

Multi step forecasting for nonlinear models of high frequency ground ozone data: a Monte Carlo approach*

Alessandro Fassò
University of Bergamo

Ilia Negri†
University of Bergamo

Abstract

Multi step prediction using high frequency environmental data is considered. The complex dynamics of ground ozone often requires models involving covariates, multiple frequency periodicities, long memory, nonlinearity and heteroscedasticity. For these reasons parametric models which includes seasonal fractionally integrated components, self exciting threshold autoregressive components, covariates and autoregressive conditionally heteroscedastic errors with heavy tails have been recently introduced. Here to obtain a h step ahead forecast for these models we use a Monte Carlo approach. The performance of the forecast is evaluated on different nonlinear models comparing some statistical indices with respect to the prediction horizon. As an application of this method, the forecast precision of a two years hourly ozone data set coming from an air traffic pollution station located in Bergamo, Italy, is analyzed.

Key Words: prediction; simulations; threshold autoregression; ARCH errors; hourly air pollution data.

*This work has been supported by MURST Cofin98 grant # 27

†Complete address of both authors: Department of Engineering, Viale Marconi, 5 – 24044 Dalmine (Italy)

1 Introduction

The aim of this paper is to present the capability of some stochastic nonlinear models to make multi step ahead forecast of hourly ozone concentration. In an environmental context nonlinear models have been recently proposed, mainly for descriptive and interpretative purposes. For modeling daily ozone and nitrogen dioxide concentration, a nonlinear model based on the one step prediction was proposed by Graff-Jaccottet and Jaunin (1998). Lewis and Ray (1997) have used multivariate adaptive regression splines for nonlinear regression models applied to daily sea temperature. The nonlinear models discussed here try to explain the complexity of ground ozone dynamics encompassing multiple seasonality, long memory, heteroscedasticity, nonlinearity, covariates and good performance on the right tail (see Fassò and Negri 2000). They can be used in order to monitor the phenomenon, detect measurements errors and capture in advance dangerous trends. Here, we focus on medium and long term prediction of future values of ozone concentration.

It is well known that for nonlinear models, multi step ahead forecasting presents various difficulties especially from a computational point of view. Despite the fact that forecasting is one of the principal purposes in time series modeling, this subject has received little attention in the literature of applied nonlinear models. This is due to the fact that, except for the one step ahead case, where the classical idea of Box and Jenkins (1970) and extended by Priestley (1988) can be used, the analytical evaluation of the prediction becomes intractable for a longer horizon (see for example Peel and Speight, 2000).

De Groijs and Kumar (1992) pointed out two main questions arising in the problem of multi step ahead forecasting for nonlinear models. The first is how to compute the exact forecast for more than two steps ahead. The second is concerned with the gaining of prediction accuracy when a nonlinear model is used rather than a linear one. Considering the last question, De Groijs and Kumar (1992) concluded that there is no clear evidence in favour of nonlinear over linear model in terms of forecasting capability. More recently, analyzing a number of self exciting threshold autoregressive models (SETAR) proposed in the econometric literature, Clements and Smith, (1999) seem to agree with these conclusions. They found that the forecasting capability of these nonlinear models is usually better when the predictions are evaluated conditionally on the state of the system but not always as good as in linear models when the state is not considered.

Computational problems are discussed in Tong (1995), where two methods are suggested to evaluate the best predictor given by the conditional expectation. One is based on the Chapman-Kolmogorov relation and con-

sists in the calculation of high order integrals that give recursive formulas for conditional expectations. In general, these integrals do not admit analytic solutions and numerical integration is adopted. The alternative method consists in a Monte Carlo study based on a long record of data simulated according to the proposed model. The last approach seems to give better results in terms of forecasting accuracy and computation costs (see Clements and Smith, 1997 and 1999). In particular, in the former reference, the authors make a useful comparison of the most popular methods for nonlinear multi step forecasting, namely the normal approximation method due to Al Qassam and Lane (1989), (see also De Groot and De Bruin, 1991, De Groot and Kumar, 1992 and references therein) the skeleton method presented in Tong (1995), the bootstrap method and the dynamic estimation method discussed by Granger and Teräsvirta (1993). Our idea starts from a revised Monte Carlo approach (see Clements and Smith, 1997) that seems to give better results in multi step ahead forecasting for SETAR models. We use a modification of this approach based on the consideration that the errors generated in the simulated paths have zero conditional mean.

The evaluation of the prediction accuracy and capability is formulated from two different points of view. The first is the classical measure given by the R^2 statistic and by the mean square or the absolute forecasting error. These global performance measures have been discussed in Bhansali (1993, 1999) and used for example by Bordignon and Lisi (2000) for nonlinear modeling of hydrological time series. The second method, which could be called *local*, goes inside to our true and predicted data and considers the global indices conditionally on the state of the system. The proportion of correct and false forecast alarms are also considered as indicators of the local accuracy in prediction of the models. This last approach permits to evaluate the forecasting performance of the models in the medium and high pollution levels which are the most important for human and environmental protection.

The rest of the paper is organized as follow. In the next section we recall the models of interest and in section 3 the problem of multi step ahead forecast for nonlinear models is analyzed. Section 4 is devoted to the presentation of the Monte Carlo method. In section 5, we give a preliminary analysis of our data set and, in section 6, we describe and discuss the results of multi step forecasting for the models presented in section 2. Conclusions are given in the final section 7, where some further developments are also proposed. In the appendix computational details are presented.

2 The SFI-SETARX-ARCH models

We consider time series model that can be acronymized by SFI-SETARX-ARCH from Seasonal Fractionally Integrated Self-Exciting AutoRegressive processes with eXogenous variables and AutoRegressive Conditionally Heteroscedastic errors. When we do not consider thresholds, or we do not use exogenous variables or both, the model becomes, respectively SFI-ARX-ARCH, SFI-SETAR-ARCH and SFI-AR-ARCH. In particular the last model, having a linear forecast frame, is considered as the base model in comparing the prediction capability of the four kinds of nonlinear models.

Let us recall the form of these models. Powers of backward operator B are defined as $B^k y_t = y_{t-k}$, $k = 1, 2, \dots$, where y_t , $t = 0, \pm 1, \pm 2, \dots$ denotes the process of interest. The vector of the exogenous and deterministic variables is denoted by $u_t = (u_{1,t}, \dots, u_{r,t})'$, $r \geq 0$, where prime denote vector transposition. Let us define the polynomials in B , $\nabla(B) = (1 - B)^{d_1} (1 - B^{24})^{d_{24}}$, $0 \leq d_1, d_{24} < 1/2$, $\alpha(B) = 1 - \alpha_1 B - \dots - \alpha_p B^p$, $p \geq 1$ and the polynomial vector $\gamma(B) = (\gamma_1(B), \dots, \gamma_r(B))$. If we introduce the vector $\beta = (\beta_0, \beta_1, \dots, \beta_q)$, $q \geq 1$, the SFI-ARX-ARCH model can be written as

$$x_t = \nabla(B)y_t, \quad (1)$$

$$\alpha(B)x_t = \gamma(B) \cdot u_t + \varepsilon_t h_{t-1}, \quad (2)$$

$$h_{t-1} = \beta_0 + \sum_{j=1}^q \beta_j |e_{t-j}|, \quad (3)$$

where $e_t = \alpha(B)x_t - \gamma(B) \cdot u_t$ and ε_t are independent and identically distributed random variables with $\mathbf{E}|\varepsilon_t| = 1$.

In the SFI-SETARX-ARCH model it is assumed that variable y_t follows a SFI-ARX-ARCH model depending on the value of lagged value of y_t . Denoting with D_1, D_2, \dots, D_m an appropriate partition of the sample space \mathcal{S} , the threshold model reads as

$$x_t = \nabla^{(j)}(B)y_t \quad (4)$$

$$\alpha^{(j)}(B)x_t = \gamma^{(j)}(B) \cdot u_t + \varepsilon_t h_{t-1}^{(j)} \quad (5)$$

$$h_{t-1}^{(j)} = \beta_0^{(j)} + \sum_{i=1}^q \beta_i^{(j)} |e_{t-i}^{(j)}|, \quad (6)$$

for $y_{t-1} \in D_j$, $j = 1, \dots, m$. Usually, when \mathcal{S} is a subset of the real line, the sets D_j , called regimes, are defined through thresholds as $D_j = \{a_{j-1} \leq y_{t-1} < a_j\}$, with all a_j belonging to \mathcal{S} . Here $e_t^{(j)} = \alpha^{(j)}(B)x_t - \gamma^{(j)}(B) \cdot u_t$ and ε_t are independent and identically distributed random variables such that $\mathbf{E}|\varepsilon_t| = 1$.

In a previous work (see Fassò and Negri, 2000) we have studied the stability conditions for these kinds of models. Identification has been carried out with classical Box-Jenkins and AIC-BIC criteria, analyzing R^2 statistics both in the sample and in the cross validation data set. The parameter estimations have been computed using a nested weighted least square approach.

3 Multi step forecast for nonlinear models

To explain the problem arising in multi step forecast for nonlinear models, let us rewrite in a different way the models presented in the previous section. From equations (1) and (2) we have

$$\alpha(B)\nabla(B)y_t = \gamma(B) \cdot u_t + \varepsilon_t h_{t-1}.$$

Taking the inverse of the convolution between the operators α and ∇ the model can be rewritten as

$$y_t = f(y_{t-1}, y_{t-2}, \dots, u_t, \alpha, \gamma, d_1, d_{24}) + \varepsilon_t h_{t-1},$$

where f is a linear function involving an infinite power series of the operator B , see Beran (1994). In a similar manner, we can rewrite the mean part of SFI-SETARX-ARCH model, given by equation (4) and (5), as

$$y_t = \sum_{j=1}^m \mathbb{I}_{D_j}(y_{t-1}) \left(f^{(j)}(y_{t-1}, y_{t-2}, \dots, u_t, \alpha, \gamma, d_1, d_{24}) + \varepsilon_t h_{t-1}^{(j)} \right),$$

where the $f^{(j)}$, $j = 1, \dots, m$ are again linear functions involving an infinite power series of the operator B and \mathbb{I}_A denotes the indicator function of a set A . In the following this function is denoted more simply with $y_{t-1} = f(y_{t-1}) + \varepsilon_t h_{t-1}$, where f is not a linear function. For the sake of simplicity we will denote (6) as $h_{t-1} = h(e_{t-1})$.

Let us denote with $y_t(h)$ the prediction of y_{t+h} when we are currently standing at time t . The best predictor in the sense of mean square error forecast is

$$y_t(h) = \mathbf{E}(y_{t+h}|y_t),$$

where conditioning on y_t stands for conditioning on the σ -algebra generated by the observations up to time t .

For $h = 1$, $y_t(1)$ can be calculated explicitly

$$y_t(1) = \mathbf{E}(y_{t+1}|y_t) = f(y_t).$$

In fact to predict the value of y_{t+1} conditionally on y_t , the regime is known exactly. Thus, when $y_t \in D_j$ we have

$$y_t(1) = f^{(j)}(y_t, y_{t-2}, \dots, u_t, \alpha, \gamma, d_1, d_{24}) = f(y_t).$$

In general, for a nonlinear model, when $h > 1$ this conditional expectation cannot be calculated explicitly. In fact to predict the value y_{t+2} the best forecast is

$$y_t(2) = \mathbf{E}(f(y_{t+1}) + \varepsilon_{t+2}h_{t+1}|y_t) = \mathbf{E}(f(y_{t+1})|y_t). \quad (7)$$

To compute this conditional expectation we need y_{t+1} that is not available. We know only $y_t(1)$ which differs from the true value by an error, that, according to the arch model, is $\varepsilon_{t+1}h_t$, where h_t follows equation (6). The best predictor of y_{t+2} is then

$$y_t(2) = \mathbf{E}(f(y_{t+1})|y_t) = \mathbf{E}(f(f(y_t) + \varepsilon_{t+1}h_t)|y_t). \quad (8)$$

In general $\mathbf{E}(f(y_{t+1})|y_t) \neq f(y_t(1))$ because f is not linear. For the general h step ahead forecast the situation described by (8) becomes even more complicate. We have

$$y_t(h) = \mathbf{E}(f(y_{t+h-1}) + \varepsilon_{t+h}h_{t+h-1}|y_t) = \mathbf{E}(f(y_{t+h-1})|y_t), \quad (9)$$

where y_{t+h-1} has to be obtained recursively going backward for $k = 1, 2, \dots, h-1$ according to the following equation

$$y_{t+h-k} = f(y_{t+h-(k+1)}) + \varepsilon_{t+h-k}h_{t+h-(k+1)}. \quad (10)$$

There is a sort of convolution h times of function f involving $h-1$ random variables $\varepsilon_{t+1}, \varepsilon_{t+2}, \dots, \varepsilon_{t+h-1}$. Written explicitly for $h = 3$ is

$$y_t(3) = \mathbf{E}(f(y_{t+2})|y_t) = \mathbf{E}(f(f(f(y_t) + \varepsilon_{t+1}h_t) + \varepsilon_{t+2}h_{t+1})|y_t).$$

To calculate this conditional expectation we can proceed by numerical calculation but, as discussed in Clements and Smith (1999), simulation method, as Monte Carlo, it gives more satisfactory results from a statistical point of view.

4 Monte Carlo forecasting

The idea is to evaluate the expected value in (8) via the Monte Carlo method by generating n values of the error ε_{t+1} . In the general case for $h > 2$

the expected value in the right hand side of (9) is calculated generating n values for each of the $h - 1$ random variables $\varepsilon_{t+1}, \varepsilon_{t+2}, \dots, \varepsilon_{t+h-1}$, appearing recursively by equation (10).

In the sequel we distinguish between the unknown value y_{t+h} and the best forecast in the sense of mean square error forecast $y_t(h)$. The $i - th$ simulation of y_{t+h} is denoted by y_{t+h}^i , the $i - th$ simulation used for the h step ahead forecast by \hat{y}_{t+h}^i . Finally the estimation of the best forecast is denoted by $\hat{y}_t(h)$ and obtained as the average over i of the \hat{y}_{t+h}^i . For $h = 1$ the estimate $\hat{y}_t(1)$ of the best forecast coincide with $y_t(1)$ and we do not need any simulation. For $h = 2$, to give a forecast of y_{t+2} we estimate the expectation in (8) simulating $i = 1, 2, \dots, n$ values for y_{t+1} . So we have

$$y_{t+1}^i = f(y_t) + \varepsilon_{t+1}^i h_t,$$

where $h_t = h(e_t)$ is known and has to be calculated from (6). Then we calculate a Monte Carlo simulation for the two step ahead forecast. This is given by

$$\hat{y}_{t+2}^i = f(y_{t+1}^i)$$

and the Monte Carlo estimate for the two step ahead best forecast is given by

$$\hat{y}_t(2) = \frac{1}{n} \sum_{i=1}^n \hat{y}_{t+2}^i.$$

For the next ahead forecast ($h=3$), we need the simulations of possible values of y_{t+2} . These are $y_{t+2}^i = f(y_{t+1}^i) + \varepsilon_{t+2}^i h_{t+1}^i$, where $h_{t+1}^i = h(e_{t+1}^i)$, with $e_{t+1}^i = \varepsilon_{t+1}^i h_t$. So a simulation for the 3 step ahead forecast is $\hat{y}_{t+3}^i = f(y_{t+2}^i)$ and the estimation of the best forecast is the average of these values. The calculation of $\hat{y}_{t+k}^i = f(y_{t+k-1}^i)$, $k > 3$, is now straightforward and we have

$$\hat{y}_t(k) = \frac{1}{n} \sum_{i=1}^n \hat{y}_{t+k}^i.$$

Note that the simulated values y_{t+k-1}^i of y_{t+k-1} are obtained recursively by

$$y_{t+k-1}^i = f(y_{t+k-2}^i) + \varepsilon_{t+k-1}^i \cdot h_{t+k-2}^i,$$

where

$$h_{t+k-2}^i = h(e_{t+k-2}^i), \quad e_{t+k-2}^i = \varepsilon_{t+k-2}^i \cdot h_{t+k-3}^i.$$

The algorithm is presented in details in appendix.

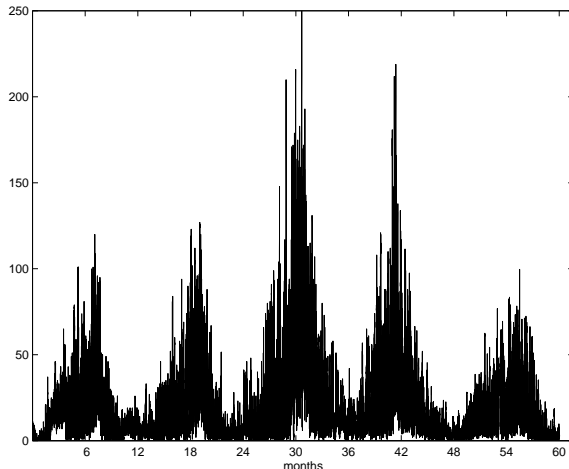


Figure 1: Series of the hourly ozone concentration in ppb

	5 years	last 2 years	summer
N	43824	17544	8640
Mean	19.0643	17.9062	27.8444
Variance	510.1663	413.4888	555.9180
Median	11	10.8	22.7
Max	249.7	218.8	218.8

Table 1: Descriptive statistics

5 Preliminary data analysis

The application data set considered is the time series of hourly ozone concentration measured in a air pollution traffic station located in the city of Bergamo near Milan in Italy. The meteorological covariates and the precursor pollutants available in the same air station that have been used in the models were nitrogen oxide and dioxide, total solar radiation, temperature, humidity and wind's velocity. The sample period runs from 1st January 1993 to 31 December 1997.

The plot of the time series (see Figure 1) displays some characteristics of the phenomenon. We see very clearly a multiple seasonal components (annual cycles are evident) and some cluster of variability seems to appear.

The analysis of the sample auto correlation function (see Figure 2) shows an evident daily cycle and a long memory component. The whole data set have been split into two subsamples. The first one (the estimation data set)

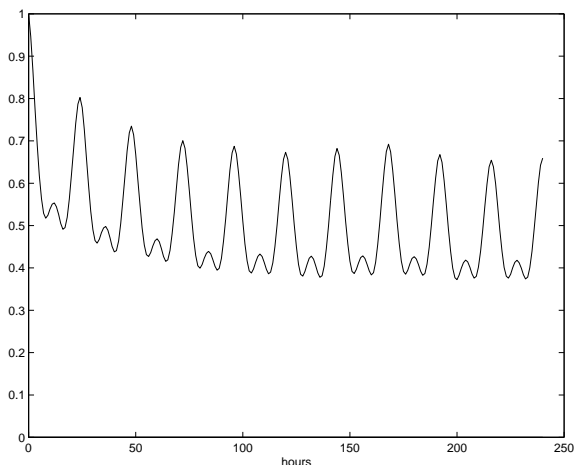


Figure 2: Autocorrelation of ozone concentration for 10 days lags

has been used for identification and estimation, the second one (the validation data set) for the analysis of the obtained model and for prediction. The data of the last two years amounts to 17544 observations and the 8640 summer data have been used for testing the forecasting performance.

Table 1 summarizes some descriptive statistics of the time series in the subsets considered.

6 Forecast accuracy comparisons

The forecasting performance of the four proposed models is carried out from two different points of view, the first called *global* and the other *local*. For each of the $N = 8640$ hours t of the validation data set, we have calculated the hourly h step ahead predictions ($h = 1, 2, \dots, 48$) based of the pre-identified models and their values have been compared with the true actual values. The Monte Carlo study has been carried out on $n = 100$ simulations for each step h . The simulated errors, according to Fassò and Negri (2000), have the $\frac{t_\nu}{m_\nu}$ distribution, where t_ν is a Student's t random variable with $\nu > 4$ degrees of freedom and m_ν is its absolute expected value.

The first criterion is based on the comparison, for each value of h , of all predicted and actual values, according to the statistics $R^2(h) = 1 - \frac{\hat{\mathbf{E}}(y_{t+h} - \hat{y}_t(h))^2}{\text{V}\hat{\mathbf{A}}\mathbf{R}(y_t)}$, Mean Absolute Error $\text{MAE}(h) = \hat{\mathbf{E}}|y_{t+h} - \hat{y}_t(h)|$ and Mean Square Error $\text{MSE}(h) = \hat{\mathbf{E}}(y_{t+h} - \hat{y}_t(h))^2$. The estimated mean $\hat{\mathbf{E}}$ and variance $\text{V}\hat{\mathbf{A}}\mathbf{R}(y_t)$ are computed over all the t in the validation data set. The

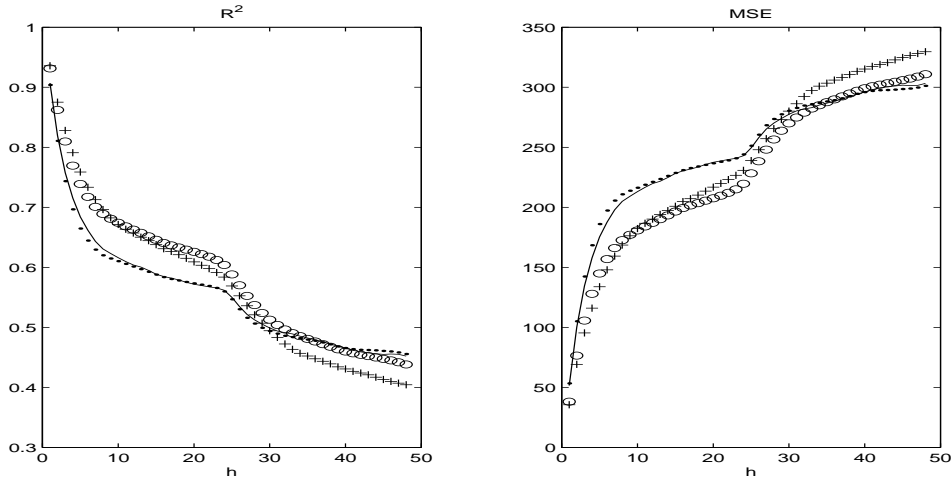


Figure 3: R^2 and MSE for SFI-AR-ARCH (\cdot), SFI-SETAR-ARCH ($-$), SFI-ARX-ARCH (\circ), SFI-SETARX-ARCH ($+$).

results are reported in Figure 3, where we see that the global $R(h)^2$ statistic decreases from 0.94 to 0.40 for the nonlinear SFI-SETARX-ARCH predictor and from 0.90 to 0.46 for the linear autoregressive SFI-AR-ARCH one. The related $R(h)^2$ curves cross around $h = 30$ hours. This reversed performance at short and long horizons is not new and agrees with mentioned results of De Groot and Kumar (1992), and Bordignon and Lisi (2000).

For the Italian protection environmental law, there is an *attention* and an *alarm* level for the ozone concentration given respectively by 90 ppb and 180 ppb. In our sample period the ozone was 204 times over the attention level and 14 times over the alarm level. From the local point of view we have calculated the three above global statistics in the two subsets defined by the attention and the alarm level. The results for the mean absolute errors are reported in Figure 4. In particular, from the global point of view of Figure 4 (a), it is confirmed that in the short term prediction SETAR and SETARX perform better than the corresponding models without thresholds. This performance is reversed in the medium and long term where the simplest model is the better. The same happens when comparing the models with covariates with respect to the corresponding pure time series models. Moreover, from the local MAE analysis of Figure 4 (b), the nonlinear pure time series SFI-SETAR-ARCH is very close to the linear plus covariate SFI-ARX-ARCH for medium to high pollution and, from Figure 4 (c), the former outperforms the latter for very high pollution. Hence the superiority of input output models over the pure time series ones is not uniform over forecast horizon and

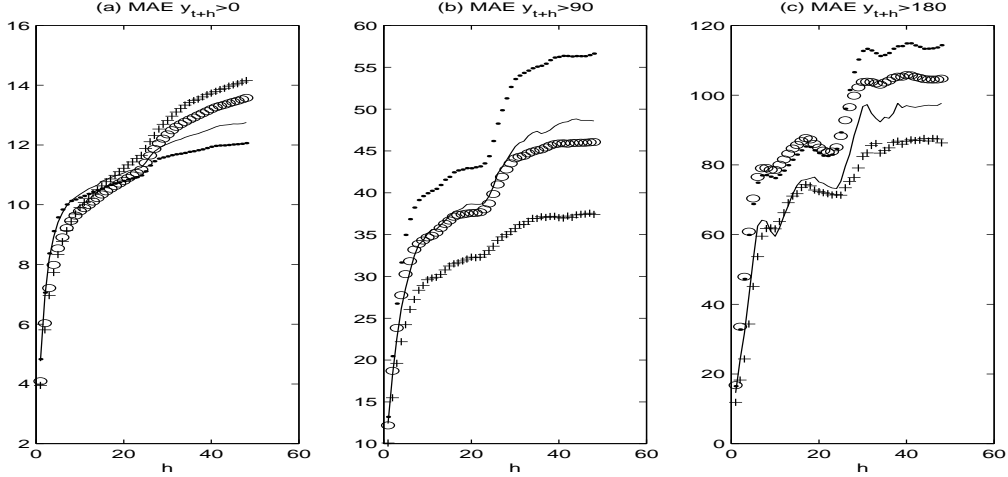


Figure 4: MAE global (a) and in the attention (b) and alarm (c) level. SFI-AR-ARCH (\cdot), SFI-SETAR-ARCH ($-$), SFI-ARX-ARCH (\circ), SFI-SETARX-ARCH ($+$).

pollution levels.

The bias in the predicted values is reported in Figure 5. In particular, note how in Figure 5 (a) is pointed out a moderate overestimation whilst local biases in Figure 5 (b) and (c) show corresponding underestimation of high pollution. This opposite behaviour is partly due to the high skew of the data distribution. Moreover, local MAE is equal to the corresponding bias at the alarm level. This is obviously due to the fact that $y_{t+h} \geq \hat{y}_t(h)$ for all $y_{t+h} > 180$ in our data set.

To go inside our data, for each step h and for each model, we have computed the proportion of correct forecast alarms

$$CA(h) = \frac{\# \text{ correctly forecast exceedings}}{\# \text{ total actual exceedings}}$$

and the proportion of false forecast alarms

$$FA(h) = \frac{\# \text{ uncorrectly forecast exceedings}}{\# \text{ total forecast exceedings}}.$$

These indices have been evaluated in the two subset given by attention and alarm levels. The results for the four models and 6 hours ahead forecast are reported in Table 2, where the first two columns report the total number of forecasting exceedings FE_i for each of the two threshold ($i = 1, 2$). Then

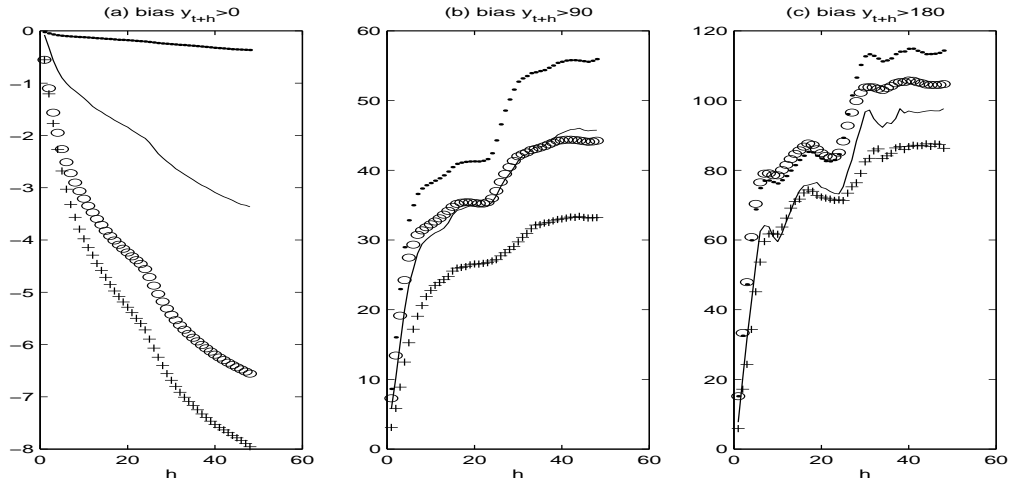


Figure 5: Bias global (a) and in the attention (b) and alarm (c) level. SFI-AR-ARCH (\cdot), SFI-SETAR-ARCH ($-$), SFI-ARX-ARCH (\circ), SFI-SETARX-ARCH ($+$).

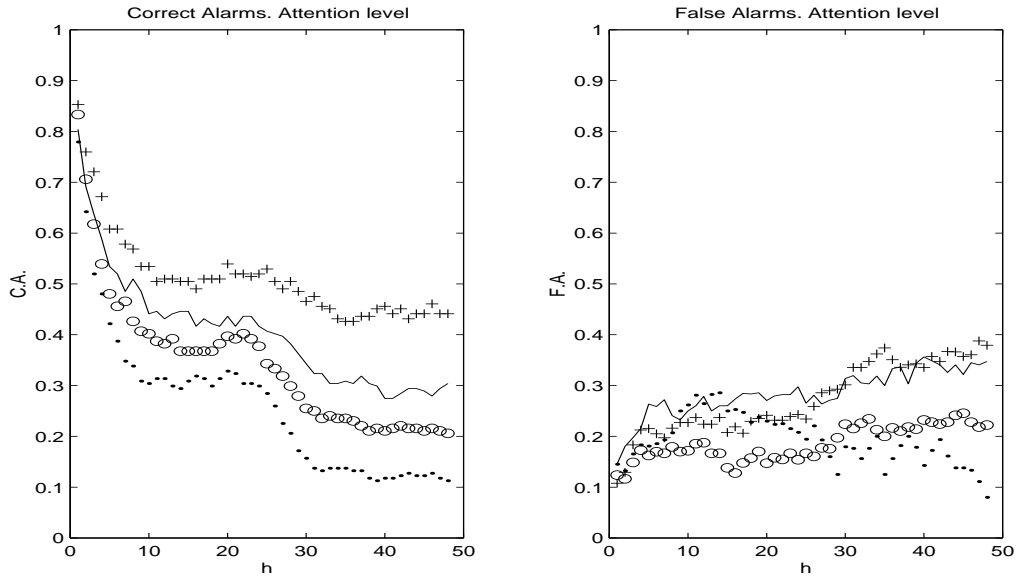


Figure 6: Correct Alarm and False Alarm in the attention level. SFI-AR-ARCH (\cdot), SFI-SETAR-ARCH ($-$), SFI-ARX-ARCH (\circ), SFI-SETARX-ARCH ($+$).

the correct (CA_i , $i = 1, 2$) and the false alarms (FA_i , $i = 1, 2$) follow. The proportion of correct and false alarms in the attention level are given in Figure 6. It clearly follows that the model which better forecasts the attention and alarm exceedings is the SFI-SETAR-ARCH. This superiority is uniform from the short to the long term.

7 Conclusions and further developments

In this paper, it has been shown how to perform multi step forecast of hourly air quality data using a general nonlinear heteroscedastic model.

The results are encouraging as the overall R^2 statistic, very high at short forecast horizons, is still about 0.50 after 30 steps. The role of nonlinear forecasts of the threshold type is especially motivated for environmental protection where correct predictions for right tail data, i.e. high pollution, are important. On the relevance of this aspect see for example Bloomfield et al. (1996). This holds true both for the short and long horizon forecasts.

The conditional heteroscedastic component is important in order to study the varying forecast precision. For one step ahead prediction it can be easily assessed through the heteroscedastic component h_t . Whenever the analytical study of the forecasting precision for multi step nonlinear predictions needs further insight, from the practical point of view, it is useful to note that the prediction standard errors can easily be computed using the same data simulated for the forecast computations.

From the operational pollution forecasting point of view one has to use forecasted covariates and this could weaken our performance results. Nevertheless from the previous discussion of Figure 4, the nonlinear pure time series model SFI-SETAR-ARCH can be used as a reference curve for models with predicted inputs.

Moreover, our models with concomitant covariates are intended to describe unknown ozone dynamics after adjusting for all the covariates and to show how this kind of complex dynamics can be handled.

An alternative and promising way for practical applications is to use a different model for each prediction horizon h . This may be done by extending the approach of Bhansali (1993) to the nonlinear case.

Appendix: The Monte Carlo algorithm

The Monte Carlo scheme can be resumed as follows:

for (t in summer period) **do**

Model	lags	FE_1	FE_2	CA_1	CA_2	FA_1	FA_2
SFI AR ARCH	1	186	9	159	9	27	0
	2	149	4	130	4	19	0
	3	131	0	111	0	20	0
	4	121	0	99	0	22	0
	5	105	0	86	0	19	0
	6	98	0	81	0	17	0
SFI SETAR ARCH	1	192	12	164	11	28	1
	2	170	10	142	6	28	4
	3	157	4	126	2	31	2
	4	156	2	120	1	36	1
	5	148	0	109	0	39	0
	6	149	0	110	0	39	0
SFI ARX ARCH	1	194	11	170	10	24	1
	2	166	3	145	3	21	0
	3	145	0	124	0	21	0
	4	137	0	113	0	24	0
	5	122	0	98	0	24	0
	6	111	0	90	0	21	0
SFI SETARX ARCH	1	195	11	174	9	21	2
	2	181	12	158	7	23	5
	3	181	4	148	3	33	1
	4	177	2	137	2	40	0
	5	161	0	129	0	32	0
	6	155	0	120	0	35	0

Table 2: Forecast Exceedings (FE_i); Correct Alarm (CA_i); False Alarm (FA_i); $i = 1$ attention level; $i = 2$ alarm level.

```

{
  compute  $y_t(1) = f(y_t)$ 
  compute  $h_t = h(e_t)$ 
  for ( $i = 1$  to length MC simulation  $n$ ) do
  {
    generate a random number  $\varepsilon_1^i$ 
    for ( $k = 2$  to horizon  $h$ ) do
    {
      compute  $e_{t+k-1}^i = \varepsilon_{t+k-1}^i h_{t+k-2}^i$ 
      compute  $y_{t+k-1}^i = f(y_{t+k-2}^i) + e_{t+k-1}^i$ 
      compute  $\hat{y}_{t+k}^i = f(y_{t+k-1}^i)$ 
      generate a random number  $\varepsilon_k^i$ 
      compute  $h_{t+k-1}^i = h(e_{t+k-1}^i)$ 
    }
    end (cycle  $k$ )
  }
  end (cycle  $i$ )
  compute  $\hat{y}_t(k) = \frac{1}{n} \sum_{i=1}^n \hat{y}_{t+k}^i = \frac{1}{n} \sum_{i=1}^n f(y_{t+k-1}^i)$ 
}
end (cycle  $t$ )

```

References

- [1] Al Qassam MS, Lane JA. 1989. Forecasting exponential autoregressive models of order 1, *Journal of Time Series Analysis*, **10**: 95-113.
- [2] Baillie RT, Bollerslev T, Mikkelsen HO. 1996. Fractionally integrated generalized autoregressive conditional heteroskedasticity, *J. Econometrics*, **74** 3-30.
- [3] Beran J. 1994. *Statistics for Long-Memory Processes*, Chapman & Hall, New York.

- [4] Bhansali R.J. 1993. Estimation of the prediction error variance and an R^2 measure by autoregressive model fitting, *Journal of Time Series Analysis*, **14**: 2, 125-146.
- [5] Bhansali R.J. 1999. Parameter estimation and model selection for multi-step prediction of a time series: a review. In *Asymptotics, nonparametrics, and time series* (Ghosh S. ed.), Marcel Dekker, New york, 201-226.
- [6] Bloomfield P, Royle JA, Steinberg LJ, Yang Q. 1996. Accounting for meteorological effects in measuring urban ozone levels and trends, *Atmospheric Environment*, **30**, 17: 3067-3077.
- [7] Bollerslev T. 1987. A conditionally heteroscedastic time series model for speculative prices and rates of return, *The review of economics and statistics*: 542-547.
- [8] Bollerslev T, Engle RF. 1986. Modelling the persistence of conditional variances, *Econom. Review*, **5**, 1: 1-50.
- [9] Bordignon S, Lisi F. 2000. Nonlinear analysis and prediction of river flow time series, *Environmetrics*, **11**: 463-477.
- [10] Box GEP, Jenkins GM. 1970. *Time Series Analysis: Forecasting and Controls*, Holden-Day, San Francisco.
- [11] Clements MP, Smith J. 1997. The performance of alternative forecasting methods for *SETAR* models, *International Journal of Forecasting*, **13**: 463-475.
- [12] Clements MP, Smith J. 1999. A Monte Carlo study of the forecasting performance of empirical *SETAR* models, *J. Applied Econometrics*, **14**: 123-141.
- [13] Cox WM., Chu S-H. 1993. Meteorologically adjusted ozone trends in urban areas: a probabilistic approach, *Atmospheric Env.*, **27B**, 4: 425-434.
- [14] De Grooijer JG, De Bruin K. 1998. On *SETAR* forecasting, *Statistic and Probability letters*, **37**: 7-14.
- [15] De Grooijer JG, Kumar K. 1992. Some recent developments in non-linear time series modelling, testing and forecasting, *International Journal of Forecasting*, **8**: 135-156.

- [16] Engle RF. 1995. *ARCH: selected readings*, (Engle Ed.); Oxford University Press, Oxford.
- [17] Fassò A, Negri I. 2000. Nonlinear statistical modelling of high frequency ground ozone data, Accepted on *Environmetrics*.
- [18] Graf-Jaccottet M, Jaunin MH. 1998. Predictive models for ground ozone and nitrogen dioxide time series, *Environmetrics*, **9**: 393-406.
- [19] Granger CWJ, Teräsvirta L. 1993. *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford.
- [20] Lewis PA, Ray BK. 1997. Modeling long-range dependence, nonlinearity, and periodic phenomena in sea surface temperatures using TSMARS, *JASA*, **92**, 439: 881-893.
- [21] Peel DA, Speight EH. 2000. Threshold nonlinearities in unemployment rates: further evidence for the UK and G3 economies, *Applied Economics*, **32**: 705-715.
- [22] Priestley M. 1988. *Nonlinear and nonstationary time series*, Academic Press, London.
- [23] Robeson SM, Steyn DG. 1990. Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations, *Atmospheric Env.*, **24**, 2: 303-312.
- [24] Tong H. 1995. *Non-linear time series*, Clarendon Press, Oxford.