
Gruppo di Ricerca per le Applicazioni
della Statistica ai Problemi Ambientali

Working paper - n.3 - May 2000

# Nonlinear statistical modelling of high frequency ground ozone data

Alessandro Fassò - Ilia Negri

# Nonlinear statistical modelling of high frequency ground ozone data

Alessandro Fassò - Ilia Negri[1]

University of Bergamo - Dept. of Engineering,
V.le Marconi 5, 24044 Dalmine, Italy,
tel +39 035 277 323 , fax +39 035 562 779,
email: fasso@unibg.it.

## Abstract

The problem of describing hourly data of ground ozone is considered. The complexity of high frequency environmental data dynamics often requires models covering covariates, multiple frequency periodicities, long memory, non linearity and heteroscedasticity. For these reasons we introduce a parametric model which includes seasonal fractionally integrated components, self exciting threshold autoregressive components, covariates and autoregressive conditionally heteroscedastic errors with high tails. For the general model, we give estimation and identification techniques.

To show the model descriptive capability and its use, we analyize a five year hourly ozone data set from an air traffic pollution station located in Bergamo, Italy. The role of meteo and precursor covariates, periodic components, long memory and non linearity is assessed.

**Key Words**: Periodical long memory; threshold autoregression; ARCH errors; hourly air pollution data.

[1]This work has been supported by MURST Cofin98 grant

# 1 Introduction

The complexity of ground ozone dynamics and of its relation to meteorological variables and precursor pollutants is well known. See e.g. Galbally et al. (1986), David Cooper and Alley (1994) and Fishman and Crutzen (1978). Statistical modelling at various time scales plays a central role and has received great attention in the past few years. Whenever, for trend analysis, low pass filter have been considered, some works discussed here consider daily averages or maxima. These are the standard daily summaries both from the statistical viewpoint and the environmental protection viewpoint. Other works focus on hourly data to analyze the interactions between precursor pollutants or meteorological variables and ozone.

Robeson and Steyn (1990) and Milionis and Davis (1994) reviewed previous forecasting models for daily maxima showing that the time series approach is better than simple regression techniques. Bloomfield et al. (1996) built up a nonlinear regression model with meteorological covariates. Galbally et al. (1986) considered nonlinear regression models for hourly data with precursors, meteorological variables and 24 hours lagged ozone concentration. Analyzing the residuals they noted some modellable autocorrelations and in their Fig. 13 we note some hints of long memory too.

In order to investigate the relationship between $NO$, $NO_2$ and $O_3$, Kuang-Jung Hsu (1992) applied $VAR$ models to hourly data. His data come from stacking up 22 partially not consecutive days and his technique provides a simple tool for preliminary multivariate analysis. Nevertheless, it seems that long memory, nonlinearity, heteroscedasticity and seasonality are underscored in this approach.

The idea of heteroscedasticity is not new in ground ozone statistical models. For example, see the Weibull heteroscedastic model discussed in Cox and Chu (1993). Moreover, Graf-Jacottet and Jaunin (1998) pointed out that homoscedastic time series models are appropriate for monthly or weekly averaging whilst, for daily averages, the use of conditionally autoregressive heteroscedastic ($ARCH$) model is appropriate.

Lewis and Ray (1997) used Time Series Multivariate Adaptive Regression Spline ($TSMARS$) to build $ASTAR$ type models for daily sea temperatures. $ASTAR$ acronimizes Adaptive spline Threshold AutoRegression and whenever the environmental case used was daily water temperatures instead of our hourly air pollution, this model has various points in common with the model to be introduced in the next section. Lewis and Ray's approach is quite flexible and has various advantages but it neither covers nor eliminates completely heteroscedasticity and long memory. Either heteroscedasticity is due to non observed covariates/dynamics or nature itself is heteroscedastic,

we choose to *model* it in some way.

In this paper we introduce a parametric model class for hourly data which encompasses covariates, multiple seasonality, long memory, heteroscedasticity, nonlinearity and good performance on the right tail. An application shows how to use these models for description and interpretation. Statistical monitoring and diagnostic may be applied to these kinds of models following e.g. Fassò (1997*c*) and (1998) to environmental protection and data validation. Forecasting ability is also discussed.

As will be apparent, the resulting models are more complex and with more parameters than most of those referenced above. This is not surprising if we compare a daily average model with an hourly data model which has richer dynamics. Hence, as long as models are only approximation to reality, we accept to increase the number of parameters as the data increase. Moreover, we will get models both able to fit the main, usually low, average pollution dynamics and to fit also the high pollution dynamics which is less frequent in time but not less important for environment and population protection.

In section 2, we introduce a class of stochastic models. In particular in subsection 2.1 we consider a *basic* autoregressive linear heteroscedastic model with *seasonal* fractional integration at lag 1 and 24 and $ARCH$ components. This model will be acronimized $SFI-ARX-ARCH$. In order to improve the approximation to the *true complex* dynamics by making our approximation *local* instead of *global*, in subsection 2.2, we extend the model to the threshold $SFI-SETARX-ARCH$ model which is piecewise linear heteroscedastic. In section 3, we discuss the estimation technique based on nested iterations of weighted nonlinear least squares. Moreover, in subsection 3.3, we define local or conditional statistics corresponding, for example, to usual autocorrelations or $R^2$ for threshold model identification and validation.

The application is described in section 4, where the model is applied to pollutants and meteorological data from a traffic air pollution station located in Bergamo, Italy. We then check the ability of the model to detect the dynamics or what we would call partial correlations in a multiple regression context.

In the concluding section 5 we discuss the results and warn about spurious correlation conclusions.

# 2 Model and notation

## 2.1 The basic model

Let us first consider a simplified version of the model which can be acronimized by $SFI-ARX-ARCH$ from Seasonal Fractionally Integrated AutoRegressive process with eXogenous variables and AutoRegressive Conditionally Heteroscedastic errors. In this case, the hourly observations $y_t$, $t = 0, \pm 1, \pm 2, ...$ obey the following defining equations:

$$x_t = (1-B)^{\delta_1}\left(1-B^{24}\right)^{\delta_2} y_t = \nabla(B)\, y_t \quad \text{say} \tag{1a}$$

$$x_t = \theta'\varphi_{t-1} + \varepsilon_t h_{t-1}, \tag{1b}$$

where $B$ is the back shift operator $B : By_t = y_{t-1}$ and prime means vector transposition.

In equation $(1b)$, the regressor vector $\varphi_{t-1}$ contains both lagged differentiated observations $x_t$ and exogenous stochastic or deterministic components $u_1, ..., u_r$, i.e.:

$$\varphi_{t-1} = \left(x_{t-1}, \ldots, x_{t-p}, u_{1,t}, \ldots, u_{1,t-s_1}, \ldots, u_{r,t-s_r}\right)'.$$

The conditional heteroscedasticity component, applied to the residuals $e_t = x_t - \theta'\varphi_{t-1}$, is given by the equation

$$h_{t-1}^k = \beta_0 + \sum_{j=1}^{q} \beta_j |e_{t-j}|^k, \tag{2}$$

where $k = 1$ or 2. Moreover, the adjusted residuals $\varepsilon_t = \frac{e_t}{h_{t-1}}$ are assumed to be independents with a common distribution such that $E|\varepsilon_t|^k = 1$ for $k = 1$ or 2.

The model parameters are denoted by $\psi = (\delta', \theta', \beta')'$ where $\delta = (\delta_1, \delta_2)'$, $\theta = (\alpha', \gamma')'$ and $\alpha$, $\gamma$ and $\beta$ are the coefficient vectors of the autoregressive, the exogenous and the conditional variance parts respectively.

**Remark 1 (*LRD*)** *The polynomial* $\nabla(B)$ *with fractional exponents* $0 \square \delta_j < \frac{1}{2}$, $i = 1, 2$, *accounts for stationary long memory* (*LRD*). *The two factors in* $\nabla(B)$ *can be written as infinite power series with* slowly *decaying terms. For example, see Beran* (1994) , *page 60,*

$$(1-B)^{\delta} = \sum_{j=0}^{\infty} \binom{\delta}{j} (-1)^k B^j$$

*where* $\binom{\delta}{j} = \frac{\Gamma(\delta+1)}{\Gamma(j+1)\Gamma\delta-j+1}$ *and* $\Gamma()$ *denotes the gamma function.*

*With abuse of terminology, the daily component* $(1-B^{24})^{\delta_2}$ *can be termed* seasonal. *Recently Arteche and Robinson* (1999) *reviewed seasonal fractional differencing and Hassler* (1994) *discussed so-called* rigid *and* flexible *models. In our case, one could use different cyclical differentiations, e.g. weekly at lag* $24 \cdot 7$, *or monthly, or properly seasonal at lag* $24 \cdot 365$. *In the last case, of course, there is a multiple of one year loss of data. We found weekly components unnecessary for our data. A similar long memory component could be used in the ARCH equation giving a* $SFI-GARCH$ *(see e.g. Baillie, Bollerlsev and Mikkelsen,* 1996*). For our data this has been unnecessary so we omit this component.*

**Remark 2** (*AR*) *The short term dynamics is given by the vector* $\theta = (\alpha', \gamma')'$. *In particular* $\alpha(B) = 1 - \sum_{j=1}^{p} \alpha_j B^j$ *is a standard autoregressive polynomial operator. The regressors* $u's$ *may include constant, trend and cyclical components, meteorological covariates, ozone precursors, and spatially lagged observations. In our model we do not use spatial modeling because of lack of data.*

**Remark 3** (*ARCH*) *The conditional heteroscedastic component is important in order to model the varying model precision or the varying forecasting precision. In finance and econometrics (see e.g. Engle,* 1995*) usually* $k = 2$ *gives the conditional variance model. Both in air pollution and in meteorological applications, the absolute errors ARCH model with* $k = 1$ *has been used (see e.g. Tol,* 1996*, and Graf-Jacottet and Jaunin,* 1998*), and, after comparing the results in our case, we agree with this practice. Gaussian parametric test of Lee and King* (1993) *or rank test described in Fassò* (1997*a and b*) *may help.*

**Remark 4** (*Kurtosis*) *The adjusted residual distribution is often assumed to be Gaussian* $N\left(0, \left(\frac{\pi}{2}\right)^{2-k}\right)$, $k = 1, 2$. *In our case, in order to account for extra kurtosis in the residuals (see e.g. Bollerslev and Engle* (1986)*, Bollerslev* (1987) *and Fassò* (1995)*), we assume that* $k = 1$ *and* $\varepsilon$ *has the distribution of* $\frac{t_\nu}{m_\nu}$ *where* $t_\nu$ *is a Student's t random variable with* $\nu > 4$ *degrees of freedom,*

$$m_\nu = E|t_\nu| = 2\sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{(\nu-1)\,\Gamma\left(\frac{\nu}{2}\right)}. \tag{3}$$

### 2.1.1 Stability conditions

Since after filtering for the regressors, our model defines a fractionally integrated $AR$ model, we see that stationarity conditions require that the poles, or the solutions $\lambda_j$ of $\alpha\left(z^{-1}\right)=0$, are inside the unit circle, i.e. $\max|\lambda_j|<1$, and $0 \square \delta_1, \delta_2<\frac{1}{2}$. Moreover, stability of the heteroscedastic component is implied by $\sum_{j=1}^{q} \beta_j<1$.

In this case the model is stable in the sense that, if fed with bounded inputs $u_{t,j}$, it gives finite outputs $y_t$. In other words, these assumptions entails that using stochastic stationary inputs gives an output stochastic process which is asymptotically second order stationary.

## 2.2 The state dependent model

As is well known, air pollution data are often nongaussian showing heavily skewed distributions. This means that a large percentage of data (e.g. 75%) is related to low pollution whilst only a small part is the high pollution data. This part of the pollutant distribution is usually the point of interest in alarm management.

>From the dynamical point of view, as long as models are only approximation to reality, the locally linear approximates may be different at different levels and using a *global* model underscores the high pollution dynamics.

In order to account for these points the general approach of state dependent models introduced by Priestley (1988) can be used. Using this approach we are led to a model where the local approximation is a function of the history of regressors up to time $t$ and the past of $y_t$, namely:

$$\psi_t=\psi_t\left(y_{t-1}, y_{t-2}, \ldots, u_{1,t}, u_{1,t-1}, \ldots, u_{r,t}, u_{r,t-1}, \ldots\right).$$

A nonparametric implementation of this approach is considered in Fassò (2000*b*). In this paper we consider a finite state $SETAR$ version of $\psi_t()$ given by the Self Exciting Threshold AutoRegression described in details in Tong (1995).

To see this let

$$\psi_t=\psi(j) \quad \text{if } y_{t-1} \in D_j,$$

where $\{D_1, ..., D_m\}$ is an appropriate partition of the observation space. Usually, sets or levels $D_j$ are defined through thresholds, *i.e.*

$$D_j=\left\{a_{j+1}<y_{t-1}<a_j\right\},$$

as we will do in the application of section 4, but generalizations include more complex subsets of the observation space, for example one may define $\{y_{t-1-i}, u_{1,t-i}, u_{r,t-i}, i = 0, ..., k\} \in D_j$.

In this way we get an invariant nonlinear model which is piecewise linear heteroscedastic and is parametrized through the finite dimensional vector $\Psi = \left(\psi(1)', ..., \psi(m)'\right)'$. With abuse of terminology we will say *submodel* $D_j$ for $y_t \in D_j$ identified by $\psi(j)$ or estimated by some $\hat{\psi}(j)$. Similarly the nonthreshold model of section 2.1 will be called *global* model.

### 2.2.1 Stability conditions

The stability conditions of section 2.1.1 are now to be evaluated for each $\psi(j)$ in order to have local stability and hence global stability, see e.g. Grillenzoni (1997). Stability may be checked by long simulation as we did for the estimated model using observed or bounded random inputs.

# 3 Estimation

In order to get estimated values $\hat{\Psi}$ for $\Psi$ one could optimize a Student $t$ conditional loglikelihood. Baillie, Chung and Tieslau (1996) follow this approach for a nonthreshold low-dimensional $ARFIMA(0,0,2) \times (0,0,2)_{12} - GARCH(1,1)$. A general discussion on estimation of seasonal long memory models is given in Arteche and Robinson (1999). Least squares $SETAR$ estimation is discussed, for example, in Petruccelli (1986) and Tong (1995).

Due to the high dimensionality of our model, we use the simplified conditional Weighted Least Squares approach ($WLS$) applied to equations $(1a)$, $(1b)$ and (2) i.e.:

$$\hat{\Psi} = \arg\min_{\Psi} \sum \left( \frac{e_t(\Psi)}{h_{t-1}(\Psi)} \right)^2 \tag{4}$$

which extends the approximate Gaussian maximum likelihood estimate of Baillie, Chung and Tieslau (1996) to the case of the seasonal fractional threshold model. Using a least squares ($LS$) estimate instead of a maximum likelihood ($ML$) estimate does not seem a severe drawback in our case thanks to the large number of observations at hand.

We can assume thresholds known a priori or estimated using *e.g.* Chan and Tong (1986) and Cheng and Liu (1995). In section 4 we use the former approach based on percentiles.

Since the error distribution parameters $\nu(j)$ and $\psi(j)$ are asymptotically orthogonal parameters (see e.g. Fassò (1995)), they are estimated separately via $ML$ with a Student $t$ likelihood for each $j = 1, ..., m$.

## 3.1 Computational details

To solve equation (4), note that, if the *true* long memory coefficients $\delta = \delta^*$ where known, the following standard iterated empirical $WLS$ Algorithm 1 would successfully apply giving $\left(\hat{\theta}(\delta^*), \hat{\beta}(\delta^*)\right)$.

**Algorithm 1** *Take initial values for* $\hat{\beta}_0$. *Then iterate steps* 1. *and* 2. *below, for* $i = 1, 2, ...,$ *up to convergence.*

1. *compute* $\hat{\theta}_i(\delta^*) = \hat{\theta}\left(\delta^*, \hat{\beta}_{i-1}\right)$ *and residuals* $e_t\left(\hat{\theta}_i\right)$ *via standard* $WLS$;
2. *compute* $\hat{\beta}_i$ *via* $LS$ *applied to equation* (2).

In order to optimize jointly for $\Psi = (\delta, \theta, \beta)$ we use nested iterations. To do this, note that for given $\theta$ and $\beta$, we have a relatively easy two dimensional optimization problem and for given $\beta = \beta^*$ the $WLS$ step 2. of Algorithm 1 above is an ordinary $LS$ problem in transformed variables. So we compute

$$\hat{\delta}(\beta^*) = \arg\min_{\delta} \left( \min_{\theta|(\delta,\beta^*)} \sum \left( \frac{e_t}{h_{t-1}} \right)^2 \right) \tag{5}$$

via numerical optimization. For joint optimization, note that $\hat{\delta}(\beta)$ slowly varies with $\beta$, so the nested iterations given in the following Algorithm 2 has been efficiently and successfully used in the next section.

**Algorithm 2** *Take initial values for* $\hat{\beta}_0$. *Then iterate steps* 1. *and* 2. *below, for* $j = 1, 2, ...$ *up to convergence:*

1. *compute* $\hat{\delta}_j = \hat{\delta}\left(\hat{\beta}_{j-1}\right)$ *via equation* (5);
2. *compute* $\left(\hat{\theta}_j, \hat{\beta}_j\right) = \left(\hat{\theta}\left(\hat{\delta}_j\right), \hat{\beta}\left(\hat{\delta}_j\right)\right)$ *via iterated steps* 1. *and* 2. *of algorithm 1.*

The error distribution parameter estimates $\hat{\nu}(j)$ are then obtained by numerical optimization of Student $t$ loglikelihood with initial values given by the moment estimator

$$\hat{\nu}(j) = \frac{4\hat{k}(j) - 6}{\hat{k}(j) - 3} \quad \text{for } \hat{k}(j) > 3,$$

where $\hat{k}(j)$ is the sample kurtosis of adjusted residuals in level $D_j$.

## 3.2 Forecasting

The model introduced here has been used in the application of the next section mainly with descriptive and monitoring purposes. Nevertheless, we discuss here from the general point of view its use as a conditional forecasting model. Conditional here means conditional to covariates $u_t$.

One step forecasting is simple thanks to the $AR$ structure of our model. Moreover, the $ARCH$ components give forecasts of the forecast precision. For example, under Gaussian assumption, approximated 95% forecast intervals are given by $\hat{y}_t \mp 2h_t$. In our case, we use intervals given by $\hat{y}_t \mp t_\alpha \frac{h_t}{m_v}$ where as usual $t_\alpha$ is the $100\,(1-\alpha)\,\%$ percentile of the Student $t$ distribution with $\nu$ degrees of freedom and $m_v$ is given in equation (3).

Multistep forecasting requires more than simply iterating one step forecasts because $SETAR$ skeletons usually have multiple fixed points. Computer intensive Monte Carlo forecasting seems to give good results, see Clements and Smith (1999). Alternatively, for $k$ steps forecasts ad hoc regression models may be used following for example Bhansali (1999) and references therein.

## 3.3 Conditional and summer moments and estimates

In order to study threshold models it is useful to use conditional moments and correlations. For example we will use $E\left(y_t y_{t-h} | y_{t-1} \in D_j\right)$ and its sample counterpart $\frac{1}{n_j}\sum_{y_{t-1}\in D_j} y_t y_{t-h}$, where $n_j$ is the number of observations in $D_j$, or the conditional determination coefficients for model $\hat{y}_t$ i.e. $R\left(D_j\right)^2 = 1 - E\left(\left(y_t - \hat{y}_t\right)^2 | y_{t-1} \in D_j\right) / Var\left(y_t | y_{t-1} \in D_j\right)$, $j = 1, ..., m$, and their sample counterparts. Note that these last quantities are important to evaluate the *local* performance of $\hat{y}_t$ and, under standard assumptions, they measure explained variances with respect to the null models given by $E\left(y_t | y_{t-1} \in D_j\right)$, $j = 1, ..., m$.

Since, as is well known, ground ozone is related to solar radiation we restrict our models to the *summer* semester, ranging from Mar-1 to Sep-30 of every year. Hence the related *summer* statistics are given, for example, by $\frac{1}{n_S}\sum_{t\in S} y_t y_{t-h}$ with obvious symbol meanings. In the sequel we will use these summer statistics extensively except for periodogram. In the same way, we get the parameter estimates using the sum in equation (4) restricted to $t \in S$.

# 4 Application to Bergamo data

## 4.1 Model building philosophy

The objective of this application is to build an interpretative model for the ozone data based on observations at the pollution station. To do this, we choose a path to the final model starting from a single time series model and then introducing meteorological and precursor covariates. Moreover, we choose to first introduce deterministic components such as linear trend, annual, weekly, daily and 12-hour cycles and then autoregressive terms and, among covariates, first meteorological variables and finally precursors. One could choose to filter out this $LS$ cyclical components and then to separately fit the other components. This is usually done in order to retain multistep forecasting ability. Since our primary objective here is an explanatory model, in order to see the joint significance, we choose to simultaneously fit both types of components; for example often observed solar radiation ruled out the daily cyclical deterministic components.

Building the long and short memory $AR$ components was based, at the initial stage, on classical looking at second moment statistics of data and residuals. At the advanced stage entering or dropping possibly lagged covariates was based on looking at classical $t$ statistics together with variation in $AIC$, $BIC$ and $R^2$ statistics in validation data set. Residual and squared residual correlations were also taken into account in the validation data set, both globally and conditionally on each level $D_j$. $ARCH$ validation has been based on summer autocorrelations of squared studentized residuals, namely $\left(\frac{e_t}{h_t}\right)^2$.

In reading approximate attained significances we kept in mind the large amount of data used so we looked not only at p-values but even at the absolute values of various correlations. Moreover, in validation data set residual correlation analysis we have take into account the increase in variance of residual correlations with respect to the independent case as explained in Fassò (2000$a$) for the standard $ARMAX$ case.

At an intermediate phase of model building we introduced a linear trend which was highly significant and physically relevant especially in high pollution levels. We dropped the linear trend in the final model presented here because it is heavily dependent on the estimation data set which shows an upward trend, whilst the validation data set shows a decreasing pattern.

## 4.2 Data set description

The observations run from Jan-1-93 to Dec-31-97. The data of the first three years amounts to 26'280 observations and the 12'960 summer data have been used for estimation while the remaining two years have been used for model selection and validation.

In Figure 1, we see $O_3$ in *ppb* with ten days summer correlogram, two years summer correlogram sampled every 10 days, and smoothed periodogram. The daily periodicity with long memory and yearly periodicity on both correlograms and periodogram with the outstanding daily harmonic frequencies $\omega_j = \frac{2\pi}{24}j$, $j = 1, 2$ is evident. The average day in Figure 2 explains the 12 hours periodicity of $O_3$ due to the complex relationship with solar radiation ($TSR$) and the precursors $NO$ and $NO_2$. Unfortunately, our data set does not include $VOC$ or benzene so some spurious correlations are possible and these cyclical components are not completely explained by the covariates introduced as some $AR$ coefficients are still needed at the corresponding lags in the final model.

Other winter pollutants like $SO_2$ and suspended particulate do not enter the model as expected. The meteorological covariates entering the model are $TSR$, temperature, humidity and wind speed. Among these, solar radiation explains the main daily periodicity. Rain and pressure did not enter the model.

## 4.3 Estimated models

After trying various numbers of subsets ($m = 1, ..., 7$) with thresholds based on percentiles, using the minimum $AIC$ criterium, we arrived at a model with $m = 3$ levels. The final thresholds are approximately the $75^{th}$ and $98^{th}$ percentiles and the corresponding ozone subsets are $D_1 = \{100 < y\}$, $D_2 = \{43 < y < 100\}$ and $D_3 = \{y < 43\}$. In this way we get a model with good performances for high pollution without loosing anything in the large 75% low pollution level. In this part of the ozone distribution we get similar results to the nonthreshold model $SFI - ARX$. Increasing the value of $m$ improves fitting, as measured by conditional residual whiteness in the low pollution level, but reduces $R^2$ statistics only beyond the forth digit and $AIC$ is flat. Moreover, it does not reduces heteroscedasticity.

To search for other kinds of nonlinearities, we used the approach of Young and Beven (1994) based on correlating the recursive least squares ($RLS$) estimate functions to the lagged regressors. To do this, we used the $RLS - ARCH$ estimates introduced in Fassò (2000*b*) . Using this approach none clear nonlinear relationship resulted. In order to further test the model complexity,

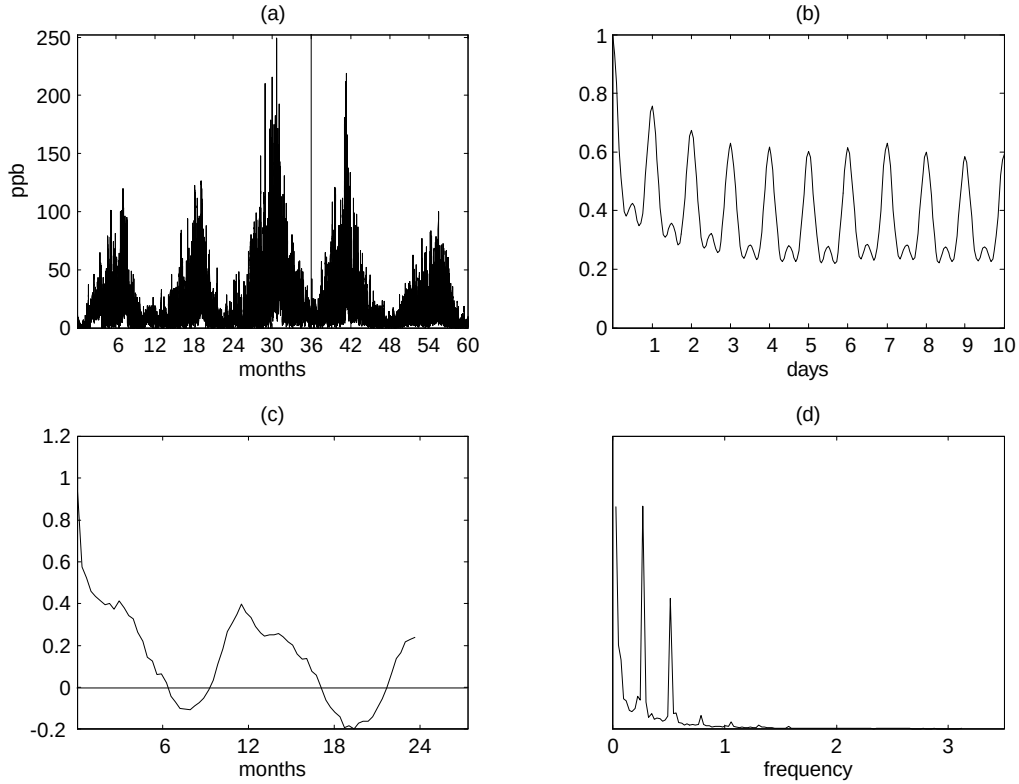


Figure 1: $O_3$ data: $(a)$ estimation and validation data set (93-97); $(b)$ 10 days summer correlations (93-95); $(c)$ 2 years summer correlations at 10 days (93-95); $(d)$ smoothed periodogram.

we directly introduced interactions among covariates given by product type regressor terms in the threshold model. In particular we used $u_{t-h,i}u_{t-k,j}$, $i \geq j$ and $h, k = 1, 2$ and, whenever they slightly improve $R^2$ in the validation data set of about 0.8%, they increase markedly the residual correlations in the validation data set. Moreover, introducing interactions between covariates and ozone, $y_{t-1}u_{t-h,i}$, $h = 1, 2$, gives unstable models. For these reasons we did not include any second and higher order interaction.

In Table 1, various $R^2$ from the validation data set are reported. In the first line we find the *naive* deterministic model

$$y_t = \gamma_1 + \gamma_2 \cos\left(\frac{2\pi t}{24 * 365}\right) + \gamma_3 \sin\left(\frac{2\pi t}{24 * 365}\right) + \gamma_4 \cos\left(\frac{2\pi\,(t-9)}{24}\right)$$
$$+ \gamma_5 \sin\left(\frac{2\pi\,(t-9)}{24}\right) + \gamma_6 \cos\left(\frac{2\pi\,(t-20)}{12}\right) + \gamma_7 \sin\left(\frac{2\pi\,(t-20)}{12}\right) + \varepsilon_t$$

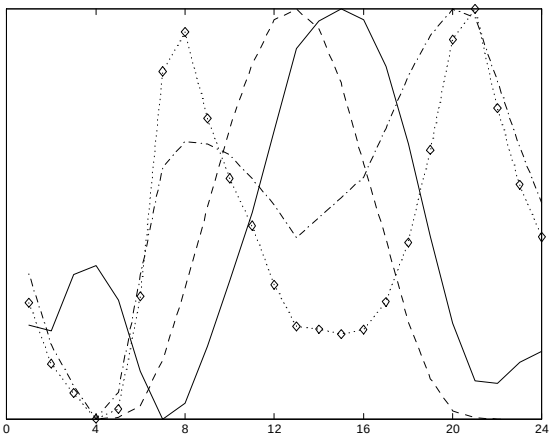


Figure 2: Time of day averages (rescaled) lines: $O_3$ solid, $TSR$ dashed, $NO$ dotted diamonds, $NO_2$ dash-dotted.

estimated via $LS$. The in sample fitting gives $R^2 = 0.400$ which falls down in the validation data set giving a useless model. Next, as a reference for the models to be discussed below, we have the persistence or random walk model,

$$y_t = y_{t-1} + \varepsilon_t.$$

| Model | $D_1$ | $D_2$ | $D_3$ | *Global* |
|---|---|---|---|---|
| *Deterministic* | – | – | – | 0.128 |
| *Random walk* | 0.241 | 0.697 | 0.904 | 0.891 |
| $SFI - AR - ARCH$ | 0.791 | 0.698 | 0.752 | 0.904 |
| $SFI - SETAR - ARCH$ | 0.821 | 0.711 | 0.755 | 0.907 |
| $SFI - ARX - ARCH$ | 0.814 | 0.782 | 0.829 | 0.931 |
| $SFI - SETARX - ARCH$ | 0.895 | 0.794 | 0.830 | 0.936 |

Table 1: $R^2$ in the validation data set.

The apparently large global fitting $R^2 = 0.891$ is essentially due to the large and smooth daily periodicity of Figure 2 and drops in the high pollution subset where $R^2(D_1) = 0.241$.

Introducing $AR$ components improves the fitting in high pollution levels as shown by the $SFI - AR - ARCH$ model whose estimated parameters are given in Table 2. This is enhanced using the threshold version $SFI - SETAR - ARCH$.

Introducing covariates gives the final models with and without thresholds, which are reported in Table 3. We see a substantial increase both of global and local fitting. Correspondingly we see a decrease of the unexplained dynamics. In particular we note a decrease of all $LRD$ components $\delta_j's$, a reduction in the number of non zero coefficients of short memory $AR$ components and a reduction of maximum poles, $\max(|\lambda_j|)$.

$LRD$ deserves some more comments. As a matter of fact, the full time series semiparametric Kunsch's estimate of $\delta_1$, discussed in Robinson (1995), being $\delta_1 = 0.56$, is outside the stationary set $|\delta_1| < \frac{1}{2}$ (see Robinson and Henry (1999) for the seasonal version of Kunsh estimate). Our $\delta_1$ and $\delta_2$ estimates decrease first to the stationary value of the global $SFI-AR$ model and further to the threshold and/or covariate models with smaller values for the $SFI-SETARX$ model of Table 3. In Table 2, note the unstable value of $\delta_1 = 0.6724$ for high pollution data set $D_1$of $SFI-SETAR$ model.

Now, let us comment on some general points which are common to models of both Table 2 and 3. The threshold models estimated show differences in subsets $D_1, D_2$ and $D_3$ in level, dynamics and heteroscedasticity but kurtosis parameter $\nu(j)$ are not significantly different and can be estimated on the pooled sample. This is confirmed by the symmetry test based on asymptotic gaussianity of estimates (see e.g. Anderson, 1971, § 5.3.6). The same testing problem could be addressed by some extension of the likelihood ratio test of Chan and Tong (1990). Not surprisingly, the 75% low pollution submodels $D_3$ are fairly similar to the related global models. Moreover, for the global and $D_3$ models, both the $AR$ and $ARCH$ components are richer than the high pollution submodels $D_1$ and $D_2$. This may be a point for a more complex dynamics of low pollution. A similar remark holds for $LRD$ which is stronger in the global and $D_3$ models. Moreover, $LRD$ is reduced by introducing covariates and its daily component vanishes in the high pollution $D_1$ data set of Table 3.

Focusing on the final models of Table 3, we see that metrological variables and cyclical components are important in $D_3$ and in the global model, whilst their role is reduced in high pollution data sets. Except $TSR$, which generally drop out daily $\sin-\cos$ terms of Table 2. Moreover, its coefficient at lag 0 does not change sign in various submodels but gets larger for high pollution $D_1$. This agrees with the nonlinear photochemical dynamics of ozone.

| | | SFI-SETAR-ARCH | | | | | | SFI-AR-ARCH | |
|---|---|---|---|---|---|---|---|---|---|
| | | $D_1$ | | $D_2$ | | $D_3$ | | | |
| Parameter | Lag | value | std | value | std | value | std | value | std |
| $SFI$ | 1 | 0.6724 | 0.0121 | 0.4038 | 0.0115 | 0.4596 | 0.0078 | 0.4531 | 0.0093 |
| | 24 | 0.1416 | 0.0029 | 0.2158 | 0.0042 | 0.2042 | 0.0032 | 0.2347 | 0.0042 |
| $AR$ | 1 | 0.5437 | 0.0563 | 0.5411 | 0.0216 | 0.5409 | 0.0123 | 0.5694 | 0.0094 |
| | 2 | -0.0949 | 0.0581 | – | – | -0.0657 | 0.0142 | -0.0518 | 0.0094 |
| | 3 | – | – | – | – | -0.0272 | 0.0135 | – | – |
| | 4 | – | – | -0.0304 | 0.0179 | -0.0387 | 0.0114 | -0.0410 | 0.0074 |
| | 7 | – | – | – | – | -0.0374 | 0.0086 | -0.0312 | 0.0075 |
| | 8 | – | – | – | – | – | – | -0.0127 | 0.0074 |
| | 9 | -0.0793 | 0.0470 | – | – | – | – | – | – |
| | 12 | – | – | – | – | – | – | – | – |
| | 15 | -0.0886 | 0.0451 | -0.0218 | 0.0166 | -0.0290 | 0.0079 | -0.0373 | 0.0061 |
| | 23 | – | – | 0.0637 | 0.0190 | 0.0488 | 0.0098 | 0.0733 | 0.0074 |
| | 24 | – | – | -0.1036 | 0.0238 | -0.0977 | 0.0131 | -0.1074 | 0.0099 |
| | 25 | – | – | 0.0734 | 0.0206 | 0.0684 | 0.0107 | 0.0920 | 0.0081 |
| $\max(\lvert\lambda_j\rvert)$ | | 0.8868 | | 0.9251 | | 0.9194 | | 0.9283 | |
| Const | – | -3.7534 | 1.9446 | -0.3914 | 0.6705 | 1.3359 | 0.0876 | 1.0882 | 0.1095 |
| Annual sin | – | – | – | – | – | -0.5225 | 0.1022 | -0.3624 | 0.1084 |
| Annual cos | – | – | – | -2.2134 | 0.6798 | – | – | -0.3225 | 0.1223 |
| 24 h. sin | – | 7.3442 | 2.1220 | 2.7125 | 0.3315 | 0.8272 | 0.0878 | 0.9239 | 0.0750 |
| 24 h. cos | – | 11.9865 | 1.6756 | 2.6457 | 0.3675 | 1.0964 | 0.0789 | 1.0570 | 0.0694 |
| 12 h. sin | – | – | – | 1.1559 | 0.3361 | 0.7425 | 0.0833 | 0.7441 | 0.0720 |
| 12 h. cos | – | – | – | -3.8894 | 0.3436 | -1.0692 | 0.0832 | -1.2460 | 0.0724 |
| ARCH | $\beta_0$ | 6.3591 | 0.7664 | 4.0317 | 0.3183 | 1.5837 | 0.0869 | 1.7535 | 0.0706 |
| | 1 | 0.1660 | 0.0429 | 0.1096 | 0.0182 | 0.1143 | 0.0107 | 0.1553 | 0.0070 |
| | 2 | 0.1673 | 0.0442 | 0.0694 | 0.0185 | 0.1409 | 0.0107 | 0.1275 | 0.0071 |
| | 3 | – | – | 0.0587 | 0.0188 | 0.0780 | 0.0106 | 0.0822 | 0.0071 |
| | 4 | – | – | 0.0381 | 0.0187 | 0.0640 | 0.0105 | 0.0590 | 0.0071 |
| | 5 | – | – | 0.0441 | 0.0188 | 0.0438 | 0.0102 | 0.0565 | 0.0071 |
| | 6 | – | – | 0.0492 | 0.0189 | 0.0597 | 0.0100 | 0.0539 | 0.0070 |
| | 24 | – | – | 0.0959 | 0.0185 | 0.1473 | 0.0096 | 0.1388 | 0.0068 |
| $\Sigma\hat{\beta}_j$ | | 0.3333 | | 0.4650 | | 0.6480 | | 0.6731 | |
| $\hat{\nu}$ | | 4.1197 (std 0.1542) | | | | | | 4.0812 | 0.1542 |

Table 2: Estimated $SFI-SETAR-ARCH$ and $SFI-AR-ARCH$ models.

Other meteorological covariates given by nonlinear functions of meteorological variables found e.g. in Robeson and Steyn (1990) and Galbally et al. (1986) have been also considered here. In particular, after introducing the lagged wind speed, the reciprocal wind component $\frac{1}{1+wind\ speed}$ was not significant nor preferable to linear components when looking both at $R^2$ and/or $AIC$ and residual correlations in the validation data set. The squared temperature component would enter the global model $SFI-ARX-ARCH$ using a global criterium ($R^2$ or $AIC$ in the validation data set improve slightly) but would not improve the high pollution fitting $R^2\left(D_j\right)$, $j=1,2$. Moreover, it would worsen the residual correlations in the validation data set.

| | | SFI-SETARX-ARCH | | | | | | SFI-ARX-ARCH | |
|---|---|---|---|---|---|---|---|---|---|
| | | $D_1$ | | $D_2$ | | $D_3$ | | | |
| Parameter | Lag | value | std | value | std | value | std | value | std |
| $SFI$ | 1 | 0.3172 | 0.0034 | 0.2516 | 0.0109 | 0.3804 | 0.0071 | 0.3776 | 0.0084 |
| | 24 | – | – | 0.1168 | 0.0017 | 0.1448 | 0.0029 | 0.1641 | 0.0032 |
| $AR$ | 1 | 0.9445 | 0.0524 | 0.7534 | 0.0226 | 0.5410 | 0.0109 | 0.5930 | 0.0077 |
| | 2 | -0.2912 | 0.0464 | -0.0772 | 0.0201 | -0.0204 | 0.0099 | – | – |
| | 4 | – | – | – | – | -0.0243 | 0.0077 | -0.0262 | 0.0062 |
| | 12 | 0.0510 | 0.0289 | – | – | – | – | – | – |
| | 15 | – | – | 0.0240 | 0.0101 | -0.0126 | 0.0059 | -0.0122 | 0.0049 |
| | 23 | – | – | – | – | 0.0317 | 0.0080 | 0.0488 | 0.0065 |
| | 24 | – | – | – | – | -0.0934 | 0.0106 | -0.0991 | 0.0086 |
| | 25 | – | – | – | – | 0.0607 | 0.0086 | 0.0785 | 0.0069 |
| $\max(\lvert\lambda_j\rvert)$ | | 0.8635 | | 0.8604 | | 0.9147 | | 0.9211 | |
| Const | – | 12.9134 | 5.7303 | 2.4639 | 0.7804 | 2.5406 | 0.4236 | 2.3062 | 0.3810 |
| Annual c. | 0 | – | – | 1.7830 | 0.6168 | 0.6684 | 0.1593 | 0.6877 | 0.1485 |
| 12 h c. | 0 | – | – | – | – | 0.5302 | 0.0917 | 0.5363 | 0.0862 |
| $TSR$ | 0 | 0.2288 | 0.0247 | 0.1451 | 0.0086 | 0.0706 | 0.0036 | 0.0684 | 0.0033 |
| | 2 | – | – | -0.0135 | 0.0088 | – | – | – | – |
| | 5 | – | – | – | – | – | – | 0.0077 | 0.0026 |
| $NO$ | 0 | -1.4724 | 0.1573 | -0.3686 | 0.0195 | 0.0160 | 0.0014 | 0.0179 | 0.0014 |
| | 1 | 1.4296 | 0.1762 | 0.3430 | 0.0320 | -0.0179 | 0.0015 | -0.0198 | 0.0014 |
| | 2 | – | – | -0.0656 | 0.0228 | – | – | – | – |
| | 3 | – | – | – | – | -0.0050 | 0.0009 | -0.0051 | 0.0009 |
| $NO_2$ | 0 | -0.2502 | 0.0660 | -0.3652 | 0.0175 | -0.3246 | 0.0062 | -0.3708 | 0.0056 |
| | 1 | 0.6121 | 0.0856 | 0.4470 | 0.0274 | 0.2814 | 0.0066 | 0.3499 | 0.0082 |
| | 2 | -0.3657 | 0.0616 | -0.0856 | 0.0220 | – | – | -0.0245 | 0.0058 |
| | 6 | – | – | 0.0203 | 0.0082 | 0.0218 | 0.0032 | 0.0242 | 0.0029 |
| Temp | 0 | – | – | – | – | 0.5405 | 0.0897 | 0.7579 | 0.0871 |
| | 1 | – | – | – | – | -0.5131 | 0.0876 | -0.7276 | 0.0862 |
| Humidity | 0 | -0.2208 | 0.0683 | – | – | -0.3431 | 0.0259 | -0.3069 | 0.0239 |
| | 1 | – | – | – | – | 0.4700 | 0.0395 | 0.4070 | 0.0367 |
| | 2 | – | – | – | – | -0.1445 | 0.0228 | -0.1172 | 0.0213 |
| Wind s. | 0 | – | – | – | – | 1.1024 | 0.1156 | 0.9422 | 0.1035 |
| | 2 | – | – | – | – | -0.6375 | 0.1153 | -0.7047 | 0.1026 |
| ARCH | $\beta_0$ | 4.0756 | 0.6087 | 3.6952 | 0.2167 | 1.3465 | 0.0761 | 1.4903 | 0.0625 |
| | 1 | 0.2287 | 0.0413 | 0.1701 | 0.0169 | 0.1639 | 0.0116 | 0.1984 | 0.0073 |
| | 2 | 0.1866 | 0.0434 | 0.0486 | 0.0170 | 0.0987 | 0.0115 | 0.1106 | 0.0074 |
| | 3 | – | – | – | – | 0.0585 | 0.0110 | 0.0720 | 0.0074 |
| | 4 | – | – | – | – | 0.0950 | 0.0104 | 0.0878 | 0.0073 |
| | 24 | – | – | 0.0853 | 0.0175 | 0.1394 | 0.0101 | 0.1209 | 0.0071 |
| | 48 | – | – | 0.0501 | 0.0191 | 0.0974 | 0.0094 | 0.0795 | 0.0070 |
| $\Sigma\hat{\beta}_j$ | | 0.4153 | | 0.3541 | | 0.6529 | | 0.6692 | |
| $\hat{\nu}$ | | 4.1871 std=0.1537 | | | | | | 4.2348 | 0.1482 |

Table 3: Estimated $SFI-SETARX-ARCH$ and $SFI-ARX-ARCH$ models.

$NO$ coefficients at lags 0 and 1 change their sign and magnitude in different submodels. This agrees with the idea that the ozone reducing effect of $NO$ is highly nonlinear and depends on other pollutants like $VOC$, which were not available for this study. To get further insight into this dynamics, let us consider the steady state effect of $NO$ on $O_3$, given by

$$\frac{\sum_{i\in no_j}\gamma_i(j)}{\sum_{i=0}^{\infty}\pi_i(j)},$$

where $\pi_j(j)$ are the coefficients of the series expansion of the full $AR$ component given by $\nabla_j(B)\alpha_j(B)$ and $no_j$ is the index set of lagged $NO$ terms using the notation of section 2.1 for submodel $D_j$. Of course this quantity has full asymptotic meaning for global models but here still summarize the input output relation in some sense. From Table 4, we see that this effect is always negative and larger in absolute value in $D_2$ than $D_1$, quite small in $D_3$ and in the global model.

The dependence of ozone on $NO_2$ seems more stable in absolute value at the various pollution levels but the variations among submodels are partially reversed having negative steady state effect smaller in $D_1$ and larger in absolute value in $D_3$. In $D_2$ the effect is positive and once more the relationship with ozone seems complex.

As shown in Figure 2 the daily variation of $NO_2$ is less deep than $NO$ and $O_3$. This explains the difference among the relationships of $NO_x'$s with ozone

| | $D_1$ | $D_2$ | $D_3$ | *Global* |
|---|---|---|---|---|
| $\left(\sum_{i=0}^{\infty}\pi_i(j)\right)^{-1}$ | 12.34 | 12.70 | 14.19 | 18.40 |
| $K(j)$ for $NO$ | -0.528 | -1.158 | -0.098 | -0.129 |
| $K(j)$ for $NO_2$ | -0.047 | 0.210 | -0.304 | -0.390 |

Table 4: Steady state effect for $SFI-SETARX$ and $SFI-ARX$ models.

## 4.4 Model validation

In Figure 3, we see the capability of Student's $t$ distribution to fit the unexplained dynamics of adjusted residuals of model $SFI-SETARX-ARCH$ of Table 3. In high pollution levels $D_1$ and $D_2$ we had approximated p-values given by 83.7% and 2.6%. Whenever a small skew arises in $D_3$, the adjusted residual variances agree with the Student $t_v$ distribution result given by

$$Var\left(\frac{t_\nu}{m_\nu}\right)=\frac{\nu}{\nu-2}\left(2\sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{(\nu-1)\,\Gamma\left(\frac{\nu}{2}\right)}\right)^{-2},$$

and the hypothesis that a common independent Student's $t$ distributed process generates the noise is roughly acceptable.

As explained before, for dynamics validation we use validation data set correlation analysis. Using the data available for the last two years, from Figure 4, we see that our model satisfies first and second order residual checks. In particular, from Figure 4.a, we see that some values exceed the

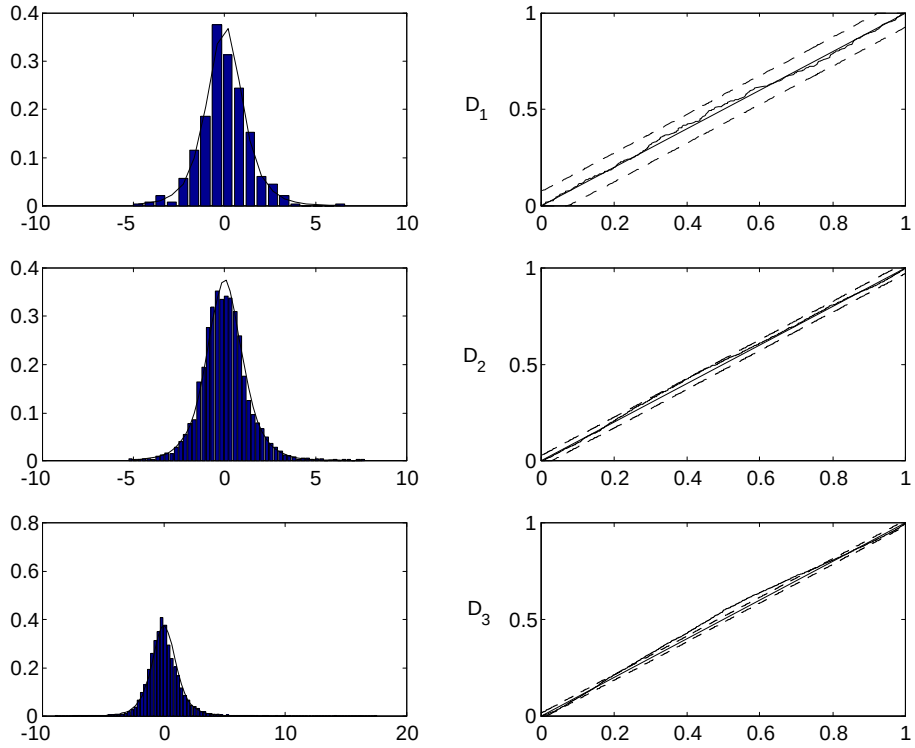


Figure 3: Student's $t$ fitting of adjusted residuals. Right pictures: probability plot with approximate Kolmogorov-Smirnov 97.5% bands.

95% band. Nevertheless, these correlations are definitely small as they do not exceed 0.065 in absolute value. Similar results hold for autocorrelation analysis in levels $D_1, ..., D_3$ and for long memory autocorrelations.

# 5 Conclusions and further developments

We have shown that, using complex nonlinear models for hourly air pollution data may enhance fitting, especially in high pollution levels. Moreover, using covariates may improve the single time series approach used, for example, for daily data in Graf-Jacottet and Jaunin (1994). In our case, the improvement pertains both fitting and stationarity.

The application was intended to illustrate the capability of the proposed model to capture the ozone time dynamics and to give a monitoring model in a specific case. The conclusions which have been drawn have an illustrative character and depend on the data available. In particular, wherever they are

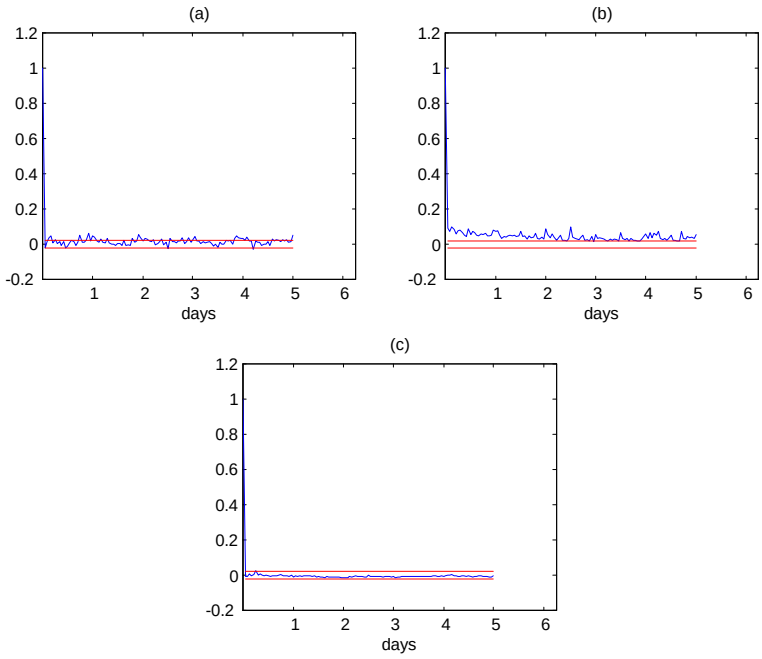


Figure 4: Validation data set residual analysis with 95% approximate bands. (a) residual correlogram; (b) squared residual correlogram; (c) squared adjusted residual correlogram.

quite good, they could be improved if other precursor pollutants like volatile organic components were available.

Using the conditional fitting approach we have shown that the fitting is quite good for high pollution, whilst, for low pollution, some unexplained complex dynamics still remains. Whenever we consider this a minor drawback for environmental protection, this point may be covered using other precursor pollutants and skewed residual distributions.

Model generalizations are possible in various directions. It is easy to cover long memory in the $GARCH$ component. This may be important for high frequency data. Moreover, multistep forecast strategies for this kind of data have to be investigated. On the one hand, we have computer intensive Monte Carlo methods; on the other hand, $k$ step forecasting ad hoc models may be used for each forecasting horizon $k$. In any case we feel that this task would not be simply a performance check on the model estimated here, as a forecasting model may be different from a descriptive model. In particular, the problem of meteorological variables and precursor forecasting would rise in forecasting applications.

# References


[1] Anderson TW. 1971. *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

[2] Arteche J, Robinson PM. 1999. Seasonal and cyclical long memory. In *Asymptotics, nonparametrics, and time series.* (Ghosh S. ed.); Marcel Dekker, New York: 115-148.

[3] Baillie RT, Bollerlsev T, Mikkelsen HO. 1996. Fractionally intergrated generalized autoregressive conditional heteroskedasticity. *J. Econometrics.* **74:** 3-30.

[4] Baillie RT, Chung C, Tieslau MA. 1996. Analysing inflation by the fractionally integrated $ARFIMA-GARCH$ model. *J.Appl.Econometrics*, **11**: 23-40.

[5] Beran J. 1994. Statistics for Long-Memory Processes, Chapmann & Hall.

[6] Bhansali RJ. 1999. Parameter estimation and model selection for multistep prediction of a tim series: a review. In *Asymptotics, nonparametrics, and time series* (Ghosh S. ed.), Marcel Dekker, New york. 201-226.

[7] Bollerslev T. 1987. A conditionally heteroscedastic time series model for speculative prces 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] Chan KS, Tong H,1986. On estimating thresholds in autoregressive models. *J. Time Series Analysis,* **7**, 3: 179-190.

[10] Chan KS, Tong H. 1990. On likelihood ratio tests for threshold autoregression. *J. R. Statist. Soc. B*, **52**, 3: 469-476.

[11] Cheng RCH, Liu WB. 1995. Confidence intervals for threshold parameters. In *Statistical modelling, proceedings of the 10th international workshop on statistical modelling.* Springer-Verlag, New York.

[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] David Cooper C, Alley FC.1994. *Air pollution control.* Waveland press. Prospect Heights.

[15] Engle RF. 1995. *ARCH: selected readings*, (Engle Ed.); Oxford University Press, Oxford.

[16] Fassò A. 1995. On some models and tests for heteroscedasticity, *Statistica*, **LV**, 1: 31-44.

[17] Fassò A. 1997*a*. *Test robusti per la rilevazione di componenti ARCH e GARCH.* Statistica, **LVII**, 3, 325-350.

[18] Fassò A. 1997*b*. *On a Rank Test for Autoregressive Conditional Heteroscedasticity.* Student, **2**, 2: 85-94.

[19] Fassò A. 1997*c*. *On some control charts for nonlinear ruptures.* Italian J. Appl. Statist., **9**, 1: 123-141.

[20] Fassò A. 1998. *One-sided Multivariate Testing and Environmental Monitoring.* Austrian Journal of Statistics, **27**, 1&2: 17-37.

[21] Fassò A. 2000*a*. Residual autocorrelation distribution in the validation data set. In printing on *J. Time Series Analysis*, **21**, 1.

[22] Fassò A. 2000*b*. *Recursive least squares with ARCH errors.* In printing.

[23] Fishman J, Crutzen PJ.1978. *The origin of ozone in the troposphere.* Nature, **274**: 855-858.

[24] Galbally IE, Miller AJ, Hoy RD, Ahmet S, Joynt RC, Attwood D. 1986. Surface ozone at rural sites in the Latrobe valley and Cape Grim, Australia. *Atmospheric Env.*, **20**: 2403-2422.

[25] Graf-Jaccottet M, Jaunin MH. 1998. Predictive models for ground ozone and nitrogen dioxide time series. *Environmetrics*, **9**: 393-406.

[26] Grillenzoni P. 1997. Optimized Adaptive Prediction. *J. Italian Statistical Society*, **6,** 1: 37-58.

[27] Hassler U. 1994. (Mis)specification of long memory in seasonal time series. *J. Time Series Analysis.*, **15**, 1:19-30.

[28] Kuang-Jung Hsu. 1992. Time series of the analysis of the interdependence among air pollutants. *Atmospheric Env.*, **26**, 4: 491-503.

[29] Lee JHH. King M. 1993. A locally most mean powerful based score test for $ARCH$ and $GARCH$ regression disturbances. *J. Business & Econ. Statist.*, 81, 819-825.

[30] Lewis PA, Ray BK. 1997. Modeling long-range dependence, nonlinearity, and periodic phenomena in sea surface temperatures using $TSMARS$. *JASA*, **92**, 439: 881-893.

[31] Milionis AE, Davis TD. 1994. Regression and stochastic models for air pollution - I. Review, comments and suggestion. *Atmospheric Env.*, **28**, 17: 2801-2810.

[32] Petruccelli JD. 1986. On the consistency of least squares estimators for a threshold $AR(1)$ model. *J. Time Series Analysis.*, **7**, 4: 269-278.

[33] Robeson SM, Steyn DG. 1990. Evaluation and comparison of statistical forecats models for daily maximum ozone concentrations. *Atmospheric Env.*, **24**, 2: 303-312.

[34] Robinson PM. 1995. Gaussian semiparametric estimation of long range dependence. The *Annals of Statistics*, **23**, 5: 1630-1661.

[35] Robinson PM, Henry M. 1999. Long and short memory conditional heteroskedasticity in estimating the memory parameter of levels. *Econometric Theory*, **15**: 299-336.

[36] Priestley M. 1988. *Nonlinear and nonstationary time series.* Academic Press, London.

[37] Tol RSJ. 1996. Autoregressive conditional heteroscedasticity in daily temperature measurements, *Environmetrics*, **7**: 67-76.

[38] Tong H. 1995. *Non-linear time series.* Clarendon Press, Oxford.

[39] Young PC, Beven KJ. 1994. Data-based mechanistic modelling and the rainfall-flow non-linearity. *Environmetrics*, **5**: 535-363.