Silvano Bordignon Michele Scagliarini

La rilevazione di cambiamenti in processi dinamici: un'applicazione ad un analizzatore di ozono

Serie Ricerche n.6

Dipartimento di Scienze Statistiche "Paolo Fortunati"
Università degli studi di Bologna
1997

Lavoro svolto nell'ambito della ricerca "Metodi statistici per il controllo dell'ambiente e la valutazione del rischio ambientale" MURST (40%) fondi 1995

Le elaborazioni dei dati S.A.R.A., forniti dal Comune di Bologna, che compaiono nel lavoro sono eseguite presso il Dipartimento di Scienze Statistiche dell'Università di Bologna sotto la supervisione della Prof.ssa Daniela Cocchi. Sono da intendersi ad uso interno e non vanno confuse con quelle previste dalla legge in materia.

Silvano Bordignon
Dipartimento di Scienze Statistiche Università degli Studi di Padova

Michele Scagliarini
Dipartimento di Scienze Statistiche Università degli Studi di Bologna

Finito di stampare nel mese di Dicembre 1997
presso le Officine Grafiche TECNOPRINT S.N.C.
Via del Legatore 3, Bologna

INDICE

1 Introduzione

I dati raccolti dalle reti di monitoraggio per il controllo della qualità dell'aria rivestono un ruolo importante nell'ambito della previsione e del controllo dell'inquinamento atmosferico. In effetti tali dati consentono: a) l'accertamento dei livelli delle concentrazioni degli inquinanti per la determinazione della qualità dell'aria, in ottemperanza alle leggi vigenti; b) la valutazione dell'impatto ambientale e degli effetti sulla salute; c) la definizione di strategie per l'abbattimento delle emissioni inquinanti. La correttezza di tali analisi e delle conseguenti politiche per il controllo della qualità dell'aria sono pesantemente condizionate dall'affidabilità dei dati raccolti. La qualità di tali informazioni purtroppo è affetta da diversi problemi, tra i quali i più rilevanti sembrano essere la presenza di dati mancanti o non validi e l'inaccuratezza delle misurazioni.

Dati mancanti e/o imprecisi derivano principalmente da guasti negli strumenti di misura, distorsioni negli analizzatori e da problemi di manutenzione, si vedano ad esempio, Davison e Hemphill (1987) e Batterman (1992). Come notato da Batterman (1992), questi problemi possono risultare critici per l'interpretazione dei dati sulla qualità dell'aria. Nonostante la loro importanza, le metodologie per affrontare questi problemi sono poco studiate e, di conseguenza, è raro che siano impiegate correntemente.

In questo lavoro si propone una procedura statistica *on-line* che può essere usata per individuare, con il minimo ritardo possibile, distorsioni nello strumento di misura, per poter così migliorare la qualità dei dati raccolti dalle reti di monitoraggio ambientale. La metodologia è basata sull'utilizzo congiunto di modelli stocastici e algoritmi per il controllo statistico di processo. Un opportuno modello stocastico viene impiegato per descrivere la dinamica dell'inquinante considerato, mentre un algoritmo del tipo GLR (Generalised Likelihood Ratio, Lorden 1971) viene utilizzato per controllare le innovazioni ottenute dal modello stocastico. La procedura proposta può essere applicata in modo automatico ed implementata come strumento ausiliario ai controlli periodici effettuati sulle stazioni di rilevamento.

La metodologia è applicata alle concentrazioni medie orarie dell'ozono rilevate in una centralina della rete di Bologna e l'ottimizzazione dell'algoritmo è effettuata tramite simulazione.

Il lavoro è strutturato come segue. La sezione 2 introduce i principali strumenti statistici necessari all'implementazione di un algoritmo di monitoraggio in sistemi stocastici. La sezione 3 contiene l'analisi descrittiva dei dati ed i risultati dell'identificazione del modello per le concentrazioni dell'ozono. La scelta dei parametri per l'algoritmo, effettuata tramite un

esperimento di simulazione, è illustrata nelle sezione 4. Infine, nell'ultima sezione sono riportati i risultati dell'applicazione e alcune considerazioni conclusive.

2 Algoritmi di monitoraggio in sistemi dinamici

Il contesto metodologico nel quale il presente lavoro si inserisce è il problema dell'individuazione di un cambio nei parametri di un sistema stocastico. Si consideri una sequenza $\{y_k\}$ di variabili casuali aventi densità p_θ dipendente da un parametro θ. Prima del tempo di cambio t_0 (non noto), θ è costante ed è uguale a θ_0, mentre dopo t_0 θ assume valore pari a θ_1. L'obiettivo è individuare il verificarsi della variazione con il minimo ritardo possibile fissato un determinato tasso di falsi allarmi.

Nel controllo statistico di processo gli algoritmi per l'individuazione di variazioni nei parametri sono ampiamente utilizzati: i più noti, come la carta di controllo di Shewart (Shewhart, 1931), la carta CUSUM (Cumulative SUM, Page, 1961) e la carta EWMA (Exponentially Weighted Moving Average, Roberts, 1959), si possono trovare, per esempio, in Montgomery (1991). Tuttavia questi algoritmi, poiché si basano sull'assunzione che $\{y_k\}$ sia una sequenza di variabili aleatorie indipendenti, risultano incapaci di operare in modo soddisfacente quando le osservazioni sono autocorrelate. In presenza di autocorrelazione nelle osservazioni, una soluzione generale al problema dell'individuazione di una variazione consiste: a) nel generare i residui del modello opportunamente adattato ai dati di partenza: infatti se il modello è correttamente specificato, i residui sono mediamente nulli prima della variazione, mentre a cambiamento avvenuto si distribuiscono intorno ad un valore significativamente diverso da zero; b) nel costruire una regola di decisione per individuare la variazione basata sull'informazione portata dai residui.

Una possibile soluzione consiste nell'utilizzare come residui di un processo stocastico $\{Y_k\}$ le innovazioni definite da $\varepsilon_k = Y_k - E_\theta(Y_k|Y_{k-1},\ldots,Y_1)$. Sotto l'ipotesi che non sia intervenuto un cambio, $\{\varepsilon_k\}$ è una sequenza di variabili aleatorie incorrelate con media nulla (indipendenti se il processo è Gaussiano). Di conseguenza algoritmi costruiti per osservazioni indipendenti o non autocorrelate possono essere utilizzati sulle innovazioni di un processo autocorrelato dopo opportune modifiche.

Si assuma che la dinamica di una sequenza di osservazioni possa essere modellata tramite un sistema stocastico lineare ed invariante. La rappresentazione *state space* del sistema assume la forma (Davis e Vinter 1984)

$$X_{k+1} = FX_k + GU_k + W_k \quad (1a)$$

$$Y_k = HX_k + JU_k + V_k \quad (1b)$$

dove X, U, Y sono rispettivamente i vettori dello stato del sistema, delle variabili di input e delle osservazioni; $\{W_k\}$ e $\{V_k\}$ sono due ***white noise*** Gaussiani indipendenti con matrici di covarianze Q e R; F è la matrice di transizione; H la matrice di osservazione; G e J le matrici di controllo.

Dato lo stato iniziale del sistema $X_0 \sim N(\mu_0, P_0)$, la sequenza delle innovazioni si può ottenere ricorsivamente utilizzando il filtro di Kalman:

$$\varepsilon_k = Y_k - HX_{k|k-1} - JU_k \quad (2)$$

dove: $X_{k+1|k} = F(X_{k|k-1} + K_k\varepsilon_k) + GU_k$ è la previsione ad un passo in avanti dello stato; $K_k = P_{k|k-1}H^T(\Sigma_k)^{-1}$ è il guadagno di Kalman ; $\Sigma_k = HP_{k|k-1}H^T + R$ è la matrice di covarianze delle innovazioni; $P_{k+1|k} = FP_{k|k}F^T + Q$ è la stima della matrice di covarianze di $X_{k+1|k}$; $P_{k|k} = (I - K_kH)P_{k|k-1}$ è la stima della matrice di covarianze di $X_{k|k}$. Segue che, sotto l'ipotesi che non sia intervenuto un cambiamento, $\{\varepsilon_k\}$ è una sequenza di variabili aleatorie indipendenti e Gaussiane con media zero e matrice di covarianze Σ_k.

Seguendo l'impostazione di Basseville e Nikiforov (1993) si possono distinguere variazioni additive e non additive. Le variazioni additive consistono in un mutamento nel segnale o nel sistema lineare ed hanno come unico effetto una variazione nel valore medio della sequenza delle osservazioni. Le variazioni non additive interessano la varianza, la correlazione e le caratteristiche dello spettro e in alcuni casi particolari sono riconducibili alla stessa logica delle variazioni additive. Poiché queste variazioni interessano le dinamiche del sistema, richiedono meccanismi di rilevazione più complessi. Nella situazione in esame è sufficiente considerare solamente il caso di variazioni additive. Tali variazioni sono introdotte nel modello *state space* (1) come segue:

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

dove Γ e Ξ sono delle matrici "guadagno" che tengono conto dell'intensità del cambio, t_0 è l'istante ignoto nel quale è intervenuta la variazione, mentre Ψ_x e Ψ_y sono vettori che rappresentano i profili dinamici delle variazioni ipotizzate. Chiaramente, se $k< t_0$, $\Psi_x = \Psi_y = 0$.

Il significato delle variazioni che si desidera monitorare grazie al modello *state space* è strettamente legato al tipo di problema che si deve risolvere. In particolare è possibile dare alle variazioni il significato presentato nel seguito. Se si considera l'equazione di stato, la variazione additiva $\Gamma\Psi_x$ generalmente corrisponde ad una distorsione degli attivatori, cioè degli input. Se invece si considera l'equazione osservazionale la distorsione che si analizza $\Xi\Psi_y$ riguarda le componenti di Y, cioè i sensori che rilevano l'output. Se per esempio $\Psi_x = 0$, Ξ è uno scalare e Ψ_y è un vettore composto da zeri con solamente la j-esima componente uguale a uno per $k \geq t_0$, allora il modello (3) ipotizza la presenza di una distorsione nel j-esimo sensore: è questa la situazione che si desidera studiare.

Come illustrato precedentemente, l'algoritmo per l'individuazione della variazione si basa sulle innovazioni: risulta quindi importante studiare il comportamento delle innovazioni in presenza di un modello specificato per il cambio. Nel modello *state space* la sequenza delle innovazioni è ottenuta grazie al filtro di Kalman. Data la linearità del modello considerato, unitamente all'effetto additivo delle variazioni ipotizzate, si può far vedere (Basseville e Nikiforov, 1993) che per lo stato, la stima dello stato e le innovazioni del modello (3) valgono le seguenti scomposizioni:

$$X_k = X_k^o + \alpha(k,t_0) \quad (4a)$$
$$X_{k|k} = X_{k|k}^o + \beta(k,t_0) \quad (4b)$$
$$\varepsilon_k = \varepsilon_k^o + \rho(k,t_0) \quad (4c)$$

dove l'esponente nullo indica le quantità relative al modello senza variazioni, a cui si somma l'effetto del cambio intervenuto al tempo $t_0 \leq k$. Le espressioni ricorsive per α, β e ρ si possono trovare in Basseville e Nikiforov (1993).

Riguardo alle innovazioni, ε_k^o corrisponde alla innovazione del modello (3) senza il cambio mentre $\rho(k,t_0)$ è il profilo dinamico della variazione.

Poiché nella presente situazione si è interessati all'individuazione di distorsioni nel dispositivo di misurazione dell'inquinante è sufficiente considerare una particolare formulazione del modello (3), data da:

$$X_{k+1} = FX_k + GU_k + W_k \quad (5a)$$
$$Y_k = HX_k + JU_k + V_k + vI_{k\geq t_0} \quad (5b)$$

dove Y_k è scalare, v è l'intensità, non nota, della variazione e $I_{k\geq t_0}$ è una funzione indicatrice. Le innovazioni del modello (5) si specializzano in

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

dove $\rho^*(k,t_0)$ è il profilo dinamico della variazione. Nell'ipotesi di comportamento *steady state* del filtro di Kalman l'espressione analitica per ρ^* è data da

$$\rho^*(k,t_0) = \mathrm{I}_{k\geq to} - \sum_{i=0}^{k-to-1} HF_*^i FK\mathrm{I}_{k-i-1\geq to} \quad (6)$$

dove $F_* = F(I - KH)$ e K è il guadagno di Kalman nell'ipotesi *steady state*. Riassumendo $\{\varepsilon_k\}$ è un *white noise* Gaussiano con media zero nell'ipotesi che non sia avvenuta la variazione, mentre si comporta come una sequenza di osservazioni indipendenti con media $v\rho^*(k,to)$ dopo il verificarsi di una variazione. Quindi il problema dell'individuazione di un cambio additivo nelle osservazioni può essere affrontato utilizzando un algoritmo per l'individuazione di variazioni che sia in grado di verificare il seguente sistema di ipotesi:

$$H_0: \varepsilon_k \sim N(0,\Sigma_k)$$
$$H_1: \varepsilon_k \sim N(v\rho^*(k,to),\Sigma_k)$$

dove $\{\varepsilon_k\}$ rappresenta la sequenza delle innovazioni del modello (5).

Per la scelta dell'algoritmo più opportuno da utilizzare occorre considerare il particolare tipo di informazioni di cui si dispone. Non essendo noti i parametri dopo la variazione, è conveniente ricorrere ad una versione generalizzata dell'algoritmo basato sul rapporto di verosimiglianza noto anche come algoritmo GLR (Generalised Likelihood Ratio, Lorden, 1971). Tale algoritmo, con riferimento al caso generale dell'individuazione di un cambiamento nel parametro θ ad un ignoto tempo t_0, si basa sulle seguenti quantità:

funzione di decisione $g_k = \max_{1 \le j \le k} \sup_{\theta 1} S_j^k(\theta_1)$;

regola di decisione $d_k = 1$ se e solo se $g_k \ge h$;

tempo di allarme $t_a = \min\{k : d_k = 1\}$

dove $S_j^k(\cdot)$, il logaritmo del rapporto di verosimiglianza calcolato per l'insieme delle osservazioni comprese tra j e k, è dato da

$$S_j^k(\theta_1) = \sum_{r=j}^{k} (\ln p_{\theta 1}(y_r) - \ln p_{\theta 0}(y_r)),$$

mentre h rappresenta un valore di soglia, da determinarsi opportunamente.

Specializzando questi risultati alla nostra situazione, descritta dal modello (5), in cui il parametro incognito è l'intensità della variazione v, la funzione di decisione diventa

$$g_k = \max_{1 \le j \le k} \sup_{v} S_j^k(v)$$

dove $S_j^k(\cdot)$ è il logaritmo del rapporto di verosimiglianza costruito questa volta sulle innovazioni del modello (5). Data la Gaussianità delle osservazioni una espressione esplicita per $\sup_v S_j^k(v)$ è data da

$$\sup_v S_j^k(v) = \hat{v}_k(j)\left(\sum_{i=j}^{k} \rho(i,j)\Sigma_i^{-1}\varepsilon_i\right) - \frac{\hat{v}_k^2(j)}{2}\left(\sum_{i=j}^{k} \rho^2(i,j)\Sigma_i^{-1}\right) \qquad (8)$$

dove Σ_i è la varianza di ε_i e $\hat{v}_k(j)$ è la stima di massima verosimiglianza di v effettuata al tempo k, assumendo che il cambio sia avvenuto al tempo j. La sua espressione risulta

$$\hat{v}_k(j) = \frac{\sum_{i=j}^{k} \rho(i,j)\Sigma_i^{-1}\varepsilon_i}{\sum_{i=j}^{k} \rho^2(i,j)\Sigma_i^{-1}}. \qquad (9)$$

L'altro parametro incognito della procedura è il tempo di cambio t_0. La ricerca del valore t_0 attraverso il calcolo della stima di massima verosimiglianza, è ottenuto considerando tutti i possibili istanti temporali che precedono il tempo k. In questo modo il numero di massimizzazioni da calcolare cresce all'infinito, al crescere di k. Per praticità t_0 è stimato cercando il massimo valore di $S_j^k(\cdot)$ all'interno di una finestra di dimensione fissata M. Si assume infatti che le variazioni antecedenti la finestra siano già state rilevate, che la finestra sia sufficientemente grande da permettere la rilevazione di tutti i guasti significativi e sufficientemente piccola per far aumentare la velocità di risposta alla variazione (Willsky e Jones, 1976). La stima di t_0 quindi risulta:

$$\hat{t}_{0k}(j) = \arg\max_{k-M+1 \le j \le k} S_j^k(\cdot) \qquad (9)$$

Una volta stimato t_0 la stima dell'intensità della variazione è data da

$$\hat{v}_k = \hat{v}_k(t_{0k}) \qquad (10)$$

per $k = t_a$, dove t_a è il tempo di allarme.

In definitiva l'algoritmo si compone dei seguenti passi: a) individuazione del cambio; b) stima del tempo di cambio e della sua intensità; c) aggiornamento delle stime iniziali dello stato e della matrice di covarianze dell'errore di stima. Mentre i primi due passi sono tipici della metodologia GLR, il terzo passo è conveniente per fornire al filtro di Kalman valori iniziali più appropriati di quelli fornitegli all'inizio del monitoraggio. La soluzione seguita nel caso in esame può essere trovata in Willsky e Jones (1976).

Per il funzionamento dell'algoritmo GLR è inoltre necessario scegliere opportunamente la soglia h e l'ampiezza della finestra M. Come sottolineato da

Lai (1995), una soluzione generale per la scelta opportuna di h e M è ancora un problema aperto. Una soluzione pratica a questo problema, adatta alla nostra situazione, è proposta nella sezione 4.

3 Dati e modello

L'ambito convenzionale di applicazione degli algoritmi di monitoraggio proposti nella sezione precedente è quello della rilevazione di guasti o rotture in sistemi dinamici che descrivono processi industriali. Un campo non convenzionale in cui queste tecniche possono essere applicate è il monitoraggio di processi di inquinamento ambientale, e più in particolare, atmosferico. Il sistema di produzione degli inquinanti, nell'ambito di una determinata area geografica, è infatti costituito principalmente dalle emissioni delle industrie, degli impianti di riscaldamento e dal traffico, a tali fonti vanno inoltre affiancate le condizioni atmosferiche che influenzano in misura sensibile le concentrazioni degli inquinanti.

La modellazione di questo sistema complesso permette di catturare l'andamento di fondo del processo e di spiegarne le cause comuni di variazione. Quindi grazie all'attivazione di algoritmi di monitoraggio si possono rilevare nel più breve tempo possibile andamenti anomali nelle concentrazioni, dovuti anche ad errori negli strumenti di rilevazione o nella trasmissione dei dati.

Nel seguito viene proposta un'applicazione finalizzata al conseguimento in tempo reale di un archivio affidabile di dati relativi alle concentrazioni di ozono (O_3) rilevate da una stazione di monitoraggio della rete di Bologna. La scelta di analizzare questo inquinante è motivata dalla tendenza generale, rilevata a partire dagli anni ottanta, alla riduzione delle emissioni di biossido di carbonio (CO_2) e anidride solforosa (SO_2), grazie soprattutto all'utilizzo di gas naturale e combustibili a basso contenuto di zolfo. In Italia come in ambito europeo, sono invece gli ossidanti fotochimici, tra i quali l'ozono, a presentare un progressivo aumento nelle concentrazioni superando spesso, nella stagione estiva, i livelli di allarme. E' quindi interessante attivare una politica di controllo per questo inquinante. La sperimentazione dell'algoritmo di controllo si basa sui dati relativi alle concentrazioni medie orarie dell'ozono rilevate presso una centralina (Giardini Margherita) del Sistema Automatico per il Rilevamento Ambientale (S.A.R.A) di Bologna. Il periodo considerato va da Aprile 1994 a Dicembre 1995. L'attenzione è rivolta al dato orario per i seguenti motivi: a) i valori orari delle concentrazioni dell'ozono sono quelli di maggiore interesse per eventuali confronti con gli standards della qualità dell'aria; b) data la connotazione *on-line* dell'algoritmo, l'utilizzo di dati rilevati con una elevata frequenza consente

l'individuazione di eventuali anomalie con maggiore rapidità. L'andamento della variabile considerata nel periodo di osservazione è riportato nella figura 1, mentre nella tabella 1 sono riportate le principali statistiche descrittive.

Media = 44.412	Mediana = 36.00	Dev. Std = 32.635	Max.= 276
% dati mancanti = 14	Skewness= 1.296	Kurtosis = 2.095	Min. = 1

Tabella 1: statistiche descrittive della variabile ozono

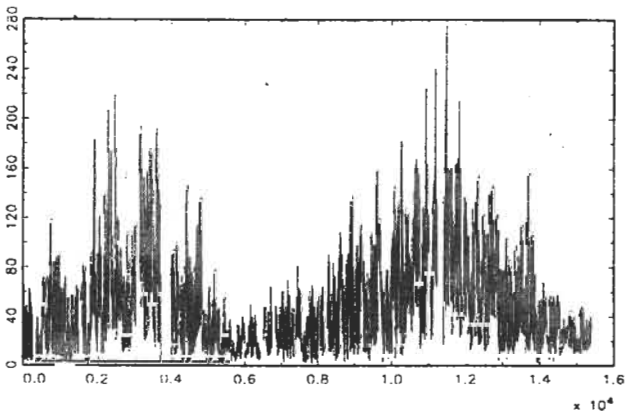

Figura 1: O_3 periodo Aprile 1994 Dicembre 1995

L'analisi preliminare della variabile è importante al fine di pervenire ad una adeguata modellazione della serie, da utilizzare poi come filtro per l'ottenimento di un processo di innovazione su cui applicare l'algoritmo GLR. Come è noto (Finzi e Brusasca, 1991), per l'inquinante in questione gli ossidi di azoto e l'irraggiamento solare ricoprono un importante ruolo esplicativo. Purtroppo la variabile irraggiamento solare non è disponibile. Tuttavia è possibile risolvere in modo soddisfacente questo problema utilizzando la temperatura come *proxy* della radiazione solare: a conferma di ciò la correlazione tra temperatura e ozono è risultata rilevante (0.703) evidenziando una similarità di comportamento tra queste due variabili. Riguardo agli ossidi di azoto emergono alcune difficoltà per un loro utilizzo. Il nostro obiettivo è quello di rilevare anomalie nell'andamento nella serie dell'ozono attribuibili a malfunzionamenti nell'analizzatore predisposto alla sua misurazione. Nel paragrafo precedente si è mostrato come questo richieda una modellazione preventiva dei dati in cui, accanto all'input stocastico, compaiano ingressi esogeni, al fine di migliorare la capacità

esplicativa del modello. Si richiede pertanto che le variabili esogene introdotte nel modello siano esenti da quegli errori (di misura) che sono proprio l'oggetto che si vuole rilevare con l'algoritmo. E' opportuno quindi inserire nel modello filtrante esclusivamente quegli inputs la cui misurazione è altamente affidabile. L'incidenza di dati mancanti nelle serie degli ossidi di azoto (intorno al 25%) porta a pensare che tali inputs presentino gli stessi problemi di rilevazione dell'ozono (se non maggiori) e quindi non è indicato utilizzarli come regressori. Questo perché la presenza di elevate percentuali di dati mancanti e di lunghe serie di osservazioni non disponibili spesso è il frutto di seri malfunzionamenti nel sensore che fanno perdere, perché cancellati dai tecnici, sia i dati rilevati, ma giudicati inattendibili, che quelli corrispondenti al tempo necessario per il ripristino del sensore. Per la temperatura questo problema invece non si presenta: la percentuale di perdita è minima (vedi la tabella 2, dove sono riportate alcune statistiche descrittive per la temperatura, e la figura 2), sembra dunque si possa ritenere tale variabile affidabile.

Media = 15.538	Mediana = 14.4	Dev. Std = 8.397	Max.= 38.5
% dati mancanti = 1.1	Skewness= .169	Kurtosis = -0.7905	Min. = -3.8

Tabella 2: statistiche descrittive della variabile temperatura

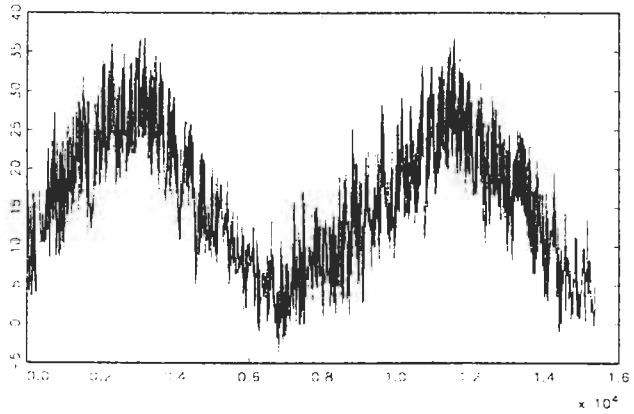

Figura 2: temperatura periodo Aprile 1994 Dicembre 1995

Per le considerazioni svolte è conveniente orientarsi verso un modello per l'ozono che considera come input la temperatura e le analisi che seguono si concentreranno solamente su queste due variabili.

Dai grafici delle figure 1 e 2 oltre alla marcata similarità di comportamento, confermata dall'alto valore di correlazione, emerge la cadenza stagionale di entrambe le variabili. Dall'esame dei *box-plots* di ozono e temperatura raggruppati per ore del giorno e dalle funzioni di autocorrelazione campionaria (figura 3 e figura 4) risulta evidente anche la presenza di un comportamento ciclico giornaliero.

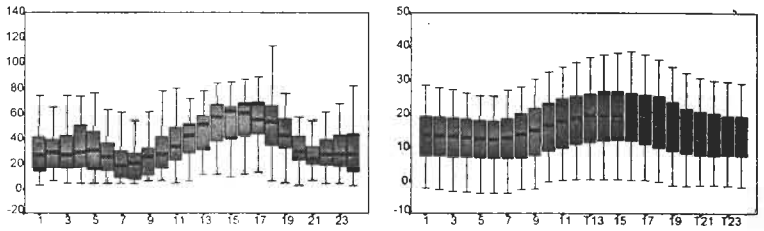

Figura 3: *box-plots* di ozono e temperatura raggruppati per ore del giorno

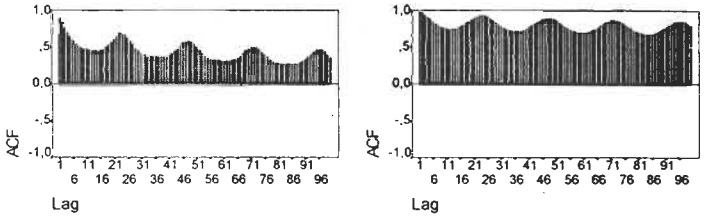

Figura 4: funzione di autocorrelazione campionaria di ozono e temperatura

Delineato il quadro dei dati a nostra disposizione occorre individuare il sottoperiodo campionario, caratterizzato da un funzionamento corretto del sensore. Infatti, per individuare correttamente eventuali anomalie nel sensore è opportuno far riferimento ad un processo identificato quando lo strumento lavora regolarmente. Ci si è quindi orientati verso periodi nei quali la presenza di dati mancanti sia poco rilevante dato che un'elevata percentuale di osservazioni non disponibili è spesso il sintomo con cui si manifesta un'anomalia di funzionamento. Il periodo scelto, inoltre, dovrebbe collocarsi nella stagione "calda", perché l'ozono presenta dei picchi generalmente nei mesi primaverili ed estivi: ne consegue che è importante controllare lo strumento di misura in questi periodi. I grafici dei logaritmi delle variabili nel periodo indicato sono riportati nella figura 5.

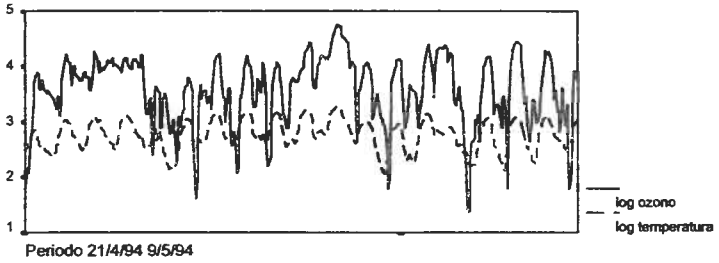

Figura 5: andamento dei logaritmi delle serie trasformate

A questo punto si identificherà sulla base dei dati disponibili un opportuno modello ARMAX per l'ozono con la temperatura come input esogeno. Sulla base delle indicazioni fornite dalle stime delle funzioni di auotocorrelazione globale e parziale, dall'andamento della funzione di autocorrelazione incrociata (figura 6) tra le variabili log-ozono e log-temperatura, opportunamente sbiancate (Pierce e Haugh, 1977), ed iterando più volte il processo di identificazione, stima e verifica, il modello migliore, utilizzando anche il criterio di informazione AIC (Akaike 1973) ed il valore della log-verosimiglianza (Hamilton, 1994), è risultato il seguente:

$$(1 - A_1 B - A_2 B^2)(1 - A_{12} B^{12})(Y_t - \alpha U_t - \beta U_{t-2}) = \xi_t \qquad (11)$$

dove Y è la variabile log-ozono, U la log-temperatura, ξ è un *white noise* Gaussiano e B è l'operatore ritardo.

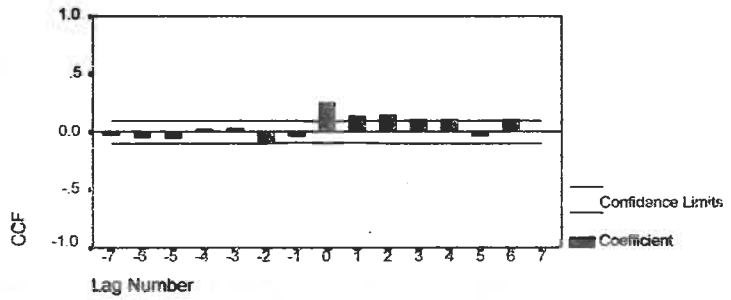

Figura 6: funzione di autocorr. incrociata tra log-ozono e log-temperatura

Le stime dei parametri con i relativi errori standard sono riportate nella tabella 3 unitamente ai valori di alcune statistiche utilizzate nella scelta del modello.

Parametro	*Stima*	*Errore Stand.*	$prob(t > t_c)$
A_1	1.0950	0.04617	0.0000
A_2	-0.3152	0.04564	0.0000
A_{12}	0.1395	0.04933	0.0049
α	0.7337	0.14117	0.0000
β	0.5619	0.14136	0.0001
Errore stand. dei residui = 0.2476			

Tabella 3: risultati della stima del modello

L'analisi dei residui ha rivelato un *fitting* soddisfacente del modello ai dati (si veda ad esempio il correlogramma dei residui di figura 7). Il modello così ottenuto viene riformulato secondo una rappresentazione *state-space* equivalente (Davis e Vinter, 1984) per poter sfruttare direttamente le procedure illustrate nel paragrafo 2.

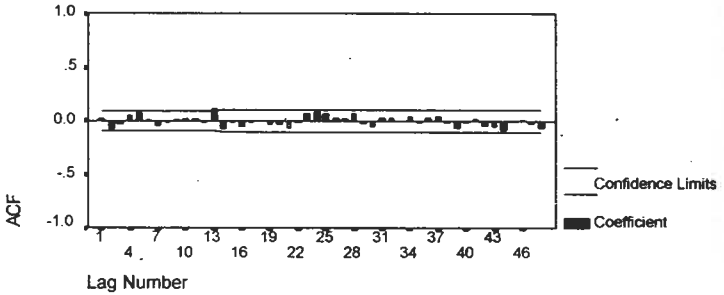

Figura 7: autocorrelazioni dei residui del modello

4 Scelta dei parametri per l'algoritmo GLR

Prima di applicare l'algoritmo GLR alle innovazioni del modello precedentemente identificato è fondamentale selezionare i valori della soglia h e della finestra M. La scelta di h rappresenta un passaggio delicato della metodologia GLR: un valore troppo basso può portare ad un numero elevato di falsi allarmi, mentre con un valore elevato si rischia di non rilevare la variazione avvenuta. Inoltre h dipende da M in quanto una variazione dovrebbe essere rilevata con un ritardo mediamente non più grande della finestra M. Un criterio generale per la scelta di questi parametri potrebbe basarsi sulla funzione Average Run Lengh, ARL (Montgomery, 1991), che definisce sotto H_0, il tempo medio

tra falsi allarmi e sotto H_1 il ritardo medio nella individuazione della variazione. La strategia generale seguita nel controllo statistico di processi industriali consiste nel costruire la regola di decisione in modo da minimizzare l'ARL sotto H_1 una volta che si è fissato il valore della stessa funzione nell'ipotesi che non sia intervenuta la variazione.

Tuttavia nella situazione in esame non è possibile derivare un'espressione analitica della funzione ARL per l'algoritmo GLR. Per risolvere questo problema si propone la seguente procedura empirica. In primo luogo si è posto M=24. Questo perché lavorando con dati orari l'ampiezza della finestra consente di individuare eventuali problemi dello strumento nell'arco di una giornata. Fissato M il valore ottimale per h è stato individuato per mezzo di un esperimento di simulazione. Le simulazioni sono state effettuate considerando diversi valori della soglia h e della variazione standardizzata δ comprendendo così situazioni con variazione e senza variazione. L'istante di rottura è stato imposto ad un tempo t_0 noto, stimando poi il ritardo $t_a - \hat{t}_0$. I risultati più significativi dell'esperimento, del quale si sono effettuate 1000 replicazioni per ogni valore di h e δ, sono riportati nella tabella 4. Sono stati presi in esame anche valori di |δ| inferiori ad uno, tuttavia i risultati hanno evidenziato che per rilevare tali variazioni sarebbero necessari valori molto bassi di h che porterebbero ad un elevato tasso di falsi allarmi rendendo inutilizzabile l'algoritmo per l'applicazione. Dall'esame della tabella 4 si rileva che il valore preferibile di soglia che consente di rilevare la variazione più piccola con un

	h=5	h=6	h=7	h=8	h=9	h=10
δ	m.d.	m.d.	m.d.	m.d.	m.d.	m.d
3.0	0.000	0.000	0.003	0.020	0.039	0.005
-3.0	0.000	0.000	0.000	0.020	0.044	0.045
2.5	0.002	0.030	0.093	0.026	0.376	0.593
-2.5	0.004	0.059	0.153	0.029	0.459	0.703
2.0	0.293	0.644	1.138	1.700	2.249	2.832
-2.0	0.401	0.760	1.231	1.768	2.234	2.955
1.5	2.517	3.626	4.843	6.070	7.468	8.779
-1.5	2.674	3.831	5.092	6.319	7.820	9.177
1.0	10.052	12.874	16.141	19.830	23.421	27.028
-1.0	10.368	13.694	17.270	20.647	24.105	27.558

Tabella 4: Risultati della simulazione per la scelta di h; δ = variazione standardizzata; m.d.= ritardo medio.

ritardo medio non superiore a 24 è pari a h = 8. In corrispondenza di tale valore si ha un tasso di falsi allarmi più che accettabile pari a 0.012.

5 Applicazione

Una volta determinati i valori "ottimali" per h e M l'algoritmo GLR è applicato ai dati partendo da Maggio 1994 fino a Ottobre 1995, sospendendone l'impiego nella stagione fredda. Prima di illustrare i risultati è utile spiegare la gestione del monitoraggio in quanto la frequente mancanza di dati inficia la possibilità di un utilizzo continuo dell'algoritmo. Nell'intento di far funzionare l'algoritmo anche in presenza di dati mancanti si è optato per la seguente strategia: a) quando vengono a mancare pochi dati (1 o 2) nella serie O_3, il calcolo del residuo e conseguentemente della statistica g_t, non viene effettuato. Nel contempo però la stima del vettore di stato viene effetuata condizionatamente sulla base dell'informazione disponibile. Con l'arrivo dei nuovi dati osservati si aggiorna l'ultima previsione a più passi e la si considera come il valore iniziale da cui far ripartire l'algoritmo. Il calcolo della statistica g_t riprende quindi dall'osservazione successiva per evitare che l'imprecisione di una previsione a più passi porti ad un falso allarme. In questo modo l'algoritmo continua a funzionare anche in presenza di isolati valori mancanti; b) quando la mancanza di dati si protrae nel tempo, l'algoritmo viene interrotto. Il monitoraggio riprende quando sono nuovamente disponibili le osservazioni dell'ozono. In questa situazione le stime iniziali dello stato relativamente alle prime osservazioni risultano imprecise. Per evitare quindi eventuali falsi allarmi il calcolo della funzione di decisione è effettivamente realizzato a partire dalla terza osservazione disponibile.

I risultati salienti dell'applicazione di questa procedura sono riassunti nella tabella 5, dove sono riportati per ciascun mese il numero degli allarmi rilevati, il ritardo medio stimato nella segnalazione dell'allarme e il valore medio delle stime degli shift segnalati.

A titolo illustrativo abbiamo riportato nelle figure 8 e 10 anche i grafici relativi all'andamento della statistica g_t per alcuni mesi (Luglio 1994 e Settembre 1995). Per una migliore interpretazione dei risultati abbiamo ritenuto utile trasferire l'informazione fornita dall'andamento della statistica g_t direttamente sulla serie dell'ozono. Nei grafici corrispondenti (figure 9 e figura 11) il tempo di rottura viene segnalato con un quadrato (▪) mentre il tempo di allarme con un cerchio (•).

Mese	*Allarmi*	*Ritardo Medio*	*Variaz. Mediat*
Maggio 94	27	3	-1.519
Giugno 94	9	6	-1.291
Luglio 94	3	14	-0.879
Agosto 94	0	0	0.000
Settembre 94	5	5	-1.366
Ottobre 94	2	2	-1.477
Maggio 95	1	0	-2.084
Giugno 95	0	0	0.000
Luglio 95	1	0	-1.648
Agosto 95	0	0	0.000
Settembre 95	1	15	-0.913
Ottobre 95	0	0	0.000

Tabella 5: Risultati dell'applicazione dell'algoritmo GLR

L'analisi dei risultati va effettuata con qualche cautela dal momento che si sta simulando ex-post il funzionamento *on-line* dell'algoritmo e non si è in grado di intervenire effettivamente sullo strumento quando viene segnalato un allarme. Tuttavia si può affermare che l'algoritmo impiegato mostra una buona capacità nel segnalare allarmi plausibilmente dovuti ad anomalie nel sensore, visto che spesso un allarme è seguito da dati mancanti. Anche la tempestività di rilevazione delle anomalie risulta buona: i ritardi medi di rilevazione si mantengono abbondantemente al di sotto del valore di finestra M=24. Infine è evidente un calo nel numero di allarmi nel 1995. Questo è probabilmente dovuto ad una migliore manutenzione degli strumenti: a sostegno di questa ipotesi si è potuto verificare che la manutenzione delle stazioni di rilevamento nel corso del 1995 è stata affidata ad una ditta diversa rispetto al 1994.

In conclusione, l'algoritmo proposto, se pure con le cautele sopra menzionate, sembra fornire risultati soddisfacenti nel rilevare, sulla base del modello identificato, andamenti anomali per l'inquinante in questione, e quindi una volta effettivamente impiegato *on-line*, può diventare un utile strumento di supporto nelle politiche di controllo dell'inquinamento atmosferico.

Ricordiamo che la rilevazione è il primo passo per l'attuazione di queste politiche: è infatti importante che i dati su cui si lavora siano il più possibile affidabili per formulare corrette conclusioni concernenti la verifica del superamento dei limiti imposti per legge, oltre che per le successive analisi.

A questo fine la verifica dell'affidabilità delle concentrazioni di un inquinante, attraverso un algoritmo di rilevazione *on-line*, permette di ottenere

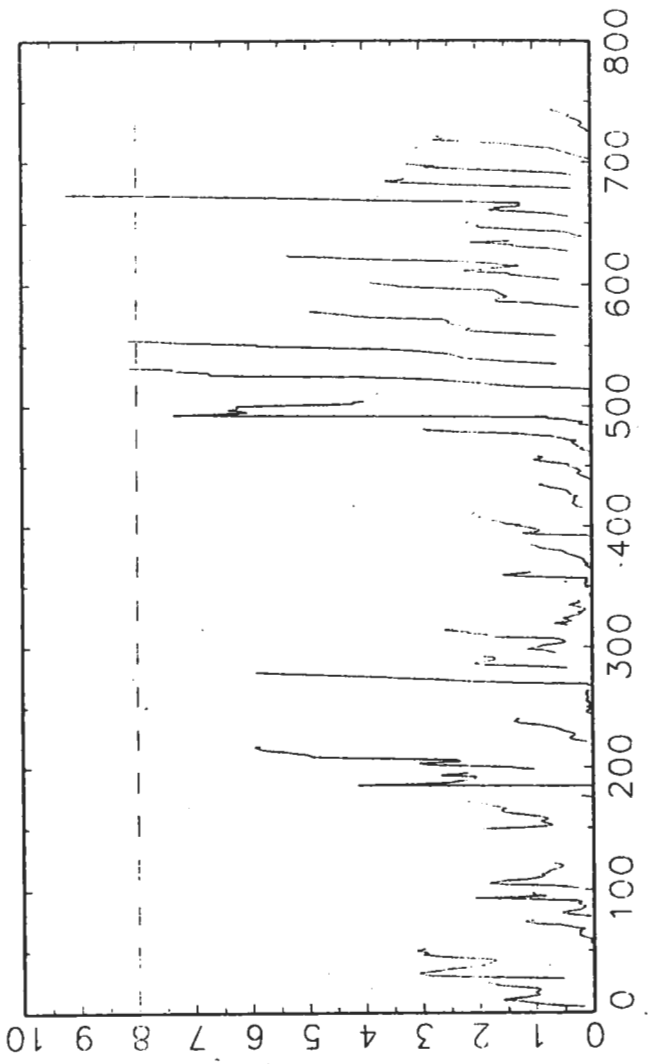

Figura 8: statistica g_t e soglia h (7/94)

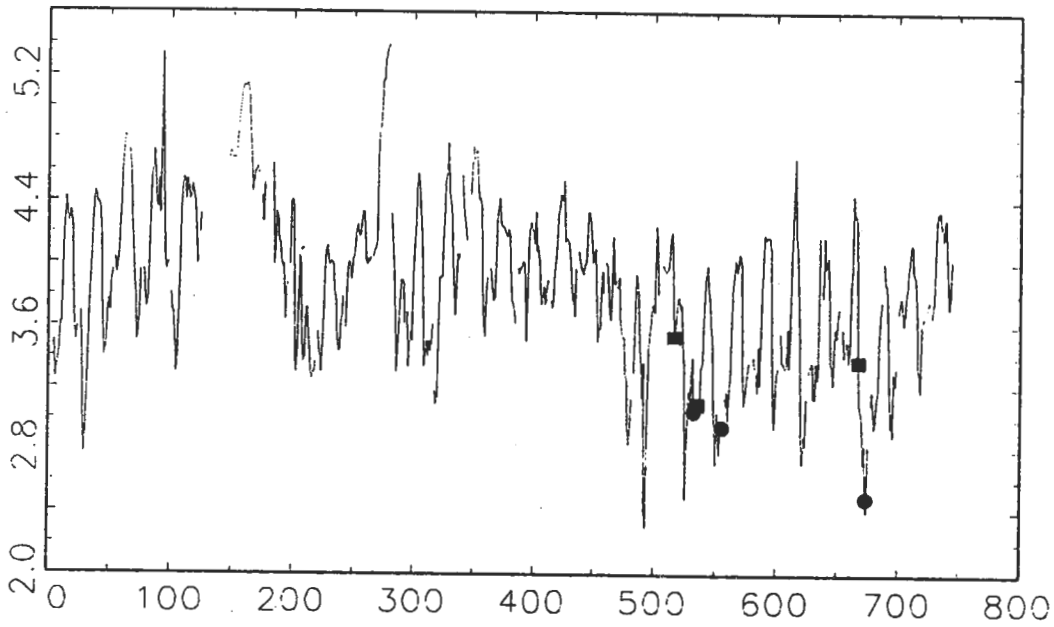

Figura 9: log-ozono con tempo di rottura e tempo di allarme (7/94)

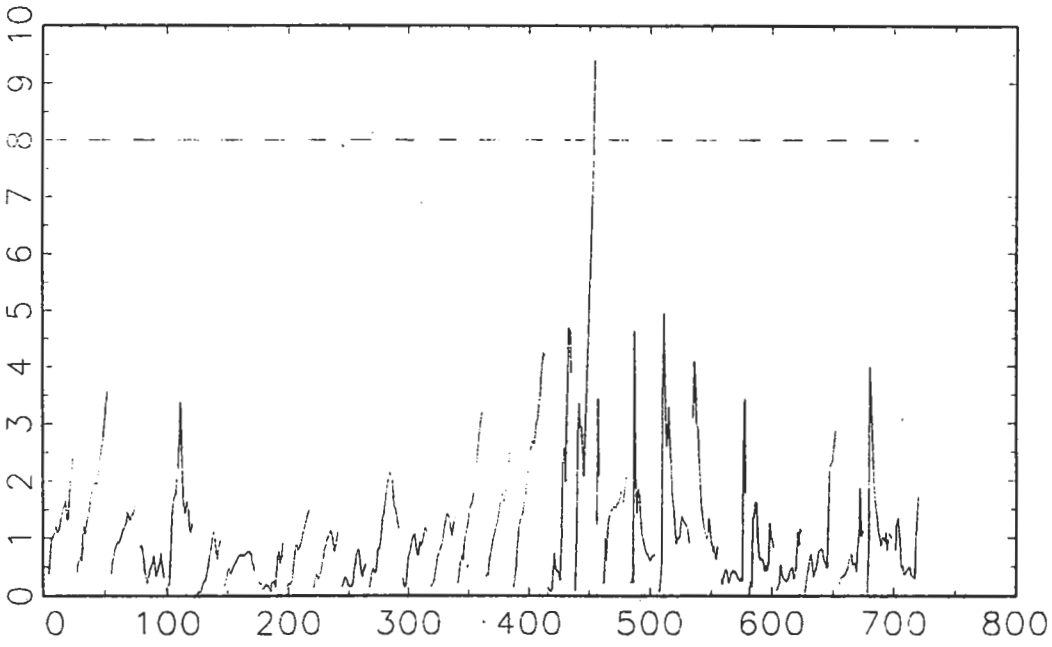

Figura 10: statistica g_k e soglia h (9/95)

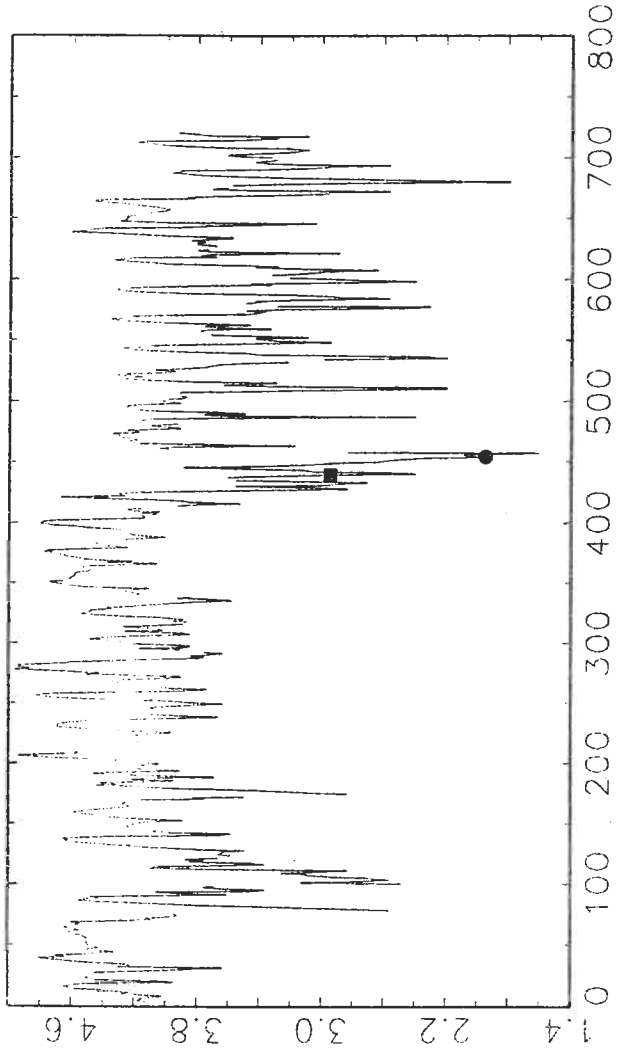

Figura 11: log-ozono con tempo di rottura e tempo di allarme (9/95)

che: *i)* il funzionamento stesso della centralina di rilevazione sia tenuto costantemente sotto controllo; *ii)* in presenza di dati anomali ci si possa recare nella centralina per verificare attraverso la taratura l'effettivo malfunzionamento della strumentazione; *iii)* mediante la stima del tempo di rottura si possano ripulire gli insiemi dei dati dalle osservazioni ottenute con strumenti non funzionanti; *iv)* infine, più velocemente è rilevato e corretto l'eventuale malfunzionamento, minore risulta il numero di osservazioni che vengono perse, ottenendo così un archivio di dati, oltre che più affidabile, anche più completo.

Riferimenti bibliografici

Basseville M., Nikiforov I.V. (1993) *Detection of abrupt changes:theory and applications* Prentice Hall, Englewood Cliffs, New Jersey.

Batterman S.A. (1992) Optimal estimators for air quality levels, *Atmospheric Environment*, vol.26A, no. 1,pp. 113-123.

Brockwell P.J., Davis R.A. (1991) *Time series:theory and methods* Springer-Verlag, New York,.

Davis M.H., Vinter R.B. (1984) *Stochastic modelling and control*, Chapman & Hall, London.

Davison A. C, Hemphill M. W. (1987) On the statistical analysis of ambient ozone data when measurements are missing, *Atmospheric Environment*, vol.21, pp. 629-639.

Lai, T.L. (1995) Sequential changepoint detection in quality control and dynamical systems, *Journal of the Royal Statistical Society,* vol.B, no 57, pp. 613-658.

Lorden G. (1971) Procedures for reacting to a change in distribution, *Annals Mathematical Statistics*, vol.42, pp. 1897-1908.

Montgomery D.C. (1991) *Introduction to Statistical Quality Control*, John Wiley & Sons, New York.

Page E. S. (1961) Cumulative Sum Charts, *Technometrics* 3, pp. 1-9

Shewhart W. A. (1931) *Economic Control of Quality of Manufactured Product.* Van Nostrand, New York.

Roberts S.W. (1959) Control Chart Tests Based on Geometric Moving Averages *Technometrics* 1, pp. 239-250

Willsky A.S., Jones H.L. (1976) A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems, *IEEE Transactions on Automatic Control*, vol.AC-21, no.1, pp.108-112.