right),
\]

where \( a_{f_1} \) and \( a_{f_2} \) are two real continuous functions, called activation function of the hidden layer \( (a_{f_1}) \) and of the output layer \( (a_{f_2}) \).
vectors are the bias terms. Neural networks learn on a training data set, tuning the parameters \(IW, OW, b\) and \(g\) by means of the well-known Levenberg–Marquardt backpropagation (BP) algorithm (Hagan et al., 1996). The choice of the best neural network weights has been done minimizing a selected cost function (MSE or WMSE, as defined in Section 3.3).

3. Materials and methods

3.1. Site characterization and data

The domain selected for this study is the Milan urban area (Fig. 3) counting three million inhabitants. It is a very industrialized area, located in northern Italy, and is characterized by high urban and industrial emissions. Furthermore, critical anthropogenic emissions, frequent stagnating meteorological conditions and Mediterranean solar radiation regularly cause high ozone level episodes in this area (Gabusi and Volta, 2005). The locations of the measurement stations considered in this work are shown in Fig. 3. Only six stations (in the study domain) where meteorological and ozone variables are available at the same time have been used. Hourly ozone concentration and temperature data are available from 2000 to 2006 for each station.

The data have been normalized between 0 and 1. In addition, spurious or unavailable data points have been eliminated. More precisely, fixing a NARX model structure defines the subset of arguments belonging to \(\{y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u)\}\) that maps to \(y(t)\). Then, for each \(t\), if either \(y(t)\) or any of the corresponding arguments of the map are not available, the input–output pair is eliminated.

In a first stage the data spanning from 2000 to 2003 and from 2005 to 2006 have been used for model identification, while validation has been performed over 2004. The year 2004 has been chosen as validation period because it represents an “average” meteorological year (Cuvelier et al., 2007). In a second stage cross-validation has also been performed, extracting each time a different year for validation, and keeping the remaining data set for identification.

3.2. Structure of the models

Statistical models have been identified to provide maximum daily 8-h average ozone concentrations (MAX8H) one day ahead. The initial hourly data set has been reformatted to obtain mean daily data of MAX8H for ozone, and maximum daily temperature (\(T\)).

In order to reconstruct the target output MAX8H\((t+1)\), a NARX model structure was assumed with the following input arguments: \(\{\text{MAX8H}(t), \text{MAX8H}(t-1), T(t+1), T(t)\}\), both for NN-based models and polynomial NARX models, briefly named NN5 and NARX in the sequel. For the latter the candidate regressor set for the model selection process has been obtained by generating all the combinations (up to cubic terms) of the five elementary inputs, while in the NN5 case the inputs are used directly to feed the network, since the neural network should be able to provide suitable fitting of the nonlinear function relating the inputs to the output.

3.3. Cost functions for model identification

The dynamics of ozone concentrations typically display abrupt peaks, which are difficult to capture with traditional modeling techniques, that optimize criteria such as the MSE or the RMSE (so averaging over all data independently of their actual values). This has typically the effect that the identified model operates a trade-off between peak regions and the remainder of the signal, which can be relevant in the presence of high nonlinearity (or nonstationarity) effects related to the signal amplitude. Such is precisely the case in the problem of interest, where scarce prediction accuracy of near-peak ozone concentrations is experienced. Unfortunately, it is precisely the correct prediction of the occurrence and value of such peaks that is vital for timely environmental monitoring and management, and, accordingly, a model’s performance is evaluated with respect to its capability to predict Over-Threshold peak ozone concentration Events (OTEs) on a day-by-day basis. This suggests that the model identification cost function should also emphasize prediction errors associated with high ozone concentrations, i.e., around a given threshold value. For this reason, this work investigates the use of a weighted version of the MSE cost function (WMSE), where the weighting factor depends on the actual value of the ozone concentration and on the applied threshold. In order to evaluate if and when such innovation results in more efficient predictive capabilities (at least in the domain of interest), the performances achievable using a standard cost function will be compared with those obtained with differently parameterized weighted functions. The WMSE is

![Fig. 2. Classical feed-forward neural network scheme.](image-url)

![Fig. 3. Case study domain with considered stations.](image-url)
defined as follows:
\[
J = \frac{1}{N} \sum_{t=1}^{N} w(t) (y(t) - \hat{y}(t))^2,
\]
where \(y(t)\) is the value of ozone concentration at time instant \(t\) \((t = 1, \ldots, N)\) and \(\hat{y}(t)\) represents the prediction of \(y(t)\) obtained through the model (either a polynomial or a NN-based NARX model). The weighting factor is designed to emphasize the errors corresponding to high values of \(y(t)\):
\[
w(t) = w_{\text{min}} + (1 - w_{\text{min}}) \frac{y(t)^h}{\bar{y}^h + y(t)^h},
\]
where \(w_{\text{min}}\) is the minimum weight, \(\bar{y}\) is the mean measured ozone value, \(\bar{y}(t)\) is the mean predicted ozone value.

Notice that the weighting factor has a sigmoidal shape for \(h > 1\), which attributes two different weights to error terms associated with low and high \(y(t)\), with a more or less gradual slope between the two values depending on \(h\) (see Fig. 4). The high weight is conventionally set to 1, whereas the low one equals \(w_{\text{min}}\). Notice that the WMSE equals the MSE if \(w_{\text{min}} = 1\). Parameter \(\bar{y}\) identifies the slope change of the curve, and should be related to the ozone threshold. The optimal values of constants \(w_{\text{min}}, h\) and \(\bar{y}\) are case-dependent and are tuned by carrying out a suitable test campaign.

Conveniently enough, model identification using index (7) requires minimal modifications of the standard optimization routines. For example, for NN-based models the stochastic gradient term used in the BP algorithm must be multiplied by \(w(t)\), since
\[
\frac{\partial J}{\partial \theta} = \frac{2}{N} \sum_{t=1}^{N} w(t) (y(t) - \hat{y}(t)) \frac{\partial \hat{y}(t)}{\partial \theta},
\]
while the NARX model is identified through standard weighted least squares.

3.4. Validation indexes

Several criteria have been considered for model validation. Conventional error measures, such as the MSE and the RMSE, that weigh equally all the available data, cannot reflect the focus of standard anti-pollution regulations on peak event forecasting, so that their use for validation is questionable. All the same, for the sake of completeness and to ease readers’ comparison, two standard prediction cost functions of this type have been used for validation purposes, namely the correlation coefficient (CC) and the explained variance (EV), or variance accounted for (VAF) together with more specific peak-oriented indicators, i.e., the detection rate (DR) and the reliability rate (RR). CC is defined as
\[
CC = \frac{\sum_{t=1}^{N} (y(t) - \bar{y}(t)) \cdot (\hat{y}(t) - \bar{y}(t))}{\sqrt{\sum_{t=1}^{N} (y(t) - \bar{y}(t))^2 \cdot \sum_{t=1}^{N} (\hat{y}(t) - \bar{y}(t))^2}},
\]
where \(\bar{y}(t)\) is the mean measured ozone value, \(\bar{y}(t)\) is the mean measured ozone value, \(\bar{y}(t)\) is the predicted ozone value and \(\bar{y}(t)\) is the mean predicted ozone value.

EV is defined as
\[
EV = 1 - \frac{\text{MSE}}{\text{Var}(y)},
\]
where \(\text{Var}(y)\) is estimated as the sample variance of the ozone concentration signal.

Irrespective of its accuracy in reproducing the target signal as well as possible, the model is rated on the basis of its capability of correctly estimating the OTEs (classified as peaks exceeding a threshold of 120 \(\mu g/m^3\)). Let \(ME\) and \(PE\) be the number of measured and predicted OTEs, respectively, and \(DE < PE\) be the number of correctly predicted OTEs (excluding false alarms). Then, the index DR defined as
\[
DR = \frac{DE}{ME},
\]
represents the fraction of the observed OTEs actually detected by the model. The reliability of the estimator can be evaluated in terms of the percentage of actual OTEs detected with respect to the total number of predicted OTEs, using the RR index:
\[
RR = \frac{DE}{PE}.
\]

4. Results and discussion

4.1. Role of the weighting function in the performance index

Although we are primarily interested with daily predictions of the ozone concentration, since environmental decisions are taken based on the daily time scale, the role of the weighting function in enhancing peak estimation is best evidenced when elaborating hourly data, the ozone formation phenomenon being highly dependent on the temperature fluctuations during the day.

To this purpose the hourly data of two stations (530 and 535, Fig. 3) over the whole 2004 have been analyzed. For each set of data the polynomial NARX identification was repeated using the WMSE with 141 different weighting functions, obtained by taking all the possible combinations of the parameters \(w_{\text{min}} \in \{0, 0.25, 0.5, 0.75\}\), \(h \in \{0.5, 1, 2, 3.5, 10, 20, 50\}\) and \(\bar{y} \in \{60, 80, 100, 120, 140\}\), as well as the unweighted MSE.

To see how this reflects on the discrete indexes, DR and RR, examine Fig. 5, which graphically shows the performance of all the identified models in terms of the two indexes. To aid the reader, circular curves have also been added, which correspond to specific values of the 2-norm \(J = \sqrt{DR^2 + RR^2}\), that provides a possibly useful aggregate performance index. In both cases (stations 530 and 535), the MSE cost function yields one of the less accurate in terms of DR, but most reliable estimators (see RR). Several dominating solutions are obtained with the WMSE, so that the DR index can be greatly improved using weighting, without significantly affecting the RR. Actually, both indexes can be simultaneously improved for station 530, although RR only slightly. In other words, it appears that the use of appropriate weighting in the cost function can consistently enhance OTE
estimation without affecting estimation reliability (sometimes even reducing the number of false alarms).

An example of the enhanced peak estimation achievable using the weighted cost function is illustrated in Fig. 6, with reference to the hourly data of station 530, where some near-threshold peaks are shown. The simulation has been obtained using the model with highest \( f \), corresponding to \( RR = 0.89882 \) and \( DR = 0.86818 \), i.e., the one obtained with weighting function parameters \( w_{min} = 0 \), \( h = 20 \) and \( \gamma = 80 \). Apparently, the quality of the prediction is improved in the peak regions (at the cost of a worsened estimation elsewhere). Actually, using suitable parameters in the weighting function, the model accuracy can be improved up to a factor of 30% in the peak regions.

Notice that different parameters are obtained for the weighting function for each separate data set. Actually, each measurement station has different characteristics, due to the location, traffic conditions, etc., and this reflects in different values of the parameters. Also, hourly and daily data (see next subsection) have been treated differently, due to the different dynamics they show. In addition, even considering a single data set, it is not trivial to say what parametrization is absolutely best, since accuracy is measured in terms of two indexes, that are partially conflicting. This is why Pareto curves are used to study this issue. Nevertheless, some regularity has been observed regarding the selected parameters of the weighting function, as explained in the sequel.

4.2. Daily prediction performances with NARX models

The daily data of the six stations are now processed, focusing on the comparison of the performances which can be obtained using different classes of NARX models and different weighting function parameterizations. As in the case of hourly data, the polynomial NARX identification was repeated using the WMSE with 141 different weighting functions, obtained by taking all the possible combinations of the parameters \( w_{min} \in \{0.25, 0.5, 0.75\} \), \( h \in \{0.5, 1.2, 3.5, 10, 20, 50\} \) and \( \gamma \in \{60, 80, 100, 120, 140\} \), as well as the unweighted MSE. The “optimal” weighting function is selected (see Fig. 7), for each station, as the one which results in minimizing a “weighted” distance, in the \( DR-RR \) space, between the so-called Utopia point (\( DR = 1 \) and \( RR = 1 \)) and the actual results.

In Fig. 8 the estimated parameter values associated with the regressors of the polynomial NARX are shown for the considered stations. The bar heights in the histogram represent the absolute values of the coefficients for the different regressors (i.e., a zero value in the graph indicates that the regressor is not selected). Some reflections are in order.

First of all, it is apparent that, for each station, the number of selected regressors (and hence the order of the dynamical models) is low, with respect to the number of available regressors. This leads to simple models and confirms that model selection is effective.

Furthermore, some types of regressors appear in all the identified models. This highlights the temperature-ozone dynamics underlying the measurements collected. For instance, linear auto-regressive terms, such as \( y(t - 1) \) or \( y(t - 2) \) have a high rate of occurrence. This implies that the “inertial” effect of ozone accumulation on the dynamics is not negligible. Furthermore, quadratic or cubic terms in \( u(t), u(t - 1), u(t - 2) \) appear very frequently. This supports the physical evidence that nonlinear models are necessary to describe the effect of the temperature on ozone concentration (and shows the advantages of using NARX models instead of linear autoregressive ones). Interestingly enough, it is also apparent that, when (nonlinear) terms appear in \( u(t) \), normally a term in \( u(t - 1) \) appears as well,
Fig. 7. Daily data: polynomial NARX prediction performance in terms of DR (x-axis) and RR (y-axis) for varying weighting function (MSE, solid black plus; WMSE, black dots; optimal WMSE, black star). Stations: 502 (a), 512 (b), 525 (c), 530 (d), 535 (e), 555 (f). The dashed lines are circles $x^2 + y^2 = r^2$.

Fig. 8. Polynomial NARX regressors for the different stations.
the corresponding parameter having opposite sign (for example, the model of the station 525 contains $1.4801u(t)^2u(t-1)-1.31u(t-1)^2$). This suggests that ozone concentration is affected not only by the absolute value of the temperature (i.e., $u(t)$), but also by its relative value with respect to the previous occurrence (e.g., by its slope).

We now focus on the optimal choice of the weighting function (7) for each station, and on the specific models that are identified thereof.

First, for stations 502, 512 and 530, the optimal triples of weighting parameters $(h, w_{\text{min}}, \tilde{y})$ are (20, 0.60), (50, 0.60), (10, 0.60), respectively. In the three cases, the weighting function (7) has a steep slope (high $h$ value), $w_{\text{min}} = 0$, and $\tilde{y}$ has a low value. The three resulting weighting functions are approximations of the Heaviside function (also referred to as step function) $H(y-60)$. In the identified models of the three stations, nonlinear terms containing $u(t)$, $u(t-1)$ and $u(t-1)$ prevail. This shows that the influence of the temperature dynamics on the ozone concentrations in these three cases is dominant.

On the other hand, for stations 525 and 535, the optimal triples of the weighting parameters $(h, w_{\text{min}}, \tilde{y})$ are (0.5, 0.120) and (1.0, 140), respectively. In both cases, the slopes $h$ of the weighting function (7) are low and $\tilde{y}$ assume high values. The identified models of these stations are simpler (lower order) than the models of the previous three stations. Specifically, linear autoregressive terms $y(t-1)$ and $y(t-2)$ are prevalent. Nonetheless, terms in $u(t)$ appear, since direct influence of temperature on ozone concentration cannot be neglected. Finally, for station 555, the optimal triple of weighting parameters $(h, w_{\text{min}}, \tilde{y})$ is (3.5, 0, 100). The model associated with such station contains both linear autoregressive terms in $y(t-1)$ and $y(t-2)$, and nonlinear terms in which ozone concentration $y(t)$ does not appear, but only terms $u(t)$, $u(t-1)$, $u(t-2)$ and $u(t-1)$. These remarks suggest that ozone accumulation (witnessed by autoregressive terms) and temperature-related effects (represented by nonlinear functions of $u(t)$) are differently balanced depending on the considered station/geographical area.

The results of the application of the presented techniques to daily data are collected in Fig. 9. Specifically, the performances of

![Fig. 9. Correlation (a), explained variance (b), detection rate (c) and reliability rate (d) of the considered models.](image-url)
the identified models are evaluated (for each station) by the indexes presented in Section 3.4, applied to the validation data set (i.e., year 2004). The NARX and NNS5 models represent polynomial and NN-based models, while the MSE and WMSE cost functions represent the standard MSE and the selected weighted cost function, respectively. When using the weighted cost function, the same parameters have been used, for each station, for both types of models. The results obtained with the presented techniques are compared also with the results obtained with a *persistent* model (where the “trivial” one-step predictor is given by the past ozone concentration value, i.e., \( y(t + 1) = y(t) \)).

By analyzing the results shown in Fig. 9, we notice that NARX and NNS5 models behave similarly. Importantly, by employing the nonstandard weighted cost function (WMSE), the quality of the peak prediction is enhanced, with respect to the case where MSE is used. This is witnessed by the values of the indexes DR and RR, previously defined in Section 3.4. From the point of view of the CC and EV indexes, the quality of the models obtained with weighted cost functions show slightly worse performances. Bad prediction performances are detected (by all models) in the application to stations 502 and 525, where the persistent model displays even better results; interestingly enough, both these stations are situated in the Milan city area, and bad prediction performances are probably caused by unmodeled factors, e.g., the *urban titration effect*, which normally affects the occurrence of OTEs. In general, the numerical results (given in terms of the statistical indexes shown in Fig. 9) are similar to the ones presented in the previous studies performed in the same domain study with different classes of stochastic models (Corani, 2005); specifically, in this previous study, CC was ranging among 0.83 and 0.86, and the value of DR was ranging between 0.67 and 0.73.

The robustness of the presented procedure has been tested through a leave-one-out cross-validation procedure, extracting each time a different validation year (from 2000 to 2006) and using the remaining data set for identification. In detail, for each of the six stations considered, seven identifications and validations have been performed for each model type. The results show a strong dependence of the statistical indexes on the selected validation year, although apparently the use of the weighted performance index always improves peak estimation with respect to unweighted estimation. More precisely:

1. The performance in terms of the indexes DR and RR strongly depends on the number of OTEs in the training set and in the validation year. For example, 2003 is a nonstandard year, in that extreme meteorological conditions were observed in the summer (with very hot temperatures, with a consequent increase in ozone pollution). Using 2003 in validation generally results in better values of DR and RR, prevalently because of the high number of peaks. The year 2004 has been selected for validation, since together with 2005, it is generally acknowledged as a meteorological average year (see, e.g., Cuvelier et al., 2007).

2. This year-induced variability in the results is experienced indifferently using the unweighted MSE or its weighted counterpart as performance index, since it is a data dependent feature. Therefore, what should really be considered important is the capability of improving the two indexes in the same training/validation conditions when using the weighted performance index.

3. This generally happens if smooth weighting functions are adopted, as done in the previous simulations for stations 525, 535 and 555. In these cases, an average improvement (up to 10%) of DR in the seven cross-validation runs is experienced for the NARX case, without reducing the RR value. On the other hand, the use of nonsmooth weightings (approxima-

5. Conclusions

Correct peak estimation of ozone concentration data is crucial in environmental monitoring for health hazards. This work concerns the use of nonlinear black-box statistical models for the modeling and prediction of these type of data, which is a hard problem due to the incomplete information delivered by the available data and the high amplitude variability of the concentration data. The main contributions of this study can be envisaged as: (a) the use of a new family of models for ozone forecasting (polynomial NARX models), (b) a performance comparison between these models and neural networks, and (c) the use of alternative cost functions for model identification. Results show that polynomial NARX models yield performances similar to NN-based NARX models (which are currently considered the state-of-the-art models for statistical ozone forecasting), and furthermore they provide information about the best regressors for a particular nonlinear phenomenon. Also, the benefits of an amplitude weighting of the identification cost function (independently of the model class adopted) in the accuracy of the peak estimation are demonstrated by the analysis. In particular, it is shown that peak estimation reliability can be enhanced by such weighting, thereby allowing for a more confident model exploitation in the monitoring process.

### Notation

**Variables**

\[
\begin{align*}
\text{w}(\cdot) & : \text{model input} \\
\nu_\text{\text{\text{

**Acronyms**

- \( Ti \) : maximum daily temperature
- \( \text{MAX} (H) \) : maximum daily 8-h average ozone concentration
- \( \text{MSE} \) : mean square error cost function
- \( \text{CC} \) : correlation coefficient
- \( \text{EV} \) : explained variance
- \( \text{VAF} \) : variance accounted for
- \( \text{OTE} \) : over threshold peak ozone concentration event
- \( \text{ME} \) : total number of measured OTE events
- \( \text{PE} \) : total number of model predicted OTE events
- \( \text{DE} \) : total number of detected measured OTE events
- \( \text{DR} \) : detection rate
- \( \text{RR} \) : reliability rate

**Acknowledgments**

Part of this work has been developed in the frame of Pilot Project QUISAT (QUalita’ dell’aria mediante l’Integrazione di misure da Terra, da Satellite e di modellistica chimica multifase e di Trasporto—contract I/035/06/0—http://www.quisit.it), sponsored and funded by the Italian Space Agency (ASI). Furthermore, the authors acknowledge Cost Action ES0602 (Towards a European Network on Chemical Weather Forecasting and Information
Systems) community. M. Farina and L. Piroddi have also been supported by The Italian National Research Project “New Techniques of Identification and Adaptive Control for Industrial Systems”.

References


