

MONITORING ALGORITHMS FOR DETECTING CHANGES IN THE OZONE CONCENTRATIONS

SILVANO BORDIGNON¹ AND MICHELE SCAGLIARINI^{2*}

¹*Department of Statistics, University of Padova, via S. Francesco 33, 35121 Italy*

²*Department of Statistics, University of Bologna, via Belle Arti 41, 40126 Italy*

SUMMARY

The quality of data collected by air pollution monitoring networks is often affected by inaccuracies and missing data problems, mainly due to breakdowns and/or biases of the measurement instruments. In this paper we propose a statistical method to detect, as soon as possible, biases in the measurement devices, in order to improve the quality of collected data on line. The technique is based on the joint use of stochastic modelling and statistical process control algorithms. This methodology is applied to the mean hourly ozone concentrations recorded from one monitoring site of the Bologna urban area network. We set up the monitoring algorithm through Monte Carlo simulations in such a way to detect anomalies in the data within a reasonable delay. The results show several out of control signals that may be caused by problems in the measurement device. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS air pollution; monitoring networks; data quality; monitoring algorithms; changes detection

1. INTRODUCTION

In the context of air pollution prediction and control, the data collected by ambient air quality monitoring networks play an important role. Actually they allow one to pursue several purposes which include: (a) the assessment of concentration levels and the compliance status with air quality standards; (b) the determination of health and environmental impacts; and (c) the selection and monitoring of emission abatement strategies. The correctness of such analyses and air quality control policies depends heavily on the reliability of the collected data. Unfortunately the quality of these data sets is affected by several problems; among the relevant ones seem to be missing or invalid data and measurement inaccuracies. Missing and inaccurate data may result mainly from instruments failure, sampler biases, calibration and maintenance problems; see Davison and Hemphill (1987) and Batterman (1992) for a discussion on these points. As noted by Batterman (1992), these problems may be critical in interpreting air quality data.

Despite their importance, few methods which address these problems exist and rarely are in common use. In this paper we propose an on-line statistical procedure which can be used to detect, as soon as possible, biases in the measurement devices, in order to improve the quality of ambient network data. The methodology is based on the joint use of stochastic models and statistical process control algorithms. A state space model is used to describe the dynamics of the

* Correspondence to: M. Scagliarini, Department of Statistics, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy.

air pollution concentration and a GLR (Generalized Likelihood Ratio; Lorden 1971) type algorithm is employed to monitor the innovations obtained from the stochastic model.

Change detection algorithms have been well established in automated fault detection of controlled dynamic systems which describe industrial processes. A non-conventional sphere where these methodologies can be usefully employed is given by monitoring of environmental pollution processes and, in particular, of the air pollution process.

In fact, in a geographic area, the pollutants emission system is made up of industrial emissions, traffic and domestic heating plants. Moreover, this system is not independent of the atmospheric conditions, which influence the pollutants' concentration. The determination of a stochastic model for this complex system allows the 'common-cause of variability' to be explained, according to the definition given by Alwan and Roberts (1988). It is then possible to use monitoring algorithms to detect anomalies in the pollutants, the so-called 'special causes of variability' in the terminology of Alwan and Roberts (1988), that may be caused by problems in the measurement device. This procedure can be made fully automatic and routinely implemented as a complementary tool of the usual periodic control procedures of the monitoring site.

In this work we illustrate an application of the methodology to mean hourly ozone (O_3) concentration data recorded from one monitoring site of the Bologna urban area network.

The choice of O_3 data is motivated by the consideration that in Italy, as well as in other European countries, the photochemical pollutants show an increasing trend in concentration, often exceeding the warning limits, especially in summer. So it is interesting and important to inspect this pollutant.

For the choice of an appropriate monitoring algorithm, a GLR type algorithm is preferred, given the lack of *a priori* information about the change. Moreover, we propose a method to select the parameters for the design of the GLR rule.

The paper is organized as follows. Section 2 introduces the main statistical tools useful for the design of a monitoring algorithm in stochastic systems. Section 3 presents a brief description of our data and the results on model identification for the ozone concentration. A method to select the parameters of the GLR algorithm, based on the joint use of Monte Carlo simulation and theoretical results, is proposed and illustrated through Section 4. Empirical results and some concluding remarks are reported in Section 5.

2. MONITORING ALGORITHMS IN DYNAMIC SYSTEMS

Our monitoring problem can be stated as the problem of detecting a change in the parameters of a stochastic system. More precisely, let us consider a sequence $\{Y_k\}$ of observed random variables with density p_θ depending upon the parameter θ . Before the unknown change time t_0 , θ is constant and equal to θ_0 . After the change, θ is equal to θ_1 . The problem is to detect the occurrence of the change as soon as possible, given a fixed rate of false alarms before t_0 .

Statistical change detection algorithms have been well established in statistical process control: the most popular of them, like Shewart, CUSUM and EWMA control charts can be found in standard quality control textbooks, e.g. Montgomery (1997). However these algorithms, based on the assumption of $\{Y_k\}$ being a sequence of independent random variables, result in poor performances when the observations are autocorrelated. In this situation one possible general approach for change detection consists of splitting the task into: (a) a generation of 'residuals', which are ideally close to zero when no change occurs, and significantly different from zero after the change; and (b) a design of decision rules which solves the change detection problem as

reflected by the residuals. Natural candidates for the residuals of a given stochastic process Y_k are the innovations defined by $\varepsilon_k = Y_k - E_\theta(Y_k|Y_{k-1}, \dots, Y_1)$. $\{\varepsilon_k\}$ is a zero mean uncorrelated (independent if the process is Gaussian) sequence under the no change hypothesis. Therefore, change detection algorithms designed for uncorrelated or independent observations can be used over the innovations of a dependent sequence after suitable modifications.

Once the detection change problem is stated in this way, let us consider a particular case which is relevant for our empirical application. Let us assume that the dynamics of an observations sequence can be modelled by a linear time invariant stochastic system represented in state space form as

$$\begin{aligned} X_{k+1} &= FX_k + GU_k + W_k, \\ Y_k &= HX_k + JU_k + V_k, \end{aligned} \quad (1)$$

where X, U, Y , are the state, input and observation vectors, $\{W_k\}$ and $\{V_k\}$ are two independent Gaussian white noises sequences with covariance matrices Q and R , respectively, F is the state transition matrix, H the observation matrix, G and J the control matrices. Given the initial state $X_0 \sim N(\mu_0, P_0)$, the innovations sequence can be obtained through the Kalman filter recursions as

$$\varepsilon_k = Y_k - E_\theta(Y_k|Y_{k-1}, \dots, Y_1) = Y_k - HX_{k|k-1} - JU_k, \quad (2)$$

where $X_{k+1|k} = F(I - K_k H) + JU_k + FK_k Y_k$ is the one-step ahead prediction of the state; $K_k = P_{k|k-1} H^T (\Sigma_k)^{-1}$ is the Kalman gain; $\Sigma_k = HP_{k|k-1} H^T + Q$ is the estimated covariance matrix of the innovations; $P_{k+1|k} = FP_{k|k} F^T + Q$ is the estimated covariance matrix of $X_{k+1|k}$; and $P_{k|k} = (I - K_k H)P_{k|k-1}$ is the estimated covariance matrix of $X_{k|k}$. Under the no changes hypothesis, $\{\varepsilon_k\}$ is a Gaussian independent sequence of random variables with zero mean and covariance matrix Σ_k .

Let us assume that a change occurs at an unknown time instant t_0 . According to Basseville and Nikiforov (1993) we distinguish between additive changes, i.e. changes in a signal or linear system that result in changes only in the mean value of the observations, and non-additive changes, where changes occur in the dynamics of the signal or system. For our purposes it is sufficient to consider only additive changes. They are introduced in the state space model (1) in the following way:

$$\begin{aligned} X_{k+1} &= FX_k + GU_k + W_k + \Gamma\Psi_x(k, t_0), \\ Y_k &= HX_k + JU_k + V_k + \Xi\Psi_y(k, t_0), \end{aligned} \quad (3)$$

where Γ and Ξ are gain matrices which account for the change magnitude, whereas Ψ_x and Ψ_y are vectors representing the dynamic profile of the assumed changes. Clearly, if $k < t_0$, $\Psi_x = \Psi_y = 0$. The specification of the gain matrices and change profiles depends on the *a priori* knowledge on the change. It is worth noting that $\Gamma\Psi_x$ in the transition equation and $\Xi\Psi_y$ in the measurement equation represent models for detecting biases in the actuators and in the sensors, respectively. For example, of $\Psi_x = 0$, Ξ is a scalar and Ψ_y is a vector, the components of which are all zero except for the j th component, which equals one for $k \geq t_0$, then the model (3) corresponds to the onset of a bias in the j th component of Y , namely the j th sensor.

As we explained previously, the change detection algorithm is based upon the innovations. Therefore, we first investigate the behavior of the innovations sequence which results from the specified model of change. Since the peculiarity of our application consists of detecting changes in the air pollutant measurement device, it is sufficient to consider a particular form of (3), given by

$$\begin{aligned} X_{k+1} &= FX_k + GU_k + W_k, \\ Y_k &= HX_k + JU_k + V_k + vI_{k \geq t_0}, \end{aligned} \quad (4)$$

where Y_k is scalar, v is the unknown change magnitude and $I_{k \geq t_0}$ is the indicator function representing a step change. It can be shown that the innovation of this model is of the form

$$\varepsilon_k = \varepsilon_k^o + v\rho^*(k, t_0), \quad (5)$$

where ε_k^o refers to the innovation obtained from (4) without the change and $\rho^*(k, t_0)$ is the dynamic profile of the change. A closed-form expression for ρ^* , assuming the steady-state behavior of the Kalman filter, is given by

$$\rho^*(k, t_0) = I_{k \geq t_0} - \sum_{i=0}^{k-t_0-1} HF_*^i FKI_{k-i-1 \geq t_0}, \quad (6)$$

where $F_* = F(I - KH)$ and K is the steady-state Kalman gain. In summary, $\{\varepsilon_k\}$ is a Gaussian white noise sequence when no change occurs and a Gaussian independent sequence with mean $v\rho^*(k, t_0)$ after the change occurrence. Therefore, the detection of an additive change in the observations is equivalent to solving the following hypothesis testing problem on the innovations of the model (4):

$$H_0 : \{\varepsilon_k\} \sim N(0, \Sigma_k) \quad H_1 : \{\varepsilon_k\} \sim N(v\rho^*(k, t_0), \Sigma_k). \quad (7)$$

Many on-line change detection algorithms are based on the log-likelihood ratio. In the case of an unknown parameter after change, it is better to employ the GLR as a generalization of the CUSUM algorithm for this situation. The interest in this algorithm is justified by its good properties and by the possibility of adapting it to more complex situations, like the ones depicted by (3). The decision rule of the GLR algorithm, adapted to the situation described by model (4), specializes to

$$g_k = \max_{1 \leq j \leq k} \sup_v S_j^k \geq h, \quad (8)$$

where S_j^k is the log-likelihood ratio of the innovations from ε_j to ε_k and h is a conveniently chosen threshold. Given the Gaussianity of the innovations, an explicit expression for the $\sup_v S_j^k$ is

$$\sup_v S_j^k = \hat{v}_k(j) \left(\sum_{i=j}^k \rho^*(i, j) \Sigma_i^{-1} \varepsilon_i \right) - \frac{\hat{v}_k^2(j)}{2} \left(\sum_{i=j}^k \rho^{*2}(i, j) \Sigma_i^{-1} \right), \quad (9)$$

where Σ_i is the variance of ε_i and $\hat{v}_k(j)$ is the maximum likelihood estimate of the change magnitude at time k , assuming a change at time j . An expression for $\hat{v}_k(j)$ is given by

$$\hat{v}_k(j) = \frac{\sum_{i=j}^k \rho^*(i, j) \Sigma_i^{-1} \varepsilon_i}{\sum_{i=j}^k \rho^{*2}(i, j) \Sigma_i^{-1}}. \quad (10)$$

The other unknown of the procedure is the change time t_0 . It can be estimated with the aid of maximum likelihood estimation, which leads to an exhaustive search of this maximum for all possible past (i.e. before k) time instants. In order not to increase linearly the size of this search, t_0 is estimated by looking for the maximum value of S inside a finite window of fixed size M :

$$\hat{t}_{0k}(j) = \arg \max_{k-M+1 \leq j \leq k} S_j^k. \quad (11)$$

This is referred to by Lai (1995) as 'window-limited GLR schemes', and the underlying intuitive idea is that older changes have already been detected. The change magnitude estimate is finally

$$\hat{v}_k = \hat{v}_k(t_{0k}) \quad (12)$$

for $k = t_a$, where t_a is the alarm time.

In summary, the algorithm consists of the following steps: (a) detection of the change; (b) estimation of the change time and magnitude; and (c) updating the initial state and error covariance estimates for the Kalman filter using the change magnitude estimate. The first two steps are basic in GLR methodology; as for the third, the reason for updating the initial estimates is to give the Kalman filter more appropriate initial values after the detection of a change than the initial values given at the beginning of processing. For an operative solution to this problem, see Willsky and Jones (1976).

The working of the GLR algorithm requires values for the threshold h and the window size M to be chosen. As pointed out by Lai (1995), general criteria for satisfactory choices of h and M are yet an open problem. A solution, suited to our case study, will be proposed in Section 4.

3. DATA AND MODEL

The data considered in this study consist of the mean hourly average of the O_3 and nitrogen dioxide (NO_2) concentrations and the hourly average temperature values (T). The data are measured at one of the sites which make up the monitoring network within the city of Bologna (Giardini Margherita). The sample was provided by the Environmental Control Office of the Municipality of Bologna from June 1993 to December 1996. The choice of the hourly frequency is motivated by the following: (a) the peak hourly ozone concentration is the value of major interest for comparison with official air pollution standards; and (b) the on-line characteristic of the monitoring algorithm is more effective with high-frequency data in order to detect the onset of possible anomalies with more timeliness.

Figures 1–3 contain the plots of O_3 , NO_2 and T values respectively. It is clear from the figures that there are a number of missing values in the data. In particular, the percentage significant of

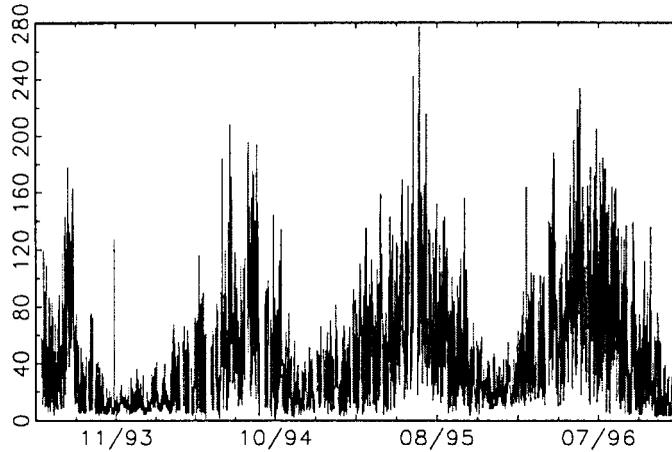


Figure 1. O_3 hourly average concentrations ($\mu\text{g}/\text{m}^3$); period 6/93–12/96

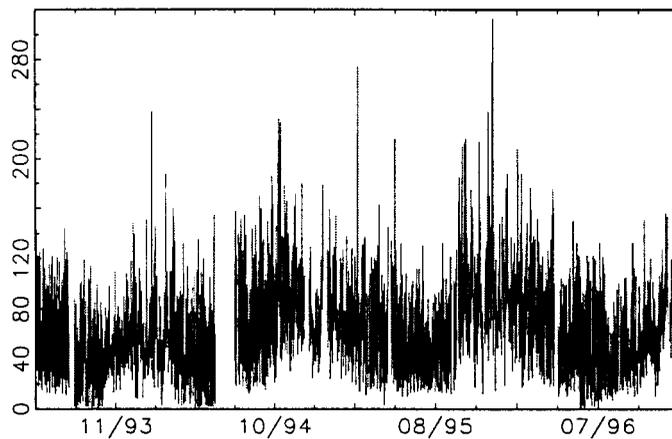


Figure 2. NO_2 hourly average concentrations ($\mu\text{g}/\text{m}^3$); period 6/93–12/96

missing hourly data are: 13.35% for O_3 ; 12.92% for NO_2 ; and 1.03% for T . Moreover, from the analysis of the data we have found the behavior of ozone concentrations for the months from April to October different when compared with the other months of the year. We therefore have split the data into these two seasons and performed the subsequent analysis mainly on the ‘ozone season’, i.e. the period of more interest from a monitoring point of view, because it is associated with higher, and thus more dangerous, values of the concentrations.

To identify a satisfactory model for the ozone data we have chosen as input variables the T values and the NO_2 hourly concentrations, which are important predictors of the ozone levels, as checked through cross-correlation analysis. Other potentially useful input variables, like wind speed or NO concentration, are not considered in this study because the related measurements are not available and/or reliable. Next, we have selected as a sufficiently ‘regular’ period for fitting the model the one from 9 July 1996 to 5 August 1996, with $N = 660$ observations.

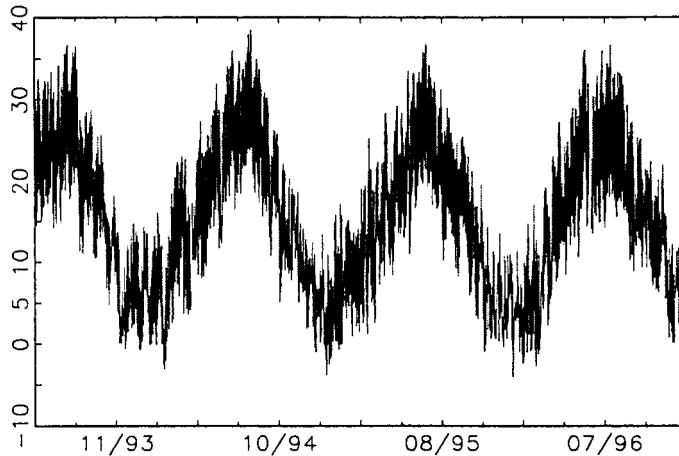


Figure 3. T hourly average values ($^{\circ}\text{C}$); period 6/93–12/96

The model fitting has been performed through the maximum likelihood approach, using the prediction error decomposition form of the likelihood function. Through a sequence of steps, which iteratively use auto- and cross-correlation analysis, maximum likelihood estimation, residuals analysis, and AIC evaluation, the most satisfactory result of the fitting procedure has led to the following model

$$Y_t = [0 \ 0 \ \dots \ \dots \ 1]X_t + V_t \tag{13}$$

$$X_{t+1} = \begin{bmatrix} 0 & 1 & 0 & \dots & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & 0 & 1 \\ f_3 & 0 & 0 & \dots & f_2 & f_1 \end{bmatrix} X_t + \begin{bmatrix} 0 & 0 & \dots \\ 0 & 0 & \dots \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ g_1 & g_2 & g_3 \end{bmatrix} \begin{bmatrix} Z_t \\ N_t \\ N_{t-1} \end{bmatrix} + W_t, \tag{14}$$

where Y_t is the log-ozone concentration at time t , X_t is the (24×1) state vector, Z_t the log-temperature, N_t the log- NO_2 hourly concentration, V_t a scalar Gaussian white noise with variance $R_t = R$, $W_t = [0 \ 0 \ \dots \ \dots \ e_t]'$ a (24×1) Gaussian white noise vector with covariance matrix

$$Q = \begin{bmatrix} 0 & 0 & 0 & \dots & \dots & 0 \\ 0 & 0 & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & \sigma_w^2 \end{bmatrix}.$$

The parameter estimates with the relative standard errors are given in Table I. The estimated variance of the innovations is 0.0212, while estimates of the remaining variances, i.e. R and σ_w ,

Table I. Estimation results relative to the model (13)–(14)

Parameter	Estimate	Standard error	<i>t</i> -value	prob > <i>t</i>
f_1	1.1102	0.0321	34.624	0.0000
f_2	-0.3662	0.0297	-12.324	0.0000
f_3	0.0149	0.0070	2.146	0.0159
g_1	0.3037	0.0294	10.339	0.0000
g_2	-0.0386	0.0159	-2.420	0.0078
g_3	0.0636	0.0162	3.934	0.0000

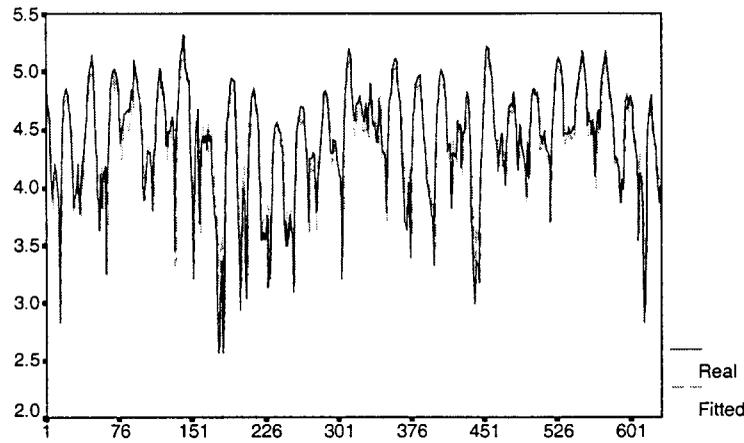


Figure 4. Real vs fitted values relative to the model (13)–(14)

are 0.0012 and 0.0185 respectively. The observed versus fitted sequence is displayed in Figure 4, with a correlation coefficient of 0.914.

4. PARAMETERS CHOICE FOR THE GLR ALGORITHM

Before applying the GLR algorithm to the innovations of the model previously identified, we have to specify the threshold h and the window size M . The choice of h may be critical, in the sense that the number of resulting alarms may be sensitive to this choice. Moreover, h depends upon M : actually it should be chosen in such a way that detecting a selected shift with an average delay not greater than M is possible. A general criterion for this choice of parameters could be based on the ARL function (Montgomery 1997), which defines, under the no change hypothesis, the mean time between false alarms and the mean delay for detection, once the change has occurred. In statistical process control the generally adopted rule consists of minimizing the mean delay for detection subject to the constraint of a fixed mean time between false alarms. Unfortunately, the explicit computation of the ARL function is not analytically tractable in our situation and the optimal choice of h and M is still an open problem, as emphasized by Lai (1995).

To solve this problem we propose the following procedure based on the joint use of Monte Carlo simulations and some known theoretical results. For the window we set the size at $M = 24$. This seems a reasonable choice with hourly data allowing the detection of possible anomalies within 1 day. This means that a failure detection with a delay greater than 24 hours should not be

Table II. Simulation results for the choice of h ; δ = normalized change; m.d. = effective mean delay

δ	h								
	11 m.d.	12 m.d.	13 m.d.	14 m.d.	15 m.d.	16 m.d.	17 m.d.	18 m.d.	19 m.d.
3.0	0	0	0	0	0	0	0	0	0
2.5	0	0	0	0	1	0	1	1	1
2.0	1	1	1	1	3	2	2	3	3
1.5	3	3	4	5	5	6	6	8	8
1	7	8	9	11	12	14	16	18	20

of any usefulness for timely air pollution control policies. Conditionally on $M = 24$, we decided to choose the optimal value of h by combining different information.

First, a simulation experiment was performed. Employing the model previously identified, the change situations were simulated, for known change instants t_0 , through 1000 replications of the experiment for several values of h and δ , where δ is a normalized version of v (see equation (4)). In this way it is possible to determine the estimated mean detection delay time $M(t_a - \hat{t}_0)$, the real mean detection delay time $M(t_a - t_0)$, and the distribution of \hat{v} and \hat{t}_0 : where \hat{v} and \hat{t}_0 are the maximum likelihood estimates of v and t_0 . Some results are reported in Table II, where for different combinations of h and δ , the effective mean delay of detection is obtained.

Further, it is possible to calculate the probability $P_D(\delta, t_0)$ of the correct detection of a change of magnitude δ at time t_0 and in this way we can determine the combination of detection delay time, $k - t_0$, and magnitude of the change, δ , which are detected with a fixed probability value for a given threshold. Following Basseville and Nikiforov (1993), under the change hypothesis the log-likelihood ratio S is a non-central χ^2 variable with one degree of freedom and non-centrality parameter J_{kt_0} ,

$$L(S_{t_0}^k) = \chi^2(1, J_{kt_0}), \quad (15)$$

where

$$J_{kt_0} = \sum_{i=t_0}^k \delta \rho^{*T}(i, t_0) \Sigma_i^{-1} \delta \rho^*(i, t_0) \quad (16)$$

is the Kullback divergence between the two joint distributions of the innovation sequence $(\varepsilon_i)_{i=j, \dots, k}$. In Table III are reported the detection delay time for which a change of magnitude δ will be detected with $P_D(\delta, t_0) > 0.8$, where $P_D(\delta, t_0)$ is evaluated according to (15).

Finally, adapting to our situation some theoretical work by Lai (1995), one can obtain, at least asymptotically, the order of magnitude of changes which are detected inefficiently by the algorithm. More precisely, Lai (1995) is able to show that a window-limited GLR scheme for detecting a change in the mean of a normally distributed variable with known variance is asymptotically optimal if one chooses the window size

$$M = \gamma$$

Table III. Detection delay time for which $P_D(\delta, t_0) > 0.8$

	δ				
	1	1.5	2	2.5	3
$h = 11$	15	1	1	1	1
$h = 12$	16	1	1	1	1
$h = 13$	18	1	1	1	1
$h = 14$	19	3	1	1	1
$h = 15$	21	4	1	1	1
$h = 16$	23	4	1	1	1
$h = 17$	25	5	1	1	1
$h = 18$	27	6	1	1	1
$h = 19$	29	7	1	1	1

and the threshold

$$h = \log \gamma + \frac{1}{2} \log(\log \gamma) + \log K + o(1),$$

where γ is the desired ARL under the no change hypothesis. Here K is

$$K = \pi^{-1/2} \int_0^{\infty} x \psi^2(x) dx, \quad (17)$$

where

$$\psi(x) = 2x^{-2} \exp \left\{ -2 \sum_{i=1}^{\infty} n^{-1} \Phi(-x\sqrt{n}/2) \right\}$$

for $x > 0$, and Φ is the distribution function of the standard normal distribution. However, the choice $M = \gamma$ for the window size still requires a heavy computational burden, since γ is usually large. Lai suggests trying values of M of the order of a constant times $\log \gamma$, i.e.

$$M \approx a \log \gamma,$$

but this window size is quite inefficient for detecting changes that are smaller than $\sqrt{2/a}$.

These arguments can be employed in our situation, where it is required that a change in the mean of a Gaussian innovation sequence is detected within a fixed time interval, given by $M = 24$, for the reason previously given. Comparing these quantities (Table IV) with the values of the standardized changes

$$\theta(k, t_0) = \delta \rho^*(k, t_0) \Sigma^{-1/2},$$

it is possible to find for each combination of $k - t_0$, δ and h values of $\theta(k, t_0)$ for which the algorithm has poor performances with fixed $M = 24$. Considering that in our situation $\theta(k, t_0)$ assumes steady-state values after a few lags, some of these values are reported in Table V.

The proposed procedure for choosing the threshold h can be summarized as follows: (a) the determination of h through Monte Carlo simulations; (b) the detection probability of a change

Table IV. a and $\sqrt{2/a}$ values with $M = 24$

h	a	$\sqrt{2/a}$
11	2.1818	0.957427
12	2	1
13	1.846154	1.040833
14	1.714286	1.080123
15	1.6	1.118034
16	1.5	1.154701
17	1.411765	1.190238
18	1.333333	1.224745
19	1.263158	1.258306

Table V. $\theta(k, t_0)$ for some change values

$ \delta $	1	1.5	2	2.5	3
$\theta(k, t_0)$	0.851237	1.276856	1.702475	2.128093	2.553712

for each combination of detection delay and threshold; and (c) the combination of change magnitude and detection delay for which the algorithm has poor performances.

It is worth noting that each step of our procedure gives different and complementary information for a reasonable and operative choice of the parameters for the change detection rule. This combination of tools is important because Lai (1995) gives asymptotic results, but our case is of a finite context. This implies that the results are approximated with a consequent degree of uncertainty. However, employing the simulation results and the probability evaluation when the three methods are in agreement we are more confident with our conclusion.

For example, with $h = 19$ a change of magnitude $\delta = 1$ is not detected in an efficient way (Tables IV and V); the detection delay time, $P_D(\delta, t_0) > 0.8$ is 29 and the estimated mean detection delay is about 20 (Table II), a value of the same order as the window size. These results suggest that $h = 19$, with $M = 24$, is not a good choice for $\delta = 1$. In general it can be noted that, with the values of the threshold considered, shifts of order $\delta = 1$ are detected with some difficulty. A shift $\delta = 1.5$, with $h = 19$, is in the 'efficient zone' of detection (Tables IV and V), the detection delay time for which $P_D(\delta, t_0) > 0.8$ is 7 and the estimated mean detection delay is about 8, so we can assert that a change of order $\delta = 1.5$ is detected efficiently.

For the application, we focus our attention on shifts of order $\delta = 1.5$. This is because to detect a change smaller than $\delta = 1.5$, with $M = 24$, a threshold value less than 11 is necessary, but as we checked through preliminary applications, these values lead to a high number of likely false alarms.

For our purposes we chose $h = 15$. In this case, for $\delta = 1.5$, the mean detection delay is about 5; at delay 4 the shift is detected with probability > 0.8 and Tables IV and V show that the variation $\delta = 1.5$ is detected efficiently.

For the performances of the algorithm under the no change hypothesis, remembering that $h \approx \log \gamma$, we can obtain an approximation of the ARL function. In our case, with $h = 15$, we get $\gamma \approx 3 \times 10^6$.

5. APPLICATION AND CONCLUDING REMARKS

With $M = 24$ for the lag window and $h = 15$ for the threshold the GLR algorithm is applied to our ozone data starting in early August 1996 and stopping at the end of the month. Before illustrating the results it is important to explain the management of the monitoring, because missing data makes the continuous functioning of the algorithm impossible. Therefore, we adopt the following strategy. When only a few data (1 or 2) are missing, multi-step ahead predictions of the state are obtained on the basis of the available information, while innovations and values of the statistic g_k are not calculated. When all information is again available, the last prediction is updated and the statistic g_k is calculated, starting from the next observation, thus avoiding the errors due to multi-step ahead predictions which lead to an increase in the false alarm rate.

Figure 5, shows the behavior of the decision statistics g_k , while in Figure 6, alarms and estimation failure times are plotted directly from the observed series data. Looking at the figures,

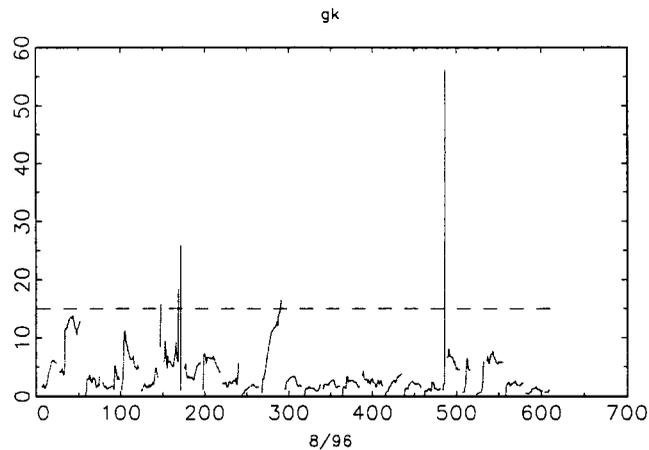


Figure 5. Decision statistics for August 1996

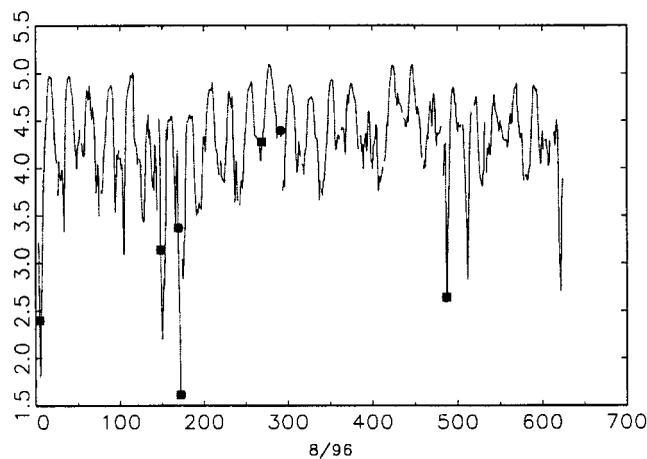


Figure 6. Time series plot of $\log O_3$ for August 1996; time alarm = circle, estimated change time = square

the algorithm seems to show a good ability to promptly signal alarms, probably due to the onset of some anomalies in the measurement instruments, given that signals are often followed by missing data. Of the six alarms detected by the algorithm in the month, four are followed by a set of missing data. Moreover, it can be noted that alarm times and estimated change times are the same for five alarms, and only in one situation is a small delay found. This shows the efficiency of the algorithm in detecting anomalies within the window size M .

Tentative conclusions based on previous results have to be taken with some caution. Actually we are simulating ex-post the on-line characteristics of the algorithm, so we cannot control directly the situation arising due to the alarm signals.

Keeping this caution in mind, we can conclude that the on-line implementation of a monitoring algorithm of the kind presented in this paper could lead to further improvements in the maintenance of air pollution monitoring sites if routinely implemented as a complementary tool to the usual periodic control procedures.

ACKNOWLEDGEMENT

Financial support by MURST ex 40 per cent and ex 60 per cent is gratefully acknowledged.

REFERENCES

- Alwan, L. C. and Roberts, H. V. (1988). 'Time series modelling for statistical process control'. *Journal of Business & Economic Statistics* **6**, 87–95.
- Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Applications*. Prentice Hall, Englewood Cliffs, NJ.
- Batterman, S. A. (1992). 'Optimal estimators for ambient air quality levels'. *Atmospheric Environment* **26**, 113–123.
- Davison, A. C. and Hemphill, M. W. (1987). 'On the statistical analysis of ambient ozone data when measurements are missing'. *Atmospheric Environment* **21**, 629–639.
- Lai, T. L. (1995). 'Sequential changepoint detection in quality control and dynamical systems'. *Journal of the Royal Statistical Society, Series B* **57**, 613–658 (with discussion).
- Lorden, G. (1971). 'Procedures for reacting to a change in distribution'. *The Annals of Mathematical Statistics* **42**, 1897–1908.
- Montgomery, D. C. (1997). *Introduction to Statistical Quality Control*. John Wiley & Sons, New York.
- Willsky, A. S. and Jones, H. L. (1976). 'A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems'. *IEEE Transactions on Automatic Control* **21**, 108–112.