UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA IN SCIENZE STATISTICHE ED ECONOMICHE

TESI DI LAUREA

IL CONTROLLO STATISTICO DI PROCESSO SU DATI MULTIVARIATI:

un caso di studio

Relatore: Prof.ssa CAPIZZI GIOVANNA

Laureanda: MARASCALCHI MARTINA

MATRICOLA: 437196

ANNO ACCADEMICO 2002-2003

Indice

Introduzione

Il controllo statistico di processo è generalmente inteso come l'insieme di particolari tecniche statistiche applicate alla sorveglianza di un processo, con lo scopo di migliorarne la produttività e la qualità. Questi metodi statistici vengono spesso utilizzati per problemi nei quali è coinvolta un'unica caratteristica e si è quindi interessati a studiare il comportamento di una singola variabile. Nella realtà industriale, tuttavia, la maggior parte dei processi produttivi vede implicate numerose variabili che devono essere analizzate in maniera congiunta al fine di ottenere più informazioni possibili sul processo dalle relazioni che intercorrono tra esse. Molte caratteristiche dei processi, infatti, risultano correlate tra loro e nel tempo: per migliorare la performance del controllo statistico multivariato, si è reso quindi necessario riuscire a cogliere la natura e l'entità del legame esistente tra le variabili e al loro interno nel tempo. Un'altra difficoltà da affrontare nel monitoraggio di un processo è legata alla ridondanza di informazioni che si origina quando il numero di variabili è troppo elevato. E' importante cercare di eliminare la presenza di informazione superflua che è all'origine di problemi di collinearità.

A partire da questi obiettivi è stata sviluppata (Jackson e Mudholkar, 1979), l'analisi delle componenti principali (PCA) applicata al controllo statistico di processo. L'utilizzo del metodo PCA porta alla creazione di un nuovo set di osservazioni tra loro incorrelate, utilizzando una trasformazione lineare delle variabili originali. La matrice di varianza e covarianza dei dati del processo viene decomposta in valori singolari dando luogo ad una matrice di autovettori, che rappresentano i legami lineari tra le variabili, ed una di autovalori, che esprime la variabilità delle caratteristiche oggetto di studio. Sulla base di questo procedimento sono state, in seguito, calcolate due statistiche: la T^2 di Hotelling, capace di misurare la variabilità "spiegata" dal modello, e la statistica Q, definita come somma dei quadrati dei residui, che rileva la quantità di varianza non individuata dalle componenti principali.

I metodi di controllo statistico multivariato, sopra menzionati, non si rivelano sempre efficienti nell'individuare eventuali cambiamenti nelle correlazioni tra le variabili, soprattutto quando le statistiche T^2 e Q restano entro i limiti di controllo.

Per un corretto funzionamento di queste tecniche di controllo, inoltre, deve valere l'assunzione di indipendenza delle osservazioni nel tempo, assunzione non sempre garantita. Per incrementare la capacità dei metodi multivariati per il controllo statistico, *Kano et al.* (2000-2002) hanno messo a punto due nuove tecniche per il monitoraggio di processo. Tali proposte prendono in esame un modello di dati di tipo dinamico, in grado di considerare la dipendenza di ogni variabile studiata dalle proprie osservazioni passate.

Il primo metodo è denominato "*Moving Principal Component Analysis*" (MPCA) ed è basato sull'idea che un cambio nella struttura di correlazione delle variabili, quindi una variazione delle condizioni operative del processo, può essere individuato monitorando la direzione delle componenti principali. A tale proposito è stato proposto l'uso di un indice che permetta di misurare l'entità del cambiamento confrontando la direzione delle variabili latenti di un insieme di riferimento con quelle del processo che si vuole studiare.

La seconda tecnica prende avvio dalla considerazione che i cambiamenti di natura specifica e non casuale del processo possono essere rintracciati analizzando la distribuzione della serie di osservazioni. La procedura consiste nel misurare la differenza tra due distribuzioni: quella di un insieme di dati ritenuto in condizioni di normalità e quella dell'insieme oggetto di studio. Anche in questo caso si è reso necessario l'uso di una misura capace di valutare quantitativamente questa differenza, a tale scopo è stato introdotto l'indice di diversità D o *dissimilarity index*.

L'obiettivo di questo studio è quello di operare un confronto tra tutte le tecniche multivariate di controllo sopra esposte; il processo che si intende analizzare non è di tipo industriale o produttivo ma deriva dal monitoraggio di alcune sostanze inquinanti e variabili meteorologiche rilevate in una contea del Texas.

Il primo capitolo illustra la necessità di passare da un controllo statistico di tipo univariato ad uno multivariato, allo scopo di tenere conto della correlazione che esiste spesso tra le molteplici variabili di un processo. Viene in seguito spiegato il metodo dell'analisi delle componenti principali dal quale prende avvio il calcolo della statistica T^2 e la costruzione della corrispondente carta di controllo; per cercare di completare l'informazione fornita da questo metodo di controllo tradizionale, viene inoltre presentata la carta dei residui basata sul calcolo della statistica *Q*.

Nel secondo capitolo, tenendo conto del fatto che le osservazioni di un processo possono essere caratterizzate dalla presenza di autocorrelazione al loro

interno, si giustifica la scelta di lavorare in un ambiente dinamico; vengono quindi introdotti i nuovi metodi di controllo multivariato basati sul calcolo dell'indice A e dell'indice di diversità D.

Il terzo capitolo si introducono le variabili ambientali, oggetto del controllo statistico, e la loro analisi descrittiva. In particolare, vengono prese in considerazione le serie di dati provenienti da tre diverse stazioni di rilevazione situate all'interno di una contea del Texas. Dopo aver spiegato brevemente le proprietà di ogni sostanza rilevata ed i legami tra gli inquinanti e le variabili meteorologiche, si procede allo studio dei principali aspetti della distribuzione di ogni variabile misurata nelle tre le centraline.

Il quarto capitolo riguarda, infine, l'applicazione al caso reale dei metodi multivariati descritti in precedenza: dopo aver considerato la correlazione esistente tra le variabili e tra le stazioni di rilevazione, vengono adottati due modelli dinamici per i quali sono state costruite le carte di controllo multivariate tradizionali e quelle degli indici A e D. Il primo modello prende in considerazione due variabili correlate tra loro in maniera positiva. Il secondo modello si basa su quello precedente ed è caratterizzato dall'introduzione di un'ulteriore variabile correlata negativamente con le precedenti. Lo scopo che ci si prefigge è quello di individuare la presenza di valori fuori controllo che segnalano cambiamenti nella struttura di relazione tra le variabili nell'arco di un dato intervallo temporale. In particolare si cercherà di capire se le procedure descritte permettono di individuare il contributo dato sia dalle diverse variabili sia dalle diverse centraline nella determinazione di valori fuori controllo. In ultima analisi sono state costruite le carte di controllo univariate per ogni variabile del modello preso in esame con lo scopo di agevolare l'individuazione delle variabili responsabili del cambiamento.

Capitolo 1

Il controllo statistico di processo

1.1 Introduzione

Il controllo statistico di processo (SPC) è generalmente inteso come un insieme di metodi e strumenti statistici in grado di assicurare che un qualsiasi processo produttivo resti qualitativamente efficiente e non sia soggetto a cause specifiche in grado di produrre cambiamenti e disturbi al suo interno. Tra gli strumenti più utilizzati per stabilire se esiste o meno uno stato di controllo del processo, quello di maggior utilizzo è sicuramente la carta di controllo che permette di distinguere le cause accidentali da quelle identificabili e suggerisce il momento in cui è necessario intervenire.

In generale, il problema da risolvere è una verifica dell'ipotesi nulla di "processo in controllo" contro l'ipotesi alternativa di un " cambiamento dovuto a cause non accidentali". Si tratta quindi di calcolare una conveniente statistica di controllo z_t tale che, fissati i due limiti superiore ed inferiore (LS e LI) della regione di accettazione, permetta di stabilire se accettare o rifiutare l'ipotesi nulla:

$$z_t \in (LI, LS) \Rightarrow accetto H_0$$
$$z_t \notin (LI, LS) \Rightarrow rifiuto H_0$$

Le carte per il controllo statistico di processo utilizzate più frequentemente , sono quelle del tipo Shewhart, tradizionalmente formate da due grafici: la carta $\bar{x}$ nella quale sono riportate le medie delle osservazioni, e la carta R delle escursioni

che permette di analizzare la variabilità del processo. L'uso di carte di controllo univariate tradizionali, come la carta Shewhart, la CUSUM e la EWMA, non risulta tuttavia appropriato quando si devono studiare processi, come quelli chimici o industriali, caratterizzati da un alto numero di variabili che risultano spesso dipendenti le une dalle altre; in tali circostanze è necessario prendere in considerazione metodi multivariati di controllo di processo (Wierda, 1994; Mason, Tracy, Young, 1995).

Nell'ambito di processi con un elevato numero di caratteristiche, infatti, può accadere che una specifica causa di variazione che interessa una variabile provochi un cambiamento nell'equilibrio dell'intero processo a causa delle relazioni che sussistono tra questa variabile e le altre. Si tratta allora di adottare un metodo di analisi multivariato che prenda in considerazione i rapporti e le correlazioni tra le diverse caratteristiche del processo e permetta di creare uno schema di controllo da applicare congiuntamente a tutte le variabili in esame. A questo proposito sono stati proposti metodi di controllo multivariati basati sulla statistica T^2 di Hotelling, che misura la distanza tra la media campionaria e il valor medio specificato sotto l'ipotesi nulla (Hotelling, 1947).

Il metodo delle componenti principali è stato quindi applicato al controllo statistico della qualità, con l'obiettivo primario di ridurre il numero di variabili maggiormente responsabili di disturbi e variazioni nel processo. Si è giunti, in tale maniera, alla costruzione di due carte di controllo: quella per la statistica T^2 basata sull'analisi delle componenti principali, in grado di cogliere l'ammontare di variazione tra le componenti del modello, e quella per la statistica Q, costruita come somma dei quadrati dei residui delle variabili latenti e capace di misurare la variazione non considerata dal modello PCA.

Il primo vantaggio che si può trarre da questa applicazione consiste nella possibilità di costruire carte di controllo come la T^2 utilizzando variabili trasformate tra loro incorrelate: per costruzione infatti le componenti principali risultano tra loro ortogonali e quindi indipendenti. Questa proprietà permette dunque di ottenere misure del processo prive delle informazioni ridondanti presenti invece nei dati originali. Da questo punto di vista l'uso della statistica T^2 contribuisce, quindi, ad una semplificazione e ad un miglioramento nell'interpretazione dei risultati ottenuti dal controllo statistico. D'altra parte questo strumento non è in grado di dare informazioni circa la natura di eventuali variazioni intercorse nel processo: essa

risulta sensibile a cambiamenti sia nella media che nella varianza di processo ma non è altrettanto efficace nel distinguere tra queste due tipologie di variazione (D. M. Hawkins, 1993). Per cercare di incrementare le informazioni fornite dalle carta di controllo appena descritte, alcuni autori hanno proposto di decomporre la statistica T^2 in maniera tale da riuscire ad individuare le variabili maggiormente responsabili di eventuali comportamenti anomali delle osservazioni (Mason, Tracy, Young, 1995).

1.2 Carte di controllo tradizionali per l'analisi multivariata

Una carta di controllo basata sulla statistica T^2 tiene conto della struttura di correlazione presente nella popolazione ottenendo un miglioramento rispetto al contributo dato dalle carte univariate al monitoraggio di un processo.

Se si considera, ad esempio, il caso di un sistema formato da due variabili, x_1 e x_2, correlate in maniera positiva, si può notare come l'utilizzo di carte di controllo univariate porta ad ottenere una regione di accettazione rettangolare definita dai limiti di controllo superiore ed inferiore, di ciascuna variabile (Figura 1.1). Le osservazioni del processo oggetto di studio si trovano tutte all'interno della regione e non segnalano quindi particolari situazioni di fuori controllo.

Tenendo in considerazione la distribuzione congiunta delle due variabili, la regione di accettazione assumerà una forma con l'esclusione di alcune osservazioni, assunte come potenziali valori anomali (Figura 1.2). Questo esempio serve a mettere in luce la povertà delle carte univariate nel rilevare situazioni di fuori controllo quando si trattano processi con più variabili che si presentano tra loro correlate.

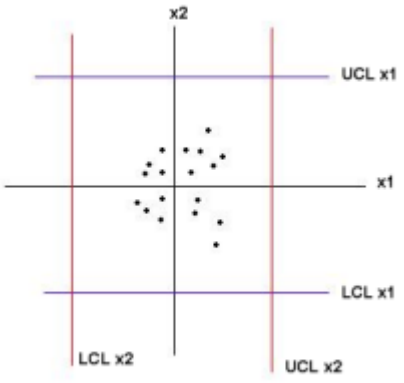


Figura 1.1: *carta di* ***controllo***

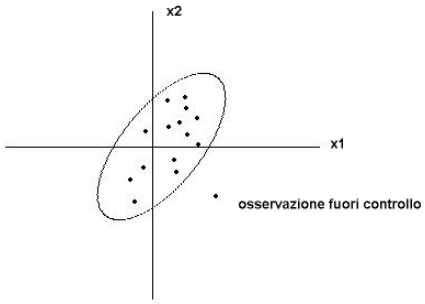


Figura 1.2: *carta di controllo*

1.2.1 Costruzione della carta T^2

La statistica sulla quale si basa questa carta di controllo è pari alla distanza di Mahalanobis tra il vettore medio delle osservazioni e quello specificato sotto l'ipotesi nulla:

$$T^2(x_i) = n(\bar{x} - \mu_0)^T S^{-1}(\bar{x} - \mu_0) \qquad (1.1)$$

dove con $\bar{x}$ ed S si indicano rispettivamente il vettore delle medie e la matrice di varianza e covarianza campionaria.

La costruzione della carta avviene in due fasi:

- *Analisi dei dati passati*

  Scopo di questa prima analisi è ottenere un insieme di dati preliminari del processo in controllo in modo da poter basare su questi la verifica delle osservazioni successive. Si considerino quindi *k* campioni di grandezza $n>p$ da una normale $N_p(\mu,\Sigma)$, la statistica T^2 calcolata per ogni campione viene riportata sulla carta di controllo e confrontata con i limiti

$$\begin{aligned} LS &= \frac{p(k-1)(n-1)}{k(n-1)-p+1} F_{\alpha,(p,k(n-1)-p+1)} \\ LI &= 0 \end{aligned} \qquad (1.2)$$

  I valori di T^2 che superano le soglie (1.2) sono esclusi dalla popolazione di riferimento e i limiti vengono ricalcolati sulla base dei campioni rimanenti.

- *Analisi dei dati correnti*

  In questa fase si tratta di verificare se il processo corrente risulta sotto controllo oppure soggetto a cambiamenti nella media o nella varianza provocati da cause determinabili. Prendendo come riferimento il vettore medio $\hat{\mu}$ e la matrice di varianza e covarianza $\hat{S}$ calcolati sulla base dei dati in controllo ottenuti nella prima parte dell'analisi, si calcola la statistica T^2 per ognuno degli *n* campioni da una $N_p(\mu,\Sigma)$ provenienti dal processo e indipendenti da quelli considerati in precedenza

(1.3)

$$T^2 = n(\bar{x} - \hat{\mu})^T \hat{S}^{-1}(\bar{x} - \hat{\mu})$$

I valori ottenuti vengono riportati sulla carta assieme ai limiti

$$LS = \frac{p(m+1)(n-1)}{m(n-1) - p + 1} F_{\alpha,(p,m(n-1)-p+1)} \qquad (1.4)$$
$$LI = 0$$

dove *m=k-a* è il numero di campioni in controllo ottenuti nell'analisi dei dati passati. Se il limite superiore viene superato significa che è sorta una specifica causa di variazione che ha portato il processo fuori controllo.

Il principale vantaggio dell'utilizzo della carta T^2 consiste nel fatto che la statistica utilizzata permette di tener conto della struttura di correlazione della popolazione, essa però ha la lacuna di non agevolare l'interpretazione dei segnali d'allarme dal momento che fornisce poche indicazioni utili per stabilire quali tra le variabili sono le maggiori responsabili della presenza di valori anomali.

Un metodo per ovviare a questa carenza è stato proposto da Mason, Tracy e Young (1995) ed è basato sull'idea di ottenere dall'insieme di dati originali due sottogruppi distinti; un primo gruppo $X_{.,(p-1)}$, al quale appartengono le prime *p*-1 caratteristiche ed un secondo gruppo $X_{.,p}$, formato solo dalla *p*-esima variabile:

$$X_i = (X_{i1}, X_{i2}, ...X_{ip}) \rightarrow X_i = (X_{i,(p-1)}, X_{i,p})$$

La statistica test calcolata per il sottogruppo delle *p*-1 variabili sarà quindi data da:

$$T^2_{(p-1)} = (\bar{x}_{i,(p-1)} - \bar{x}_{(p-1)})' S^{-1}_{XX} (\bar{x}_{i,(p-1)} - \bar{x}_{(p-1)})$$

dove $\bar{x}_{(p-1)}$ e S_{XX} sono rispettivamente il vettore delle medie campionarie calcolato sulle prime (*p*-1) caratteristiche e la corrispondente matrice di varianza e covarianza, dato che vale:

$$S = \begin{bmatrix} S_{xx} & s_{xX} \\ s_{xX} & s_x^2 \end{bmatrix} \qquad (1.5)$$

La statistica T^2 può quindi essere decomposta nel seguente modo:

$$T^2 = T^2_{(p-1)} + T^2_{p,....p-1} \qquad (1.6)$$

dove

$$T^2_{p,...p-1} = \frac{X_{ip} - \overline{X}_{p,...p-1}}{s_{p,...p-1}}$$

rappresenta la statistica corrispondente alla *p*-esima variabile di X_i che è stata standardizzata rispetto alla propria media e deviazione standard condizionate alle precedenti *p*-1 distribuzioni ($X_{1,1}, X_{i,2}...X_{i,(p-1)}$). In particolare, definito b_p il vettore che stima i coefficienti di regressione della *p*-esima variabile sulle precedenti *p*-1

$$b_p = S^{-1}_{XX} s_{xX}$$

la media di X_{ip} condizionata alle altre *p*-1 caratteristiche è data dalla relazione:

$$\overline{X}_{p,...p-1} = \overline{X}_p - b'_p (X_{i,(p-1)} - \overline{X}_{(p-1)})$$

Un secondo metodo per identificare quali tra le variabili in esame hanno causato la presenza di fuori controllo nel processo è stato messo a punto da Jackson (1991) ed è basato sulla teoria delle componenti principali che, essendo combinazioni lineari delle variabili originali, permettono di ridurre la dimensionalità del problema. Infatti, uno dei maggiori problemi che si incontra lavorando con un elevato numero di variabili è quello di avere una ridondanza di informazioni che anziché favorire la comprensione delle relazioni esistenti tra le caratteristiche esaminate, ostacolano l'efficienza di uno schema di controllo statistico.

1.3 Analisi delle componenti principali

L'obiettivo di questo metodo di analisi è quello di ridurre la dimensione di un problema multivariato permettendo di passare, senza perdere troppe informazioni, da p variabili correlate a $k<p$ componenti incorrelate, combinazioni lineari delle variabili originali.

Il punto di partenza dell'analisi per componenti principali consiste nella possibilità di tradurre una matrice simmetrica e non singolare, come quella di varianza e covarianza S, in una matrice diagonale L, pre-moltiplicandola e post-moltiplicandola per una matrice U ortonormale:

$$U'SU=L$$

Le colonne della matrice U sono detti "vettori caratteristici" o "autovettori" di S mentre gli elementi sulla diagonale di L sono chiamati "radici caratteristiche" (l_1, l_2, …l_p) o "autovalori" e possono essere determinati tramite l'*equazione caratteristica*:

$$\left|S - lI\right| = 0 \qquad (1.7)$$

dove I è la matrice identità.

Geometricamente questa procedura si traduce in una rotazione degli assi principali della matrice di covarianza. Le direzioni dei coseni dei nuovi assi rispetto ai vecchi sono rappresentate proprio dagli elementi dei vettori caratteristici (Figura1.3)

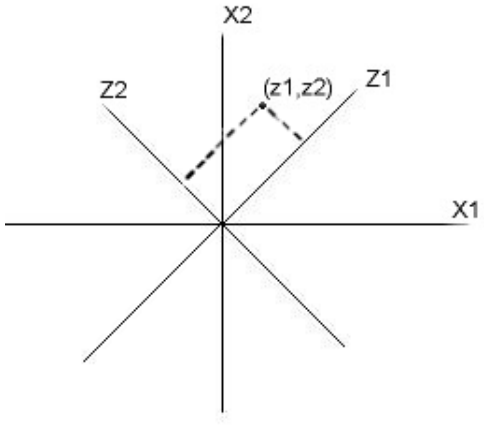


Figura 1.3: *Rotazione degli assi*

Consideriamo un problema *p*-variato in cui la matrice di varianze e covarianze sia rappresentata da

$$S = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{12} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & . \\ s_{1p} & s_{2p} & \cdots & s_p^2 \end{bmatrix}$$

Naturalmente se le covarianze non sono nulle significa che esiste una qualche relazione tra le variabili la cui forza del legame può calcolata grazie al coefficiente di correlazione ($r_{ij}=s_{ij}/s_i s_j$).

La rotazione degli assi porterà le *p* variabili correlate a trasformarsi in nuove *p* variabili incorrelate le cui coordinate saranno espresse negli autovettori u_i appartenenti alla matrice *U*:

$$x_1, x_2, ...x_p \rightarrow z_1, z_2, ...z_p$$

Le variabili trasformate sono chiamate componenti principali o variabili latenti di *x*

$$z = U'(x - \bar{x}) \qquad (1.8)$$

in particolare z_i rappresenta la *i*-esima componente principale e può essere scritta come combinazione lineare di variabili x_i^* centrate rispetto alla media:

$$z_i = u_i' x^* = u_{1i} x_1^* + u_{2i} x_2^* + ...u_{pi} x_p^*$$

Ciascuna di queste componenti avrà media nulla e varianza rappresentata dagli autovalori l_i, calcolati risolvendo l'equazione caratteristica (1.7). In particolare, gli autovalori vengono ordinati in maniera decrescente $l_1 > l_2 ... > l_p$ e rappresentano la porzione di varianza del processo che può essere spiegata da ogni componente. Ogni variabile latente è quindi in grado di spiegare una parte sempre più piccola della varianza del processo: la prima componente è rappresentata dalla combinazione lineare delle variabili originali in grado di spiegare la massima

varianza, la seconda da una combinazione, ortogonale alla prima, con varianza subito inferiore e così via. Complessivamente la variabilità delle *p* variabili originali è spiegata da tutte le *p* componenti calcolate; alcune variabili latenti però risultano più rappresentative delle altre in quanto sono in grado di spiegare una percentuale maggiore di varianza totale. E' estremamente importante a questo punto, per soddisfare un criterio di parsimonia e di non ridondanza delle informazioni, saper scegliere un numero di componenti principali che sia adeguato a rappresentare l'intero processo portando ad un decisivo miglioramento dell'interpretazione dei dati e ad una perdita limitata di informazione. Il metodo delle componenti principali si propone proprio di proiettare l'insieme di dati originali in uno spazio ortogonale di dimensioni $k<p$ sufficienti a descrivere la maggior parte della variabilità del processo.

Esistono diversi criteri per la selezione del numero di componenti principali, in generale si tende a prendere in considerazione tutte le variabili latenti in grado di spiegare tra il 70% e il 90% della varianza totale; una regola comunemente usata stabilisce di escludere le componenti associate ad autovalori minori di uno (per osservazioni scalate). La procedura SCREE (Jackson, 1979) è un metodo grafico che permette di scegliere il numero di variabili latenti da considerare è di seguito descritto:

- Si sceglie *k*, numero di variabili latenti, in modo tale che

$$\sum_{i=1}^{k} l_i \geq 0.9 \sum_{i=1}^{p} \sigma_i^2$$

dove σ_i^2 rappresenta la varianza di x_i

- Si incrementa *k* fino a che l_i risulta inferiore della varianza media di x_i
- Si disegna il grafico di $\sum_{i=1}^{j} l_i$ e si individua il valore *k* in corrispondenza del quale il grafico presenta una curva più accentuata.

1.3.1 La carta T^2 costruita con il metodo PCA

Jackson (1980) suggerisce un metodo di controllo statistico di un processo basato sulle componenti principali. In primo luogo, egli suggerisce di riscalare le variabili trasformate dividendole per le loro radici caratteristiche:

$$w_i = u_i / \sqrt{l_i} \Rightarrow y_i = w_i'(x - \bar{x}$$

in questo modo si ottengono variabili che, oltre ad essere incorrelate e centrate rispetto alla media , hanno anche varianza unitaria.

Procedendo in questo senso, la statistica T^2 (1.1) può essere scritta come la somma dei quadrati di componenti principali indipendenti

$$T^2 = \sum_{i=1}^{p} y_i^2 = y'y \qquad (1.9)$$

dal momento che la matrice di varianza e covarianza delle variabili latenti scalate è una matrice identità.

La statistica T^2 inoltre è legata alla distribuzione *F* dalla seguente relazione:

$$T_{p,n,\alpha}^2 = \frac{p(n-1)}{n-p} F_{p,n-p,\alpha} \qquad (1.10)$$

per questa ragione i limiti utilizzati nella carta di controllo T^2 per componenti principali sono uguali a quelli costruiti per le variabili originali del processo. E' interessante osservare che la carta T^2 costruita per le variabili originali porta agli stessi risultati di quella costruita per tutte le componenti principali. Se si prendono invece in considerazione solo le *k* variabili più rappresentative la statistica assume la forma

$$T_k^{\ 2} = \sum_{i=1}^{k} y_i^2$$

Dopo aver costruito la carta di controllo risulta molto interessante indagare su quale delle variabili latenti è la maggior responsabile nel determinare uno stato di fuori controllo. Anche in questo caso è possibile procedere alla decomposizione della statistica: tale procedimento può avvenire in maniera più facile rispetto a quanto

visto in precedenza per le variabili originali dal momento che, per costruzione, le componenti principali che compongono la statistica T^2 sono fra loro indipendenti e non è quindi necessario tener conto delle relazioni esistenti tra le caratteristiche in esame.

Scrivendo la statistica (1.9) in maniera estesa, come:

$$T^2 = (y_1)^2 + (y_2)^2 + ... + (y_p)^2$$

si giunge alla decomposizione della statistica T^2 che assume la forma:

$$T^2 = T_1^2 + T_2^2 + ... + T_i^2 + ... + T_p^2$$

dove la generica T_i^2 rappresenta la statistica basata solo sulla *i*-esima componente principale. Nel caso in cui si consideri la statistica costruita solo sulle prime k variabili latenti, la decomposizione è del tipo

$${}_kT^2 = (y_1)^2 + (y_2)^2 + ... + (y_k)^2 = {}_kT_1^2 + {}_kT_2^2 + ... + {}_kT_i^2 + ... + {}_kT_k^2$$

1.3.2 Carta di controllo Q per i residui

L'uso esteso del metodo delle componenti principali come strumento per ridurre la dimensionalità dei dati e la frequente applicazione di questa procedura alla regressione e al controllo statistico della qualità hanno posto il problema della bontà della stima dei modelli ottenuti grazie a tale tecnica. Quando le componenti principali vengono usate come metodi di riduzione, uno strumento importante per il controllo della qualità della stima raggiunta può essere rappresentato dai residui associati alle variabili latenti, che risultano utili anche per testare la presenza di eventuali valori anomali. A questo proposito Jackson e Mudholkar (1980) proposero l'utilizzo di una carta di controllo per i residui, carta *Q*, nella quale la statistica test è rappresentata dalla differenza tra i dati originali e le osservazioni stimate mediante le componenti principali più significative. Dal momento che l'analisi delle componenti principali può essere usata per ridurre la dimensione della matrice dei valori originali, il numero di variabili latenti usato per stimare i dati del processo è generalmente più piccolo di quello delle caratteristiche originali ($k<p$). L'utilizzo dei

residui ha dunque lo scopo principale di catturate l'ammontare di variazione del processo che non viene colta dal modello per componenti principali.

La trasformazione in componenti principali consente di riscrivere i dati originali nella forma:

$$x = \bar{x} + Uy$$

dove $\bar{x}$ rappresenta il vettore delle medie campionarie, U la matrice degli autovettori che permette di invertire l'equazione (1.8) per effetto della sua ortonormalità $(U^{-1} = U')$ ed y sono i punteggi delle componenti principali:

$$y = U'x$$

Naturalmente ciò e possibile quando si considera un numero di componenti pari a quello delle variabili originali; se invece si prendono in esame $k<p$ variabili trasformate, nella ricostruzione dei dati originali si dovrà tenere conto di un termine residuo:

$$x = \bar{x} + U_k y_k + (x - \hat{x})$$

In questo caso U_k è la matrice dei vettori caratteristici costituita dalle prime k colonne, y_k sono i punteggi delle k variabili latenti e $(x - \hat{x})$ è il vettore dei residui, con:

$$\hat{x} = U_k y_k$$

pari al vettore di stima delle osservazioni originali.

La statistica Q, proposta da Jackson e Mudholkar, viene costruita come somma dei quadrati dei residui:

$$Q_k = (x - \hat{x})'(x - \hat{x}) = x'(I - U_k U_k')x$$

I limiti della carta di controllo sulla quale verranno riportati i valori di Q sono calcolati mediante la seguente relazione :

(1.11)

$$Q_\alpha = \vartheta_1 \left[\frac{c_\alpha \sqrt{2\vartheta_2 h_0^2}}{\vartheta_1} + \frac{\vartheta_2 h_0 (h_0 - 1)}{\vartheta_{10}^2} + 1 \right]^{\frac{1}{h_0}}$$

Nella (1.11), c_α è il quantile della distribuzione normale a livello (1-α) e:

$$\vartheta_1 = \sum_{k+1}^{p} l_i \qquad \vartheta_{2=} \sum_{k+1}^{p} l_i^2 \qquad \vartheta_3 = \sum_{k+1}^{p} l_i^3 \qquad h_0 = 1 - \frac{2\vartheta_1 \vartheta_3}{3\vartheta_2^2}$$

Un valore della statistica *Q* significativamente elevato potrebbe essere dovuto alla presenza di una variabilità casuale estremamente alta oppure alla possibilità che le componenti principali considerate nel modello non siano riuscite ad individuare e a spiegare tutte o nuove fonti di instabilità; una strada da intraprendere, in questo caso, per indagare sulla natura dei fuori controllo potrebbe essere quella che prevede di analizzare i residui di ciascuna delle variabili presenti nel modello.

Le carte di controllo possono talvolta funzionare in maniera poco efficiente: una prima giustificazione è data dal fatto che le statistiche T^2 e Q, quando risultano entro i limiti di controllo, non sono in grado di individuare cambiamenti nelle correlazioni tra le variabili di processo. Il secondo ostacolo può essere rappresentato dalla presenza di autocorrelazione all'interno delle variabili: a tale proposito, nel secondo capitolo, vengono descritti due metodi innovativi capaci di tener conto della dipendenza temporale delle osservazioni.

Capitolo 2

Nuovi metodi per il controllo statistico di un processo multivariato

2.1 Introduzione

L'ipotesi di base nella costruzione delle carte di controllo è che le osservazioni generate dal processo siano indipendenti e quindi incorrelate nel tempo; molte delle variabili coinvolte nei processi invece risultano correlate a vari istanti temporali e con altre caratteristiche. Un modo di procedere potrebbe allora essere quello di identificare la struttura di correlazione dei dati tramite un apposito modello serie storiche e applicare le carte di controllo tradizionali ai residui del modello stimato; le carte di controllo potranno essere considerate efficienti se i residui risulteranno tutti entro i loro limiti di controllo dimostrandosi così indipendenti. Questo procedimento purtroppo risulta abbastanza complicato e poco economico se applicato a processi dove sono implicate numerose variabili.

In questo capitolo verranno presentati due metodi innovativi per cercare di apportare un miglioramento nei metodi MSPC in caso si trattino variabili autocorrelate: il primo, detto *moving principal component analysis* (MPCA) identifica eventuali cambiamenti nella direzione delle componenti principali o nel sottospazio da queste individuato (Kano et al., 2000), il secondo, DISSIM (Kano, 2002) misura in grado di diversità tra due insiemi di dati multivariati.

2.2 Il metodo PCA dinamico

Quando si prende in considerazione una matrice di dati X di dimensioni $n \times p$ da un processo in cui le risultano incorrelate nel tempo, ci si appresta a lavorare su un modello statico poiché tutte le osservazioni sono dipendenti dal solo istante temporale t nel quale sono state rilevate. In questo caso l'analisi per componenti principali applicata alla matrice X porta a lavorare con combinazioni lineari "statiche" perché riferite solo al tempo t.

Una versione dinamica dell'analisi per componenti principali si è resa necessaria per cercare di spiegare efficacemente l'ammontare di autocorrelazione presente nei dati e nelle variabili latenti. Lavorare in un sistema dinamico permette di considerare le relazioni temporali esistenti tra le osservazioni e di conoscere quindi quanto i valori correnti dipendano dal passato. A questo proposito è necessario identificare i legami lineari sussistenti tra le variabili, cioè trovare lo spazio nullo della matrice dei dati X attraverso la soluzione della seguente equazione:

$$Xb = 0$$

Supponiamo, per esempio, di voler identificare almeno una relazione di primo ordine cioè un legame tra due istanti temporali consecutivi, la soluzione è del tipo:

$$[X_{(t)}, X_{(t-1)}]b = 0$$

Se consideriamo invece un caso generale, indicando con l l'ammontare di ritardi riscontrabili nelle osservazioni del processo, si lavora con una matrice composta da l+1 colonne. La prima è costituita dalle osservazioni iniziali, la seconda dalle stesse ritardate di un passo, fino ad arrivare alla l-esima colonna che contiene i dati originali ai quali sono stati aggiunti l ritardi.

$$X(l) = \left[X_{(t)}, X_{(t-1)}, \ldots, X_{(t-l)}\right] = \begin{bmatrix} x'_{(1)} & x'_{(0)} & \cdots & x'_{(1-l)} \\ x'_{(2)} & x'_{(1)} & \cdots & x'_{(2-l)} \\ \cdots & \cdots & \cdots & \cdots \\ x'_{(m)} & x'_{(m-1)} & \cdots & x'_{(m-l)} \end{bmatrix}$$

In questo caso la soluzione è data da:

$$X(l)b = 0$$

Per individuare l'ordine del sistema ed il numero di relazioni dinamiche presenti fra i dati è possibile adottare la seguente procedura:

I) Si pone $l = 0$ e si indica con p il numero delle colonne della matrice dei dati X.

II) Si costruisce la matrice

$$X = \left[X_{(t)}, X_{(t-1)}, \ldots, X_{(t-l)}\right]$$

III) Si procede all'analisi delle componenti principali e al calcolo dei punteggi

IV) Indicando con j il numero di componenti principali e con r(l) il numero di relazioni lineari dei dati, si pone $j = n \times (l+1)$ e r(l)=0.

V) Si verifica se la j-esima componente principale evidenzia una relazione lineare tra le variabili; se ciò avviene si prosegue al passo VI, altrimenti si va al punto VII.

VI) Si pone j=j-1 e r(l)=r(l)+1.

VII) Si calcola il numero di nuove relazioni

$$r_{new}(l) = r(l) - \sum_{i=0}^{l-1}(l-i+1)r_{new}(i)$$

VIII) Se $r_{new}(l) \leq 0$ si passa al punto X altrimenti si prosegue al IX.

IX) Si aggiorna l=l+1 e si torna al punto II.

X) STOP.

2.3 Moving Principal Component Analysis

L'idea principale sulla quale è basato il metodo MPCA è che un cambiamento delle condizioni operative del processo, ossia una variazione nelle correlazioni delle variabili che lo descrivono, può essere individuato monitorando le direzioni delle componenti principali.

L'obiettivo primario è quello di controllare lo stato del processo in maniera continua per verificare se intercorrono cambiamenti nel sottospazio determinato dalle variabili latenti. Il metodo MPCA prevede che vengano calcolate le componenti principali sulla base di sottomatrici di dati, generate di volta in volta facendo scorrere di un passo lungo l'asse del tempo la matrice delle osservazioni originali, scalate rispetto alla propria media e varianza. Dopo aver eseguito la procedura PCA su un insieme di dati in controllo preso come insieme di riferimento, viene definito un indice in grado di individuare cambiamenti nelle direzioni delle variabili latenti. Tale indice misura la differenza tra le componenti principali calcolate e quelle in uno stato di controllo, ricavate dalle sottomatrici dei dati che si intendono studiare. Indicata con $u_i(k)$ l'*i*-esima componente principale, calcolata al passo *k,* e con u_{i0} il corrispondente autovettore dell'insieme dei dati in controllo, si ottiene

$$A_i(k) = 1 - \left| u_i(k)^T u_{i0} \right| \qquad (2.1)$$

L' indice (2.1) può assumere valori compresi tra zero e uno: in particolare assume valore nullo quando l'*i*-esima variabile latente dei dati da analizzare è equivalente a quella di riferimento, e raggiunge l'unità se i due vettori in questione sono tra loro ortogonali. L'indice $A_i(k)$ calcolato ad ogni passo viene poi rappresentato graficamente assieme ad un limite di controllo: se qualche valore supera il limite di riferimento, il processo viene giudicato fuori controllo.

Il metodo MPCA permette di scoprire cambiamenti nella correlazione tra le variabili del processo che sono difficilmente individuabili utilizzando le statistiche T^2 e Q. Tuttavia se le varianze dei punteggi si rivelano troppo simili tra loro, anche questo indice può risultare poco efficace. Per ovviare a tale difficoltà è opportuno sorvegliare un cambiamento nel sottospazio determinato dalle componenti con varianza simile piuttosto che le variazioni di ogni variabile latente (Kano et al., 2001).

2.3.1 Procedura MPCA

La procedura descritta qui di seguito permette di applicare il metodo MPCA ai dati di un processo.

A] Selezione della matrice di riferimento e del limite di controllo

Si individuano in primo luogo gli autovettori di riferimento su cui basare il controllo delle successive osservazioni

1) Si osserva una matrice $X(n\times p)$, di dati provenienti da un processo operante in condizioni di controllo;
2) Le colonne di tale matrice vengono standardizzate ottenendo osservazioni di media nulla e varianza unitaria;
3) Alla matrice scalata si applica la procedura PCA al fine di ottenere i p autovettori (u_{i0}):
4) Si sceglie un'adeguata dimensione della finestra temporale w;
5) Facendo scorrere lungo l'asse temporale la matrice scalata, ottenuta al passo 2) della procedura, si ottengono (n-w+1) sottomatrici di dimensione $(w\times p)$, (Figura 2.1);
6) Si calcolano gli autovettori $u_i(k)$ per ogni sottomatrice costruita al passo precedente;
7) Vengono calcolati gli (n-w+1) valori dell'indice A_i;
8) Si calcola un limite di controllo tale che l'1% degli indici A_i sia fuori dal limite;
9) Si riportano su un grafico i valori della statistica A_i.

B] Controllo del processo corrente

Al fine di individuare eventuali cambiamenti nelle condizioni operative del processo, si procede nel modo seguente:

1) Viene acquisita una matrice Y, di dimensioni $(n \times p)$, costituita dalle osservazioni del processo che si intende studiare;
2) Si standardizzano le colonne di Y rispetto alla media e alla varianza della matrice di riferimento;
3) Vengono generate (n-w+1) sottomatrici, di dimensione $(w \times p)$,dalla matrice appena ottenuta, facendo scorrere la finestra lungo l'asse del tempo (Figura 2.1);
4) Il metodo delle componenti principali è applicato ad ogni sottomatrice e si individuano gli autovettori $u_i(k)$;
5) Si calcolano i valori dell'indice A_i utilizzando come riferimento per il confronto i vettori u_{i0} ottenuti precedentemente;
6) Viene tracciata la carta di controllo per gli (n-w+1) valori dell'indice calcolati al punto 5) utilizzando come limite quello ottenuto nella fase A];

Se qualche osservazione supera il limite, il processo viene considerato fuori controllo.

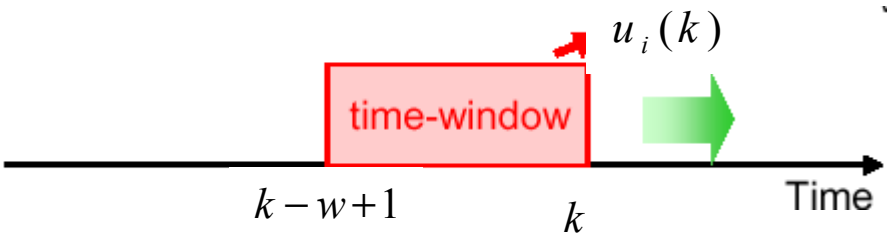


Figura 2.1: *generazione delle sottomatrici facendo scorrere la finestra lungo l'asse del tempo*

2.4 Indice di diversità

I concetti di "somiglianza" e di "diversità" sono spesso usati nei problemi di classificazione, Kano (2002) cerca di migliorare la *performance* dei metodi statistici di controllo multivariato prendendo proprio spunto dal concetto di diversità. Il metodo proposto si basa sull'idea che un cambiamento nelle condizioni operative del processo può essere osservato individuando il grado di diversità della distribuzione di un processo rispetto a quella di un insieme di dati di riferimento.

Una tecnica, equivalente all'analisi delle componenti principali, utile al fine di individuare la differenza tra le distribuzioni di due distinti insiemi di dati è l'espansione di Karhunen-Loeve (KL) (Ku, Storer e Georgakis, 1995). Tale metodo viene utilizzato, alla pari delle PCA, allo scopo di ottenere una riduzione della dimensionalità dei dati.

Si considerino due matrici di dati, X_1 e X_2, costituite entrambe da N osservazioni su p variabili, centrate rispetto alla media. La matrice di varianza e covarianza di X_i (per i=1,2) è data da:

$$R_i = \frac{1}{N_i - 1} X_i' X_i$$

mentre la covarianza totale per le due matrici è espressa come:

$$R = \frac{N_1 - 1}{N_1 + N_2 - 1} R_1 + \frac{N_2 - 1}{N_1 + N_2 - 1} R_2 = \frac{1}{N_i - 1} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}^T \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \qquad (2.2)$$

Una volta ottenuta la matrice (2.2) si procede alla sua decomposizione in valori singolari, individuando le matrici P_0 e Λ che soddisfano le seguenti equazioni

$$RP_0 = P_0\Lambda \qquad\qquad P = P_0\Lambda^{-\frac{1}{2}}$$

su questa base, i dati delle matrici X_i vengono trasformati in base alla relazione

$$Y_i = \sqrt{\frac{N_i - 1}{N - 1}} X_i P_0 \Lambda^{-\frac{1}{2}} \qquad \text{per } i=1,2$$

Le matrici di varianza e covarianza per le nuove Y_i assumono la forma

$$S_i = \frac{1}{N_i - 1} Y_i' Y_i = \frac{N_i - 1}{N - 1} P' R_i P \qquad \text{per } i=1,2$$

con:

$$S_1 + S_2 = I \qquad (2.3)$$

Procedendo, come in precedenza, alla decomposizione in valori singolari delle nuove matrici di varianze e covarianze si ottiene

$$S_i w_j^{(i)} = \lambda_j^{(i)} w_j^{(i)} \qquad \text{per } i=1,2$$

Si può notare che, per effetto della relazione (2.3), gli autovettori di S_1 *e* S_2 sono gli stessi e che sono verificate le seguenti due relazioni:

$$S_2 w_j^{(i)} = (1-\lambda_j^{(i)}) w_j^{(i)} \qquad 1-\lambda_j^{(1)} = \lambda_j^{(2)} \qquad (2.4)$$

Dal momento che gli autovettori di S_i rappresentano le direzioni delle variabili latenti e gli autovalori sono equivalenti alla loro varianza, dopo la trasformazione lineare le matrici Y_1 e Y_2 risultano avere le stesse componenti principali ma posizionate in ordine inverso per effetto delle relazioni (2.4). Questa considerazione porta alla conseguenza che la correlazione più forte per il primo insieme di dati corrisponde alla minor correlazione del secondo insieme e viceversa. Infatti per la relazione (2.4), se le serie di dati risultano tra loro abbastanza simili, gli autovalori risultano vicini a 0.5. D'altra parte, se gli insiemi delle osservazioni sono differenti un valore elevato dell'autovalore $\lambda_j^{(1)}$ determina un valore molto basso dell'autovalore $\lambda_j^{(2)}$.

Sulla base di quanto detto fino ad ora, viene costruito un apposito indice capace di misurare il grado di somiglianza tra le due distribuzioni di dati. Tale indice, D, è detto *Dissimilarity Index* o *Indice di Diversità* ed è espresso dalla relazione:

$$D = \frac{4}{p} \sum_{j=1}^{p} (\lambda_j - 0.5)^2$$

dove p indica il numero di variabili e λ_j il j-esimo autovalore della matrice di varianza e covarianza dei dati trasformati. Anche in questo caso l'indice assume valori compresi tra zero e uno: quando i due insiemi di dati sono tra loro simili l'indice risulterà vicino al valore nullo, viceversa, se le serie di dati si presentano differenti D si troverà prossimo all'unità.

2.4.1 Procedura DISSIM

Anche in questo caso, come per la procedura MPCA, è necessario adottare una finestra temporale di misura adeguata ed un insieme di dati provenienti da un processo che risulti in stato di controllo. Su questo insieme di si basa il confronto con la distribuzione dei dati del processo che si intende analizzare. La procedura descritta qui di seguito permette di calcolare l'indice di diversità.

A] Selezione della matrice di riferimento e del limite di controllo

Vengono calcolati, in questa fase, gli autovettori di riferimento su cui poter basare il controllo del processo corrente

1) Si osserva una matrice $X, (n \times p)$, utilizzando dati provenienti da un processo operante in condizioni di normalità;
2) Le colonne di tale matrice vengono standardizzate in modo da ottenere osservazioni aventi media nulla e varianza unitaria;
3) Si sceglie un'adeguata dimensione della finestra *w;*
4) Facendo scorrere la finestra lungo l'asse temporale si generano (n-w+1) sottomatrici di dimensione (Figura 2.2);
5) La matrice ottenuta al passo 4) viene adottata come matrice di riferimento;
6) Si calcolano, per ogni matrice ottenuta al passo 4), i valori dell'indice D e si crea un vettore colonna di dimensioni ($(n-w+1)\times p$) contenente tali misure;
7) Viene individuata la mediana dei valori di D;
8) Si considera l'indice D_m in corrispondenza della mediana e si adotta come matrice di riferimento X_{ref}, quella utilizzata per il calcolo del valore D_m;
9) Vengono nuovamente calcolati i valori dell'indice D per ogni matrice ottenuta al passo 4), assumendo come matrice di riferimento quella ricavata al punto 8) ;
10) Il limite di controllo dei valori è calcolato in modo tale da lasciare fuori solo l'1% dell'intero campione;

11) Si disegna la carta di controllo per i valori D_i utilizzando il limite calcolato al punto precedente;

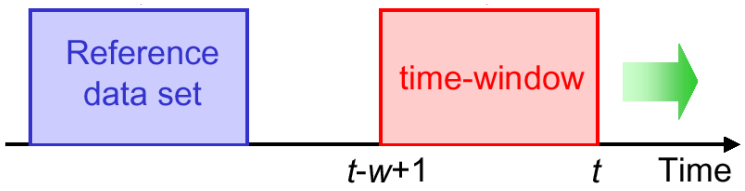


Figura 2.2: *generazione delle sottomatrici facendo scorrere la finestra lungo l'asse del tempo*

B] Controllo del processo corrente

Per individuare eventuali differenze tra la distribuzione dei dati in controllo e quella dell'insieme che si vuole controllare, si procede ora alla costruzione dell'indice D per il processo corrente:

1) Si acquisisce una matrice Y, avente dimensioni $(n \times p)$, contenente i dati del processo che si vuole analizzare;
2) Le p colonne di Y vengono standardizzate rispetto alla media e alla varianza della matrice di riferimento ottenuta nella fase precedente;
3) Viene fatta scorrere la finestra lungo l'asse del tempo generando $(n\text{-}w+1)$ sottomatrici di Y, aventi dimensione $(w \times p)$, (Figura 2.2);
4) Si calcola il valore dell'indice di diversità per ogni sottomatrice ottenuta al punto 4) adottando come matrice di riferimento la X_{ref} ottenuta al punto 8) della fase A];
5) La carta di controllo per gli $(n\text{-}w+1)$ valori dell'indice D, calcolati al punto precedente, viene tracciata utilizzando il limite di controllo ottenuto nella fase A].

Se alcuni valori dell'indice D superano il limite tracciato, il processo verrà ritenuto fuori controllo.

Le piccole variazioni che con i metodi di controllo multivariato tradizionali erano difficili da individuare, possono ora essere identificati grazie alla tecnica

DISSIM poiché i cambiamenti intercorsi nel processo influenzano la correlazione tra le variabili e le distribuzioni dei dati. Dal momento che viene adoperata una finestra temporale, l'indice D cambia in maniera più graduale rispetto alle statistiche T^2 e Q. La grandezza della finestra temporale influenza infatti l'effetto di lisciamento dell'indice, migliorando la sua capacità di scoprire eventuali malfunzionamenti nel processo. La scelta della costante *w*, che dipenderà dall'ampiezza dell'intervallo campionario studiato e dal numero di variabili considerate, diventa così critica: un valore troppo elevato, infatti, potrebbe ridurre la velocità nell'individuare cambiamenti nelle condizioni operative del processo.

Capitolo 3
Applicazione al Monitoraggio Ambientale

3.1 Introduzione

In questo capitolo si intende confrontare l'efficienza dei metodi statistici di controllo della qualità basati sull'Analisi delle Componenti Principali (statistiche T^2 e Q) con quella di due nuove metodologie: la *Moving Principal Component Analisys* (indice A) e la *Dissim* (indice D) (Kano et al.,2000-2002).

Vengono utilizzati dati provenienti dal monitoraggio ambientale dell'aria effettuato da tre diverse centraline poste in prossimità di Houston, Texas.

In un primo momento ci si è dedicati alla creazione di un conveniente insieme di riferimento. Si è quindi proceduto all'individuazione di particolari legami tra gli agenti inquinanti e le variabili meteorologiche, mettendo in luce sia le relazioni esistenti all'interno di ogni centralina sia quelle presenti tra siti di rilevazione differenti.

Dall'analisi delle correlazioni si è giunti a considerare un particolare modello che mette in relazione l'ozono con la radiazione solare misurata in ciascuna centralina . Su questi dati vengono costruite le carte di controllo descritte nei capitoli precedenti.

3.2 Inquinamento atmosferico

La questione dell'inquinamento atmosferico è diventata negli ultimi anni di estrema rilevanza sia a livello mondiale che a livello locale: basti pensare all'importanza di garantire la qualità ambientale all'interno delle città.

Le sostanze responsabili dell'inquinamento atmosferico sono numerose e diversificate in termini di caratteristiche chimico-fisiche e di effetti su salute e ambiente; normalmente si distinguono gli inquinanti primari, sostanze direttamente immesse nell'atmosfera a causa di attività antropiche o fenomeni naturali, dagli inquinanti secondari, formatisi per reazioni chimiche o fisiche dagli inquinanti primari.

Le principali fonti di emissione possono essere individuate negli impianti a combustione e nei processi produttivi industriali, nel traffico autoveicolare e in tutte le attività naturali che regolano l'ambiente. Conoscendo gli effetti dannosi dell'inquinamento atmosferico sulla salute umana è necessario raccogliere e organizzare un elevato numero di informazioni su ambiente e risorse naturali al fine di sviluppare adeguate politiche di tutela in materia ambientale. In particolare sarà necessario prendere in considerazione tre aspetti: le caratteristiche di ogni composto inquinante, il tipo di relazione con le diverse condizioni climatiche e i possibili effetti sull'ambiente e sulla vita dell'uomo.

Riportiamo, a tale scopo, una breve descrizione delle variabili inquinanti e meteorologiche prese in considerazione in questo studio.

3.2.1 Ossidi di Azoto

Con il termine "ossidi di azoto" viene generalmente indicata la somma pesata di due sostanze presenti spesso congiuntamente nell'atmosfera: il monossido di azoto (NO) e il biossido di azoto (NO_2). Entrambi i gas si originano dalla reazione tra azoto e ossigeno e dalla presenza di elevate temperature e radiazioni solari che favoriscono la combustione: è logico dunque aspettarsi che quanto più elevata è la temperatura atmosferica tanto più alta sarà la presenza di ossidi di azoto nell'aria. Per questo

particolare motivo le sorgenti maggiori di ossidi di azoto sono da ricercarsi nel traffico veicolare e nell'attività industriale. In presenza di sostanze ossidanti quali l'ozono, gli idrocarburi e i radicali liberi gli ossidi di azoto (in particolare il biossido la cui tossicità è molto più elevata di quella del monossido) possono trasformarsi e innescare delle reazioni chimiche tali da portare alla formazione di smog fotochimico e acido nitrico il principale responsabile, assieme all'acido solforico, del fenomeno delle piogge acide. E' importante, a questo proposito, sottolineare che gli ossidi di azoto tendono a restare nell'atmosfera più a lungo rispetto ad altri composti, pertanto i fenomeni meteorologici possono incidere pesantemente sulla distribuzione e sul trasporto di queste sostanze.

Per quanto riguarda i possibili effetti sulla salute umana, l'inquinamento dovuto agli ossidi di azoto può portare all'insorgere di gravi difficoltà respiratorie e malattie polmonari quali bronchiti e enfisemi e asma, aggravando talvolta anche patologie cardiache già presenti nell'individuo.

3.2.2 Ozono

L'ozono è un gas formato da tre atomi di ossigeno (O_3) che si combinano tra loro per azione delle radiazioni solari e dei fulmini che sono in grado di fornire l'elevata energia richiesta per la reazione. In natura si trova in concentrazioni rilevanti negli strati alti dell'atmosfera terrestre dove costituisce una fascia protettiva nei confronti della radiazione ultravioletta del sole. In questa zona dell'atmosfera, detta "stratosfera", l'ozono è dunque indispensabile alla vita sulla terra poiché assorbe le radiazioni dannose per la saluta umana. Negli ultimi anni, lo scudo di protezione formato dall'ozono nella stratosfera ha subito una parziale distruzione a seguito dell'azione di sostanze inquinanti quali ossidi di azoto e clorofluorocarburi generando il fenomeno del "buco dell'ozono". Negli strati bassi dell'atmosfera, nella fascia denominata "troposfera", l'ozono è presente tipicamente in basse concentrazioni ma la presenza di sostanze chimiche inquinanti, soprattutto in corrispondenza delle aree urbane, può favorire un aumento nelle sue concentrazioni. Al livello del suolo la molecola di ozono si forma quando gli inquinanti, principalmente ossidi di azoto e composti organici volatili, reagiscono favoriti dalla

presenza della luce solare e delle radiazioni UV; in sostanza, il biossido di azoto (NO_2) si dissocia in monossido di azoto (NO) e in ossigeno atomico (O) che, a sua volta, si combina con l'ossigeno molecolare (O_2) a formare la molecola di ozono (O_3). Le concentrazioni di ozono sono influenzate da diverse variabili meteorologiche come l'intensità della radiazione solare, la temperatura e la direzione e velocità del vento. Per tale ragione è facile registrare le più elevate concentrazioni di ozono nei periodi tardo-primaverili ed estivi, caratterizzati da alte temperature e poca ventilazione.

I motivi che rendono necessari il monitoraggio dell'ozono e la riduzione delle sue concentrazioni in atmosfera sono numerosi. La presenza di elevati livelli di ozono danneggia in primo luogo la salute umana, degli animali e delle piante e provoca inoltre un deterioramento dei materiali e degli edifici. Vari studi hanno evidenziato che l'esposizione all'inquinamento dovuto all'ozono induce nell'uomo irritazioni agli occhi, mal di testa, difficoltà e malattie respiratorie, crisi asmatiche.

3.2.3 Polveri atmosferiche

Le polveri atmosferiche sono costituite da particelle solide e liquide che rimangono sospese nell'aria e che, a seconda del processo di formazione si differenziano per dimensioni, composizione e provenienza. Si possono distinguere diverse classi di polveri a seconda della grandezza del diametro (generalmente variabile tra 0.005 e 100 μm); in particolare si dicono "grossolane" le particelle con diametro che varia da 2.5 μm a 30 μm e "fini" quelle di dimensione inferiore. Le prime nascono principalmente da combustioni incontrollate e dalla disgregazione ed erosione dei suoli; le seconde derivano dalle emissioni del traffico veicolare, dall'attività industriale e dalla combustione di residui agricoli e sono ritenute causa di difficoltà respiratorie e dell'aggravarsi di malattie cardiovascolari negli individui. Tra le polveri più pericolose troviamole PM_{10} o polveri inalabili, perché sono in grado di penetrare nel tratto superiore dell'apparato respiratorio, e le $PM_{2.5}$ o polveri respirabili, in grado di raggiungere il tratto inferiore dell'apparato respiratorio. Queste ultime, in particolare, risultano, potenzialmente pericolose per la presenza di un certo numero di sostanze dannose quali ad esempio i solfati, il carbonio e talvolta

metalli tossici che possono provocare nell'uomo gravi disturbi respiratori e malattie di natura cancerogena.

3.3 Origine dei dati studiati

Nel corso degli ultimi anni in Texas si è rivolta particolare attenzione alla salvaguardia ambientale e al controllo dell'inquinamento. A questo proposito, nel 1993, è nata un'importante agenzia per la protezione ambientale, *la Texas Natural Resource Conservation Commission* (TNRCC) con il compito di sorvegliare le concentrazioni dei principali inquinanti nell'ambiente in modo da poter adottare politiche più adeguate per la difesa delle risorse naturali.

La TNRCC rileva quotidianamente informazioni sulla composizione dell'aria nelle varie zone del Texas ed in particolare in quattro aree urbane, più densamente popolate, che si rivelano critiche data la frequente violazione degli standard fissati per le concentrazioni di inquinanti. Le aree in questione sono:

- Beaumont / Port Arthur
- Dallas / Fort Worth
- El Paso
- Houston / Galveston

In questo contesto si restringe l'analisi all'area di Houston / Galveston (Figura 4.1). All'interno di questa zona si sono selezionate tre centraline scelte in base alla loro disposizione geografica e alla disponibilità di informazioni. Due centraline sono poste nella contea di Harris: la C35 è situata all'interno di un centro urbano in prossimità della città di Houston e la C26 nella sua periferia. La terza centralina, C34, si trova nella contea di Galveston in prossimità del Golfo del Messico (Figura 4.2 e 4.3). Nella Tabella A sono riportate le collocazioni di ogni centralina in termini di longitudine e latitudine.

Le stazioni di rilevazione considerate raccolgono continuamente dati sulle concentrazioni degli inquinanti e sulle caratteristiche delle variabili meteorologiche: i parametri rilevati nelle diverse centraline sono riassunti nella Tabella B. Per la

variabile relativa alle polveri atmosferiche è stato possibile raccogliere informazioni solo nelle centraline C34 e C35. In queste stazioni, ogni cinque minuti, viene calcolato il valor medio dei campioni relativi agli agenti inquinanti e alle variabili meteorologiche: le medie orarie sono quindi il frutto dei 12 valori medi ottenuti in un'ora. Per questo studio, in particolare, sono state prese in considerazione le medie orarie delle diverse variabili nelle tre centraline, prendendo come riferimento temporale gli anni 2001 e 2002.

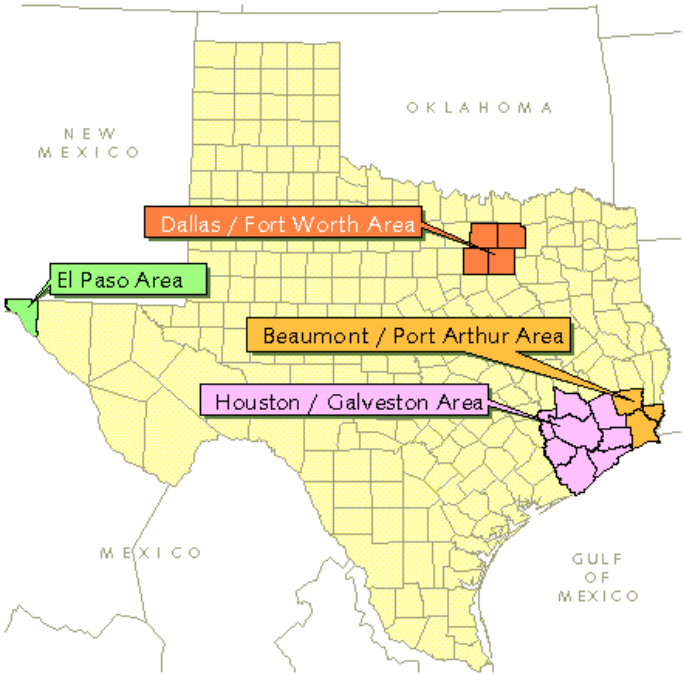


Figura 4.1: *Aree del Texas dove le concentrazioni di inquinanti superano i livelli fissati.*

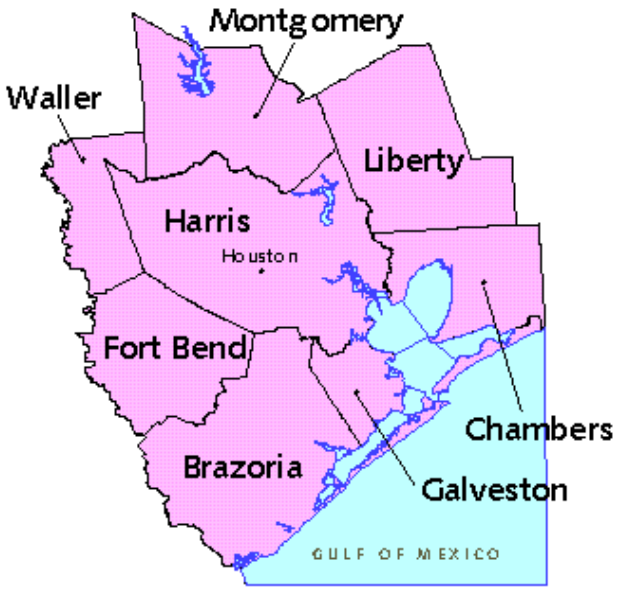


Figura 4.2: *Contee della regione di Houston/Galveston*

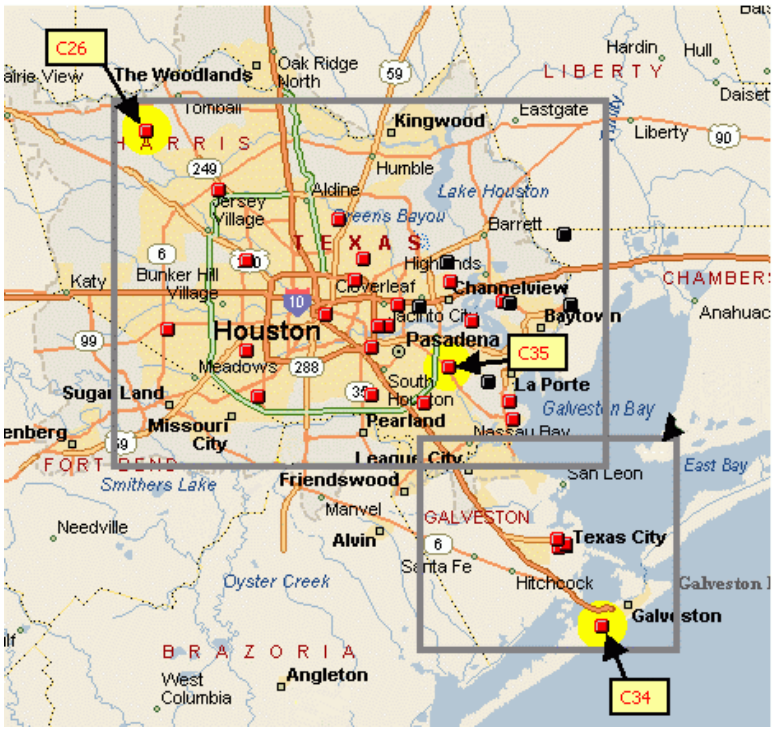


Figura 4.3: *Disposizione delle centraline*

Tabella A

Centralina	Contea	Longitudine	Latitudine	Altezza (s.l.m)
26	Harris	95° 40' 26" W	30° 02' 22" N	55 m
34	Galveston	94° 51' 23" W	29° 15' 47" N	0 m
35	Harris	95° 07' 43" W	29° 40' 11" N	6 m

Tabella B

SOSTANZE INQUINANTI	UNITA' DI MISURA
Ossidi di Azoto	Ppb
Ozono	Ppb
Polveri respirabili	µg/m3

VARIABILI METEOROLOGICHE	UNITA' DI MISURA
Velocità del vento	mph
Direzione del vento	deg
Deviazione standard della velocità del vento	deg
Massima raffica di vento	mph
Temperatura esterna	deg F
Radiazioni solari	Ly/minI

3.4 Elaborazione dei dati

Le osservazioni orarie raccolte per ogni centralina e per ogni parametro considerato sono state elaborate ed aggregate allo scopo di ottenere informazioni di tipo giornaliero. Per ogni giorno del biennio preso in esame sono state calcolate tutte le possibili medie mobili su otto ore, considerando nulle le medie ottenute su un numero di dati inferiore a cinque (quindi con più di tre osservazioni mancanti sulle otto di riferimento). Tale soluzione si è resa necessaria per evitare di avere dati poco rappresentativi a causa di una perdita troppo alta di informazioni, inevitabile se si considera l'effetto di lisciamento operato dalle medie mobili. Una volta ottenute queste medie si è scelta, per l'intera giornata, quella con il valore massimo, In tal modo si è giunti ad ottenere, per ogni parametro preso in esame, due serie storiche composte da 365 dati, una riferita all'anno 2001 e l'altra al 2002. Allo scopo di conoscere la stazionarietà in media ed in varianza delle serie, su ciascuna di esse è stata effettuata un'analisi volta ad individuare la struttura di autocorrelazione parziale e totale, delle osservazioni e dei loro quadrati.

3.4.1 La stima dei dati mancanti

Procedendo all'elaborazione dei dati e allo studio delle serie storiche ci si è trovati ad affrontare un problema di mancanza di rilevazioni in alcune centraline. Le cause più frequenti che hanno impedito alle stazioni di ottenere i dati sono da ricercarsi innanzitutto nel cattivo funzionamento degli strumenti di rilevazione e nell'interruzione volontaria del loro uso per permetterne la taratura; in altri casi invece, i dati non risultavano disponibili per volontà degli stessi membri della TNRCC che non hanno ritenuto valide le osservazioni misurate dalle centraline. Nella tabella C vengono riportate le variabili studiate nelle tre centraline con il relativo numero di dati mancanti.

Per stimare i dati mancanti si è ritenuto opportuno utilizzare il metodo della regressione lineare. Sia y_m una variabile per la quale non si disponga di una o più osservazioni nell'arco temporale *m* (pari a un mese). Sia y_{m1} la stessa variabile osservata in un intervallo m1 avente andamento temporale il più possibile omogeneo a quello evidenziato nel periodo *m*. Scelta y_{m1} come variabile dipendente e come

Tabella C

	VARIABILE	CENTRALINA		
		26	34	35
2001	Ossidi di Azoto	15	1	17
	Ozono	3	-	9
	Polveri	-	-	5
2002	Ossidi di Azoto	8	13	5
	Ozono	18	14	-
	Polveri	-	15	3
	Velocità del vento	-	-	-
	Direzione del vento	-	31	-
	Dev.Std.della velocità	-	31	-
	Massima raffica di vento	-	4	-
	Radiazioni solari	-	5	-
	Temperatura esterna	-	5	-

variabili esplicative le osservazioni misurate, per lo stesso periodo di tempo, sulle altre variabili inquinanti o meteorologiche ($x_{1,m1}, x_{2,m1}, ...x_{p,m1}$), si procede all'adattamento di un modello di regressione. I coefficienti ottenuti dalla regressione, vengono in seguito utilizzati per la stima di y_m come illustrato dalla (4.1).

$$y_m = \beta_1 x_{1,m} + \beta_2 x_{2,m} + ... + \beta_p x_{p,m} \qquad (4.1)$$

3.5 Analisi preliminare dei dati

Prima di procedere alla scelta del modello su cui costruire le carte di controllo, è necessario studiare le caratteristiche generali dei dati a disposizione. L'analisi descrittiva parte dall'osservazione dei grafici delle serie storiche di ogni variabile per tutte le centraline allo scopo di coglierne l'andamento generale. Si passa poi alla costruzione dei diagrammi a scatola e degli istogrammi ed al calcolo delle principali statistiche di base per cercare di conoscere meglio le distribuzioni delle variabili in esame.

3.5.1 Sostanze inquinanti nel 2001

L'analisi descrittiva è stata compiuta osservando il comportamento di ogni variabile in prossimità delle tre centraline al fine di operare un confronto tra i dati raccolti dalle diverse stazioni e verificare l'esistenza di particolari legami.

Nella tabella D sono riportate le principali statistiche di base relative alle variabili considerate

Tabella D

	Centralina	Min	Max	Media	Mediana	Varianza	Dev.Std.
Ossidi di azoto	26	2,88	72,13	17,15	13,54	131,49	11,47
	34	0,01	61,12	10,63	6,83	127,79	11,30
	35	1,63	217,4	29,25	21,18	722,30	26,88
Ozono	26	4	137	43,02	39	387,87	19,69
	34	0	88	36	34	277,98	16,67
	35	2	114	40,26	37	358,81	18,94
Polveri	34	2,59	41,06	11,54	10,55	33,18	5,76
	35	2,76	35,42	12,15	11,08	34,7	5,89
Velocità del vento	26	2,64	18,16	8,015	7,28	9,513	3,08
	34	3,63	26,33	13,57	12,75	15,23	3,90
	35	3,21	16,64	8,678	8,32	7,086	2,66
Massima raffica	26	5,67	33,85	15,75	14,78	25,1	5,01
	34	9,93	44,49	20,64	19,19	34,36	5,86
	35	8,1	30,44	16,98	16,23	10,20	3,19
Radiazioni solari	26	0,043	1,195	0,686	0,717	0,087	0,29
	34	0,059	1,21	0,736	0,785	0,093	0,30
	35	0,039	1,156	0,631	0,643	0,077	0,28
Temperatura esterna	26	36,81	98,64	77,63	79,60	169,90	13,03
	34	36,56	87,54	73,52	74,98	120,20	10,96
	35	36,23	93,01	74,66	76,89	144,29	12,01

3.5.1.1 Ossidi di Azoto

Osservando le serie relative agli ossidi di azoto è possibile notare la presenza di valori abbastanza elevati e soggetti ad una discreta variabilità nei mesi più freddi dell'anno. Durante il periodo estivo, invece, si ha un netto abbassamento delle concentrazioni rilevate. La centralina 35 situata all'interno del centro urbano presenta valori mediamente più elevati di quelli registrati dalle altre centraline soprattutto rispetto alla centralina 34 che è situata in prossimità del golfo del Messico. Tale risultato è abbastanza comprensibile se si considera il fatto che la fonte principale di produzione degli ossidi di azoto è il traffico veicolare e l'attività industriale. Per quanto riguarda le distribuzioni delle tre serie, osservando gli istogrammi e i diagrammi a scatola è facile notare una forte asimmetria positiva causata dalla presenza di code pesanti a destra e l'esistenza di valori anomali segnalati per mezzo di un cerchietto dai box-plot. Quanto illustrato dai grafici delle serie storiche trova conferma nei box-plot mensili, utili per meglio evidenziare l'evoluzione delle serie nel corso dell'anno. E' facile infatti vedere come si abbia un abbassamento delle concentrazioni degli ossidi di azoto durante i periodi tardo-primaverile ed estivo.

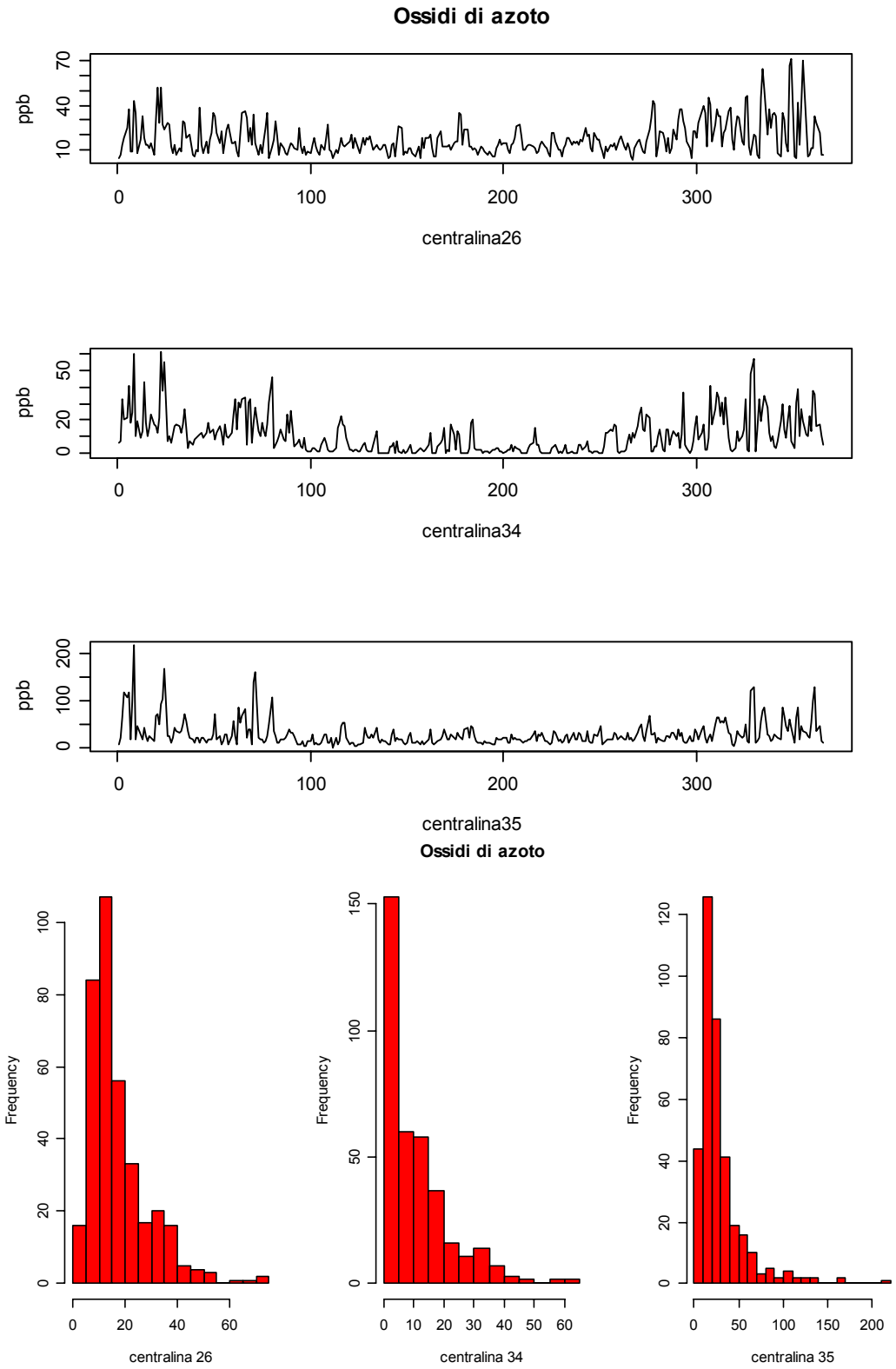


Figura 4.4 *Grafici ed istogrammi delle tre serie di ossidi di azoto per il 2001*

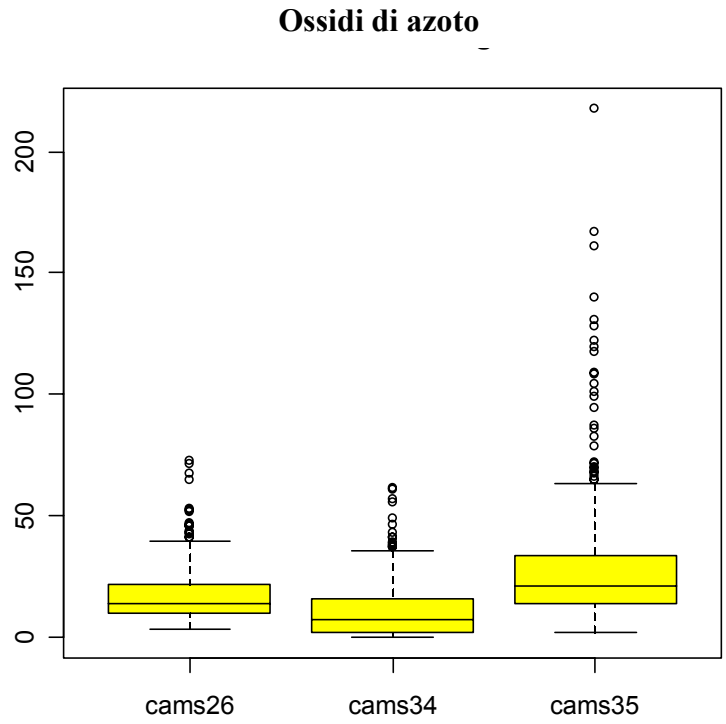


Figura 4.5: *Diagrammi a scatola delle tre serie di ossidi di azoto per il 2001*

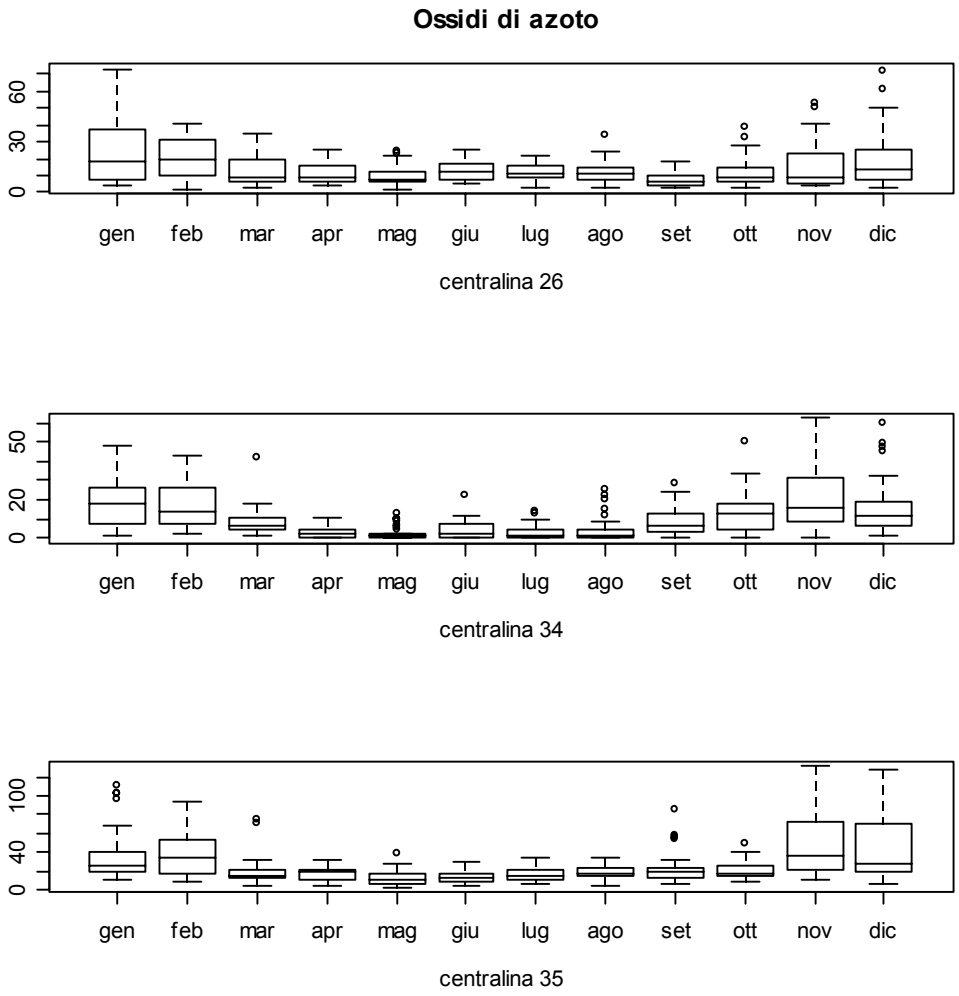


Figura 4.6: *Diagrammi a scatola delle mensili tre serie di ossidi di azoto per il 2001*

3.5.1.2 Ozono

Per quanto riguarda l'ozono e l'ammontare della sua concentrazione in prossimità delle tre stazioni di rilevamento è necessario fare una importante considerazione: per sua stessa origine e natura, tale sostanza è strettamente correlata all'andamento di variabili meteorologiche quali temperatura e la radiazione solare risulta pertanto naturale aspettarsi un aumento dei valori rilevati in corrispondenza dei periodi primaverile ed estivo. I diagrammi a scatola mettono il luce come le distribuzioni di dati nelle tre stazioni siano abbastanza simili anche se i valori risultano leggermente più elevati nella centralina 26. Dai grafici degli istogrammi è possibile notare anche in questo caso la presenza di code pesanti a destra. L'asimmetria positiva rilevata per l'ozono risulta, tuttavia, meno marcata rispetto a quanto visto per gli ossidi di azoto; tale considerazione è valida in maggior misura per la centralina 34 posta in prossimità del Golfo del Messico.

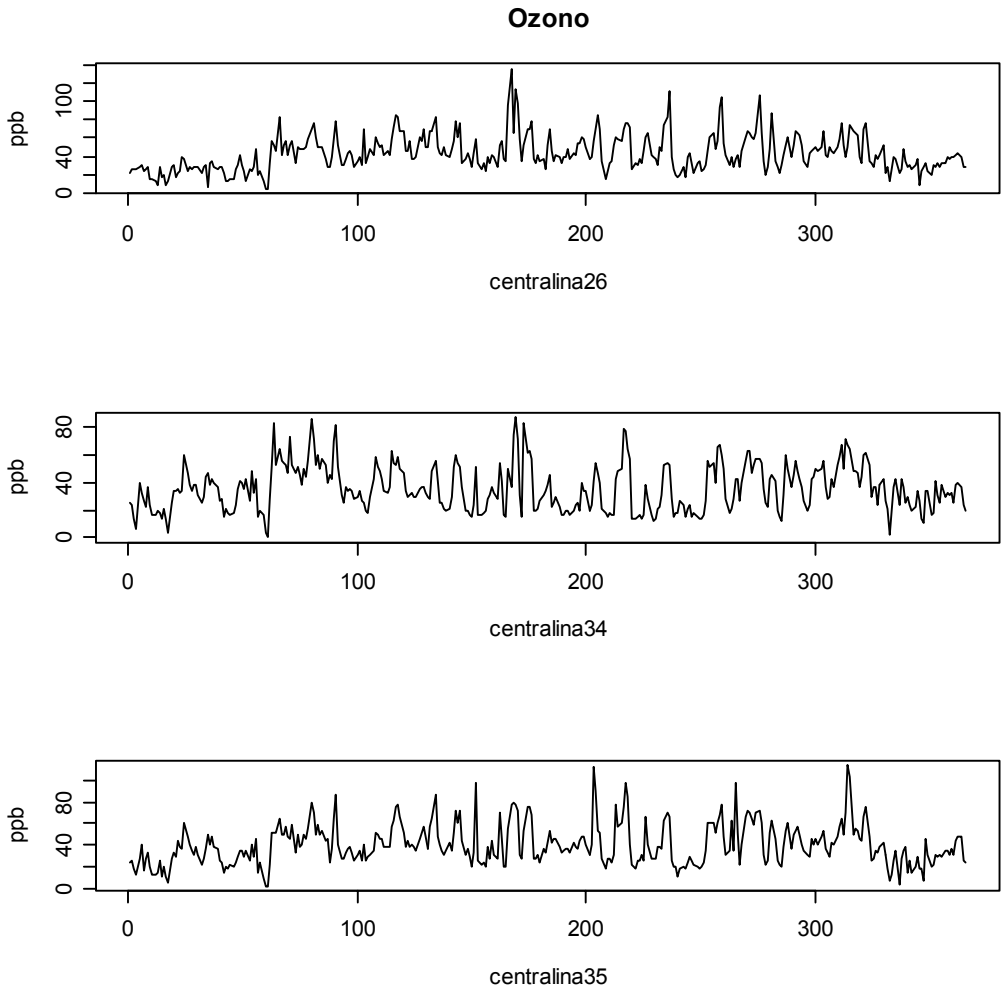


Figura 4.7: *Grafici delle tre serie di ozono per il 2001*

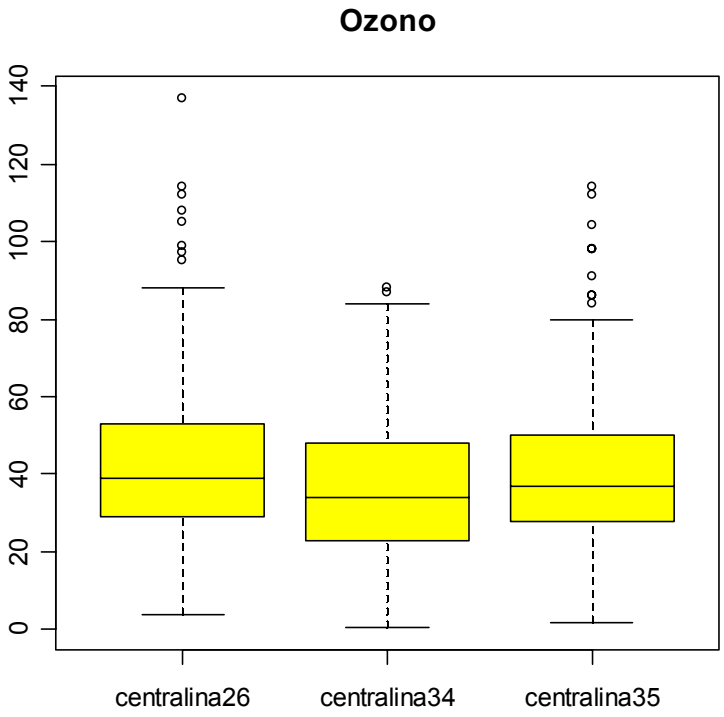


Figura 4.8: *Diagrammi a scatola delle tre serie di ozono per il 2001*

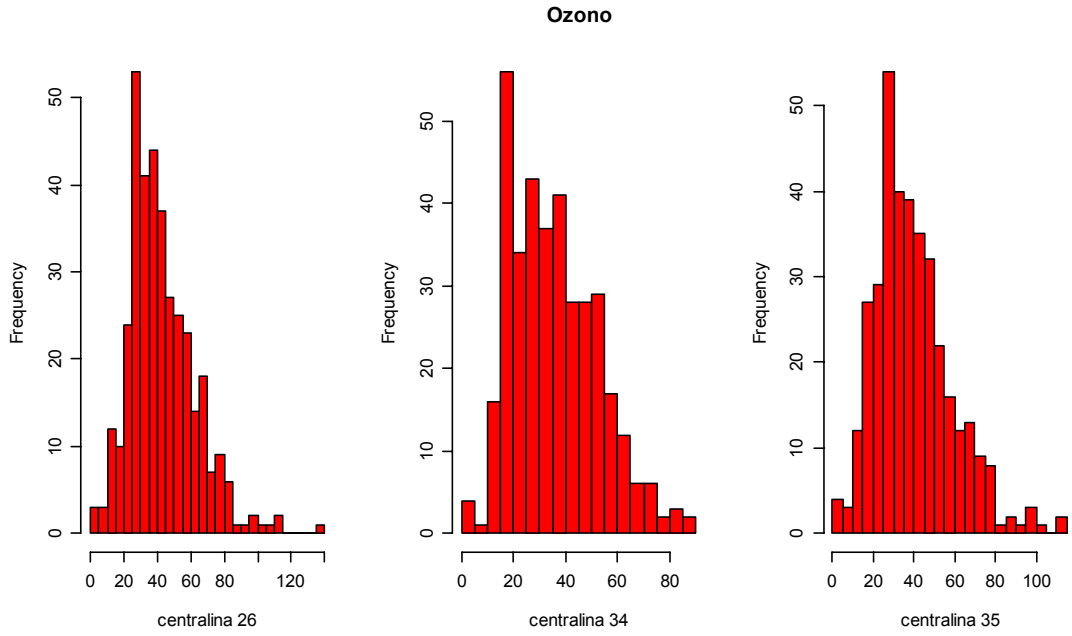


Figura 4.9: *Istogrammi delle tre serie di ozono per il 2001*

3.5.1.3 Polveri atmosferiche

Anche se non è stato possibile registrare i valori delle polveri atmosferiche in tutte le centraline, si è scelto di studiare l'andamento delle concentrazioni di tale inquinante, considerando il fatto che risultano particolarmente pericolose per la salute umana. In base ad un'analisi delle autocorrelazioni i valori registrati risultano abbastanza stazionari sia in media che in varianza; osservando i grafici delle serie, le osservazioni presentano un andamento non troppo irregolare e assumono valori sempre inferiori alla soglia d'allarme standard di 65 $\mu g/m^3$. Le distribuzioni presentano code pesanti sulla destra, come illustrano gli istogrammi e i box-plot che segnalano la presenza di valori anomali.

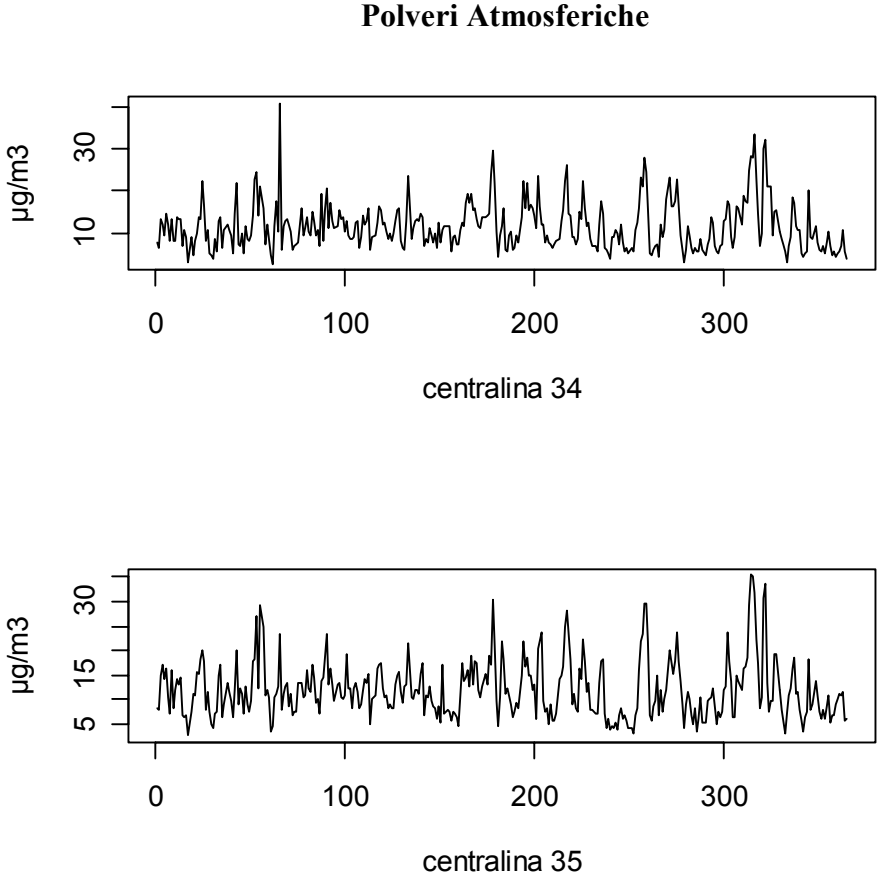


Figura 4.10: *Serie delle polveri atmosferiche per il 2001*

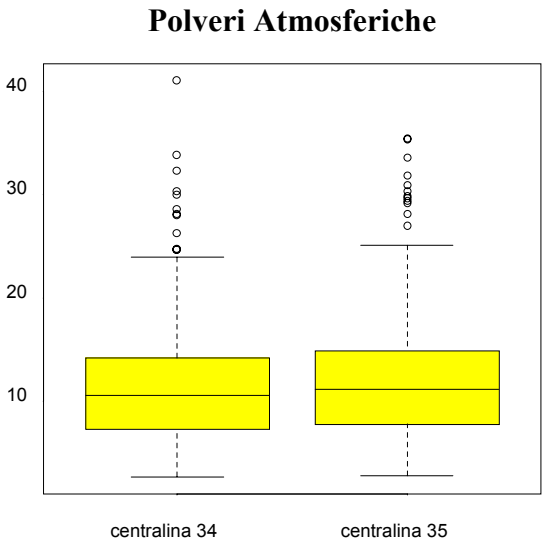


Figura 4.11: D*iagrammi a scatola delle serie delle polveri*

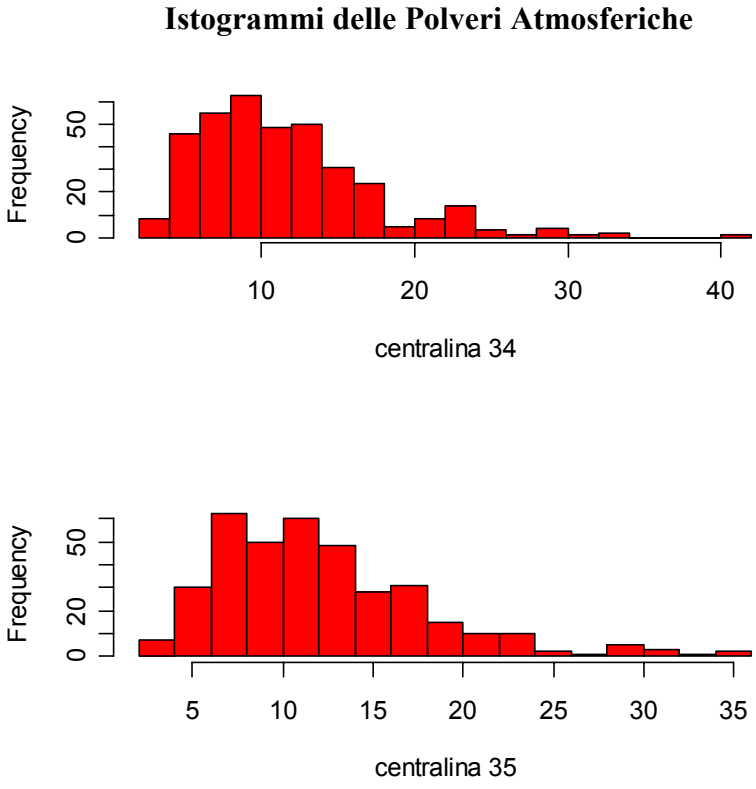


Figura 4.12: *Istogrammi delle serie delle polveri*

3.5.2 Variabili meteorologiche nel 2001

La concentrazione delle sostanze inquinanti nell'atmosfera terrestre dipende fortemente dalle variabili meteorologiche che regolano le condizioni climatiche sia all'interno di un giorno solare che nell'arco di un'intera stagione. Studiando le serie di tutte le variabili rilevate dalle tre centraline della TNRCC è stato possibile accorgersi di quanto la presenza di ossidi di azoto, ozono e polveri atmosferiche fosse strettamente correlata alla direzione e alla velocità del vento e di come le concentrazioni fossero altamente influenzate dall'aumento o dalla diminuzione dei valori della temperatura e della radiazione solare.

L'analisi descrittiva della "velocità del vento" e della "massima raffica di vento", misurate in miglia orarie, mette in luce una forte irregolarità dei valori registrati ed un abbassamento nell'andamento delle serie esaminate in prossimità del periodo estivo. Questo risultato è più che comprensibile se si considera la natura e la forte instabilità del vento ed il fatto che tende ad essere nell'anno meno frequente nei periodi più caldi. Si può inoltre notare, osservando i diagrammi a scatola, che le serie rilevate nella centralina 34 assumono valori mediamente più elevati di quelli ottenuti nelle altre centraline. Ciò può essere giustificato dal fatto che tale centralina è situata in un luogo aperto che si affaccia sul golfo del Messico.

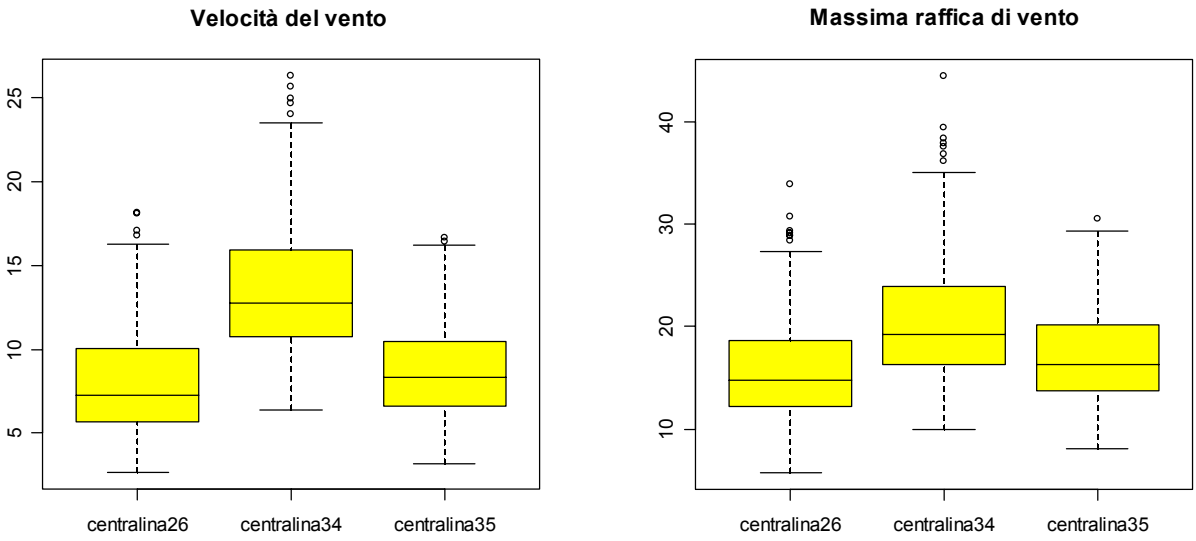


Figura 4.13: *Box-plot della velocità e della massima raffica di vento*

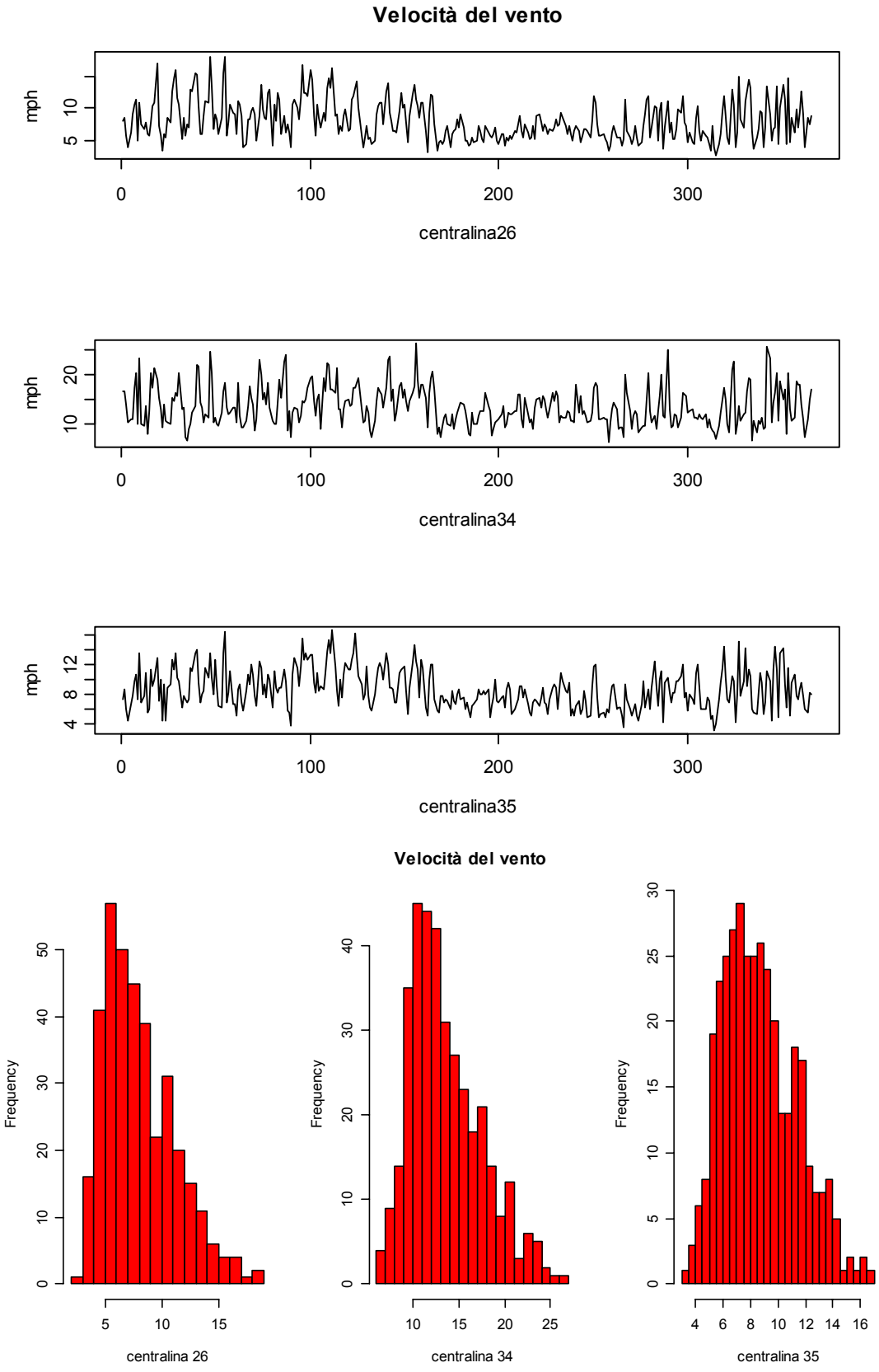


Figura 4.14: *Serie ed istogrammi della velocità del vento per il 2001*

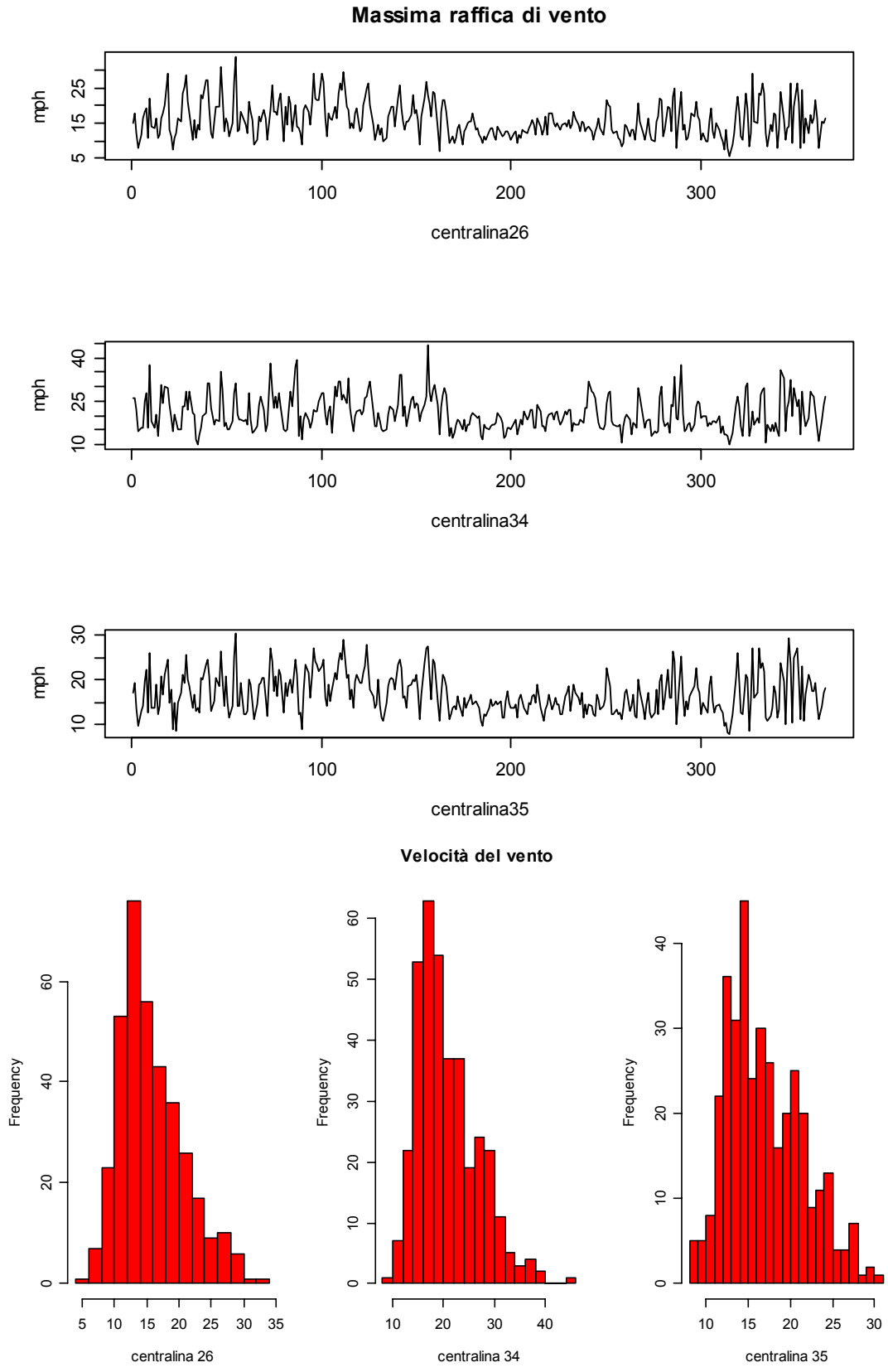


Figura 4.15: *Serie ed istogrammi della massima raffica di vento per il 2001*

Le serie dei dati riguardanti la radiazione solare e la temperatura esterna presentano caratteristiche simili tra le varie centraline; i grafici illustrano chiaramente la non stazionarietà in media dei valori registrati. Si può, in particolare, notare un trend crescente che interessa i mesi da marzo a maggio 2001 ed uno decrescente nel periodo che va da agosto a novembre. Queste tendenze sono più che giustificate considerando il passaggio dalla stagione primaverile a quella estiva, con un conseguente miglioramento delle condizioni atmosferiche, ed il sopraggiungere dell'autunno e dei mesi più freddi.

I diagrammi a scatola evidenziano che la centralina 34 ha registrato valori della radiazione solare mediamente più alti rispetto alle altre; per quanto riguarda invece la variabile temperatura, sono le osservazioni della serie relativa alla centralina 26 ad avere media e quantili più elevati in confronto a quelli delle altre.

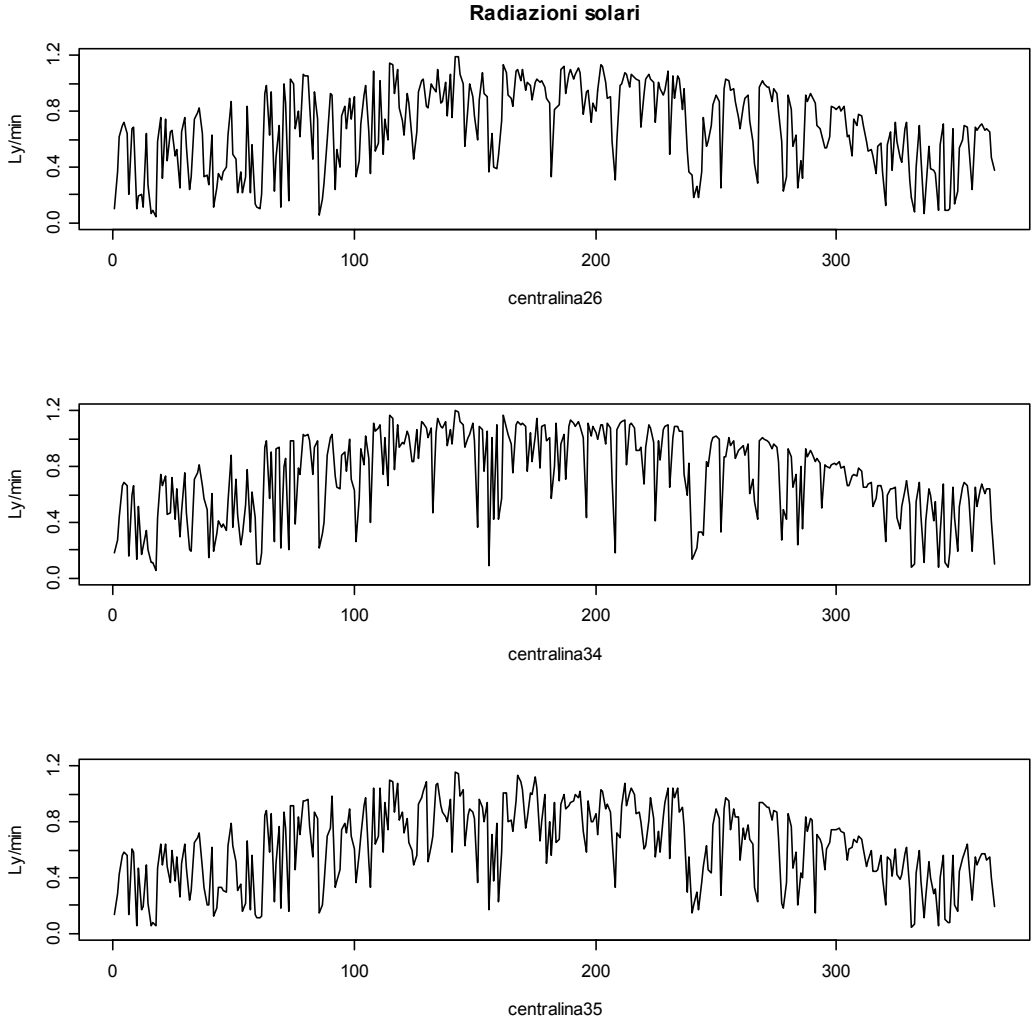


Figura 4.16: *Serie della radiazione solare*

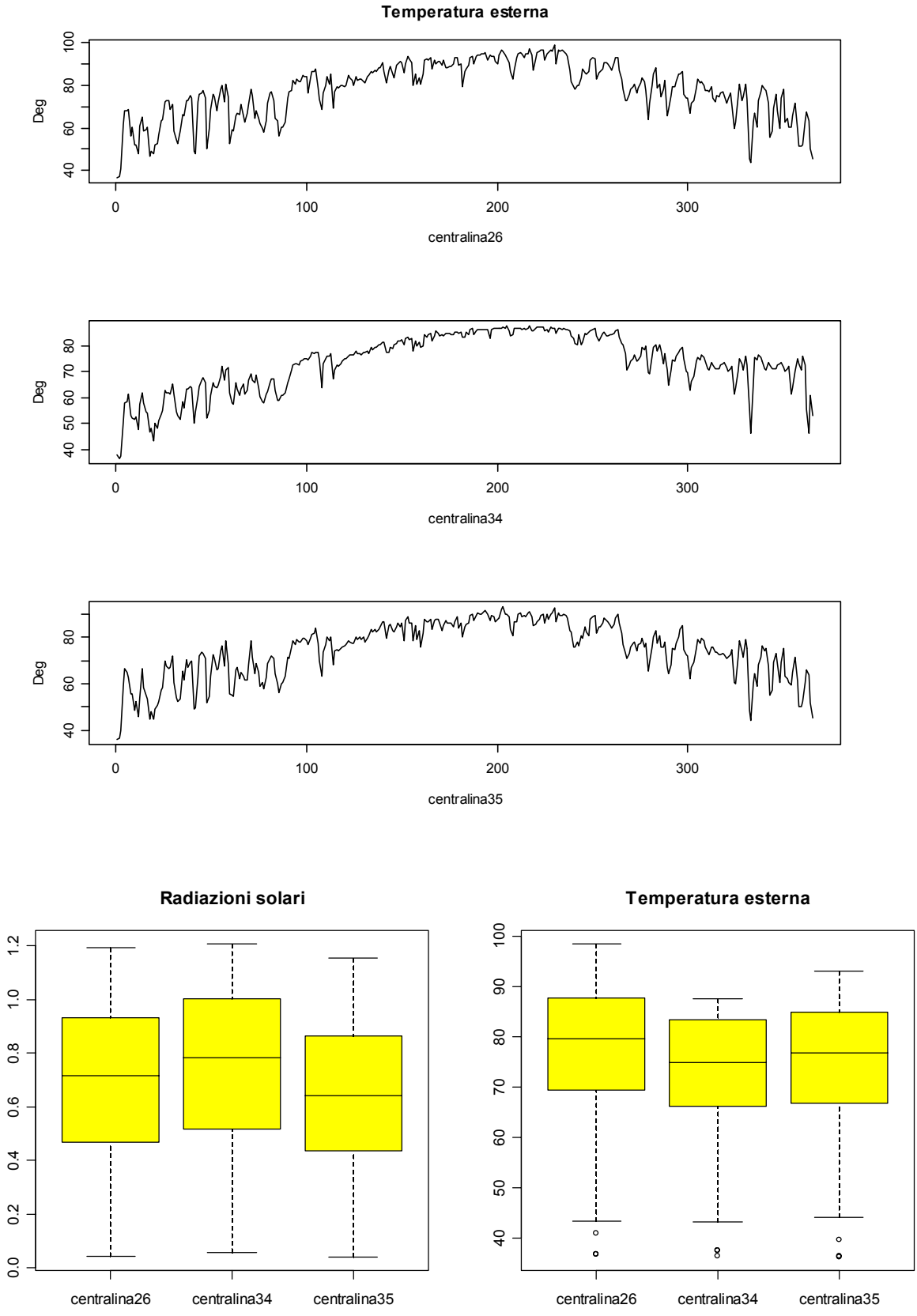


Figura 4.17: *Diagrammi a scatola della radiazione solare e della temperatura esterna*

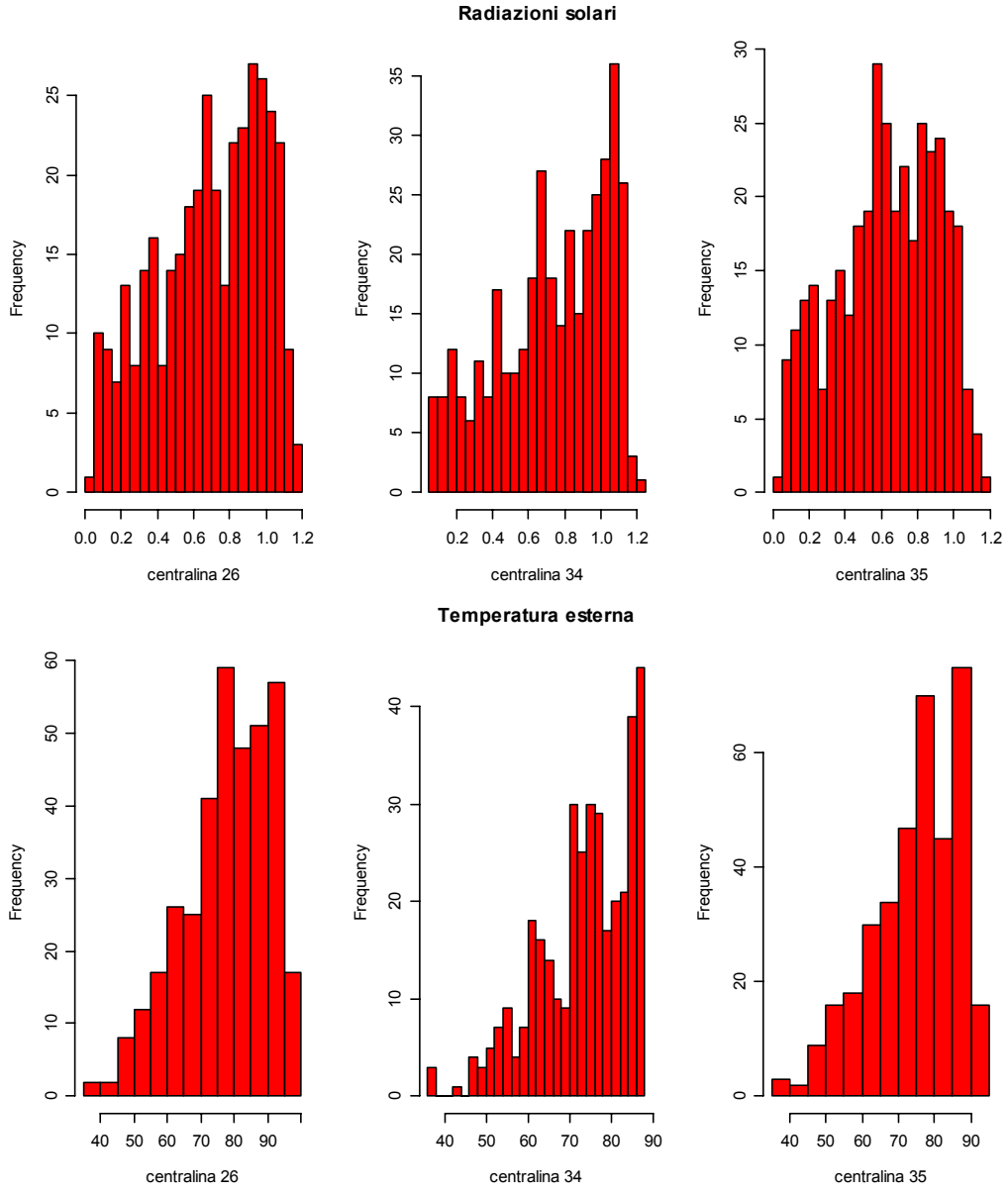


Figura 4.18: *Istogrammi della radiazione solare e della temperatura esterna*

Dall'osservazione degli istogrammi è possibile riscontrare la presenza di asimmetria in tutte le distribuzioni della radiazione solare e della temperatura: la centralina 34 in particolare, evidenzia questa tendenza in maniera più marcata rispetto alle altre stazioni di rilevazione.

3.5.3 Correlazioni

Nella Tabella E sono state riportate le correlazioni tra tutte le variabili rilevate nelle tre centraline allo scopo di poter verificare l'esistenza o l'assenza di una qualsiasi relazione tra i parametri misurati.

Come era facile aspettarsi, è evidente la presenza di un legame importante tra le concentrazioni di inquinanti: i valori relativi agli ossidi di azoto della centralina 34 e quelli della centralina 35 risultano legati da una forte correlazione positiva ed altrettanto può dirsi per le polveri atmosferiche; gli ossidi di azoto, in particolare, sembrano risentire dell'influenza esercitata dalla "temperatura esterna", in altre parole valori elevati di questa variabile tendono a far diminuire la concentrazione dell'inquinante nell'atmosfera favorendone la separazione in monossido di azoto e ossigeno. L'ozono misurato in ogni centralina risulta, inoltre, strettamente correlato con quello delle altre centraline e fortemente dipendente dalla radiazione solare. Questa relazione trova naturale giustificazione nel processo stesso di formazione dell'ozono stratosferico: l'aumento della radiazione solare e della temperatura favorisce le reazioni fotochimiche che portano gli inquinanti primari a dissociarsi, contribuendo alla formazione di ozono dannoso per la salute umana.

Ad influire positivamente su ossidi di azoto e, soprattutto, sull'ozono sono le polveri inalabili prodotte in gran parte dal traffico veicolare e dai processi di combustione industriale: questi miscugli di particelle solide e liquide sono in gran parte composte anche da molecole del tipo NO_x che assumono una funzione rilevante nella formazione dei principali inquinanti.Un ruolo importante nell'entità delle sostanze rilevate giocano la velocità del vento, la sua entità e direzione. Osservando la tabella delle correlazioni si vede come queste variabili, positivamente correlate per loro stessa natura, contribuendo alla dispersione degli ossidi di azoto, dell'ozono e delle polveri, producono plausibilmente un abbassamento delle loro concentrazioni nell'atmosfera.

Osservando le correlazioni tra le centraline, si nota che il legame tra la C26 e la C34 è più debole di quello registrato tra queste e la centralina 35. Questa relazione è valida per ognuna delle variabili considerate, tranne la velocità del vento, e la ragione di questo comportamento potrebbe essere identificata nella distanza che separa le centraline. La C34 e La c26 sono infatti separate da 117,15 Km di distanza e si trovano a circa un grado di latitudine e di longitudine l'una dall'altra. La stazione di rilevazione 35 invece occupa posizione intermedia in termini di distanza e di latitudine e longitudine (52,39 Km dalla C34 e 66,85 Km dalla C26).

	NOx 26	NOx 34	NOx 35	O3 26	O3 34	O3 35	PM 34	PM 35	WS 26	WS 34	WS 35	RWD 26	RWD 34	RWD 35	MWG 26	MWG 34	MWG 35	SWD 26	SWD 34	SWD 35	OT 26	OT 34	OT 35	SR 26	SR 34	SR 35
NOx 26	1,00																									
NOx 34	0,320	1,000																								
NOx 35	0,378	0,743	1,000																							
O3 26	-0,031	-0,024	0,033	1,000																						
O3 34	0,152	0,273	0,246	0,691	1,000																					
O3 35	0,003	0,088	0,115	0,791	0,815	1,000																				
PM 34	0,127	0,165	0,112	0,415	0,424	0,441	1,000																			
PM 35	0,146	0,178	0,182	0,459	0,480	0,561	0,861	1,000																		
WS 26	-0,268	-0,180	-0,239	-0,331	-0,242	-0,352	-0,255	-0,293	1,000																	
WS 34	-0,324	-0,203	-0,296	-0,303	-0,299	-0,367	-0,314	-0,395	0,658	1,000																
WS 35	-0,171	-0,284	-0,277	-0,156	-0,173	-0,232	-0,178	-0,226	0,816	0,600	1,000															
RWD 26	-0,216	0,288	0,233	-0,077	0,019	0,005	-0,065	-0,022	0,093	0,088	-0,180	1,000														
RWD 34	-0,110	0,107	0,219	-0,057	-0,077	-0,028	-0,136	-0,089	0,106	0,029	-0,075	0,593	1,000													
RWD 35	-0,209	0,208	0,228	-0,050	-0,006	0,024	-0,064	0,005	0,090	0,033	-0,124	0,791	0,736	1,000												
MWG 26	-0,303	-0,227	-0,286	-0,295	-0,236	-0,330	-0,264	-0,313	0,973	0,683	0,812	0,056	0,073	0,039	1,000											
MWG 34	-0,288	-0,175	-0,289	-0,321	-0,265	-0,367	-0,340	-0,423	0,624	0,959	0,557	0,055	0,002	-0,014	0,671	1,000										
MWG 35	-0,259	-0,266	-0,309	-0,238	-0,212	-0,298	-0,276	-0,338	0,846	0,738	0,922	-0,055	-0,025	-0,071	0,880	0,727	1,000									
SWD 26	0,126	0,135	0,335	0,350	0,228	0,349	0,190	0,277	-0,649	-0,484	-0,576	0,135	0,198	0,201	-0,599	-0,454	-0,579	1,000								
SWD 34	-0,122	0,145	0,083	0,065	0,045	0,056	-0,059	-0,038	-0,039	-0,019	-0,147	0,286	0,270	0,298	-0,008	0,039	-0,038	0,267	1,000							
SWD 35	-0,006	0,171	0,232	0,209	0,152	0,283	0,139	0,193	-0,586	-0,324	-0,631	0,242	0,220	0,277	-0,535	-0,268	-0,538	0,694	0,365	1,000						
OT 26	-0,228	-0,552	-0,338	0,381	0,013	0,280	0,185	0,162	-0,163	-0,213	-0,013	-0,221	0,029	-0,079	-0,113	-0,267	-0,109	0,197	-0,062	0,125	1,000					
OT 34	-0,236	-0,525	-0,346	0,341	-0,034	0,215	0,098	0,073	-0,186	-0,135	-0,082	-0,148	0,075	-0,035	-0,116	-0,167	-0,111	0,217	-0,007	0,206	0,913	1,000				
OT 35	-0,263	-0,546	-0,341	0,379	0,026	0,293	0,173	0,151	-0,160	-0,195	-0,032	-0,167	0,061	-0,039	-0,103	-0,243	-0,099	0,208	-0,025	0,159	0,988	0,930	1,000			
SR 26	-0,169	-0,137	0,022	0,595	0,317	0,513	0,176	0,201	-0,258	-0,245	-0,192	0,020	0,093	0,047	-0,248	-0,337	-0,269	0,314	-0,024	0,179	0,561	0,471	0,547	1,000		
SR 34	-0,156	-0,234	-0,055	0,578	0,317	0,502	0,171	0,196	-0,192	-0,188	-0,060	-0,076	0,006	-0,019	-0,175	-0,289	-0,161	0,266	-0,085	0,105	0,603	0,520	0,603	0,866	1,000	
SR 35	-0,185	-0,150	-0,017	0,617	0,375	0,560	0,191	0,208	-0,247	-0,211	-0,107	-0,027	0,043	0,023	-0,234	-0,299	-0,195	0,281	-0,022	0,153	0,547	0,460	0,553	0,922	0,906	1,000

Tabella F: Correlazioni tra le serie di ogni variabile per tutte le centraline per l'anno 2001

3.5.4 Sostanze inquinati nel 2002

I risultati ottenuti nell'analisi descrittiva sui dati del 2002 non sono molto lontani da quanto visto per i valori dell'anno precedente; molte serie relative alla stessa variabile presentano infatti andamento e caratteristiche simili.

Nella Tabella F sono riportate le principali statistiche di base relative alle variabili considerate nel 2002.

Tabella F

VARIABILE	Centralina	Min	Max	Media	Mediana	Varianza	Dev.Std.
Ossidi di azoto	26	1,94	73,03	14,06	10,59	126,86	11,26
	34	0,01	62,83	10,01	6,13	133,87	11,57
	35	3,53	131,3	26,03	18,59	534,06	23,11
Ozono	26	8	123	41,59	39	285,45	16,90
	34	9	98	39,44	36	263,86	16,24
	35	5	115	39,72	36	323,71	17,99
Polveri	34	1,55	52,69	11,42	9,86	46,43	6,81
	35	2,61	54,68	12,93	10,88	59,52	7,72
Velocità del vento	26	3,26	21,29	8,73	8,08	11,05	3,32
	34	3,23	30,83	14,14	13,53	21,06	4,59
	35	3,54	17,04	9,00	8,53	7,82	2,80
Massima raffica	26	7,44	34,94	17,01	16,15	28,29	5,32
	34	0,97	47,26	21,32	20,23	48,02	6,93
	35	7,94	32,14	17,72	17,14	22,93	4,79
Radiazioni solari	26	0,04	1,181	0,677	0,723	0,082	0,29
	34	0,034	1,195	0,704	0,749	0,093	0,30
	35	0,023	1,149	0,632	0,661	0,079	0,28
Temperatura esterna	26	41,05	94,07	76,22	78,99	166,76	12,91
	34	37,88	89,94	72,93	74,05	127,92	11,31
	35	40,69	95,15	75,91	77,55	162,90	12,76

3.5.4.1 Ossidi di azoto

E' facile notare come nella prima parte dell'anno e nel periodo autunnale la serie assuma valori molto elevati e soggetti a maggiore variabilità rispetto alle misurazioni raccolte durante i mesi primaverili ed estivi. Tale fenomeno potrebbe essere attribuito alla maggior intensità dell'attività industriale ma soprattutto del traffico veicolare durante il periodo in questione. Anche in questo caso, come nel 2001, la distribuzione dei dati raccolti nella centralina 35, posta all'interno del centro urbano, assume valori mediamente più elevati di quelli misurati nelle altre stazioni poste più in periferia o in prossimità di spazi aperti (centralina 26) e che si affacciano

sul mare, come la centralina 34. Osservando gli istogrammi si notano, per tutte le centraline, pesanti code sulla destra delle distribuzioni, in corrispondenza di concentrazioni elevate che si manifestano come possibili valori anomali nei diagrammi a scatola.

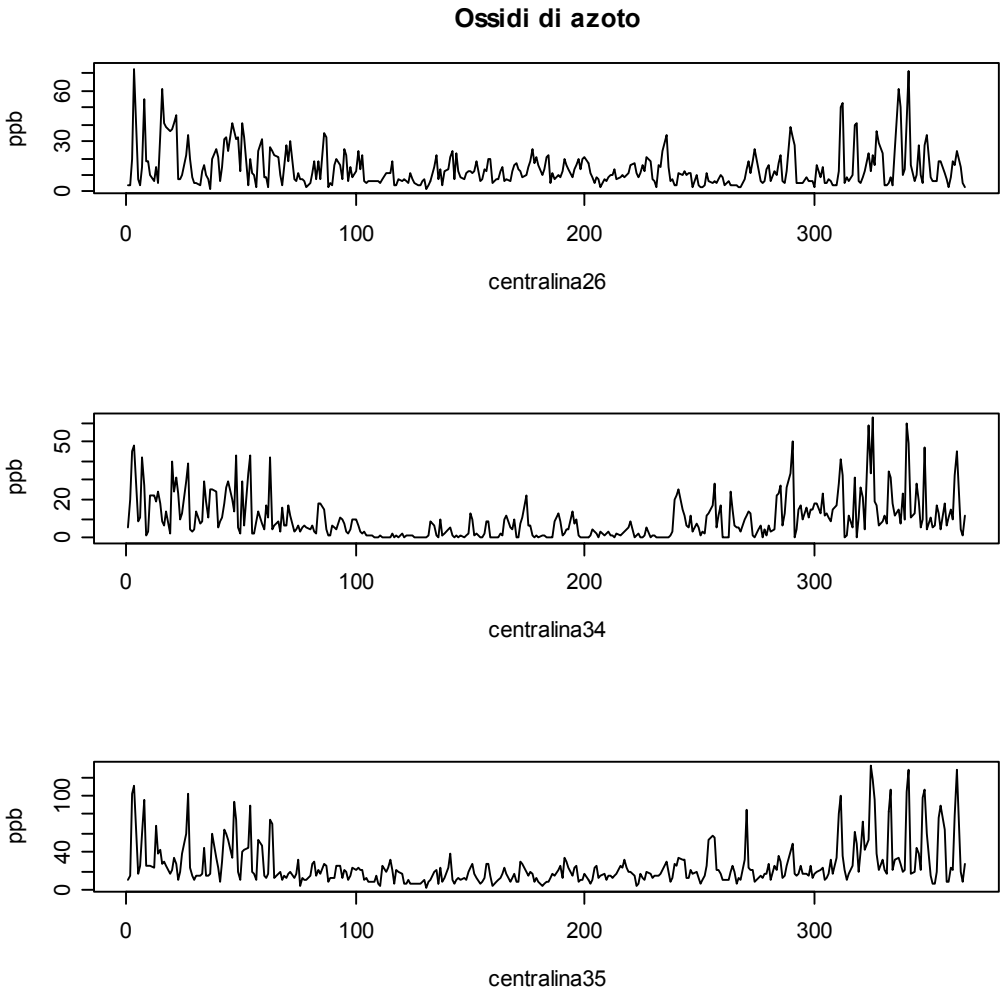


Figura 4.19: *Serie degli ossidi di azoto 2002*

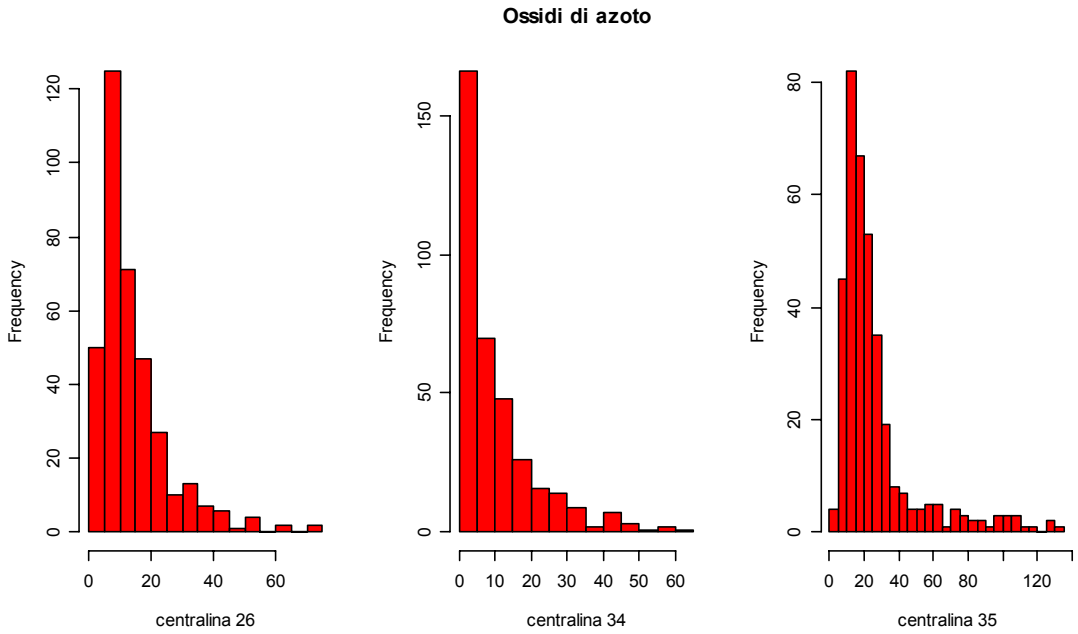


Figura 4.20: *Istogrammi degli ossidi di azoto 2002*

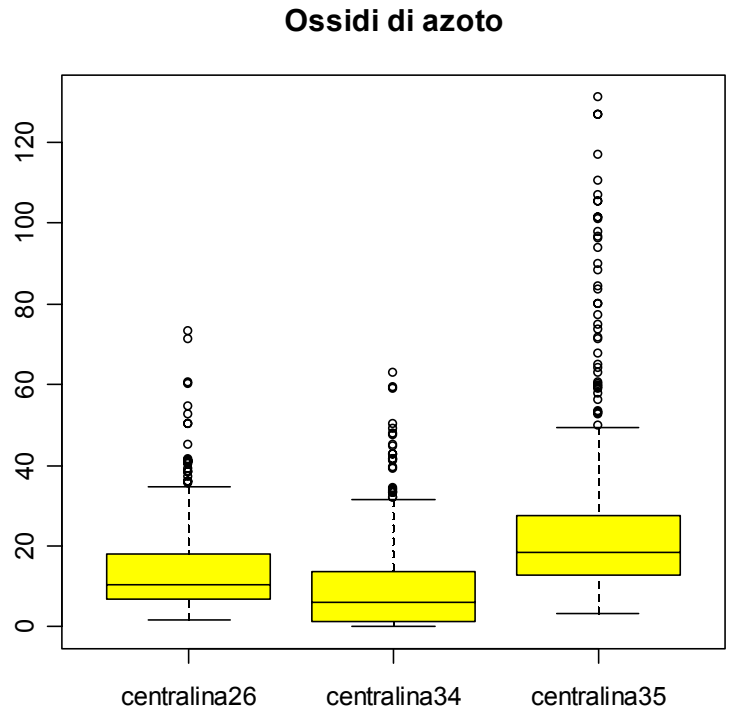


Figura 4.21: *Box-plot degli ossidi di azoto 2002*

3.5.4.2 Ozono

L'osservazione delle serie riguardanti la concentrazione di ozono nel 2002 mette in luce la presenza di una discreta variabilità dei dati che caratterizza tutto l'anno e che aumenta nei mesi centrali quando, per effetto dell'aumento della temperatura e della radiazione solare dovute all'arrivo della stagione estiva, si registrano valori più elevati della variabile studiata.

Le distribuzioni dei dati rilevati dalle tre diverse centraline si presentano, ancor più che nel 2001, abbastanza simili. Sono anche evidenti osservazioni che superano i limiti superiori dei diagrammi a scatola rivelandosi così come potenziali *outliers*.

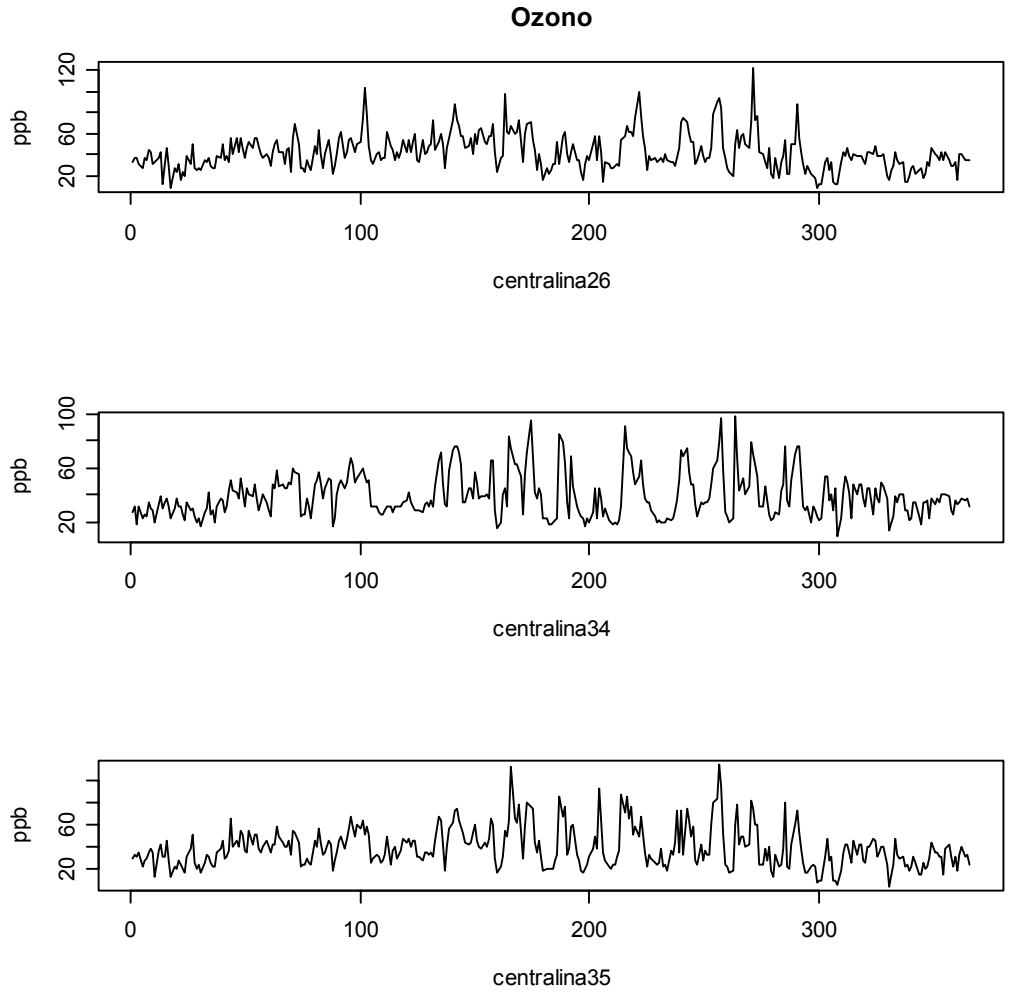


Figura 4.22: *Serie delle concentrazioni di ozono nel 2002*

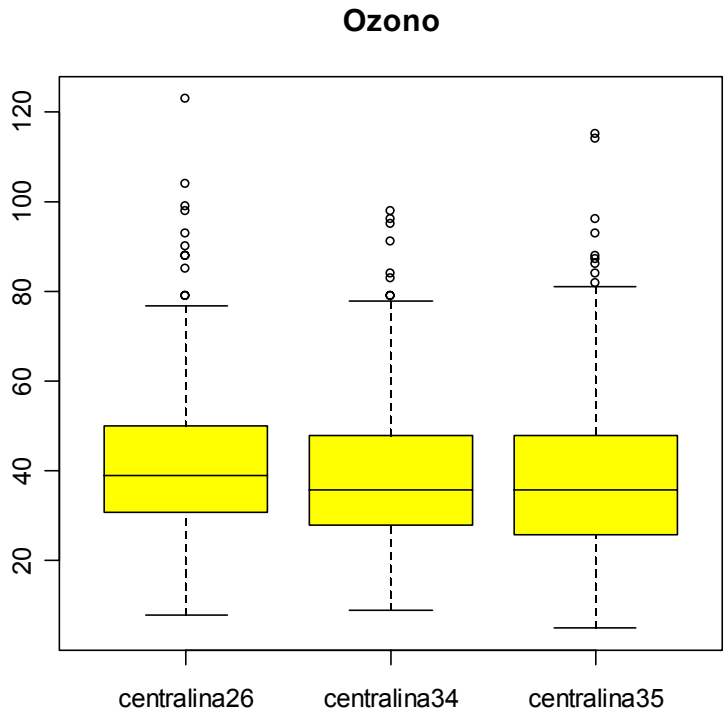


Figura 4.23: *Box-plot delle concentrazioni di ozono nel 2002*

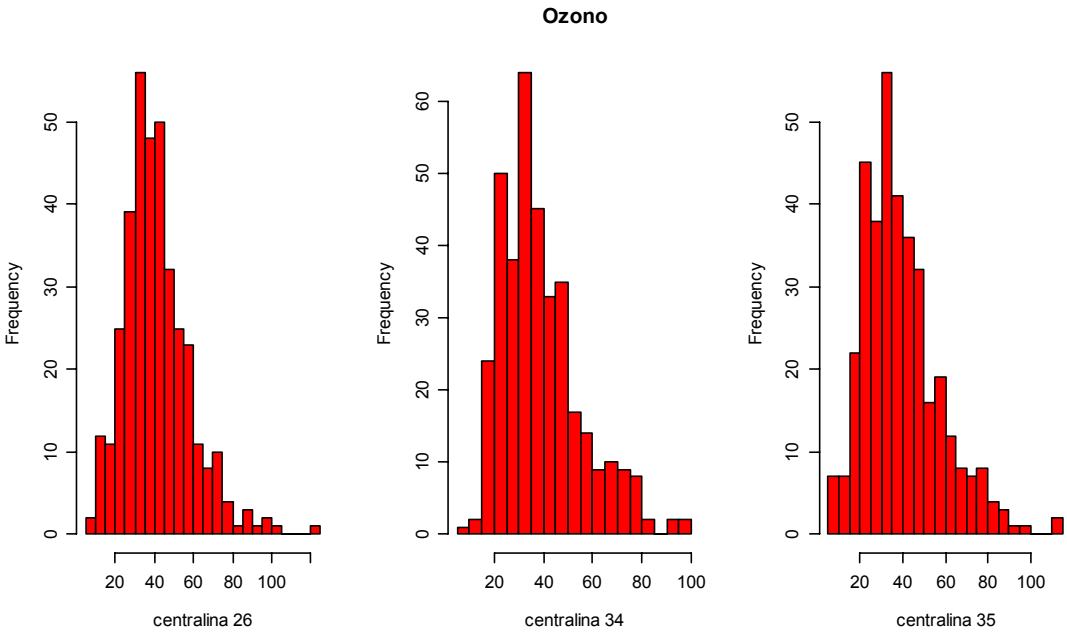


Figura 4.24: *Istogrammi delle concentrazioni di ozono nel 2002*

3.5.4.3 Polveri atmosferiche

Entrambe le serie storiche delle polveri atmosferiche, rilevate dalle centraline 34 e 35, riflettono l'esistenza di picchi soprattutto nei mesi di luglio e settembre.

Dallo studio dei diagrammi a scatola e degli istogrammi è possibile rilevare l'esistenza di valori anomali che superano i limiti superiori dei box-plot andando a formare delle code pesanti a destra. Come si era già osservato relativamente all'anno 2001, anche in questa occasione, le misurazioni effettuate nella centralina 35, situata nel centro urbano, si rivelano in media leggermente più elevate di quelle ottenute nella centralina 34.

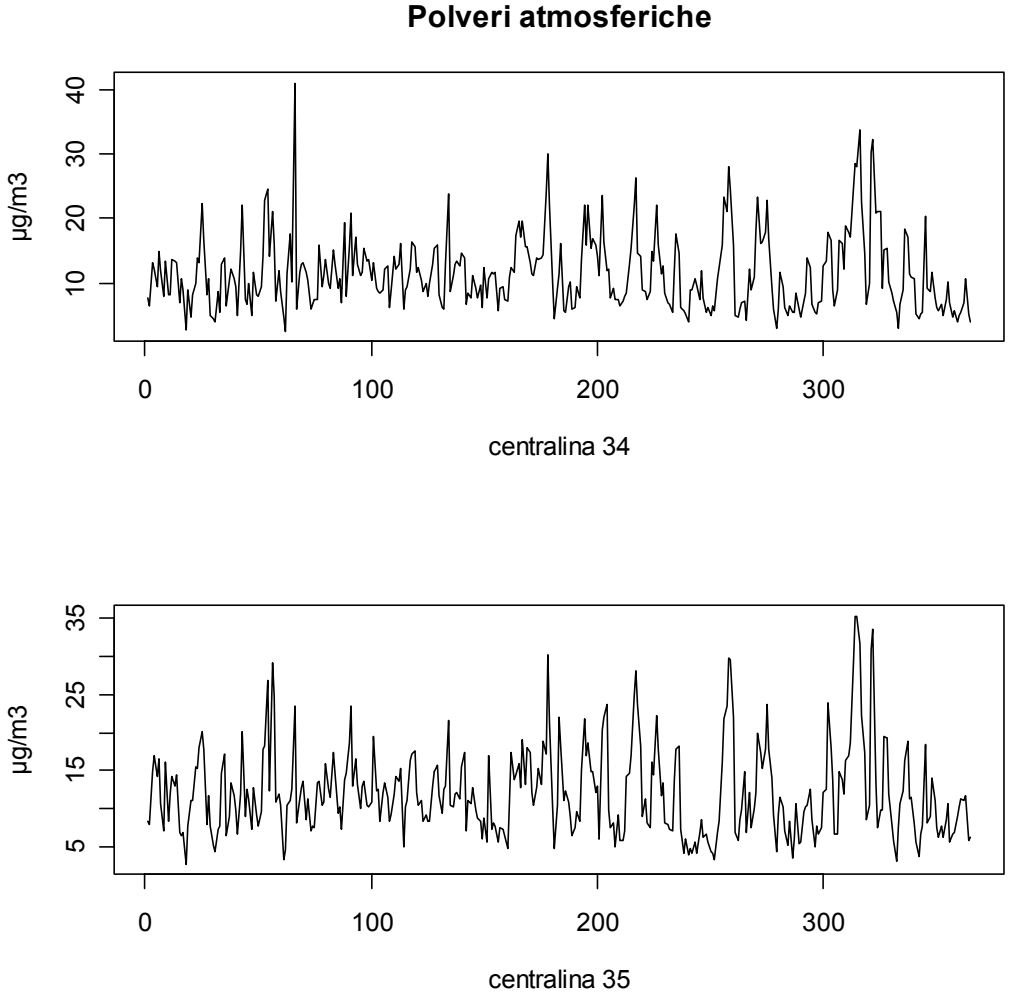


Figura 4.25: *Serie delle polveri atmosferiche nel 2002*

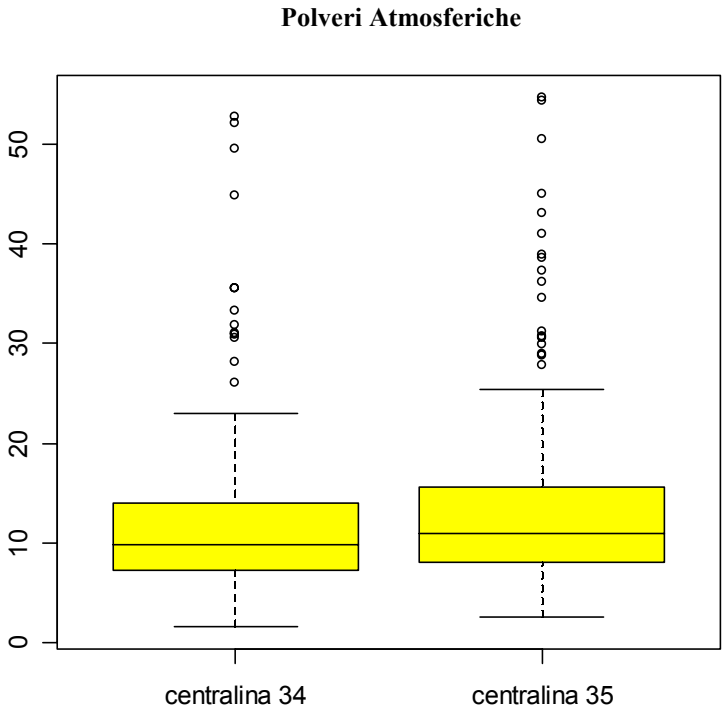


Figura 4.26: *Diagrammi a scatola delle polveri atmosferiche nel 2002*

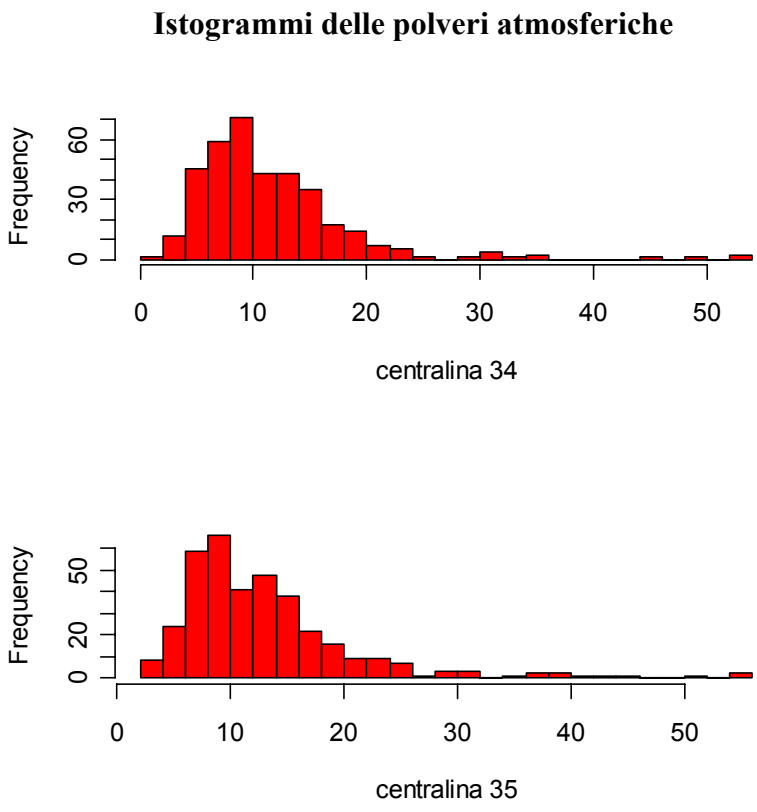


Figura 4.27: *Istogrammi delle polveri atmosferiche nel 2002*

3.5.5 Variabili meteorologiche nel 2002

I risultati ottenuti dall'analisi descrittiva per le variabili meteorologiche misurate nel 2002 non si discostano molto da quelli acquisiti nell'anno precedente.

Anche in questo caso, le serie relative alla "velocità del vento" e alla "massima raffica di vento" si rivelano molto variabili, soprattutto nei mesi invernali, e alcune distribuzioni appaiono gravate da code pesanti. Gli istogrammi mettono in

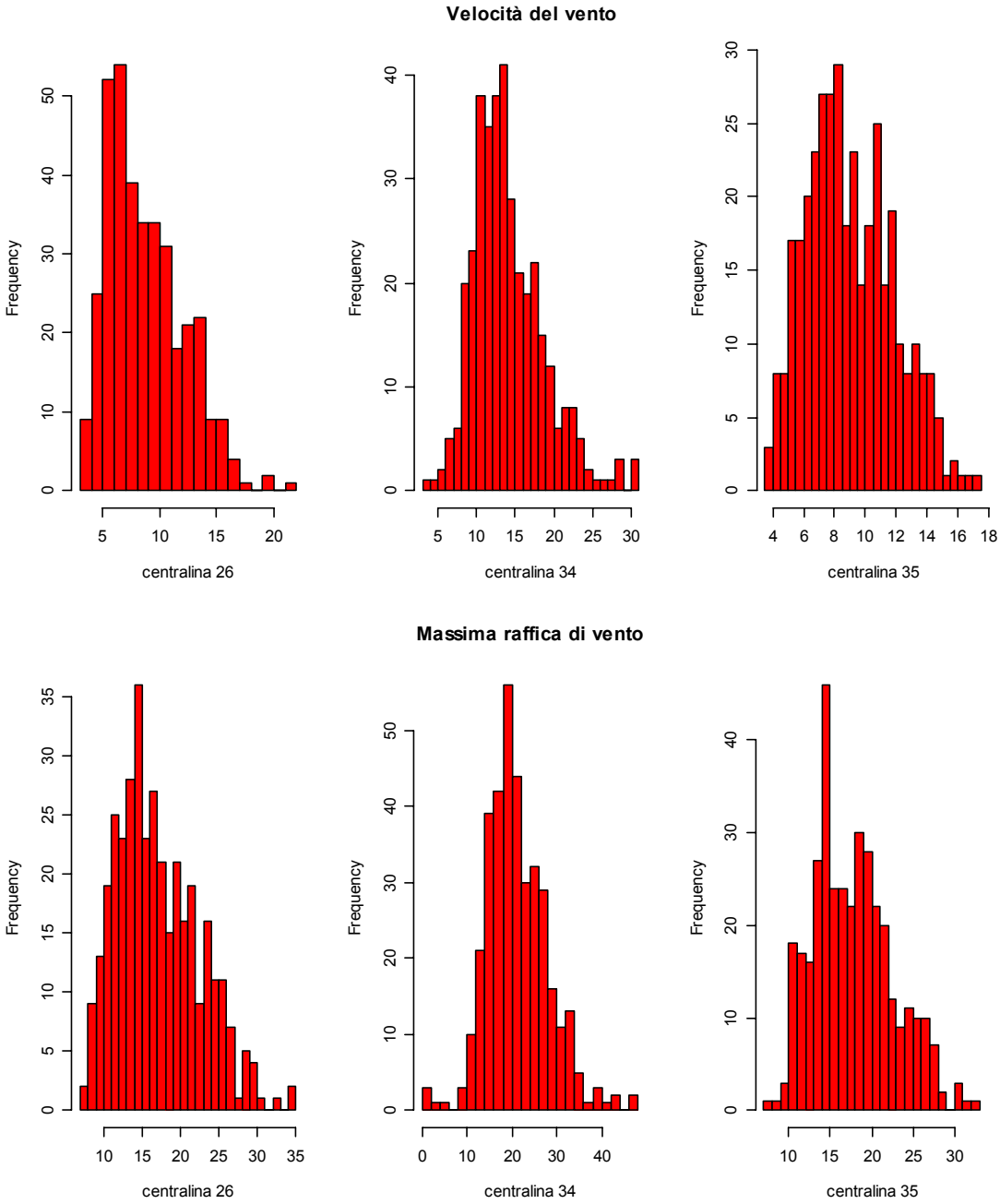


Figura 4.28: *Istogrammi della velocità e della massima raffica di vento nel 2002*

luce una leggera asimmetria positiva in tutte le distribuzioni: le osservazioni della centralina 34, tuttavia, presentano un'asimmetria meno marcata rispetto alle altre rivelando un comportamento più stabile. Osservando i diagrammi a scatola è inoltre

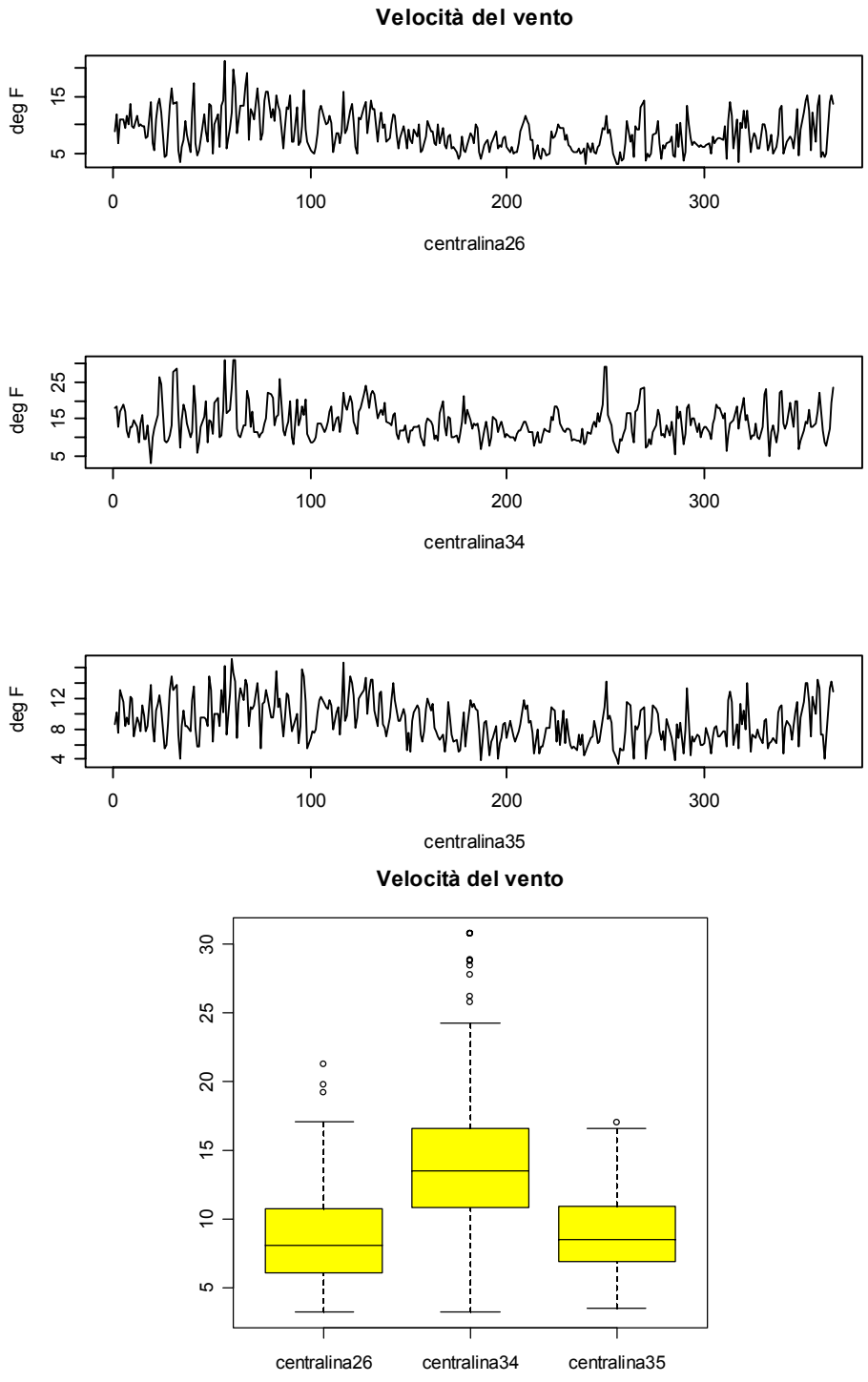


Figura 4.29: *Serie e box-plot della velocità nel 2002*

possibile notare che le misurazioni ottenute nella centralina 34 assumono valori più elevati rispetto alle rilevazioni delle altre centraline, ciò può essere condizionato dal fatto che la stazione di rilevazione in questione è situata in uno spazio aperto vicino al mare.

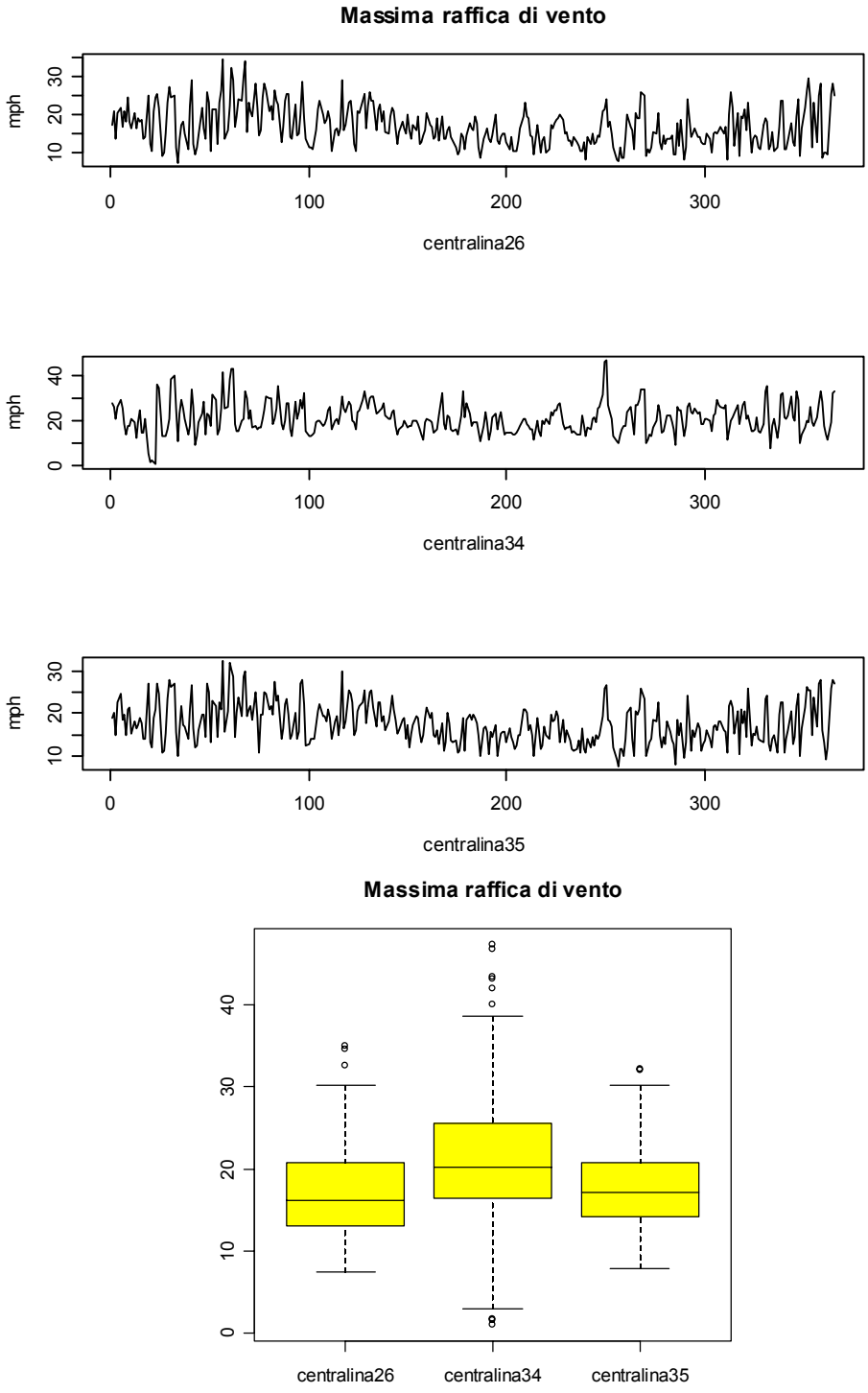


Figura 4.30: *Serie e box-plot della massima raffica di vento nel 2002*

Anche per quanto riguarda la temperatura esterna e la radiazione solare non si notano particolari differenze rispetto al 2001. Come è possibile vedere dai grafici l'andamento delle serie storiche segue il naturale susseguirsi delle stagioni

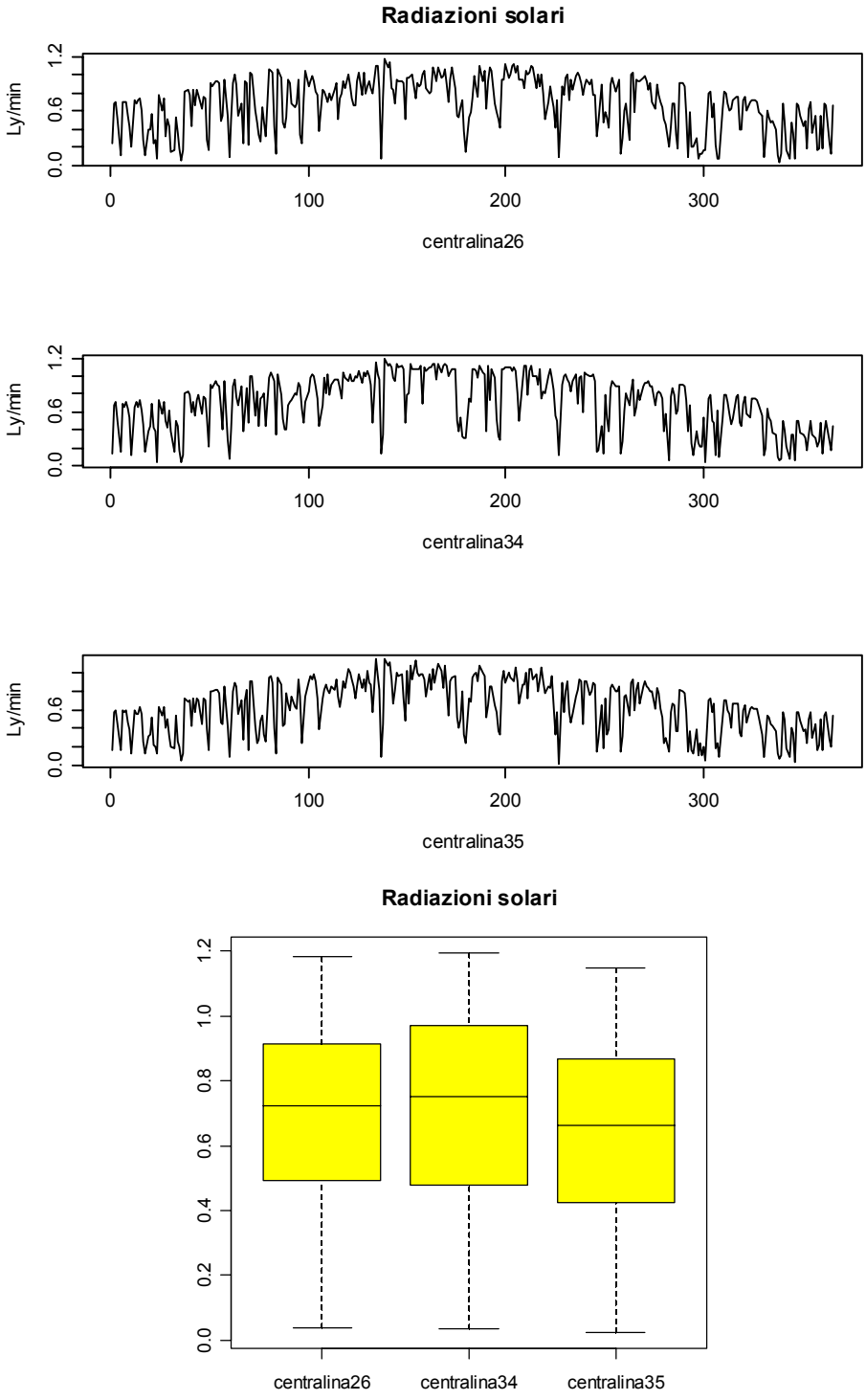


Figura 4.31: *Serie e box-plot della radiazione solare nel 2002*

attestandosi su valori elevati durante il periodo primaverile-estivo e su valori inferiori nei mesi invernali.

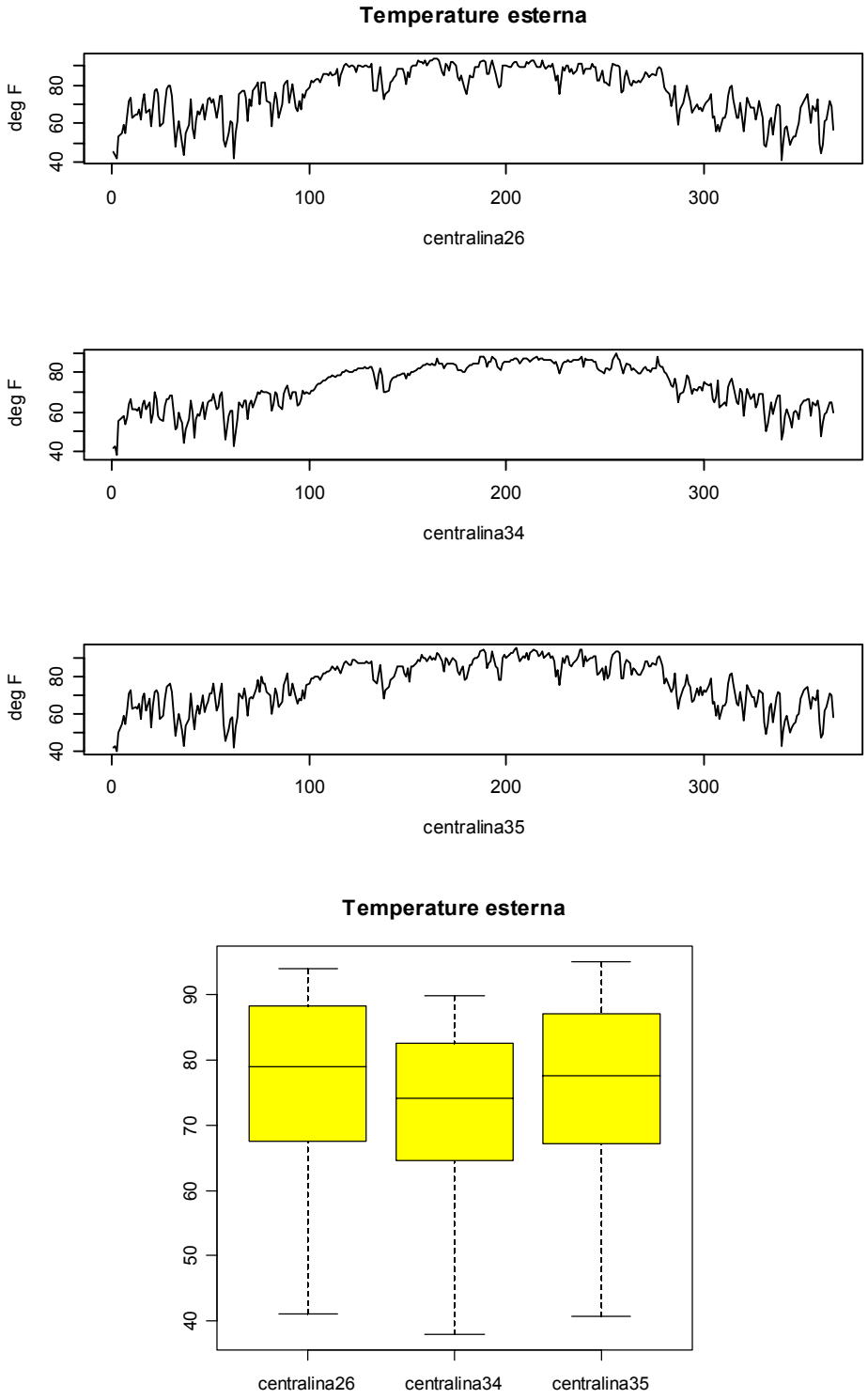


Figura 4.32: *Serie e box-plot della temperatura esterna nel 2002*

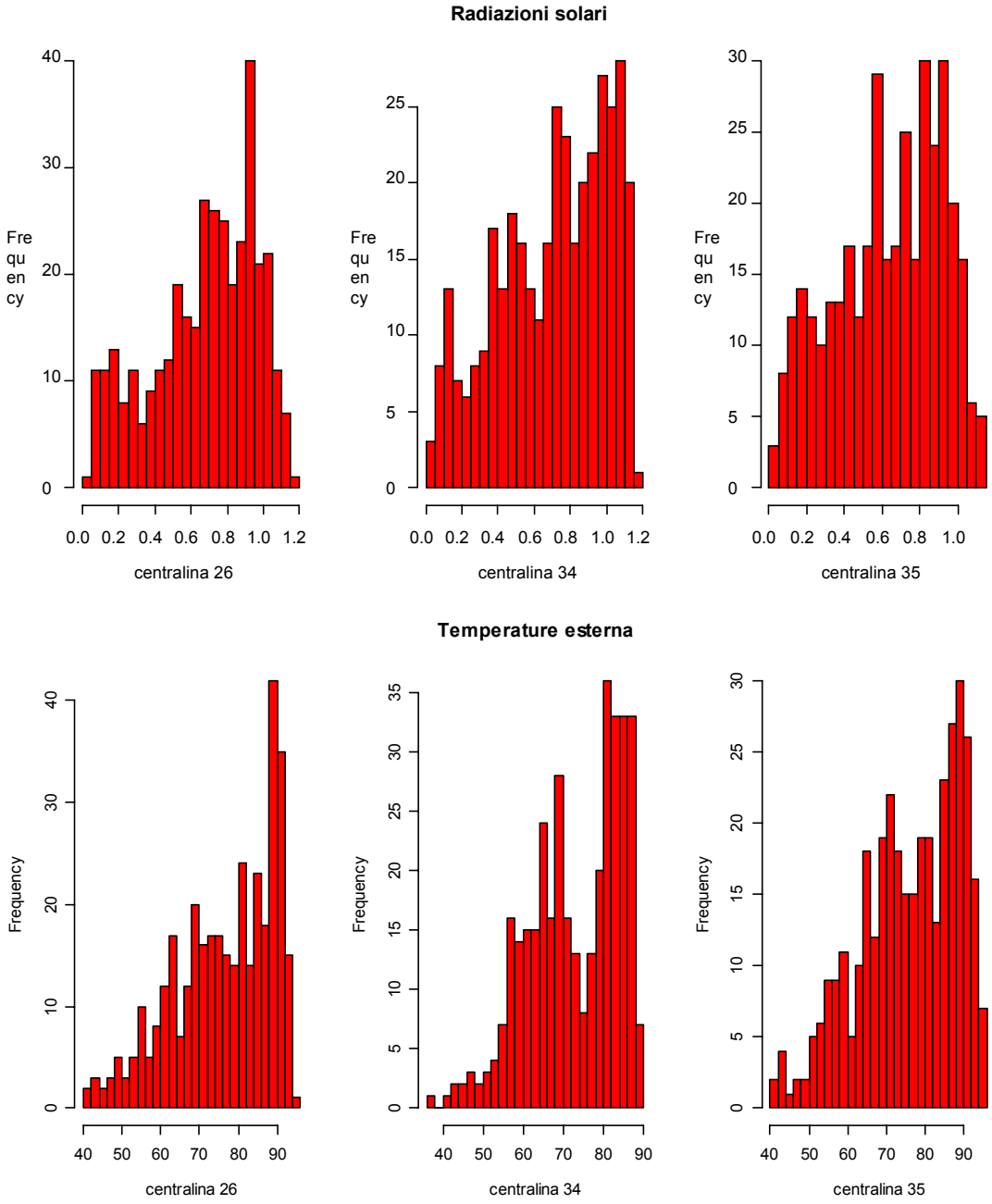


Figura 4.33: *Istogrammi temperatura esterna e della radiazione solare per il 2002*

3.5.6 Correlazioni 2002

Al fine di studiare la presenza e l'entità di un qualsiasi legame tra i parametri misurati, nella Tabella G sono state riportate le correlazioni tra tutte le variabili

rilevate nelle tre centraline. Anche in questo caso, come per i dati del 2001, esistono legami significativi tra le concentrazioni di inquinanti. Tali relazioni riguardano soprattutto gli ossidi di azoto e le polveri inalabili che risultano positivamente correlati ai valori delle stesse variabili negli altri siti.

Per quanto concerne l'ozono si può notare, tra le centraline 26 e 35, una correlazione più forte rispetto al legame esistente tra ciascuna di queste e la stazione 34. Ciò probabilmente è dovuto alla loro collocazione geografica che presenta alcune somiglianze per quanto riguarda l'ambiente circostante e si differenzia dalla realtà in cui è immersa la centralina 34, posta in uno spazio più aperto e in prossimità del mare. Appare inoltre evidente il legame tra le variabili "temperatura esterna" e "radiazione solare", per natura strettamente correlate, e l'inquinante "ozono" la cui formazione sappiamo essere fortemente favorita dalla presenza di alte temperature e dall'azione dei raggi UV. La temperature esterna, oltre ad avere una buona correlazione positiva con l'ozono, è legata negativamente agli ossidi di azoto: questa relazione, riscontrabile anche nella tabella relativa ai dati del 2001, può essere giustificata dal fatto che alte temperature rendono più facile la dissociazione della molecola di ossido di azoto in una di monossido di azoto e una di ossigeno monoatomico (che poi andrà a formare l'ozono legandosi ad una molecola di ossigeno).

Si rivela ancora una volta fondamentale l'azione esercitata dalla velocità del vento e dalle variabili "massima raffica di vento" e "direzione": la tabella delle correlazioni mette in luce la relazione inversa esistente tra queste e gli inquinanti; l'entità delle concentrazioni di ossidi di azoto e polveri, e quindi anche di ozono, tende infatti a diminuire per l'effetto di dispersione causato dalla presenza di forte vento, e ad aumentare in caso di una ventilazione debole o assente. Soffermandosi ad analizzare le correlazioni tra le diverse centraline, si nota inoltre che per ogni variabile, esclusa la velocità del vento, il legame tra la C26 e la C34 è leggermente più debole di quello registrato tra queste e la stazione 35. La ragione di questo comportamento potrebbe essere identificata nella maggior distanza che separa queste centraline: esse si trovano infatti a circa un grado di latitudine e di longitudine l'una dall'altra e a 117,15 Km di distanza; la stazione di rilevazione 35 invece è situata in posizione intermedia sia in termini di latitudine e longitudine sia in termini di chilometri (52,39 Km dalla C34 e 66,85 Km dalla C26).

	NOx 26	NOx 34	NOx 35	O3 26	O3 34	O3 35	PM 34	PM 35	WS 26	WS 34	WS 35	RWD 26	RWD 34	RWD 35	MWG 26	MWG 34	MWG 35	SWD 26	SWD 34	SWD 35	OT 26	OT 34	OT 35	SR 26	SR 34	SR 35
NOx 26	1,000																									
NOx 34	0,395	1,000																								
NOx 35	0,479	0,728	1,000																							
O3 26	0,009	-0,046	0,044	1,000																						
O3 34	-0,277	-0,514	-0,427	0,301	1,000																					
O3 35	0,042	0,012	0,088	0,797	0,300	1,000																				
PM 34	-0,074	0,080	0,025	0,284	0,291	0,354	1,000																			
PM 35	-0,029	0,093	0,101	0,334	0,293	0,437	0,875	1,000																		
WS 26	-0,205	-0,208	-0,246	-0,206	-0,244	-0,269	-0,223	-0,290	1,000																	
WS 34	-0,325	-0,166	-0,268	-0,245	-0,234	-0,309	-0,194	-0,254	0,627	1,000																
WS 35	-0,121	-0,331	-0,304	-0,112	-0,148	-0,216	-0,218	-0,287	0,850	0,580	1,000															
RWD 26	-0,047	0,338	0,278	-0,041	-0,274	0,039	0,011	0,049	0,137	0,143	-0,108	1,000														
RWD 34	-0,148	0,214	0,167	-0,041	-0,114	0,043	0,000	-0,026	0,165	0,135	-0,024	0,690	1,000													
RWD 35	-0,089	0,349	0,292	-0,031	-0,248	0,042	0,051	0,070	0,105	0,155	-0,166	0,844	0,708	1,000												
MWG 26	-0,236	-0,266	-0,299	-0,187	-0,175	-0,253	-0,215	-0,295	0,979	0,651	0,848	0,096	0,119	0,071	1,000											
MWG 34	-0,350	-0,192	-0,252	-0,222	-0,199	-0,269	-0,207	-0,256	0,538	0,931	0,499	0,114	0,083	0,135	0,582	1,000										
MWG 35	-0,223	-0,304	-0,309	-0,181	-0,207	-0,255	-0,226	-0,301	0,889	0,720	0,929	0,012	0,044	0,005	0,912	0,665	1,000									
SWD 26	0,190	0,246	0,344	0,245	0,082	0,350	0,187	0,255	-0,610	-0,456	-0,607	0,162	0,121	0,178	-0,604	-0,380	-0,601	1,000								
SWD 34	-0,024	0,252	0,146	0,129	-0,055	0,219	0,134	0,160	-0,213	-0,140	-0,428	0,381	0,361	0,463	-0,206	-0,055	-0,304	0,395	1,000							
SWD 35	0,074	0,180	0,237	0,208	0,151	0,326	0,128	0,188	-0,596	-0,342	-0,661	0,172	0,099	0,227	-0,570	-0,269	-0,577	0,619	0,471	1,000						
OT 26	-0,214	-0,530	-0,416	0,400	0,217	0,373	0,257	0,271	-0,155	-0,291	-0,033	-0,299	-0,152	-0,313	-0,104	-0,290	-0,148	0,032	-0,108	0,067	1,000					
OT 34	-0,277	-0,514	-0,427	0,301	0,198	0,300	0,291	0,293	-0,244	-0,234	-0,148	-0,274	-0,114	-0,248	-0,175	-0,199	-0,207	0,082	-0,055	0,151	0,932	1,000				
OT 35	-0,248	-0,516	-0,404	0,371	0,212	0,364	0,286	0,303	-0,215	-0,297	-0,099	-0,259	-0,114	-0,260	-0,158	-0,279	-0,192	0,078	-0,082	0,120	0,980	0,967	1,000			
SR 26	-0,031	-0,099	0,038	0,586	0,409	0,556	0,177	0,225	-0,203	-0,215	-0,166	0,004	0,068	0,007	-0,201	-0,241	-0,240	0,251	-0,003	0,235	0,514	0,409	0,484	1,000		
SR 34	-0,048	-0,195	-0,075	0,566	0,493	0,510	0,225	0,258	-0,177	-0,228	-0,063	-0,078	0,016	-0,074	-0,175	-0,279	-0,186	0,177	-0,085	0,145	0,595	0,493	0,573	0,851	1,000	
SR 35	-0,060	-0,168	-0,039	0,610	0,461	0,570	0,232	0,271	-0,216	-0,203	-0,082	-0,057	0,029	-0,055	-0,207	-0,228	-0,190	0,227	-0,059	0,169	0,570	0,461	0,554	0,920	0,903	1,000

Tabella F: Correlazioni tra le serie di ogni variabile per tutte le centraline per l'anno 2002

Capitolo 4
Carte di controllo

4.1 Carte di controllo multivariate

Valutando quanto visto nell'analisi descrittiva dei dati e considerando, in particolare, le relazioni messe in luce dalle correlazioni tra variabili e centraline si è giunti a conoscenza del particolare legame che interessa le concentrazioni di ozono e la radiazione solare in tutte le centraline oggetto di studio. Per questo motivo si è scelto di approfondire un modello dinamico tra queste due variabili, in modo tale da tenere in considerazione sia la relazione esistente tra esse, sia la loro evidente dipendenza con le osservazioni passate.

Il modello oggetto di studio è quindi del tipo:

$$[X_{ozono26}(t-1), X_{ozono34}(t-1), X_{ozono35}(t-1), X_{radiazioni26}(t-1), X_{radiazioni34}(t-1), X_{radiazioni35}(t-1), X_{ozono26}(t), X_{ozono34}(t), X_{ozono35}(t), X_{radiazioni26}(t), X_{radiazioni34}(t), X_{radiazioni35}(t)]$$

Su questo modello verranno applicate le carte di controllo descritte in precedenza: in primo luogo saranno costruite le carte multivariate tradizionali utilizzando le statistiche T^2 e Q per verificare se sono presenti cambiamenti importanti nella media delle osservazioni; in un secondo momento verranno calcolati

gli indici D e A, che ci permetteranno di identificare cambiamenti significativi nella struttura di correlazione tra i diversi istanti temporali e tra le varie centraline.

La statistica T^2 è una misura della variazione all'interno del modello per componenti principali mentre la statistica Q valuta l'ammontare di variazione non spiegata dalle variabili latenti. Queste carte tuttavia non sono sempre in grado di identificare un cambiamento nella correlazione delle variabili finché gli indici T^2 e Q giacciono all'interno dei limiti, risulta quindi interessante studiare il miglioramento ottenibile per mezzo dei metodi "DISSIM" e "MPCA" che si concentrano proprio sui cambiamenti nelle strutture relazione tra variabili nel tempo.

4.1.1 Definizione dell'insieme di riferimento

Per l'analisi sul modello "ozono-radiazione solare" nelle tre centraline si è scelto di effettuare il controllo sui dati relativi al 2002, adottando come insieme di riferimento le osservazioni appartenenti al 2001.

Il procedimento prevede la creazione di un insieme di riferimento. Si costruiscono le carte T^2 e Q per i dati originali del 2001 e si eliminano tutte le osservazioni che superano i limiti di controllo calcolati secondo la (1.2). Dal momento che non è possibile distinguere la natura dei fuori controllo a causa della scarsità di informazioni si è deciso di escludere tutti i valori più grandi dei limiti della carta T^2.

Nella Tabella G vengono presentate tutte le osservazioni escluse dall'insieme di riferimento

Tabella G

Osservazioni fuori controllo
71,132,133,152,153,160,166,167,168,169
173,174,203,204,220,237,265,266,291

4.1.2 Carte di controllo tradizionali

Si procede, ora alla costruzione delle carte tradizionali multivariate T^2 e Q basandosi sul metodo delle componenti principali; di seguito vengono riportate le

informazioni più significative ottenute con il metodo PCA; in particolare sono presentate:

* la media e le deviazioni standard di ogni variabile del processo in controllo
* gli autovalori del processo in controllo
* la matrice delle componenti scalate del processo in controllo

a)

	Xozono26(t-1)	Xozono34(t-1)	Xozono35(t-1)
Media	41,944	35,369	38,836
Dev.Std.	18,776	16,284	17,327
	Xozono26(t)	**Xozono34(t)**	**Xozono35(t)**
Media	41,684	35,041	38,532
Dev.Std.	18,512	16,015	16,896
	Xradiazioni26(t-1)	**Xradiazioni34(t-1)**	**Xradiazioni35(t-1)**
Media	0,678	0,727	0,625
Dev.Std.	0,297	0,305	0,278
	Xradiazioni26(t)	**Xradiazioni34(t)**	**Xradiazioni35(t)**
Media	0,675	0,729	0,624
Dev.Std.	0,296	0,308	0,278

b)

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Autovalori	2,6930156	1,4310317	1,1457575	0,643886	0,516263	0,380144
Proporzione di var. spiegata	0,6061334	0,1711548	0,1097175	0,03465	0,022276	0,012078
Prop. di var.spiegata cumulata	0,6061334	0,7772882	0,8870057	0,921656	0,943932	0,95601
	Comp.7	**Comp.8**	**Comp.9**	**Comp.10**	**Comp.11**	**Comp.12**
Autovalori	0,3738708	0,3307415	0,3190899	0,263665	0,239873	0,219761
Proporzione di var. spiegata	0,0116824	0,0091426	0,0085097	0,00581	0,004809	0,004036
Prop. di var.spiegata cumulata	0,9676921	0,9768347	0,9853444	0,991155	0,995964	1

c)

	Autovettori											
	Comp.1	**Comp.2**	**Comp.3**	**Comp.4**	**Comp.5**	**Comp.6**	**Comp.7**	**Comp.8**	**Comp.9**	**Comp.10**	**Comp.11**	**Comp.12**
OZ26t1	-0.306	-0.184	-0.228	-0.434	-0.404	0.132	0.555	0.363				
OZ34t1	-0.245	-0.451	-0.121	-0.252	0.436		-0.383	0.207	-0.305	-0.416	0.101	
OZ35t1	-0.301	-0.286	-0.208	-0.340	0.103	-0.223	-0.120	-0.487	0.380	0,306	-0.136	
SR26t1	-0.294	0.240	-0.354	0.233		-0.419	0.217	-0.135	-0.408	-0.106	-0.301	0.407
SR34t1	-0.282	0.249	-0.388	0.209		0.515	-0.311	0.202	0.416		-0.159	0.230
SR35t1	-0.306	0.206	-0.359	0.255					-0.145	0.164	0.473	-0.626
OZ26	-0.313	-0.141	0.230	0.101	-0.693		-0.492		-0.281	0.107		
OZ34	-0.247	-0.403	0.243	0.440	0.290	0.211	0.257	0.253	-0.129	0,332	-0.135	
OZ35	-0.306	-0.228	0.270	0.372			0.222	-0.312	0.390	-0.565		
SR26	-0.280	0.324	0.320	-0.123	0.101	-0.425	-0.100	0.456	0.235		-0.365	-0.318
SR34	-0.284	0.304	0.286	-0.301	0.175	0.497	0.140	-0.393	-0.306		-0.256	-0.190
SR35	-0.291	0.294	0.341	-0.147	0.146					0.128	0.633	0.491

Tabella PCA: a) medie e deviazioni standard; b) autovalori; c) autovettori

La variabilità totale dei dati è spiegata da tutte le 12 componenti principali ciascuna delle quali spiega a sua volta una proporzione decrescente di varianza. Un criterio per la scelta del numero adeguato di componenti principali da tenere in considerazione, cercando di non perdere troppe informazioni, consiste nell'includere il numero di componenti in grado di spiegare una percentuale abbastanza grande della variabilità totale (usualmente si scelgono quelle che spiegano tra il 70 ed il 90 per cento della varianza). Quale criterio per la scelta delle componenti principali, si è scelto di adottare quello che stabilisce di prendere in considerazione le componenti con varianza superiore ad uno; nel nostro caso si tratta quindi di utilizzare tre componenti principali. La prima componente individua il contributo dato da ciascuna variabile al processo in esame, la seconda invece opera una distinzione tra l'influenza esercitata dall'ozono e quella data dalla radiazione solare; la terza componente principale infine pone l'attenzione sul contributo dato dalle variabili al tempo t come contrapposto a quello offerto dalle stesse al tempo t-1.

Tabella H

Osservazioni fuori controllo
102,141,142,166,167,174,175,216,217,255
256,257,258,264,271,272,291,309

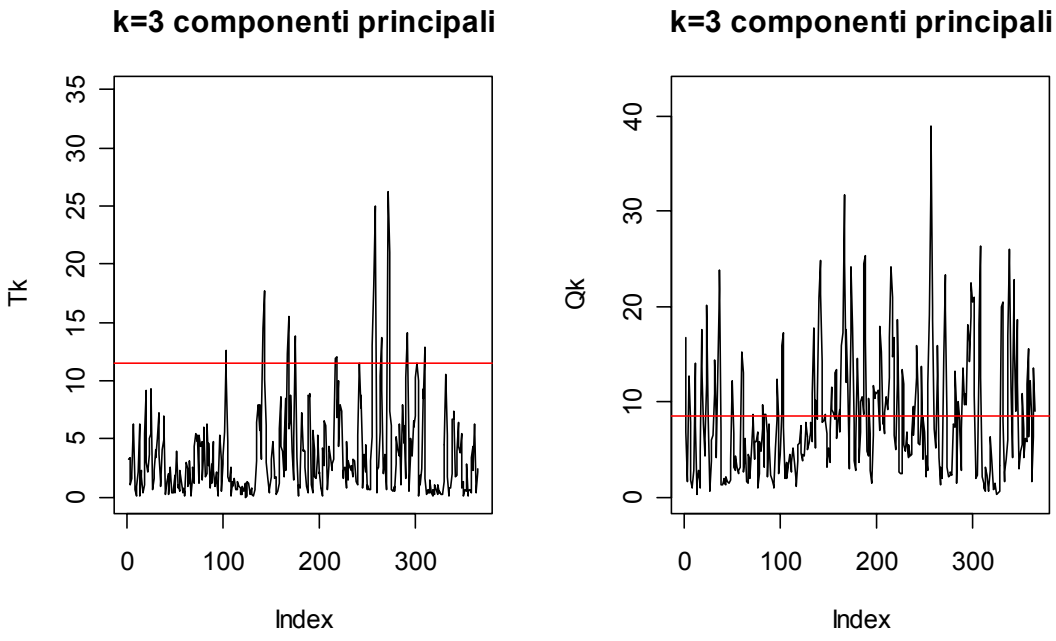


Figura 4.1: *Carte T^2 e Q costruite per tre componenti principali*

Dall'osservazione della carta T^2 si possono notare 18 valori fuori controllo, anche di entità rilevanti concentrati soprattutto nella parte centrale dell'anno. Anche nel grafico della statistica Q ci sono numerose osservazioni che superano il limite superiore, ciò significa che gran parte della varianza non viene spiegata dalle tre componenti principali considerate. I limiti delle carte di controllo sono riassunti nella tabella sottostante:

	limiti
T	11,57201
Q	8,475935

Per cercare di capire quale delle tre componenti principali è la maggior responsabile dei fuori controllo si procede alla scomposizione della statistica T^2 nelle sue componenti.

Tabella I

Oss	[,1]	[,2]	[,3]	[,4]
102	12,66978	0,42818	8,35208	3,88952
141	14,60401	0,05476	9,44484	5,10442
142	17,69268	0,54565	7,60720	9,53982
166	12,32022	0,01270	11,21328	1,09424
167	15,48786	2,44463	6,56854	6,47469
174	13,78688	0,00084	4,58116	9,20488
175	13,10247	1,16622	5,23116	6,70509
216	11,91372	0,00144	8,89565	3,01663
217	12,11774	0,43648	5,64288	6,03838
255	13,01855	0,05389	8,95871	4,00595
256	18,08634	0,36409	14,56556	3,15669
257	25,04240	0,00003	14,70835	10,33403
258	23,15513	4,23497	4,18159	14,73858
264	13,71686	0,03933	10,90714	2,77039
271	26,22346	1,70569	17,48717	7,03060
272	21,03509	1,54840	9,67713	9,80955
291	14,17754	0,61227	3,21156	10,35370
309	12,93892	4,18885	2,61752	6,13256

Come si può facilmente notare (Tabella I) sono solo tre le osservazioni (167, 258, 309) fuori controllo imputabili congiuntamente a tutte e tre le componenti. La maggior responsabile delle osservazioni anomale sembra essere la seconda componente principale cui sono attribuibili 4 valori anomali singolarmente e 7 fuori controllo se si considera l'influsso congiunto della terza variabile latente,

responsabile di tre fuori controllo. La componente principale che meno contribuisce in termini di fuori controllo è quindi la prima; osservando la carta T^2 relativa ad essa è evidente la presenza di un solo valore che supera il limite superiore.

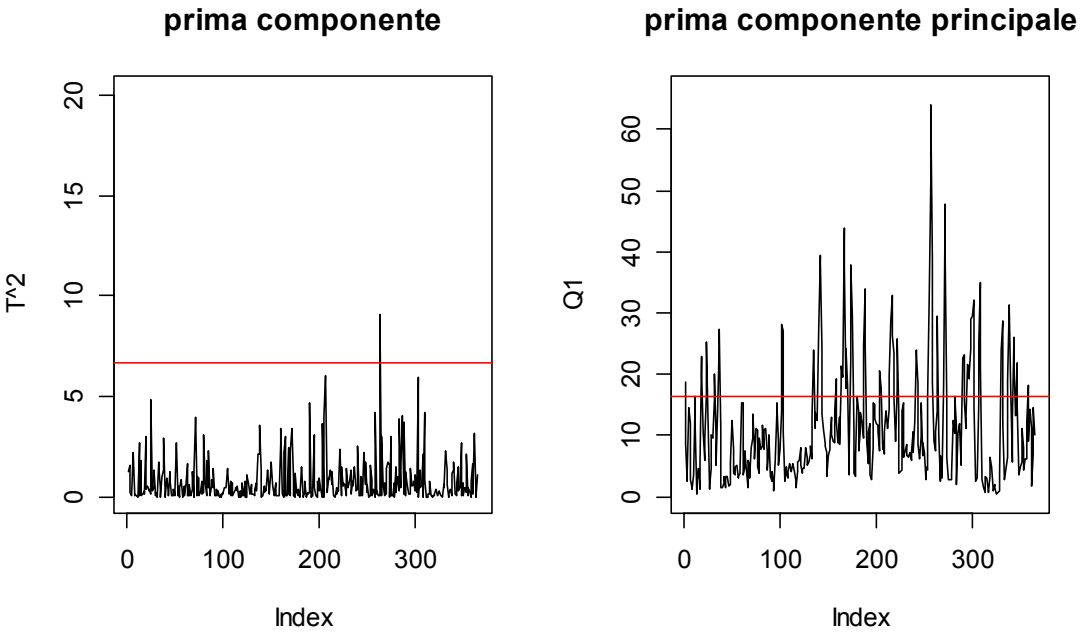


Figura 4.2: *Carte T^2 e Q costruite per la prima componente principale*

4.1.3 Statistiche *D* e *A*

Applichiamo ora i due metodi suggeriti da Kano (2001) per sorvegliare insiemi di dati multivariati e autocorrelati. Per disegnare le carte di controllo ci serviamo della costruzione di due statistiche: l'indice di diversità *D*, per quanto riguarda la procedura *dissim*, e l'indice *A* per la procedura *delle moving principal component analysis*.

Considerando la numerosità dell'insieme dei dati oggetto di studio e conoscendo l'importanza di utilizzare una dimensione della finestra temporale adeguata per facilitare l'interpretazione degli indici *A* e *D*, si è scelto si costruire le carte di controllo per tre diversi valori di *w*: 50,100,150.

✔ *Indice di diversità D*

Osservando le carte, così disegnate, relative all'indice di diversità (Figura 5) si può notare in tutte la presenza di un trend decrescente che interessa i valori di *D*. La carta costruita per la finestra temporale *w*=50, in particolare, mette in luce la presenza di un picco in corrispondenza dell'osservazione 250 nella parte finale dell'indice *D*. Per *w*=100 si ha un comportamento nella parte iniziale dei dati, seguito da un trend crescente.

Per *w*=150 si nota l'evidente trend decrescente che caratterizza tutti i valori dell'indice. Tale andamento potrebbe risultare rilevante per l'identificazione dei cambi nella struttura di relazione tra le variabili in esame.

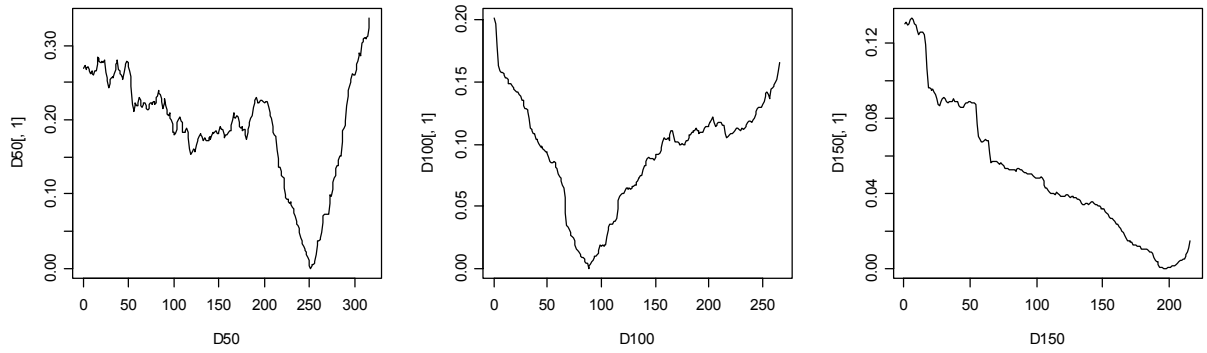


Figura 4.3: *Indice D calcolato per w=50,100,150*

✔ Indice A

Diversamente dall'indice *D*, questa statistica analizza la variazione subita da ciascuna componente principale. Sono riportate di seguito tutte le carte costruite per ogni componente principale in corrispondenza dei diversi valori della finestra temporale (w=50, 100,150).

La carta relativa alla prima componente si presenta notevolmente diversa da quella della seconda e terza variabile latente che presentano andamenti più simili. Questo conferma quanto visto in precedenza con la decomposizione della statistica T^2 : il contributo dato dalla prima componente principale è minimo rispetto a quello dato dalle altre due. Nelle carte disegnate per la prima variabile latente si notano alcuni picchi che interessano soprattutto la parte centrale dei valori dell'indice calcolato e che portano segnalare la presenza di fuori controllo. Questo comportamento rispecchia l'andamento dell'ozono e della radiazione solare che, durante il periodo estivo, presentano valori del 2002 tendenzialmenmte più elevati di quelli rilevati nel 2001. Le carte relative alla seconda e terza componente costruite per w=50 mettono in evidenza numerose osservazioni, all'inizio, in centro e alla fine del processo, che identificano la presenza di cambiamenti tra il 2002 ed il 2001, preso come riferimento. Calcolando l'indice *A* per le stesse componenti ma con una finestra temporale più ampia si ottiene un leggero effetto di lisciamento che porta ad osservare, per w=100, due diverse situazioni caratterizzate da un trend decrescente iniziale seguito da uno crescente. Per w =150 la statistica tende ad assumere valori prossimi allo zero tranne che per le primissime osservazioni che individuano una situazione anomala.

L'andamento dell'indice *A* relativo alle altre componenti principali risulta piuttosto irregolare a mano a mano che si passa dalla quarta variabile latente all'ultima; in particolare per w=50 e w=100 si nota la propensione dell'indice ad attestarsi su valori elevati. E' evidente, in tutte le carte l'effetto di lisciamento ottenuto utilizzando un valore della finestra temporale più grande.

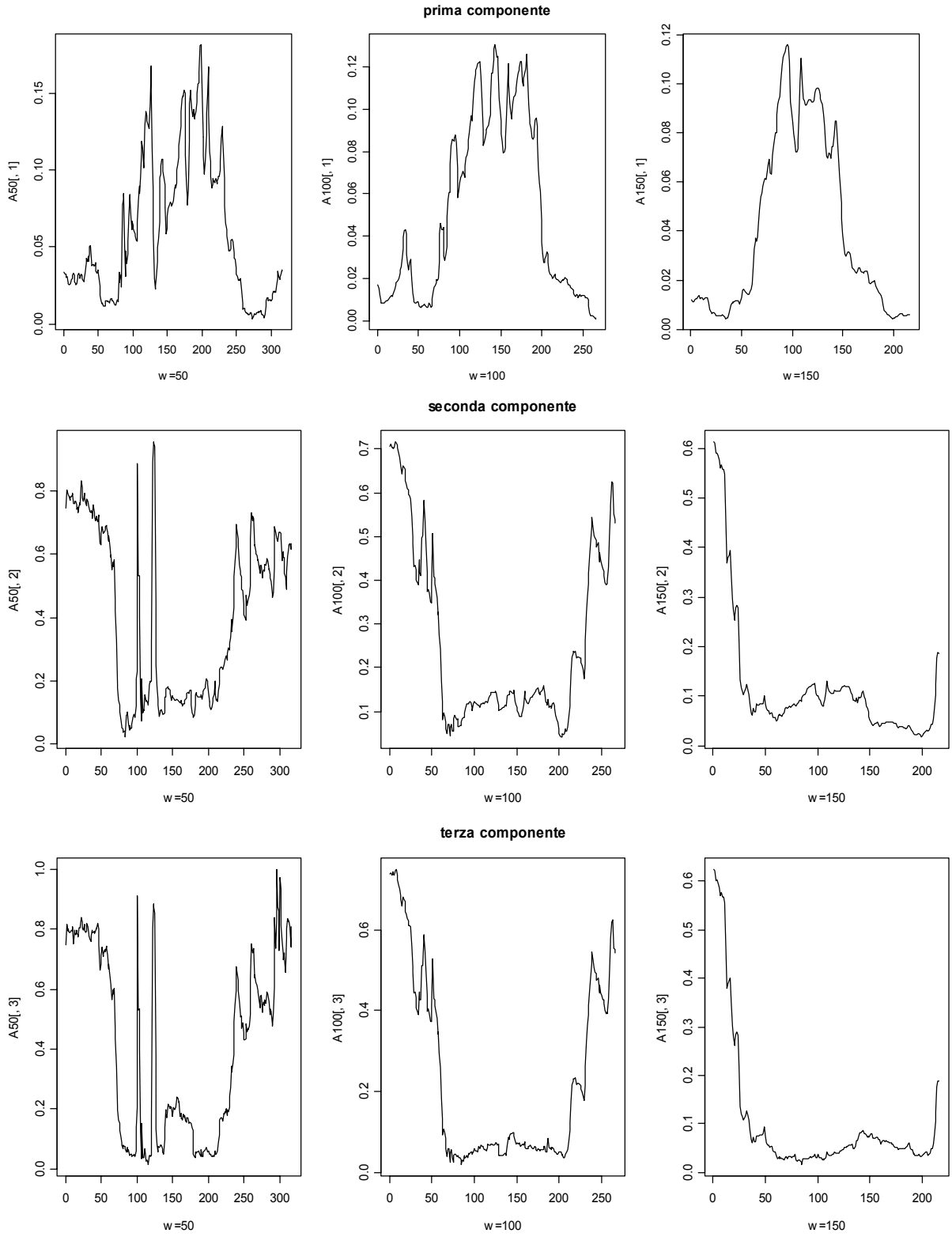


Figura 4.4: *Carte di controllo dell'indice A per w=50,100,150*

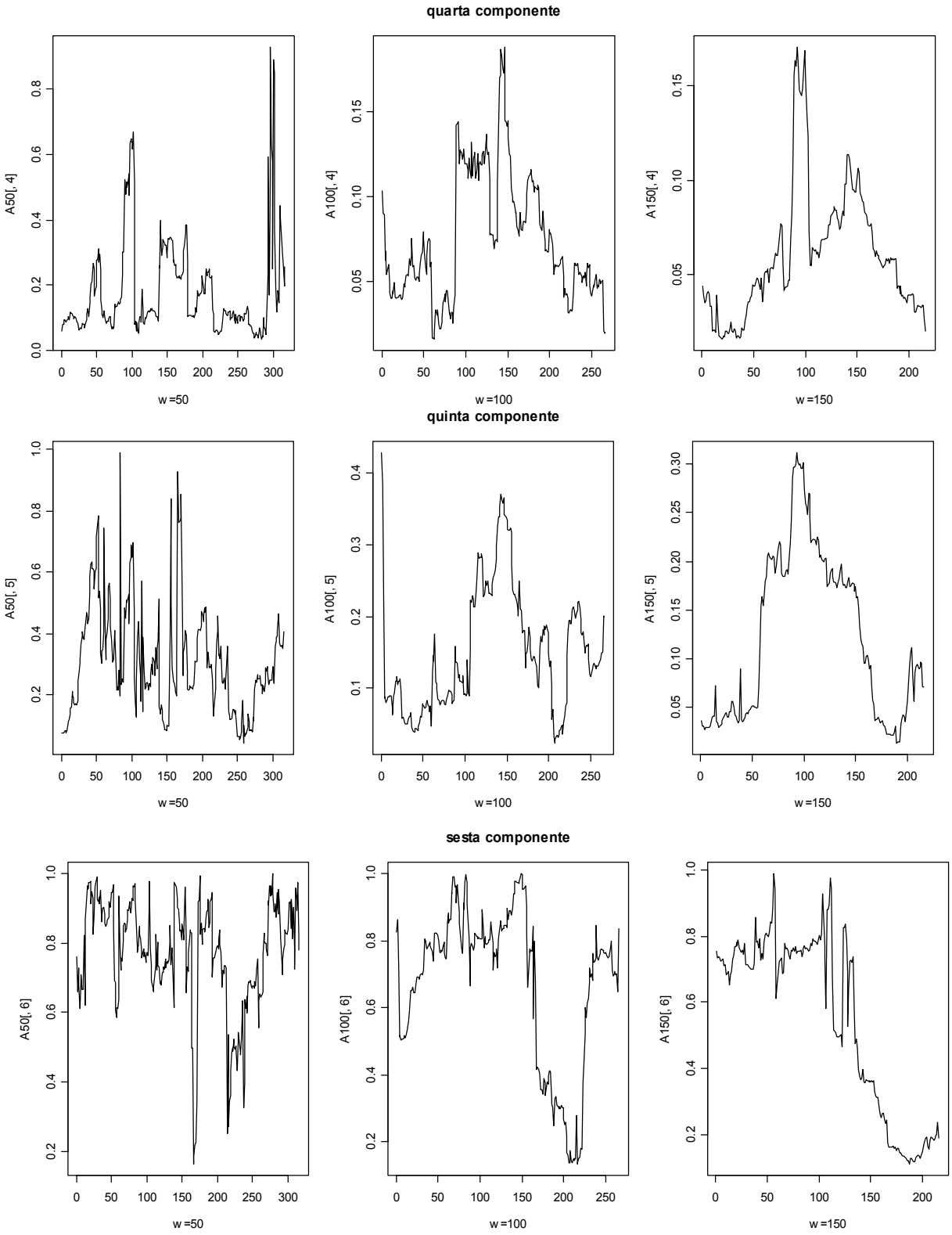


Figura 4.5: *Carte di controllo dell'indice A per w=50,100,150*

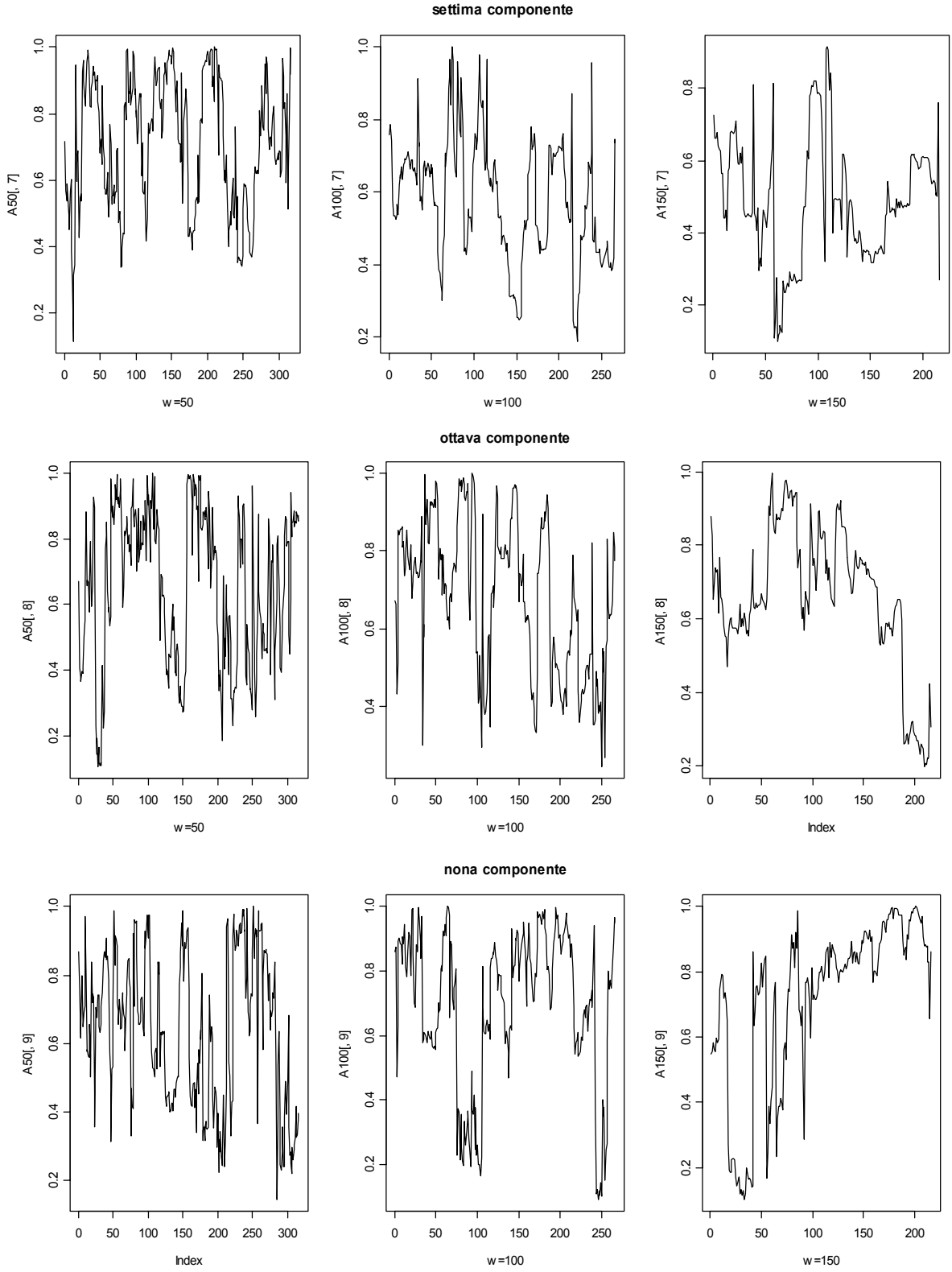


Figura 4.6: *Carte di controllo dell'indice A per w=50,100,150*

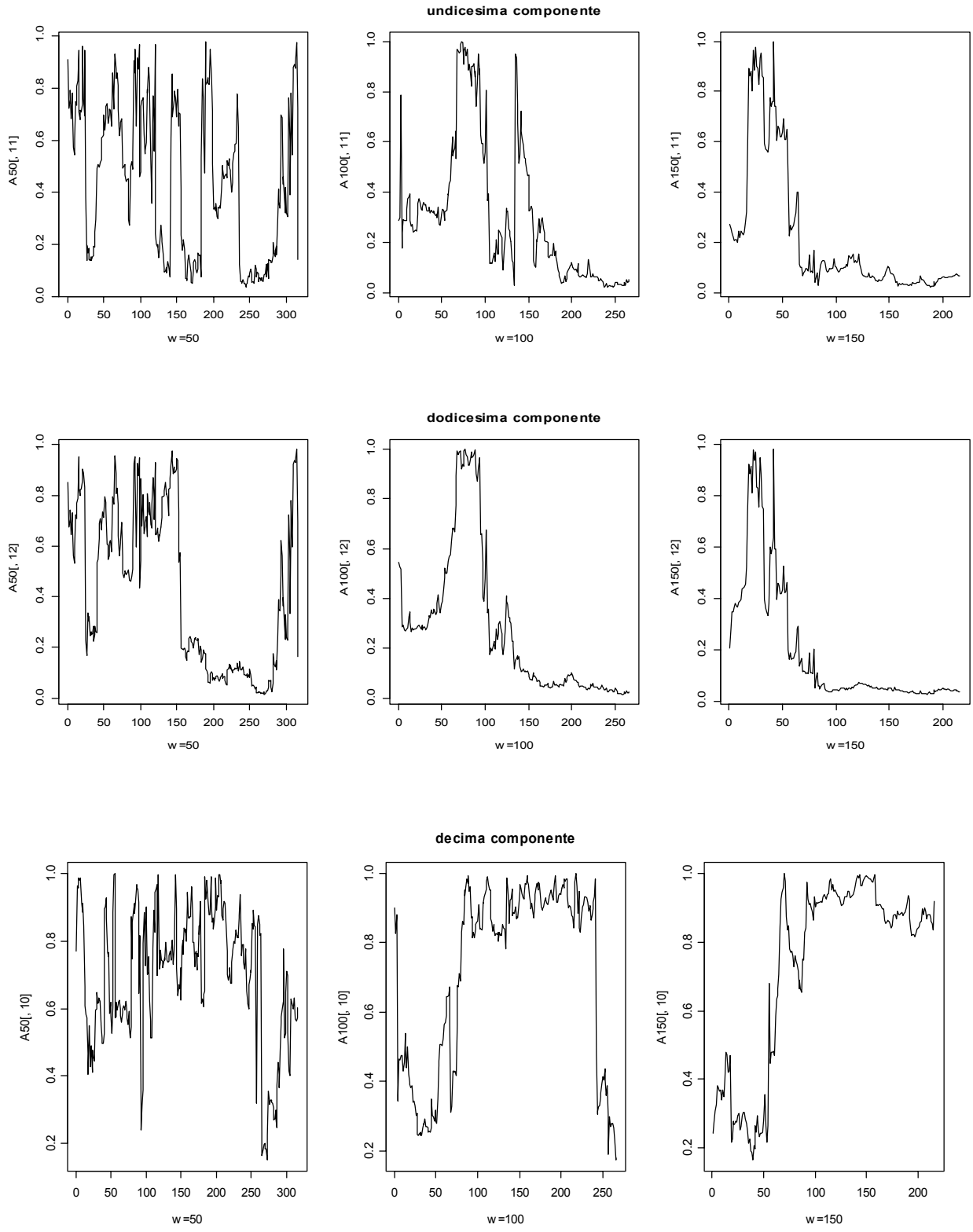


Figura 4.7: *Carte di controllo dell'indice A per w=50,100,150*

4.1.3.1 Confronto tra centraline

Studiando attentamente le componenti principali del modello "ozono-radiazione solare" (Tabella PCA c)) si possono individuare alcune variabili latenti in grado di descrivere l'influenza esercitata sui dati da ciascuna centralina. La sesta componente, ad esempio, aiuta a distinguere il contributo dato dalla centralina 34 rispetto a quello delle altre stazioni. I coefficienti delle variabili nella C34 risultano infatti di segno opposto e più elevati di quelli delle centraline 26 e 35. Per avere un sostegno ulteriore, nell'interpretare l'influenza della centralina 34 sul modello, si può prendere in considerazione la nona componente che distingue il contributo dato da questa stazione rispetto a quello congiunto delle altre due.

E' possibile fare un ragionamento analogo per la centralina 26: nella quinta componente infatti i coefficienti relativi all'ozono risultano di segno contrario rispetto a quelli relativi alle altre centraline ed il valore attribuito alla radiazione solare nella stessa centralina è più basso di quelli delle C34 e C35. Anche in questo caso la settima variabile latente aiuta a capire qual è l'influenza data dalla centralina 26 in confronto a quella data congiuntamente dalle altre due.

Analizzando la decima componente principale, è possibile intuire, infine, il contributo dato dalla stazione di rilevazione 35. I coefficienti relativi alle variabili misurate in questa centralina presentano infatti segno contrario rispetto a quelli calcolati per i parametri della C34 e della C26 mettendone in evidenza l'influenza esercitata sul modello.

Sulla base di questa analisi si può procedere ad una interpretazione più approfondita dei grafici dell'indice *A* per le componenti principali appena considerate; l'obiettivo è scoprire se esistono degli effetti di specifiche centraline sull'andamento crescente o decrescente dell'indice *A*.

Osservando i grafici della quinta componente (Figura 4.6), che descrive l'influsso sui dati della centralina 26, si nota la presenza di valori anomali nella parte centrale dell'anno. Ciò è indice del fatto che, in tale periodo, è intercorsa una variazione tra il 2002 ed il 2001, variazione dovuta principalmente all'influenza esercitata dalle osservazioni della centralina 26. Inoltre, dal momento che per questa componente è presente un trend decrescente nella seconda parte dell'indice, è

possibile affermare che il cambiamento va gradatamente attenuandosi nella parte finale dell'anno.

Nell'analisi descrittiva si è potuto osservare una differenziazione tra le serie di ogni variabile rilevate nelle tre diverse centraline. Alcuni valori della radiazione solare infatti presentavano per la centralina 34 un trend decrescente, nelle osservazioni finali, più accentuato rispetto alle altre stazioni. Per capire se la centralina 34 ha un effetto particolare sul modello studiato, come nel caso della C26, si procede ad interpretare l'andamento dell'indice A in corrispondenza della sesta componente. Si nota che i valori della statistica nella prima parte della serie, sono in prevalenza mediamente assestati attorno allo 0.8 per poi subire una brusca diminuzione circa tra le osservazioni 150 e 200 (per w=100). Questo comportamento è indice del fatto che il è dovuto in linea di massima all'influenza della centralina 34; il contributo di tale stazione di rilevazione tende in seguito ad attenuarsi e a conformarsi a quello delle altre centraline come testimonia la presenza del trend decrescente.

La centralina 35 sembra contribuire, principalmente nella parte finale dell'anno, ai cambiamenti rilevati tra il 2002 e il 2001. Osservando i grafici dell'indice A costruiti per la decima componente principale si nota un brusco salto nella media della statistica in corrispondenza circa dell'osservazione 70 (w=100); a partire da essa l'indice assume valori mediamente prossimi allo 0.9 segnalando la presenza di un forte cambiamento nelle correlazioni probabilmente imputabile alle variabili rilevate nella centralina 35.

Se si studia, infine, l'andamento della serie di A nei grafici dell'undicesima e della dodicesima componente, si nota che l'indice assume valori elevati relativamente alla prima parte dell'anno. Le componenti "undici" e "dodici" aiutano a distinguere il contributo fornito dalla radiazione solare misurata nella centralina 35 rispetto a quello dato, per la stessa variabile, in modo congiunto dalla C34 e dalla C26. Il picco osservabile nei grafici mostra come, nel periodo iniziale dell'anno, siano presenti cambiamenti di correlazione imputabili principalmente all'effetto delle stazioni 34 e 26. Mentre per la centralina 35 il valor medio della variabile "radiazione solare" resta invariato nel passaggio dal 2001 al 2002 (Tabelle D e F), per le stazioni C34 e C26 si registra una diminuzione del valore medio delle serie; tale abbassamento potrebbe essere la causa della variazione che si registra nei grafici delle ultime due componenti, dovuta congiuntamente alle stazioni 34 e 26.

4.1.4 La scelta dell'anno di riferimento

Dal momento che le sostanze inquinanti e le variabili meteorologiche difficilmente presentano le stesse caratteristiche col passare del tempo, la decisione dell'anno di riferimento sul quale basare lo studio e la costruzione delle carte di controllo diventa una scelta delicata: potrebbe capitare, infatti, il caso in cui un anno base sia adeguato per alcune variabili ma si riveli sbagliato per le altre. A questo proposito, si vogliono ora confrontare le carte *A* costruite per la terza e la quarta componente prendendo prima come anno base il 2001 (A_{2001}) e poi il 2002 (A_{2002}).

Nei grafici costruiti per la terza componente principale (Figura 4.8) non si notano sostanziali differenze nell'andamento dell'indice: per una finestra temporale pari a w=100 si notano due diverse situazioni di fuori controllo dovute alla presenza di un trend decrescente, nella parte iniziale della serie, e crescente in quella finale . Utilizzando un valore della finestra temporale più elevato, w =150, si vede che le

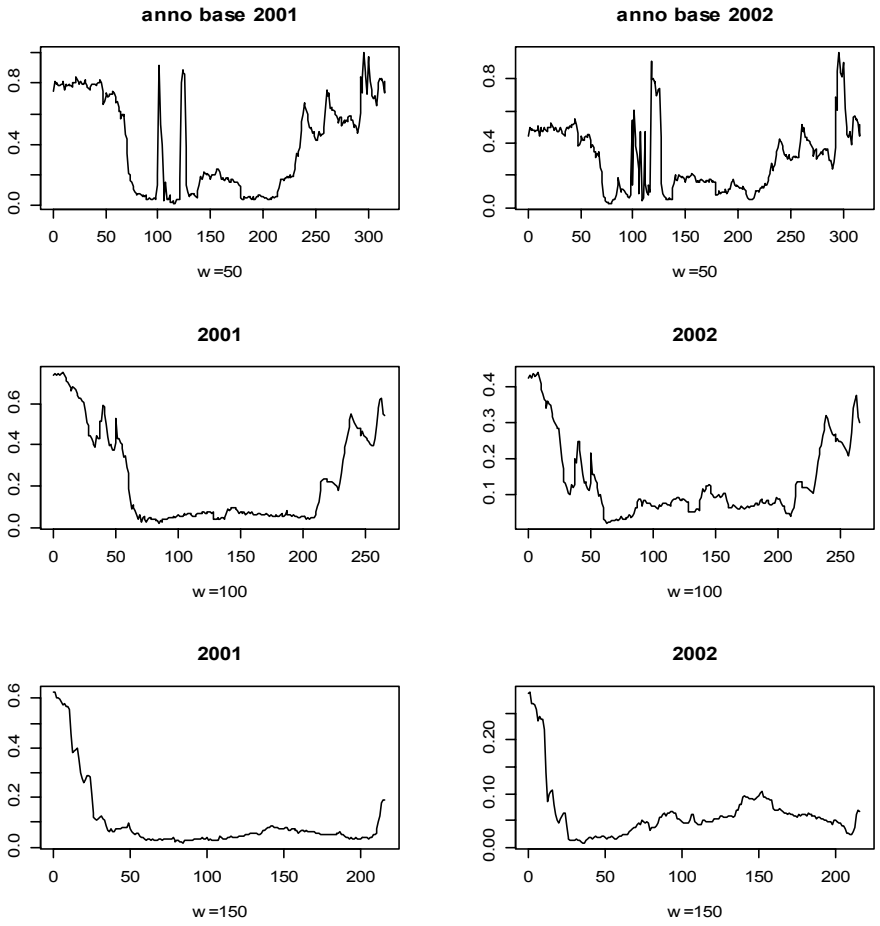


Figura 4.8: *confronto tra le carte A costruite per la 3° componente adottando due diversi anni di riferimento*

prime osservazioni della serie mettono in luce la presenza di un cambiamento. In seguito tuttavia, l'indice *A* assume valori molto prossimi allo zero. Anche per quanto riguarda la quarta componente (Figura 4.9), l'indice calcolato prendendo come riferimento il 2001 presenta un andamento simile a quello costruito utilizzano il 2002 come anno base. Per w=50 la serie di A_{2001} sembra avere un comportamento un po' meno irregolare rispetto a quella di A_{2002} presentando dei valori anomali sia in corrispondenza delle osservazioni centrali che in quelle finali. Cosiderando un valore di w leggermente più ampio (w=100) si possono notare delle osservazioni fuori controllo (tra la 90 ela 160) nell'indice calcolato in riferimento al 2001. L'esistenza di un cambiamento è riscontrabile anche nella carta che prende come anno base il 2002. In entrambe le carte è visibile un trend decrescente che interessa i dati a partire circa dalla metà della serie. Se si osservano infine le carte costruite per w=150 si nota la presenza di un picco nella serie di A_{2001} in corrispondenza dei valori centrali. Nella carta che fa riferimento al 2002 questo comportamento assume prevalentemente l'aspetto di un trend crescente che coinvolge i dati a partire dall'osservazione 70.

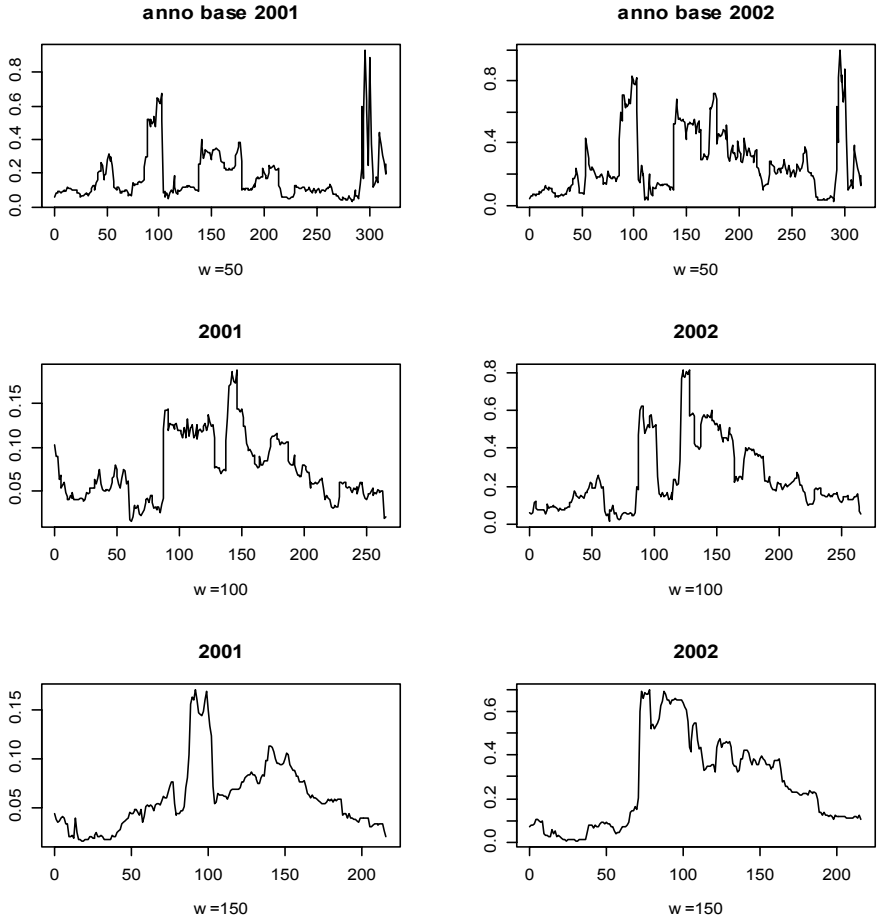


Figura 4.9: *confronto tra le carte A costruite per la 3° componente adottando due diversi anni di riferimento*

4.2 Carte di controllo univariate

Si effettua ora l'analisi univariata delle sei variabili considerate nel modello studiato allo scopo di individuare eventuali caratteristiche responsabili dei valori anomali che sono stati evidenziati dalle carte di controllo multivariate.

In questo caso, come in molti processi, l'ipotesi di indipendenza delle osservazioni viene violata, si hanno quindi dei dati che risultano autocorrelati e quindi dipendenti dalle misure rilevate in tempi precedenti. La presenza di autocorrelazione ha un profondo effetto sulle carte di controllo: si può infatti osservare un aumento della frequenza con la quale vengono lanciati segnali di falsi allarmi o, viceversa, l'assenza di segnalazioni quando invece sono presenti nella serie dei fuori controllo. Per questa ragione, l'autocorrelazione può causare diversi disturbi alle carte di controllo portando a fare delle conclusioni erronee sullo stato del processo. Diversi metodi per studiare carte di controllo per dati autocorrelati sono stati proposti da numerosi autori fra i quali Lian, Wu ed Ermer (1982); per eliminare le autocorrelazioni fra i dati è necessario ricavare un modello adeguato della serie temporale in esame e applicare la carta di controllo ai residui di questo modello.

I modelli lineari di serie storiche usati più frequentemente sono quelli autoregressivi integrati a media mobile, tipicamente detti modelli ARIMA(*p,d,q*)

$$\Phi_p(B)\nabla^k X_t = \Theta_q(B)\varepsilon_t$$

- Il polinomio autoregressivo di ordine *p* descrive il valore corrente del processo come combinazione lineare dei *p* valori precedenti ed è rappresentato da

$$\Phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)$$

- Il polinomio a media mobile di ordine *q* descrive il processo come linearmente dipendente da *q* precedenti valori di ε ed è rappresentato da

$$\Theta_q(B) = (1 - \vartheta_1 B - \vartheta_2 B^2 - \ldots - \vartheta_q B^p)$$

- Con il termine ε_t si vuole indicare il termine di errore del processo distribuito come

$$\varepsilon \sim N(0, \sigma_p^2)$$

- L'operatore indicato con la lettera *B* è detto *backword shift operator* o operatore ritardo ed è definito come

$$Bx_t = x_{t-1} \qquad B^p x_t = x_{t-p}$$

- L'operatore differenza di ordine *k* opera la "differenziazione "della serie, quando necessaria, sottrae cioè le osservazioni, le une alle altre

$$\nabla x_t = x_t - x_{t-1} \qquad \nabla^k x_t = x_t - x_{t-k}$$

I residui del modello stimato per i quali verranno costruite le carte di controllo sono calcolati secondo la relazione

$$r_t = x_t - \hat{x}_t$$

VARIABILE	MODELLO ARIMA
OZONO	
CAMS 26	$\hat{x}_t = 41.586 + 0.655x_{t-1} + \varepsilon_t$
CAMS 34	$\hat{x}_t = 39.438 + 0.739x_{t-1} - 0.129x_{t-2} + \varepsilon_t$
CAMS 35	$\hat{x}_t = 39.723 + 0.633x_{t-1} + \varepsilon_t$
RADIAZIONI SOLARI	
CAMS 26	$\hat{x}_t = 0.677 + 1.498x_{t-1} + 0.352x_{t-2} + \varepsilon_t$
CAMS 34	$\hat{x}_t = 0.704 + 1.5485x_{t-1} + 0.2934x_{t-2}\varepsilon_t$
CAMS 35	$\hat{x}_t = 0.625 + 1.735x_{t-1} + 0.397x_{t-2} + \varepsilon_t$

4.2.1 Carta delle escursioni mobili

Per studiare la variabilità dei residui si procede alla costruzione di una serie di misure dette "escursioni mobili" date dalla differenza tra ogni valore e quello precedente. Si calcola poi la variabilità media come

$$\overline{MR}(2) = \frac{\sum_{i=1}^{N} |x_i - x_{i-1}|}{N-1}$$

Sulla carta di controllo vengono riportate le misure delle escursioni mobili, il limite centrale corrispondente alla variabilità media ed i limiti superiore (UCL) ed inferiore (LCL) così ottenuti:

$$UCL = \overline{MR}(2) + 3 \cdot \frac{d_3}{d_2} \cdot \overline{MR}(2)$$

$$LCL = \overline{MR}(2) - 3 \cdot \frac{d_3}{d_2} \cdot \overline{MR}(2)$$

dove i valori delle costanti sono d_2=1.128 e d_3=0.853.

4.2.2 Carta per misure singole

Sulla carta vengono tracciati i residui e_t, il limite centrale dato dalla loro media e i limiti di controllo superiore ed inferiore dati da:

$$UCL = \bar{x} + 3 \cdot \frac{\overline{MR}(2)}{d_2}$$

$$LCL = \bar{x} - 3 \cdot \frac{\overline{MR}(2)}{d_2}$$

dove d_2=1.128 e MR(2) è la media delle escursioni mobili.

4.2.3 Carta per gli errori di previsione della carta EWMA

La carta di controllo EWMA (Roberts, 1959) può essere considerata come una media pesata di tutte le osservazioni passate e correnti:

$$z_t = \lambda \cdot x_t + (1 - \lambda) \cdot z$$

Essa si rivela molto robusta alla non normalità delle osservazioni e permette di considerare l'influenza esercitata dalla presenza di autocorrelazione tra le osservazioni passate.

La costante λ determina la memoria della statistica EWMA cioè la frazione di decadimento dei pesi e quindi l'aumento di informazione dei dati storici: minore è il valore della costante maggiore è l'influenza esercitata dalle osservazioni passate. E'

$$\min_{\lambda} \sum_{i=1}^{N} e_i^2$$

possibile stimare il parametro λ usando un metodo iterativo servendosi della procedura dei minimi quadrati:

$$e_t = x_t - z_{t-1}$$

Dopo aver stimato la costante più opportuna per la memoria della carta e aver calcolato il valore delle statistiche, si procede alla determinazione degli errori di previsione dati dalla differenza tra le osservazioni al tempo t e la statistica al tempo t-1

I valori ottenuti vengono riportati sulla carta di controllo assieme ai limiti superiore (LS), inferiore (LI) e centrale (LC):

$$\begin{aligned} LC &= 0 \\ LCL &= -3\hat{\sigma}_p \\ UCL &= +3\hat{\sigma}_p \end{aligned}$$

dove $\hat{\sigma}_p = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N}}$

4.2.4 Ozono

4.2.4.1 Centralina 26

Le carte di controllo univariate riferite alla variabile ozono rilevata nella centralina 26 mostrano alcuni residui che superano i limiti di controllo superiore sia nella carta delle medie che in quella delle escursioni. Tali valori (osservazioni 102, 163, 271 e 290) sono riscontrabili anche nelle carte di controllo multivariate; questo risultato conferma la forte influenza della variabile "ozono" nella determinazione dei valori anomali sull'intero modello multivariato.

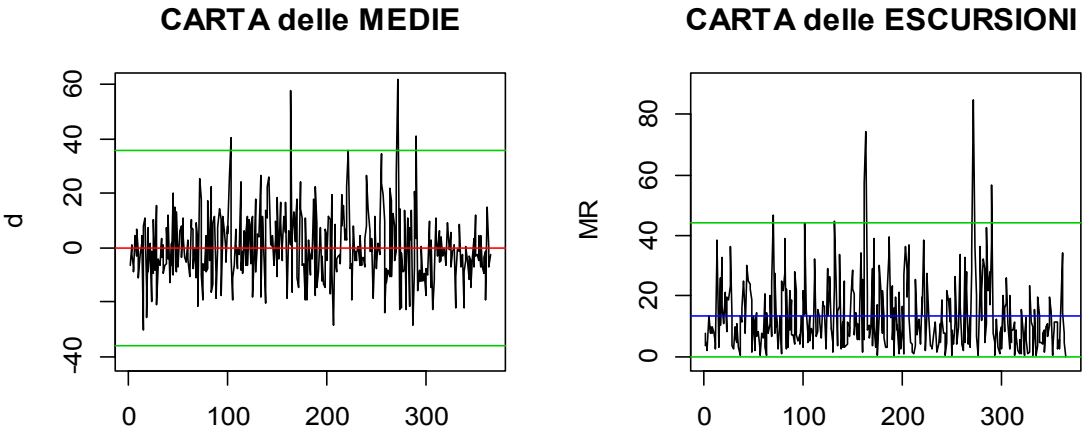


Figura 4.10: Carte di controllo per i residui dell'ozono misurato nella centralina 26

La carta degli errori di previsione basati sulla statistica EWMA è stata calcolata per un valore della costante λ pari a 0.02; la memoria della carta e quindi il peso delle osservazioni passate diventa pertanto molto elevato.

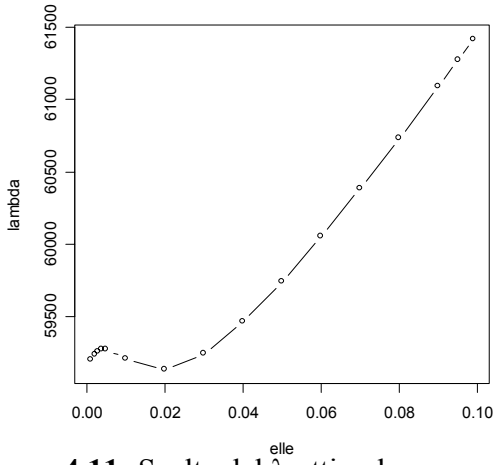


Figura 4.11: Scelta del λ ottimale

λ	$\sum_{i=1}^{N} e_i^2$
0.001	59198.98
0.002	59235.55
0.003	59258.48
0.004	59269.66
0.005	59271.31
0.010	59208.01
0.020	59129.97
0.030	59242.19
0.040	59462.73
0.050	59740.20
0.060	60051.34
0.070	60384.42
0.080	60732.70
0.090	61091.92
0.095	61274.71
0.099	61422.21

Le osservazioni risultate anomale in precedenza superano, ancora una volta, il limite di controllo superiore (Figura 4.12).

Il fatto che i fuori controllo determinati da tutte le carte univariate siano situati prevalentemente nella parte centrale delle osservazioni conferma quanto visto nel caso multivariato con l'analisi dell'indice A per la quinta componente: la centralina 26 risultava infatti influenzare proprio la parte centrale della serie.

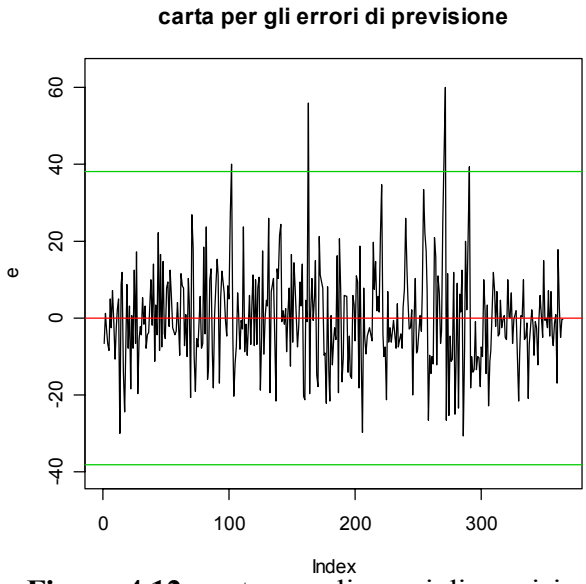


Figura 4.12: carta per gli errori di previsione

4.2.4.2 Centralina 34

Per la centralina 34 la variabile "ozono" presenta fuori controllo sia nella carta delle medie sia in quella delle escursioni mobili. Anche in questo caso i fuori controllo corrispondono per la maggior parte a quelli rilevati dall'analisi multivariata. L'ozono relativo alla centralina 34 risulta quindi molto influente e particolarmente responsabile nella determinazione di valori anomali quando si considera il modello multivariato che prende in considerazione questo inquinante e la radiazione solare nelle tre diverse stazioni. Tali considerazioni trovano conferma nella carta per gli errori di previsione per la statistica EWMA, nella quale si notano le stesse osservazioni fuori controllo. Una testimonianza del peso influente esercitato dalla variabile ora in esame nel modello multivariato si ottiene osservando i coefficienti della seconda componente principale, la maggior responsabile nelle determinazione dei valori anomali: la costante che esprime l'effetto dell'ozono misurato nella centralina 34 è infatti abbastanza grande.

$$Y_{2°cp} = -0.184x_{oz26t-1} - 0.451x_{oz34t-1} - 0.286x_{oz35t-1} + 0.240x_{sr26t-1} + 0.249x_{sr34t-1} + 0.260x_{sr35t-1} - 0.141x_{oz26t} - 0.403x_{oz34t} - 0.228x_{oz35t} + 0.324x_{sr26t} + 0.304x_{sr34t} + 0.294x_{sr35t}$$

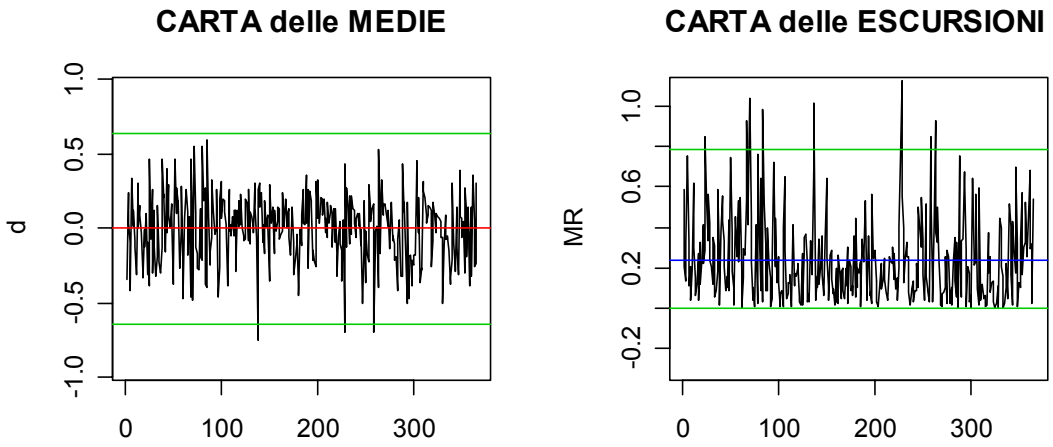
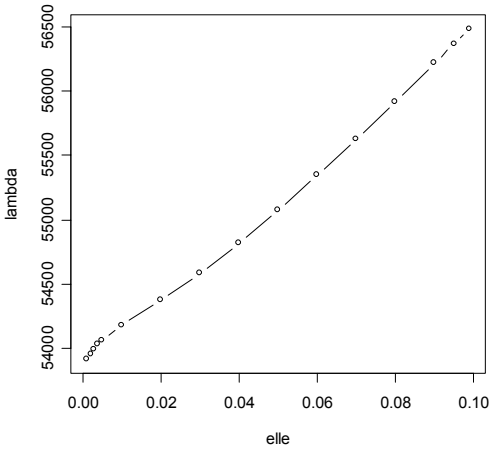


Figura 4.13: Carte di controllo per l'ozono rilevato dalla centralina 34



λ	$\sum_{i=1}^{N} e_i^2$
0.001	53913.89
0.002	53958.69
0.003	53998.29
0.004	54033.40
0.005	54064.69
0.010	54183.70
0.020	54372.57
0.030	54582.98
0.040	54819.03
0.050	55074.57
0.060	55345.09
0.070	55627.08
0.080	55917.72
0.090	56214.85
0.095	56365.33
0.099	56486.50

Figura 4.14: Scelta del λ ottimale

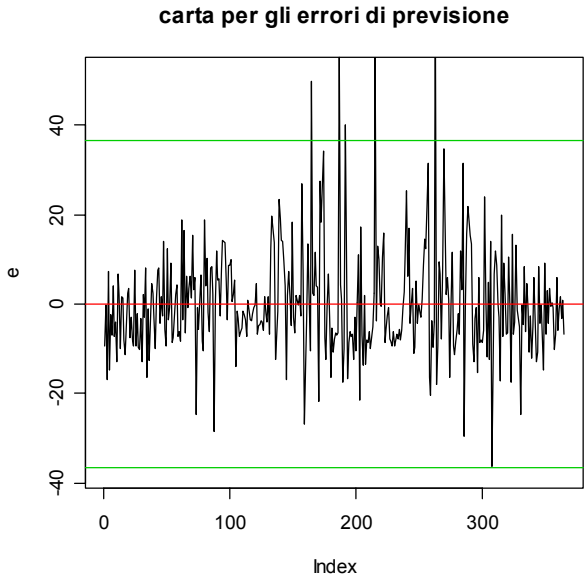


Figura 4.15: Carta per gli errori di previsione

4.2.4.3 Centralina 35

Le considerazioni fatte in precedenza riguardo all'ozono misurato nelle centraline 26 e 34 posso essere estese anche alla centralina 35; le carte presentano fuori controllo più o meno all'altezza delle osservazioni risultate oltre il limite superiore sia negli schemi univariati che in quelli multivariati. Appare ormai certa, quindi, la presenza di cause specifiche di disturbo che hanno interessato, in particolari, giorni tutte le stazioni di rilevazione oggetto di studio. E' interessante notare che i valori fuori controllo si concentrano soprattutto nella parte centrale e finale della serie. L'influenza di questi valori anomali era stata già riscontrata nell'analisi multivariata dell'indice A calcolato per la decima componente.

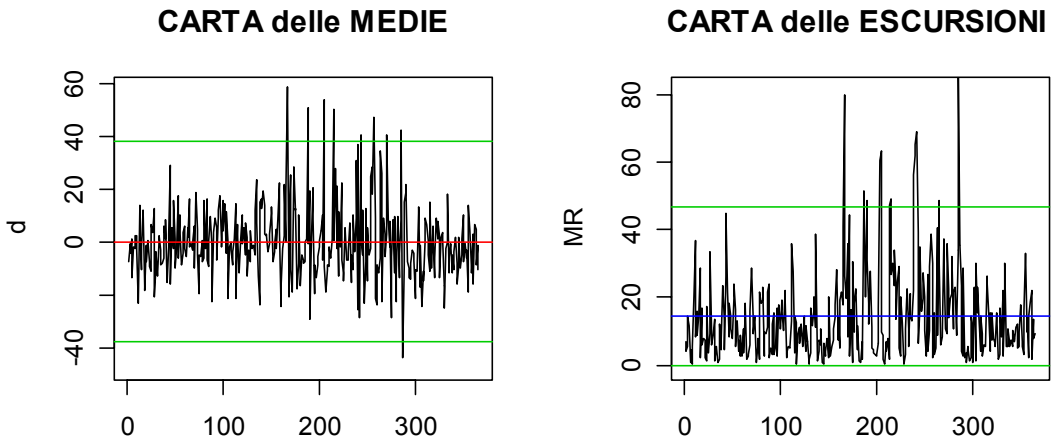
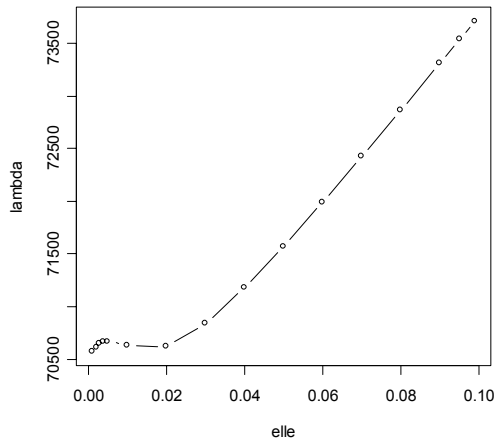


Figura 4.16: Carte di controllo per l'ozono rilevato dalla centralina 35



λ	$\sum_{i=1}^{N} e_i^2$
0.001	70567.96
0.002	70614.06
0.003	70645.35
0.004	70663.82
0.005	70671.81
0.010	70632.68
0.020	70619.66
0.030	70837.43
0.040	71177.82
0.050	71572.58
0.060	71993.05
0.070	72426.79
0.080	72867.87
0.090	73313.13
0.095	73536.73
0.099	73715.93

Figura 4.17: Scelta del λ ottimale

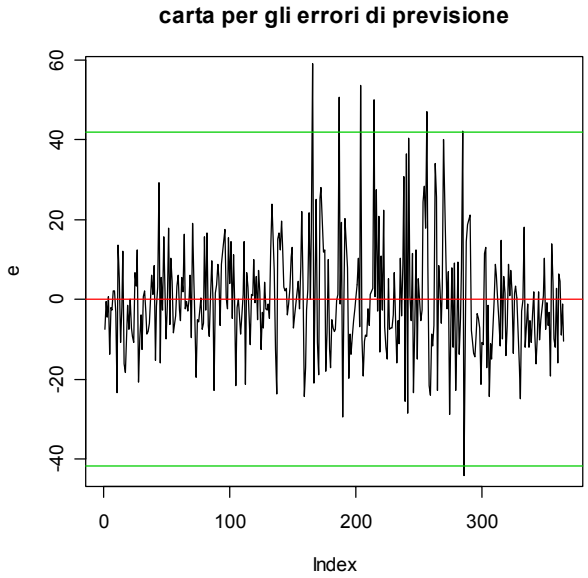


Figura 4.18: Carta per gli errori di previsione

Questa variabile latente, in grado di spiegare l'influenza della centralina 35 sul modello dei dati, aveva messo in luce un cambio nella struttura di correlazione proprio nella seconda parte della serie dell'indice.

4.2.5 Radiazione solare

4.2.5.1 Centralina26

Cerchiamo ora di analizzare il comportamento delle radiazione solare rilevata dalla stazione 26 servendoci delle carte per escursioni e misure singole, e della carta EWMA per gli errori di previsione. Questi validi strumenti che permettono di identificare cambiamenti significativi all'interno del processo mettono in luce la presenza di tre valori fuori controllo nella carta delle medie e di diversi fuori controllo nella carta delle escursioni mobili. A parte l'osservazione 258, gli altri valori anomali non corrispondono a quelli ottenuti dall'analisi del modello multivariato, né dall'analisi dell'ozono.

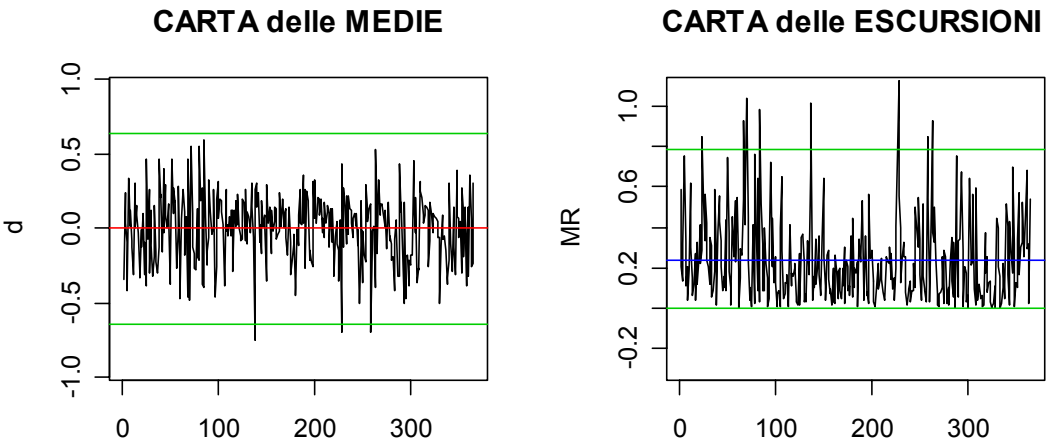


Figura 4.19: Carte per misure singole per le radiazioni solari della centralina 26

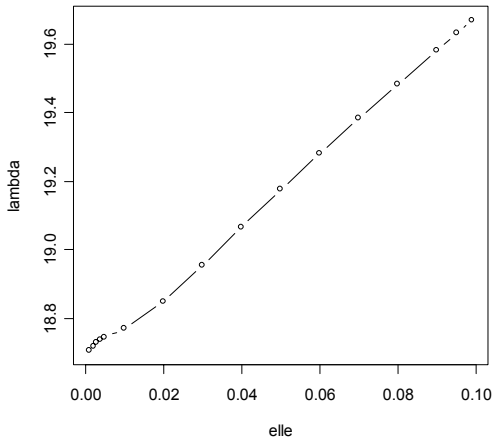


Figura 4.20: Scelta del λ ottimale

λ	$\sum_{i=1}^{N} e_i^2$
0.001	18.70422
0.002	18.71757
0.003	18.72804
0.004	18.73633
0.005	18.74311
0.010	18.77034
0.020	18.84905
0.030	18.95453
0.040	19.06585
0.050	19.17532
0.060	19.28121
0.070	19.38379
0.080	19.48399
0.090	19.58279
0.095	19.63195
0.099	19.67125

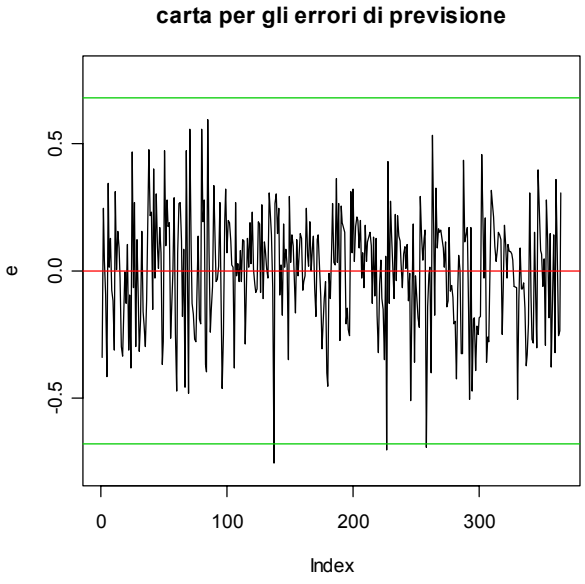
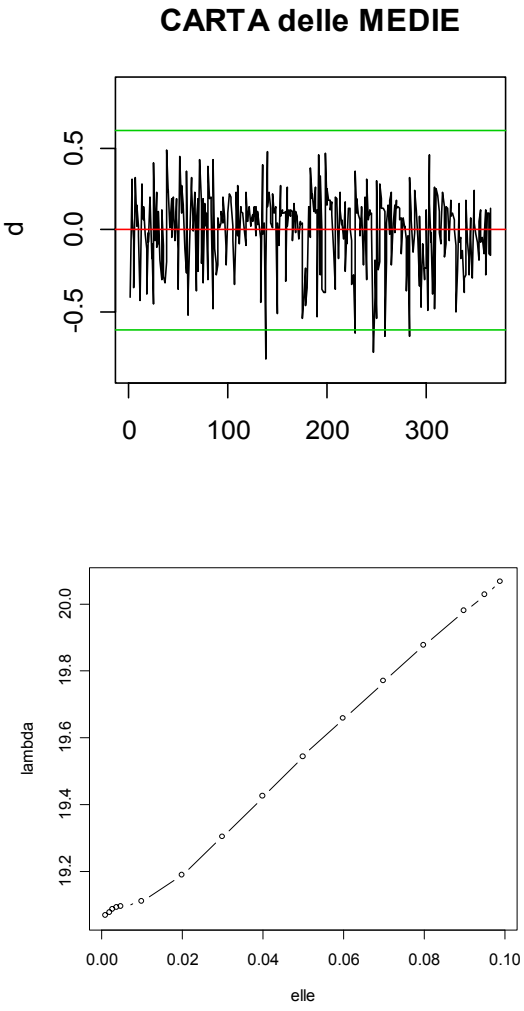


Figura 4.21: Carta per gli errori di previsione

4.2.5.2 Centralina 34

Nella figura 4.22 sono riportate le carte di controllo univariate relative ai residui del modello ARIMA stimato per la radiazione solare della centralina 34. Grazie alla correlazione con la stessa variabile rilevata nelle altre centraline, si possono individuare dei valori anomali che superano il limite di controllo inferiore. Gli *outliers*, evidenziati dalle carte, corrispondono infatti alle osservazioni 227 e 258 ovvero a due dei tre fuori controllo riscontrabili nelle carte relative alla centralina 26. E' importante osservare, inoltre, che tali valori anomali rispecchiano l'aumento dei valori della radiazione solare, durante il periodo centrale dell'anno, che era stato individuato anche nell'analisi descrittiva.



λ	$\sum_{i=1}^{N} e_i^2$
0.001	19.06667
0.002	19.07817
0.003	19.08592
0.004	19.09109
0.005	19.09470
0.010	19.11085
0.020	19.18902
0.030	19.30315
0.040	19.42450
0.050	19.54368
0.060	19.65828
0.070	19.76821
0.080	19.87411
0.090	19.97679
0.095	20.02718
0.099	20.06712

Figura 4.22: Carte per misure singole e scelta del λ ottimale per la radiazione solare della centralina 34

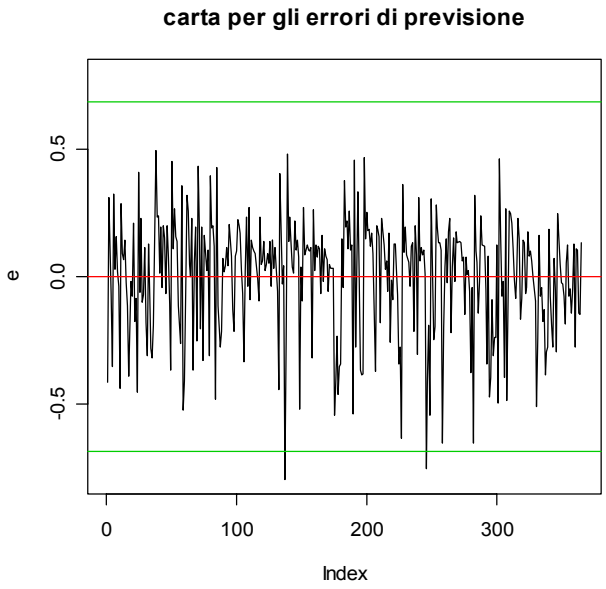


Figura 4.23: Carta per gli errori di previsione

4.2.5.3 Centralina 35

Vengono presentate, in ultimo, le carte di controllo per la radiazione solare rilevate all'interno della stazione 35; questa variabile è altamente correlata con quella misurata nelle altre centraline e presenta un andamento dei residui molto simile ad esse. Le differenze tra i valori originali e quelli del modello di serie storiche stimato risultano quasi tutte entro i limiti di controllo, fatta eccezione per quelle corrispondenti alle osservazioni 137, 227 e 258 che non rispettano il limite inferiore.

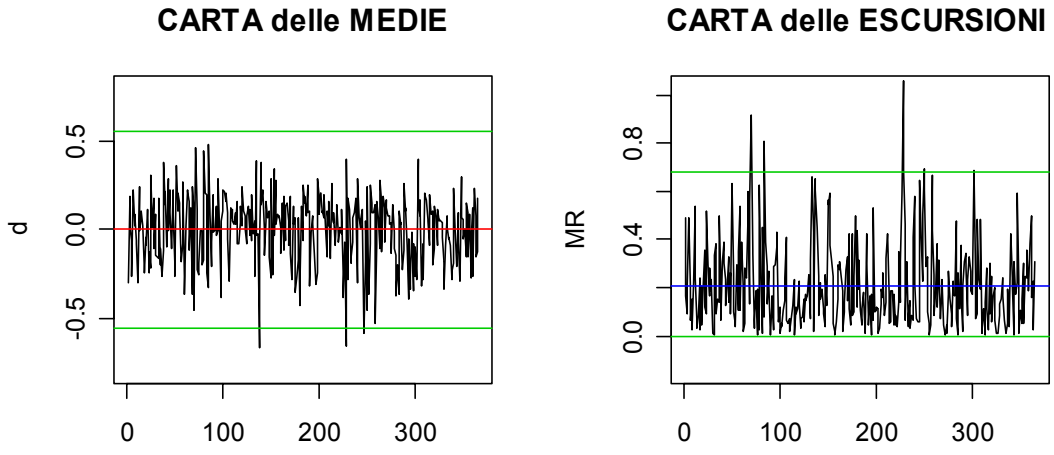
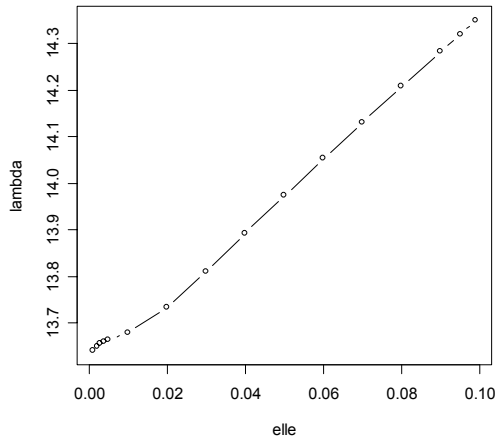


Figura 4.24: Carte per misure singole delle radiazioni solari per la centralina 35

Osservando attentamente la carta delle escursioni ci si accorge che la serie della radiazione solare misurata nella C35 presenta un andamento meno variabile rispetto a quelle misurate nelle altre centraline. Questa distinzione era stata introdottaanche dall'analisi dell'indica *A*: le ultime due componenti principali del modello permettono infatti di distinguere l'influenza della stazione 35 da quella delle centraline 26 e 34.



λ	$\sum_{i=1}^{N} e_i^2$
0.001	13.63961
0.002	13.64848
0.003	13.65492
0.004	13.65964
0.005	13.66325
0.010	13.67837
0.020	13.73336
0.030	13.81036
0.040	13.89260
0.050	13.97426
0.060	14.05388
0.070	14.13153
0.080	14.20773
0.090	14.28310
0.095	14.32064
0.099	14.35065

Figura 4.25: Scelta del λ ottimale

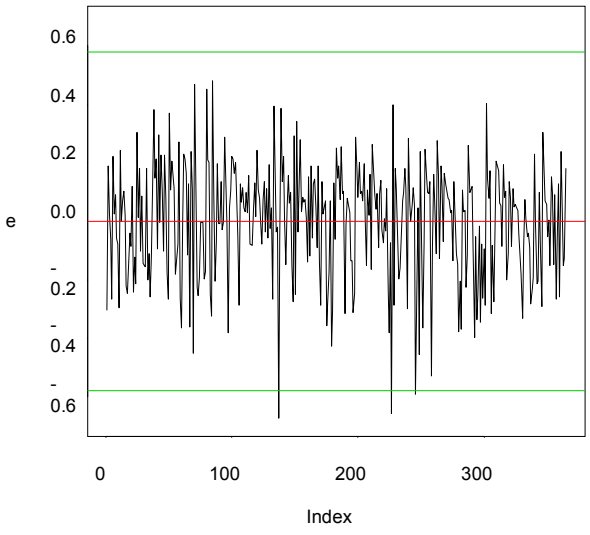


Figura 4.26: Carta degli errori di previsione

Considerando tutti i risultati ottenuti dalle carte di controllo univariate, si possono riscontrare diverse affinità con l'analisi multivariata nella determinazione dei valori anomali. In particolare si è visto come l'andamento crescente delle

variabili in prossimità del periodo centrale dell'anno, che era stato riscontrato nell'analisi descrittiva, abbia in seguito trovato conferme e spiegazioni sia per merito dell'analisi multivariata che di quella univariata. Il contributo dato da queste metodologie ha inoltre consentito di attribuire a ciascuna centralina la propria responsabilità nella determinazione di fuori controllo.

4.3 Nuovo modello

4.3.1 Carte tradizionali

Nel modello "ozono-radiazione solare" si è proceduto allo studio di due parametri tra loro positivamente correlati. Si vuole ora studiare brevemente un modello del tipo "ozono-velocità del vento-radiazione solare" utilizzando sempre un approccio dinamico basato su una relazione temporale fra i dati di primo ordine. Lo scopo è osservare come le carte tradizionali e i metodi dinamici reagiscono all'introduzione nel modello di una variabile negativamente correlata con le altre.

L'analisi delle componenti principali del modello porta ai risultati espressi dalla tabella L. Utilizziamo come criterio per la scelta delle componenti principali quello che prevede di prendere in considerazione le variabili latenti con varianza superiore ad uno. Il numero di componenti scelte per descrivere il modello è dunque quattro. La prima variabile latente distingue il contributo positivo dato dalla velocità del vento da quello negativo di ozono e radiazione solare; la seconda componente prende in considerazione l'effetto congiunto del vento e della radiazione solare mentre la terza contrappone la variabile "radiazione solare" alla velocità del vento e all'ozono. La quarta componente principale infine, distingue il contributo positivo dell'inquinante da quello negativo delle altre due variabili meteorologiche.

La carta di controllo per la statistica T^2, costruita adottando come insieme di riferimento l'anno 2001, presenta complessivamente diciassette valori fuori controllo mentre nella carta per i residui Q è possibile notare numerosi valori anomali (Figura 4.27). Procedendo alla decomposizione della statistica T^2 per le quattro componenti principali utilizzate nella stima del modello, si nota che la maggior parte dei valori anomali sono attribuibili congiuntamente alla prima e alla quarta variabile latente. Queste componenti sono complessivamente responsabili di ben 11 valori anomali; la

scarsa influenza delle altre variabili latenti nella determinazione dei fuori controllo è evidente dall'osservazione delle carte T^2 costruite per ciascuna di esse (Figura4.28).

Tabella L

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
OZ26t1	-0.293		-0.228	-0.104	-0.131		0.502	-0.230	
OZ34t1	-0.242		-0.436		-0.107		0.158	0.428	-0.195
OZ35t1	-0.289		-0.323				0.323	0.183	
WS26t1	0.155	-0.391		0.266	-0.312	-0.217	-0.100	0.275	0.259
WS34t1	0.120	-0.330	0.175	0.330	-0.265	0.479		-0.107	-0.629
WS35t1	0.112	-0.431		0.158	-0.358	-0.356	0.162	-0.190	0.248
SR26t1	-0.276	-0.113	0.147	-0.347	-0.189		-0.236	0.106	
SR34t1	-0.255	-0.170	0.128	-0.365	-0.243		-0.197		
SR35t1	-0.285	-0.138	0.107	-0.321	-0.249		-0.233		
OZ26	-0.303			0.229				-0.565	0.228
OZ34	-0.241		-0.306	0.360			-0.436	0.146	
OZ35	-0.299		-0.147	0.286			-0.349	-0.170	
WS26	0.157	-0.362	-0.249	-0.160	0.348	-0.158	-0.172	0.211	
WS34	0.150	-0.343	-0.133	-0.170	0.191	0.696			0.485
WS35		-0.385	-0.270	-0.229	0.346	-0.219		-0.322	-0.360
SR26	-0.258	-0.149	0.349	0.136	0.239			0.211	
SR34	-0.258	-0.191	0.290		0.281	-0.126	0.235		
SR35	-0.268	-0.175	0.301	0.146	0.281			0.122	

	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16	Comp.17	Comp.18
OZ26t1	0.205	0.520		-0.341	-0.260			0.126	
OZ34t1	-0.137	-0.373	0.155		-0.259	0.308	0.371		
OZ35t1		-0.132		0.351	0.492	-0.370	-0.329	-0.129	
WS26t1	0.313		0.410	-0.201		-0.297		0.115	-0.200
WS34t1					0.104				
WS35t1	-0.324		-0.291	0.282		0.253		-0.129	0.193
SR26t1	0.297	0.272	0.300	0.243		0.219	0.158	-0.342	0.399
SR34t1	-0.244	-0.349	-0.247	-0.474		-0.365	0.140		0.179
SR35t1				0.186		0.259	-0.304	0.413	-0.544
OZ26	0.354	-0.468	0.170			0.238	-0.114		
OZ34	-0.241	0.258		-0.237	-0.154		-0.484	-0.221	0.106
OZ35		0.239	-0.226	0.251	0.211	-0.198	0.581	0.150	-0.146
WS26	0.395		-0.439	-0.178	0.170	0.249		0.180	0.148
WS34	-0.172								
WS35	-0.174		0.356	0.131	-0.172	-0.240		-0.175	-0.148
SR26	0.239	-0.116	-0.307	0.116	-0.411	-0.113		-0.436	-0.328
SR34	-0.322	0.136	0.193	-0.307	0.502	0.309		-0.127	-0.153
SR35	-0.101		0.124	0.187	-0.211	-0.196	-0.136	0.548	0.468

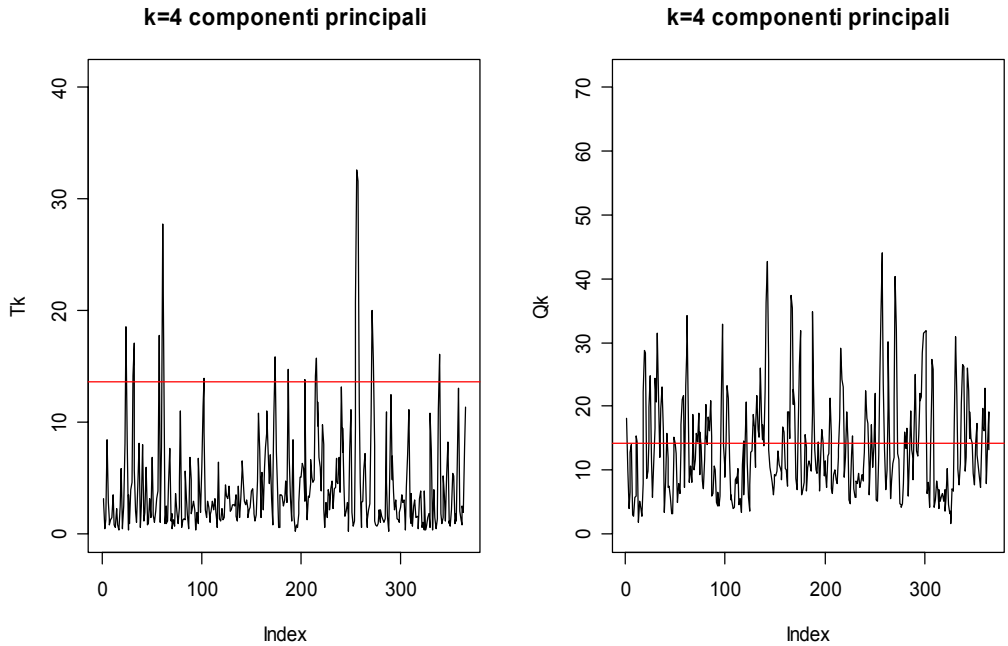


Figura 4.27: *Carte T^2 e Q costruite per quattro componenti principali*

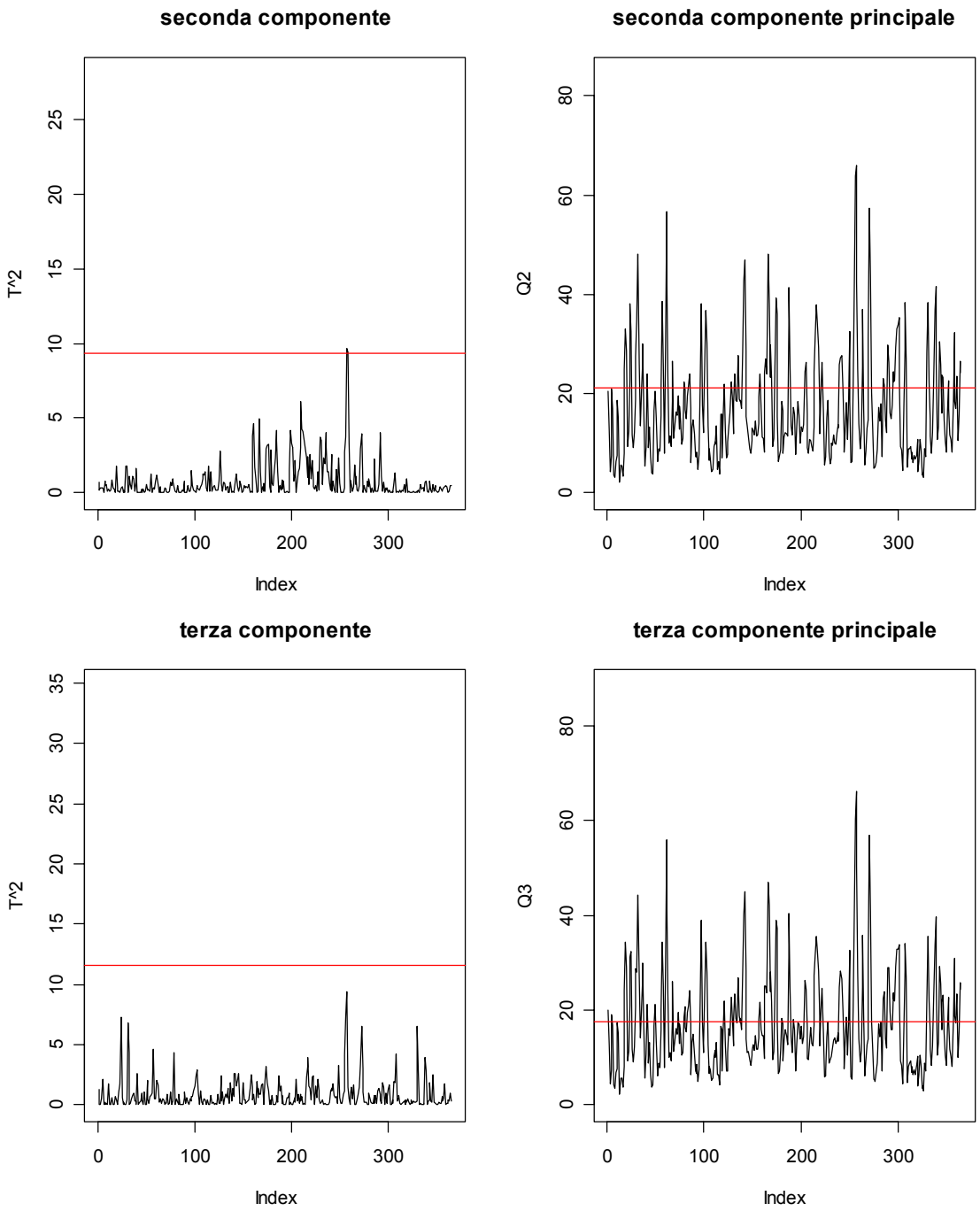


Figura 4.28: *Carte T^2 e Q costruite per la seconda e per la terza componente principale*

4.3.2 Indici D e A

✔ *Indice D*

Se si osservano le carte dell'indice di diversità costruite per un valore della finestra temporale *w=50* e *w*=100 si nota la presenza di un leggero trend crescente

che interessa la prima parte della serie segnalando la presenza di un cambiamento nelle condizioni operative del processo. La parte finale dei valori dell'indice D è invece caratterizzata da un trend decrescente che testimonia come il divario tra la distribuzione dell'insieme di riferimento e quella delle variabili del modello vada diminuendo abbastanza velocemente. La carta D costruita per *w*=150, a causa dell'effetto lisciamento dovuto all'aumento della finestra temporale, presenta un unico trend decrescente che interessa tutte le osservazioni.

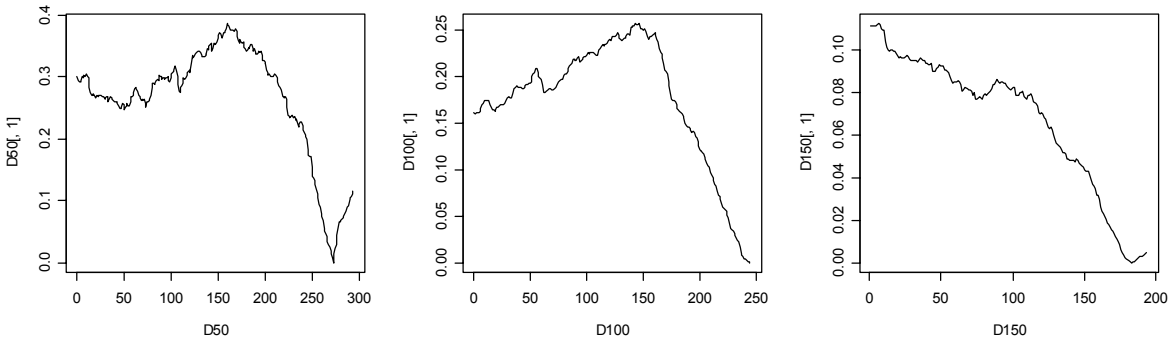


Figura 4.29: *Carte dell'indice D*

✔ *Indice A*

Vengono ore riportate le carte di controllo dell'indice A costruite per ogni componente principale utilizzando come valore della finestra temporale *w*=50,100,150. La statistica *A* misura il contributo dato da ogni variabile latente alla realizzazione di situazioni anomale. La carta costruita per la prima componente presenta, per *w*=100 e *w*=150, un trend crescente che interessa la prima metà delle osservazioni e segnala l'esistenza di numerosi valori anomali. Dato che questa variabile contrappone il contributo dato dall'ozono e dalla radiazione solare all'influenza esercitata dalla variabile" velocità del vento", l'andamento della serie dell'indice testimonia che le due variabili positivamente correlate sono responsabili di cambiamenti nelle condizioni operative nella seconda metà dei dati. Per quanto riguarda la seconda componente invece si nota un unico trend negativo che coinvolge tutti i valori dell'indice. La terza componente presenta per *w*=100 un picco tra le osservazioni 50 e 150; dato che questa variabile latente spiega l'influenza sui dati del modello esercitata dalla radiazione solare, il cambio nelle condizioni operative del processo segnalato da tutte le carte può attribuirsi a tale variabile. La quarta

componente, infine, presenta un trend crescente che porta ad individuare cambiamenti nelle condizioni operative che interessano la parte finale dei dati.

Segnaliamo anche in questo caso alcune componenti che più di altre identificano l'effetto che ciascuna centralina esercita sul modello. A questo proposito si nota che la sesta e la nona componente testimoniano l'influenza che la centralina 34 opera sui dati contrapponendola al contributo dato dalle stazioni C26 e C35. Le carte *A* per queste componenti presentano un andamento crescente; questo è indice del fatto che i cambiamenti nelle condizioni operative del processo che interessano la parte finale delle osservazioni possono essere in parte attribuiti all'influenza dei dati misurati dalla centralina 34.

L'effetto dovuto alla centralina 26 può essere invece individuato osservando le componenti dalla dieci alla dodici, con particolare attenzione verso quest'ultima. La serie dell'indice della dodicesima variabile latente presenta un andamento abbastanza irregolare che segue un trend decrescente nella parte iniziale e crescente in quella finale. I cambiamenti della struttura di correlazione per i dati corrispondenti alla parte non centrale dell'anno possono essere in parte attribuiti proprio all'influenza esercitata dalle centralina 26.

Le ultime tre componenti del modello sono quelle in grado di descrivere l'effetto dovuto alla centralina 35. La diciottesima variabile latente è quella che meglio spiega l'influenza esercitata dai dati rilevati in questa stazione; è possibile notare l'esistenza di un picco che porta a lanciare una segnalazione di fuori controllo per le prime cento osservazioni (w=150).

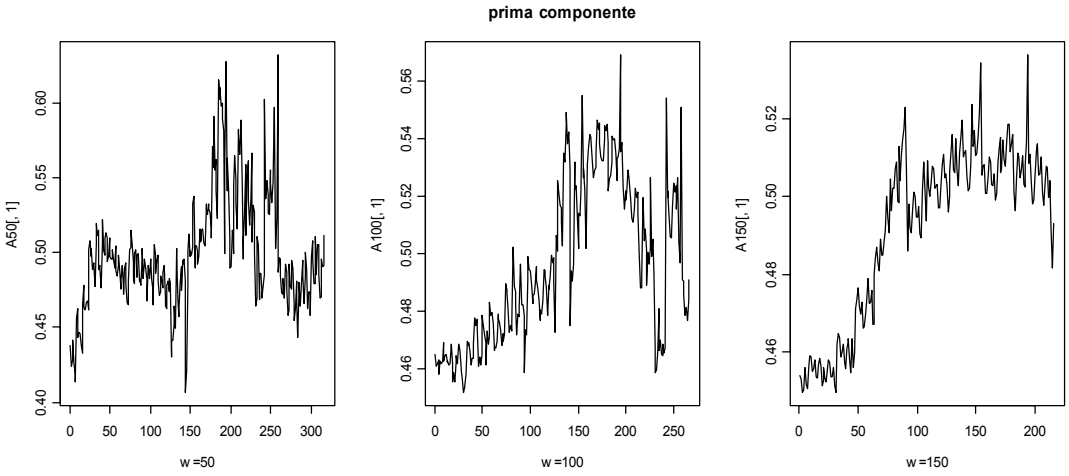


Figura 4.30: *Carta A per w=50,100,150*

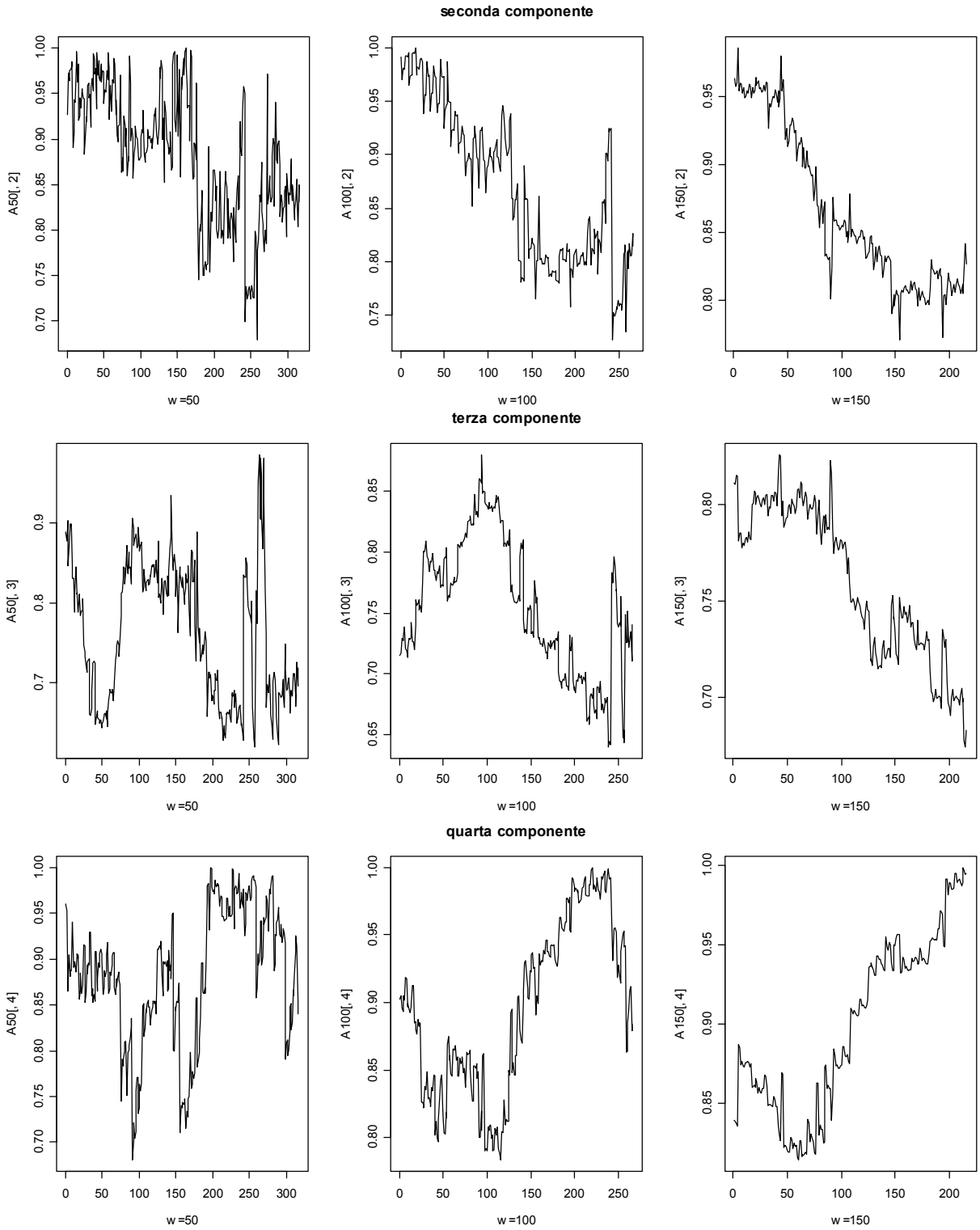


Figura 4.31: *Carta A per w=50,100,150*

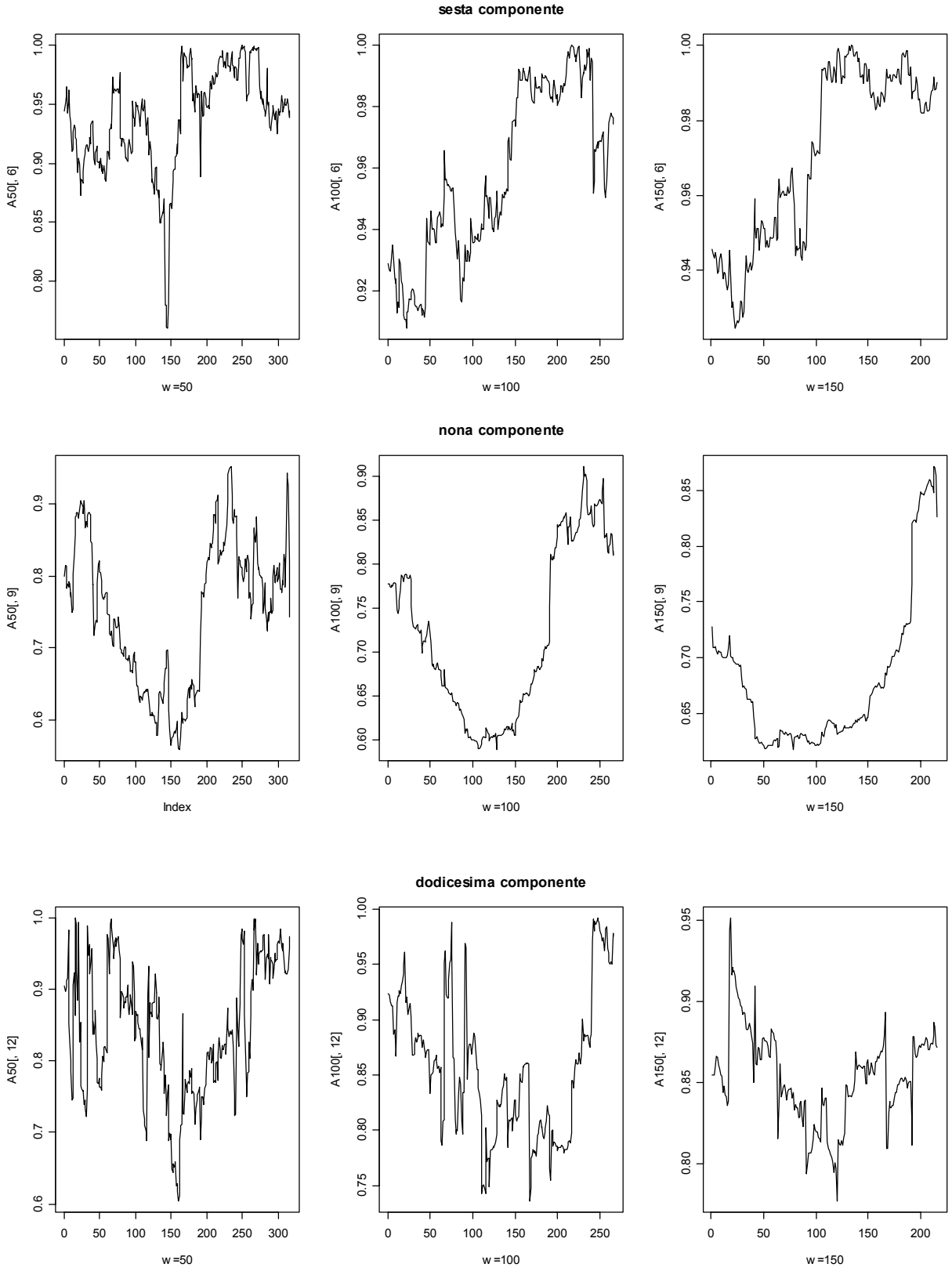


Figura 4.32: *Carta A per w=50,100,150*

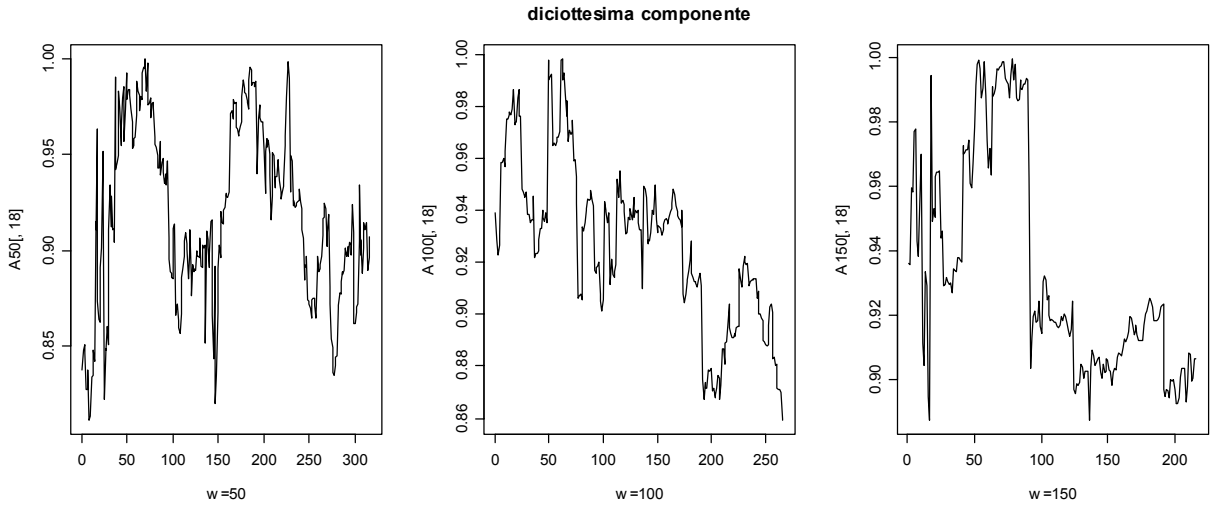


Figura 4.33: *Carta A per w=50,100,150*

4.4 Carte univariate

Vengono ora presentate le carte univariate per la variabile "velocità del vento" misurata nelle tre centraline. Questa analisi viene effettuata utilizzando i residui dei modelli ARIMA stimati per le tre serie della variabile in questione. I modelli di serie storiche identificati sono riportati nella tabella sottostante.

VARIABILE	MODELLO ARIMA
VELOCITA' DEL VENTO	
CAMS 26	$\hat{x}_t = 8.740 + 1.731x_{t-1} + 0.357x_{t-2} + \varepsilon_t$
CAMS 34	$\hat{x}_t = 39.437 + 0.360x_{t-1} + \varepsilon_t$
CAMS 35	$\hat{x}_t = 0.002x_{t-1} + \varepsilon_t$

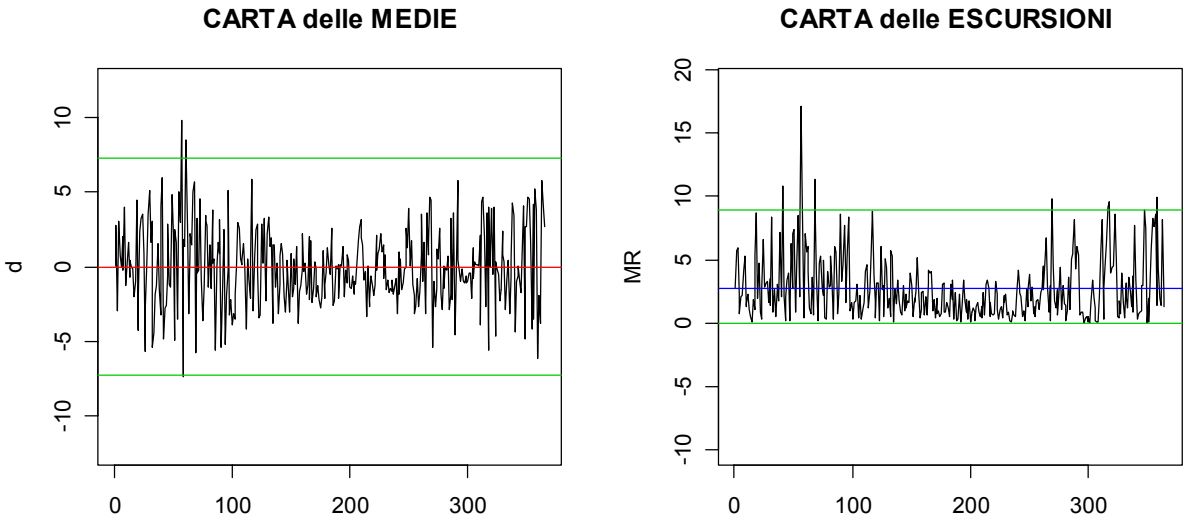
4.4.1 Velocità del vento

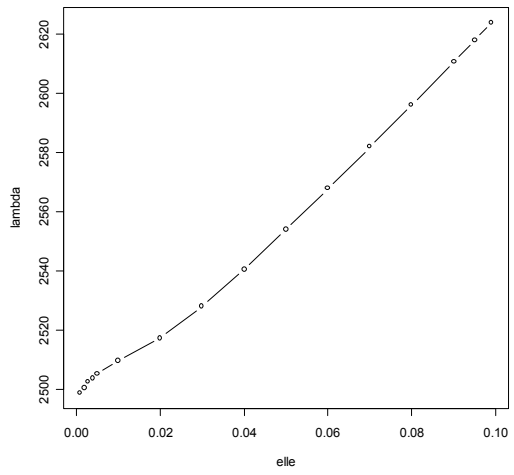
4.4.1.1 Centralina 26

I residui del modello ARIMA per la serie della velocità del vento nella centralina 26 mostrano pochi valori anomali sia nella carte delle medie che in quella

delle escursioni. I fuori controllo appartengono tutti alle osservazioni iniziali: questo risultato non stupisce se si considera quanto emerso in precedenza dall'analisi multivariata. Infatti, dall'analisi dell'indice A per la dodicesima componente principale, in grado di descrivere l'influenza esercitata sui dati dalla centralina 26, è emerso che l'effetto della stazione di rilevazione, ed in particolare della variabile "velocità del vento", interessa prevalentemente il periodo iniziale e quello finale dell'anno (Figura 4.32).

I grafici della carta per gli errori di previsione presentano anch'essi un numero esiguo di valori anomali concentrati in prossimità delle osservazioni iniziali.





λ	$\sum_{i=1}^{N} e_i^2$
0.001	2498.607
0.002	2500.599
0.003	2502.308
0.004	2503.781
0.005	2505.058
0.010	2509.612
0.020	2517.386
0.030	2528.008
0.040	2540.573
0.050	2553.981
0.060	2567.773
0.070	2581.813
0.080	2596.085
0.090	2610.598
0.095	2617.948
0.099	2623.873

Figura 4.34: *Carta delle medie e delle escursioni per la velocità del vento nella centralina 26. Scelta del λ ottimale.*

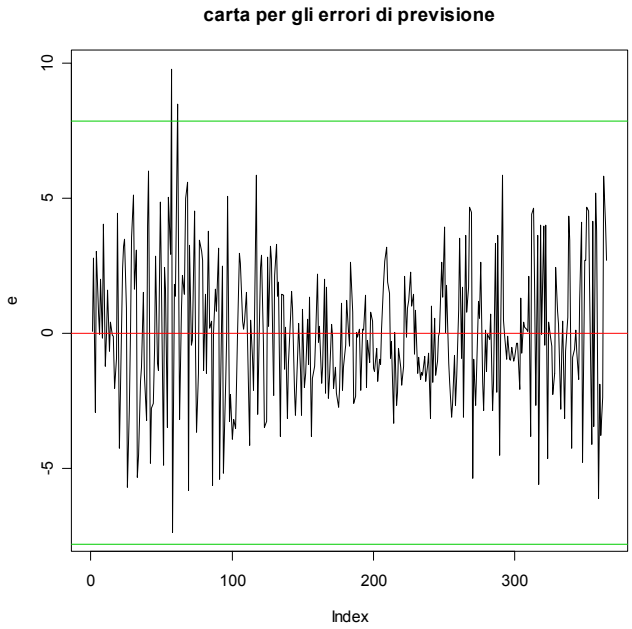


Figura 4.35:*Carta per gli errori di previsione*

4.4.1.2 Centralina 34

Le carte di controllo delle medie e delle escursioni, costruite per i residui della centralina 34, presentano diversi fuori controllo che interessano la seconda parte dell'anno e corrispondono solo in parte ai valori anomali messi in luce dalla carta multivariata di T^2. Se consideriamo però le informazioni fornite dall'indice A

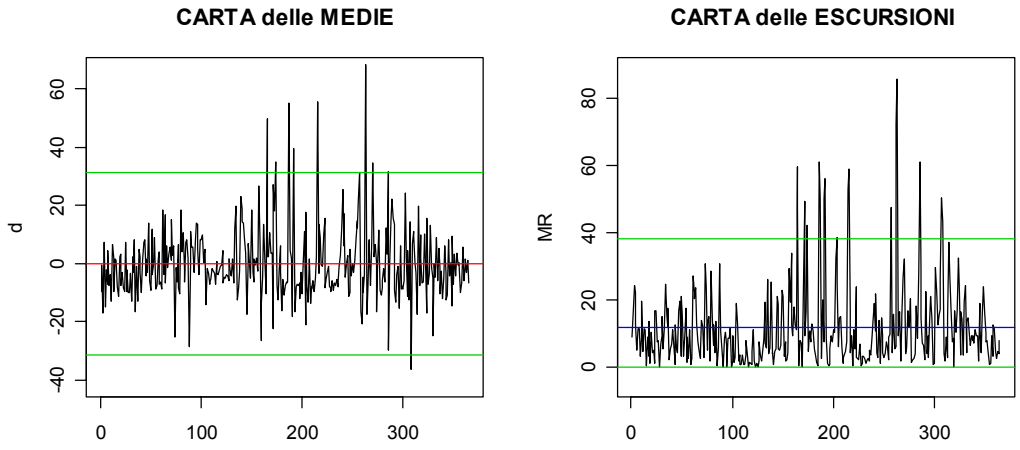
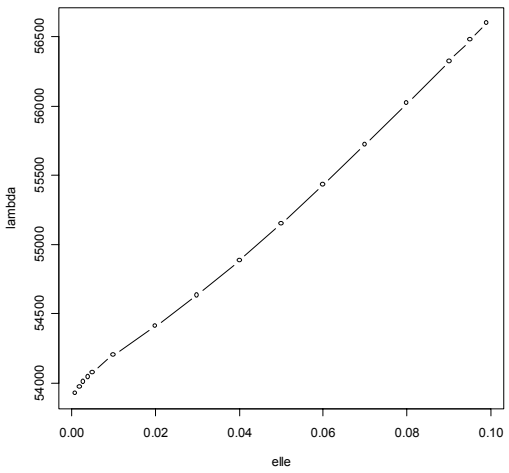


Figura 4.36: *Carta delle medie e delle escursioni per la velocità del vento nella centralina 34*

ed in particolare dalla serie relativa alla sesta componente (Figura 4.32), si nota che l'influenza esercitata dalla centralina 34 ed in particolare dalla velocità, del vento viene esercitata soprattutto nella parte finale del periodo considerato. La carta per gli errori di previsione conferma i risultati emersi dalle carte per misure singole costruite sui residui.



λ	$\sum_{i=1}^{N} e_i^2$
0.001	53922.22
0.002	53967.49
0.003	54007.83
0.004	54043.91
0.005	54076.35
0.010	54202.93
0.020	54408.95
0.030	54634.72
0.040	54884.06
0.050	55151.14
0.060	55431.71
0.070	55722.48
0.080	56020.81
0.090	56324.67
0.095	56478.20
0.099	56601.66

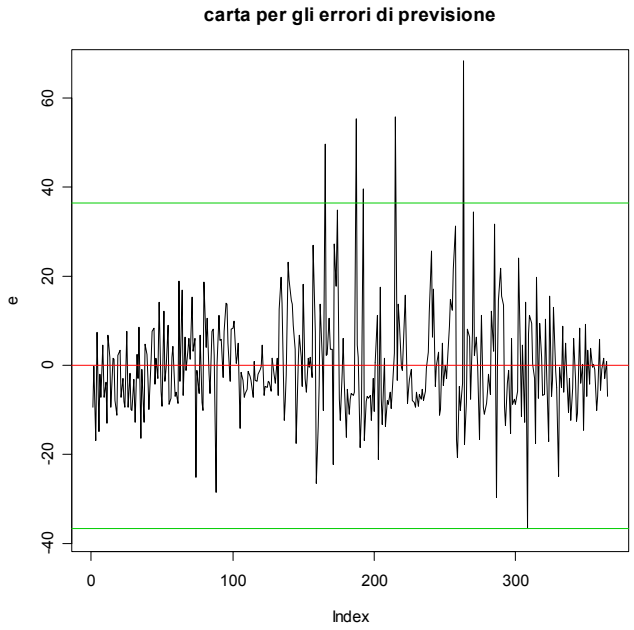


Figura 4.37: *Scelta del λ ottimale e carta per gli errori di previsione*

4.4.1.3 Centralina 35

Le carte di controllo univariate riferite alla variabile "velocità del vento" rilevata nella centralina 35 presentano alcuni valori dei residui fuori controllo. I valori anomali riscontrati sia nella carta delle medie che in quella delle escursioni riguardano principalmente il periodo centrale dell'anno. Questo risultato conferma quanto visto in precedenza nell'analisi multivariata. La prima componente principale del modello contrappone il contributo dato dal vento a quello delle altre due variabili considerate: osservando l'andamento dell'indice A per la prima variabile latente ci si accorge che proprio a partire dai valori centrali della serie si ha un cambiamento nelle condizioni operative del processo. Appare ormai probabile che tale cambiamento possa essere attribuito parzialmente alla velocità del vento . Osservando la carta per gli errori di previsione si può notare come le osservazioni fuori controllo corrispondono abbastanza ai valori anomali messi in luce dalla carta per le medie dei residui.

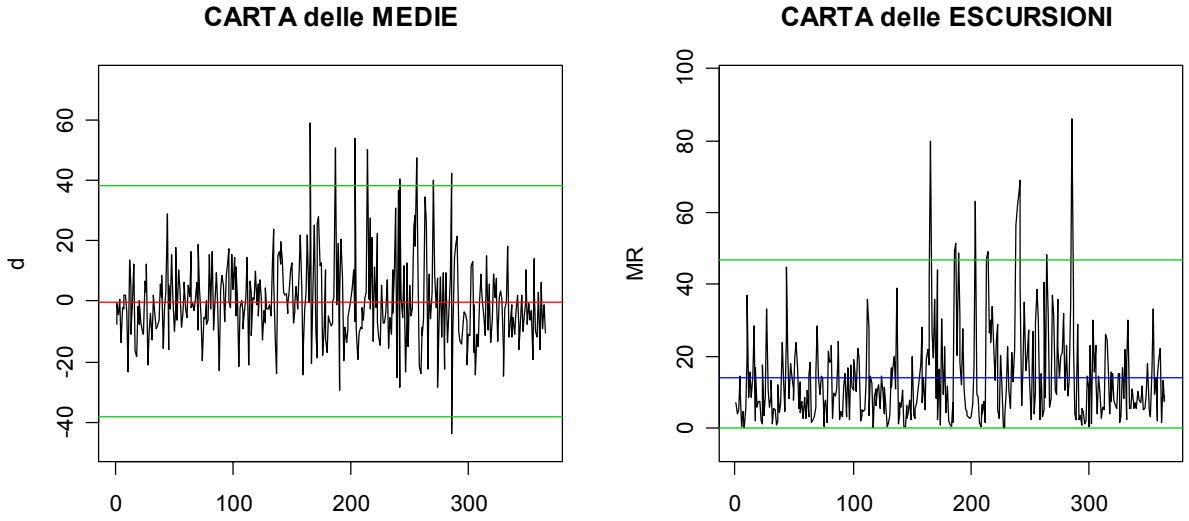
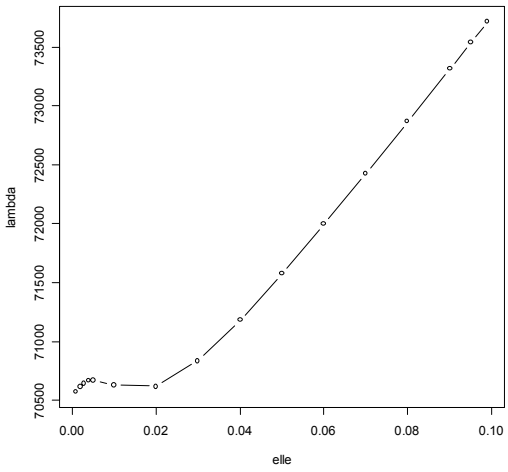


Figura 4.38: *Carta delle medie e delle escursioni per la centralina 35*



λ	$\sum_{i=1}^{N} e_i^2$
0.001	70567.96
0.002	70614.06
0.003	70645.35
0.004	70663.82
0.005	70671.81
0.010	70632.68
0.020	70619.66
0.030	70837.43
0.040	71177.82
0.050	71572.58
0.060	71993.05
0.070	72426.79
0.080	72867.87
0.090	73313.13
0.095	73536.73
0.099	73715.93

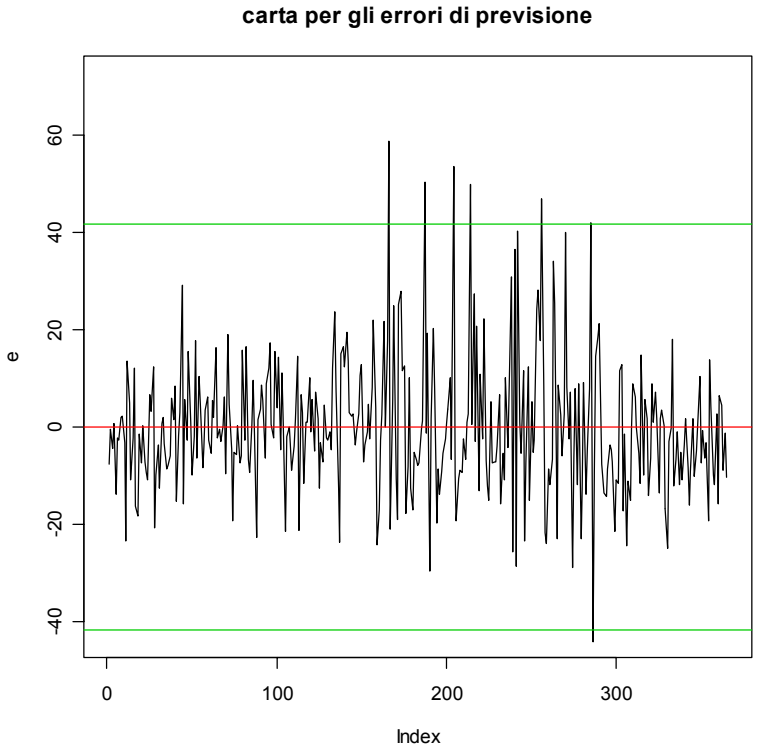


Figura 4.39: *Scelta del λ ottimale e carta per gli errori di previsione*

Conclusioni

In questo studio si è scelto di confrontare diversi metodi per il controllo multivariato di processo. Partendo dalla costruzione delle carte tradizionali per la statistica T^2 e per i residui, si è passati al calcolo di due indici, in grado di tenere maggiormente in considerazione la correlazione esistente tra le variabili in successivi intervalli temporali (*Kano et al*., 2001). L'obiettivo è quello di tener conto del carattere "dinamico" dei dati , cioè dal fatto che le osservazioni rilevate risultano autocorrelate e quindi dipendenti dal passato. Nel modello studiato per i dati ambientali è stato preso in esame, in particolare, il legame esistente tra i dati al tempo corrente X_t e quelli all'istante immediatamente precedente X_{t-1}.

Entrambe le metodologie, *DISSIM* e *MPCA*, prevedono di applicare l'analisi delle componenti principali a sottoinsiemi dei dati di ampiezza pari ad una determinata finestra *w*, originati facendo scorrere la matrice delle osservazioni lungo l'asse temporale. L'indice di diversità *D* viene calcolato, per ogni sottomatrice, misurando la differenza tra la distribuzione dei dati da controllare e quella di un insieme di riferimento. L'indice *A*, invece, è costruito come misura del cambiamento di direzione avvenuto tra le componenti principali di un insieme corrente di osservazioni e quelle di un insieme dato. Un accorgimento da adottare quando si utilizzano queste tecniche consiste nel valutare attentamente il valore della finestra temporale *w*. Dal momento che un valore elevato di questo parametro produce un effetto di lisciamento negli indici, le due procedure potrebbero infatti subire un rallentamento nella capacità di identificare cambiamenti operativi all'interno del processo.

Nello studio di un caso reale con dati derivanti dal monitoraggio ambientale, entrambe le procedure si sono dimostrate efficienti nel rintracciare cambiamenti nella struttura di relazione tra le variabili ambientali rilevate 2001 e quelle rilevate nel 2002.

Per il modello "ozono-radiazione solare" la carta multivariata tradizionale T^2 ha rivelato *in primis* la presenza di valori anomali nella parte centrale del periodo considerato, proprio in corrispondenza della "stagione dell'ozono (estiva)", durante

la quale si ha un incremento delle concentrazioni di questo inquinante a causa dell'aumento della radiazione solare e della temperatura.

L'analisi multivariata condotta utilizzando le statistiche *A* e *D* ha portato ad un miglioramento nell'interpretazione di questo risultato. Nella fase di analisi descrittiva, dall'osservazione dei grafici delle serie relative alle variabili studiate, è stato possibile ipotizzare, per alcune centraline ed in particolare per la C34, un aumento dei valori medi di "ozono" e "radiazione solare", in corrispondenza dei mesi centrali dell'anno. Gli indici *A* e *D* hanno confermato tale ipotesi evidenziando una variazione della relazione tra l'ozono e la temperatura proprio in corrispondenza di questo periodo.

Quando nel modello da analizzare è stata introdotta la variabile "velocità del vento", negativamente correlata con le altre, si sono ottenuti risultati interessanti. La carta T^2 ha infatti evidenziato valori fuori controllo soprattutto nella parte iniziale e finale dell'arco temporale considerato. Le informazioni ottenute dagli indici *A* e *D* hanno confermato la presenza di cambiamenti nella relazione tra le variabili misurate nel 2002 e quelle misurate nel 2001. In particolare è stato messo in evidenza il fatto che tali variazioni interessano principalmente le osservazioni finali e sono dovute ad un effetto congiunto di ozono e radiazione solare contrapposto all'azione esercitata dal vento.

Grazie alle nuove procedure ed al contributo dato dall'analisi univariata, è stato inoltre possibile individuare l'influenza di ogni centralina nella determinazione dei fuori controllo. Nel caso multivariato, osservando i risultati dell'analisi delle componenti principali, si è riusciti ad identificare le variabili latenti capaci di spiegare l'influenza esercitata da ogni centralina all'interno del modello. Lo studio delle serie dell'indice *A* fatto per ogni componente principale, ha permesso, in parte, di attribuire alle diverse centraline la responsabilità di alcuni valori anomali e quindi di cambiamenti.

Per il modello "ozono-radiazione solare", la centralina 34 è quella che fornisce la maggiore evidenza di un cambiamento nell'intervallo di tempo considerato. Le centraline 26 e 35 sembrano invece interessate ai cambiamenti nella struttura di relazione dei dati soprattutto all'inizio ed alla fine dell'arco di tempo considerato. I risultati ottenuti dall'approccio multivariato hanno in seguito trovato conferma nell'osservazione delle carte univariate costruite per ogni variabile del modello.

Il modello “ozono- radiazione solare- velocità del vento” porta a delle conclusioni differenti. Le misurazioni effettuate dalla centralina 34 sembrano interessate da un cambiamento soprattutto nella parte centrale ed in quella finale delle osservazioni, quelle della centralina 35 verso la fine dell’arco temporale considerato, quelle della stazione di rilevazione 26 infine sia sui dati iniziali che su quelli finali.

I metodi *MPCA* e *DISSIM* si sono dunque rivelati efficaci nel determinare variazioni nella relazione temporale tra le variabili e nel mettere in risalto particolari *pattern* spaziali. Tali tecniche hanno infatti permesso di individuare sia i periodi di tempo in cui sono più evidenti le differenze tra i due insiemi posti a confronto sia il sito geografico in cui tali differenze sembrano più rilevanti.

Appendice

Procedura per il calcolo dell'indice *A*

A] Determinazione degli autovettori di riferimento

Valori in ingresso:

x: matrice di dati in controllo $X_{(n\times p)}$ con *n* pari al numero di campioni e *p* pari al numero di variabili.

npc: numero massimo di autovettori.

Valori in uscita:

v: matrice degli autovettori di riferimento.

mean: vettore delle medie ottenute dalla matrice di dati in controllo.

std: vettore delle deviazioni standard ottenute dalla matrice di dati in controllo.

```
proc(3)=refer(x,npc);
local rowx,colx,n,mean,std,scal,v,cov,va,ve;

rowx=ROWS(x);          #inizializzazione delle variabili#
colx=COLS(x);
mean=MEANC(x);
std=STDC(x);
n=1;                   #standardizzazione della matrice X#
scal=zeros(rowx,colx);
  do while n<=colx;
       scal[.,n]=(x[.,n]-mean[n])/std[n];
       n=n+1;
  endo;
cov=vcx(scal);
{va,ve}=eigrs2(cov);
va=rev(va);
print"autovalori con dati std" va;
v=zeros(8,8);
v[.,1]=ve[.,8];
v[.,2]=ve[.,7];
v[.,3]=ve[.,6];
v[.,4]=ve[.,5];
```

```
v[.,5]=ve[.,4];
v[.,6]=ve[.,3];
v[.,7]=ve[.,2];
v[.,8]=ve[.,1];

retp(v,mean,std);
endp;
```

B] Determinazione della statistica *A*

Valori in ingresso:

x: matrice da analizzare $Y_{(n\times p)}$ con *n* pari al numero di campioni e *p* pari al numero di variabili.

npc: numero massimo di autovettori.

ref: matrice degli autovalori di riferimento.

mean: vettore delle medie ottenute dalla matrice di dati in controllo.

std: vettore delle deviazioni standard ottenute dalla matrice di dati in controllo.

window: dimensione della finestra temporale

Valori in uscita:

A: indice *A*.

```
proc(1)=mpca(x,ref,mean,std,window,npc);
local rowx,colx,tstart,tmonit,tend,scal,A,
step,cl,i,U,S,P,T,n,va,ve,xcrf,v,cov;

rowx=ROWS(x);              #inizializzazione delle variabili#
colx=COLS(x);
tstart=window;
tmonit=rowx-tstart+1;
tend=tstart+tmonit-1;
n=1;                      #standardizzazione della matrice Y#
scal=zeros(rowx,colx);     #rispetto a media e varianza di X#
   do while n<=colx;
       scal[.,n]=(x[.,n]-mean[n])/std[n];
       n=n+1;
   endo;
step=tstart;
i=1;
A=zeros(tmonit,npc);
```

```
do while step<=tend;
    xcrf=scal[step-window+1:step,.]; #sottomatrice dei dati#
    cov=vcx(xcrf);
    {va,ve}=eigrs2(cov); #determinazione degli autovettori#
    va=rev(va);
    v=zeros(8,8);
    v[.,1]=ve[.,8];
    v[.,2]=ve[.,7];
    v[.,3]=ve[.,6];
    v[.,4]=ve[.,5];
    v[.,5]=ve[.,4];
    v[.,6]=ve[.,3];
    v[.,7]=ve[.,2];
    v[.,8]=ve[.,1];
    do while i<=COLS(v);
       if v[.,i]'*ref[.,i]<0; v[.,i]=-v[.,i]; endif;
       i=i+1;
    endo;
A[step-tstart+1,1]=(1-abs(diag(v[.,1]'*ref[.,1])))';
A[step-tstart+1,2]=(1-abs(diag(v[.,2]'*ref[.,2])))';
A[step-tstart+1,3]=(1-abs(diag(v[.,3]'*ref[.,3])))';
A[step-tstart+1,4]=(1-abs(diag(v[.,4]'*ref[.,4])))';
A[step-tstart+1,5]=(1-abs(diag(v[.,5]'*ref[.,5])))';
A[step-tstart+1,6]=(1-abs(diag(v[.,6]'*ref[.,6])))';
A[step-tstart+1,7]=(1-abs(diag(v[.,7]'*ref[.,7])))';
A[step-tstart+1,8]=(1-abs(diag(v[.,8]'*ref[.,8])))';
step=step+1;
endo;
retp(A);
endp;
```

✔ Procedura per il calcolo dell'indice *D*

A] Determinazione della matrice di riferimento

Valori in ingresso:

x: matrice di dati in controllo X con n pari al numero di campioni

window: dimensione della finestra temporale

Valori in uscita:

xref: matrice di riferimento X_{ref}.

```
proc(1)=ref(x,window);
local
xref,npc,numpc,xorg,rowx,colx,tstart,tmonit,tend,mean,sdev,
scal,i,d,nref,nall,rref,step,xcrt,rmix,pmix,smix,sss,ttt,c,ind
ex,dsort,
ycrt,scrt,reftim;

xorg=x;                    #inizializzazione delle variabili#
rowx=ROWS(x);
colx=COLS(x);
npc=colx;
tstart=window;
tmonit=rowx-tstart+1;
tend=tstart+tmonit-1;

mean=MEANC(x);
sdev=STDC(x);
numpc=1;                 #standardizzazione della matrice X#
scal=zeros(rowx,colx);

do while numpc<=npc;
    scal[.,numpc]=(x[.,numpc]-mean[numpc])/sdev[numpc];
    numpc=numpc+1;
endo;

d=zeros(tmonit,1);
xref=scal[tstart-window+1:tstart,.]; #matrice di riferimento#
nref=ROWS(xref);                        #provvisoria#
nall=nref+window;
rref=xref'*xref/nall;         #matrice di covarianza di X_ref#
step=tstart;

do while step<=tend;
     xcrt=scal[step-window+1:step,.]; #matrice da confrontare#
  /* rmix=rref+xcrt'*xcrt/nall; */        # con X_ref#
     {smix,pmix}=eigrs2(rref+xcrt'*xcrt/nall); #autovalori e#
     sss=rev(smix);              #autovettori della matrice #
     pmix=pmix';                #di covarianza totale R#
     pmix=rev(pmix);
     pmix=pmix';
     ttt=sss^(-1/2);
     i=1;
     do while i<=rows(sss);
```

```
            if sss[i]<(1e-10); ttt[i]=0; endif;
            i=i+1;
        endo;
        c=zeros(npc,npc);
        ttt=diagrv(c,ttt);
        ycrt=sqrt(window/nall)*xcrt*pmix*ttt;
        scrt=ycrt'*ycrt/window;
        d[step-tstart+1,1]=4*meanc((eig(scrt)-0.5)^2);
        step=step+1;
    endo;

    index=sortind(d);   #ordino i vettori in senso crescente#
    dsort=d[index];

    if tmonit==1;
        reftim=tstart;
    else;
        reftim=tstart+index[floor(tmonit/2)]-1; #determinazione#
    endif;                                      #della mediana#


    xref=xorg[reftim-window+1:reftim,.];  #matrice di riferimento#
    retp(xref);                             #definitiva#

    endp;
```

B] Determinazione della statistica *D*

Valori in ingresso:

x: matrice da analizzare $Y_{(n\times p)}$ con *n* pari al numero di campioni e *p* pari al numero di variabili.

window: dimensione della finestra temporale.

xref: matrice di riferimento X_{ref}.

Valori in uscita:

d: indice *D*.

```
    proc(1)=d(x,xref,window);
    local
npc,npc0,numpc,xorg,row,row0,col,col0,tstart,tmonit,tend,meanx,sdevx
,scalref,i,d,
    nref,nall,rref,step,xcrt,rmix,pmix,smix,sss,ttt,c,scalx,
    ycrt,scrt,cl;
```

```
row0=ROWS(xref);        #inizializzazione delle variabili#
col0=COLS(xref);
npc0=COLS(xref);
row=ROWS(x);
col=COLS(x);
npc=col;
tstart=window;
tmonit=row-tstart+1;
tend=tstart+tmonit-1;

meanx=MEANC(xref);      #standardizzazione della matrice X_ref #
sdevx=STDC(xref);
numpc=1;
scalref=zeros(row0,col0); #standardizzazione della matrice Y#
                          #rispetto a media e varianza di X#

do while numpc<=npc0;
    scalref[.,numpc]=(xref[.,numpc]-
meanx[numpc])/sdevx[numpc];
    numpc=numpc+1;
endo;

scalx=zeros(row,col);
numpc=1;

do while numpc<=npc;
    scalx[.,numpc]=(x[.,numpc]-meanx[numpc])/sdevx[numpc];
    numpc=numpc+1;
endo;

d=zeros(tmonit,1);
nref=ROWS(scalref);
nall=nref+window;
rref=scalref'*scalref/nall;   #matrice di covarianza di X_ref#
step=tstart;

do while step<=tend;
     xcrt=scalx[step-window+1:step,.]; #matrice da confontare#
  /* rmix=rref+xcrt'*xcrt/nall; */          #con di X_ref#
     {smix,pmix}=eigrs2(rref+xcrt'*xcrt/nall); #autovalori e#
     sss=rev(smix);               #autovettori della matrice #
     pmix=pmix';                   #di covarianza totale R#
```

```
      pmix=rev(pmix);
      pmix=pmix';
      ttt=sss^(-1/2);
      i=1;
      do while i<=rows(sss);
          if sss[i]<(1e-10); ttt[i]=0; endif;
          i=i+1;
      endo;
      c=zeros(npc,npc);
      ttt=diagrv(c,ttt);
      ycrt=sqrt(window/nall)*xcrt*pmix*ttt;
      scrt=ycrt'*ycrt/window; #matrice di covarianza della#
                              #matrice trasformata#
      d[step-tstart+1,1]=4*meanc((eigrs(scrt)-0.5)^2);

      step=step+1;
endo;

retp(d);
endp;
```

Bibliografia

DI FONZO T. & LISI F. (2000) *"Complementi di statsistica economica: analisi delle serie storiche univariate"*. CLEUP, Padova.

HAWKINS D. M. (1991*) "Multivariate Quality Control Based on Regression-Adjusted Variables"*. Technometrics, Vol. 33, No.1, 61-75.

HAWKINS D. M. (1993) *"Regression Adjustment for Variables in Multivariate Quality Control"*. Journal of Quality Technology, Vol. 25, No. 3, 170-182.

JACKSON E. & MUDHOLKAR G. (1979) *"Control Procedures for Residuals Associated With Principal Component Analysis"*. Technometrics, Vol. 21, No. 3, 341-349.

JACKSON E. (1980) *"Principals Components and Factor Analysis: Part I-Principal Components"*. *J*ournal of Quality Technology, Vol. 12, No. 4, 201-213.

JACKSON E. (1959) *"Quality Control Methods for Several Related Variables"*. Technometrics, Vol. 1, No. 4, 590-377.

KANO M., NAGAO K., OHNO H., HASEBE S., HASHIMOTO I., STRAUSS R. & BAKSHI B. (2000) *"Comparison of statistical process monitoring methods: application to the Eastman challenge problem"*. Computers and Chemical Engineering, Vol. 24, 175-181.

KANO M., NAGAO K., OHNO H., HASEBE S., HASHIMOTO I., STRAUSS R. & BAKSHI B. (2002) *"Comparison of multivariate statistical process monitoring methods with applications to the Eastman challenge problem"*. Computers and Chemical Engineering, Vol. 26, 161-174.

KANO M., OHNO H., HASEBE S., & HASHIMOTO I. (2001) *"A new multivariate statistical process monitoring method using principal components analysis"*. Computers and Chemical Engineering, Vol. 25, 1103-1113.

KANO M., OHNO H., HASEBE S., & HASHIMOTO I. (2002) *"Statistical Process Monitoring Based on Dissimilarity of Process Data"*. AIChE, Vol. 48, No. 6, 1231-1240.

KOURTI T. & MacGREGOR J.F. (1996) *"Multivariate SPC Methods for Process and Product Monitoring"*. Journal of Quality Technology, Vol. 28, No. 4, 409-428.

KU W., STORER R. & GEORGAKIS C. (1995) *"Disturbance detection and isolation by dynamic principal components analysis"*. Chemometrics and Intelligent Laboratory System, Vol. 30, 179-196.

MASON R., TRACY N. & YOUNG J. (1995) *"Decomposition of T^2 for Multivariate Control Chart Interpretation"*. Journal of Quality Technology, Vol. 27, No. 2, 99-108.

MASON R., TRACY N. & YOUNG J. (1995) "*Monitoring a Multivariate Step Process"*. Journal of Quality Technology, Vol. 28, No. 1, 39-50.

MASON R., TRACY N. & YOUNG J. (1995) "*Multivariate Control Charts for Individual Observation"*. Journal of Quality Technology, Vol. 24, No. 2, 88-95.

MURPHY B.J. (1987) *"Selecting out of control variables with the T^2 multivariate quality control procedure"*. The Statistician, Vol. 36, 571-583.

NOMIKOS P. MacGREGOR J.F. (1995) *"Multivariate SPC Charts for Monitoring Batch Processes"*. Technometrics, Vol. 37, No. 1, 41-59.

WIERDA S.J. (1994) *"Multivariate statistical process control – recent results and directions for future research"*. Statistica Neerlandica, Vol. 48, No. 2, 147-168.

WEBSTER R., OLIVIER M.A. (2000) *"Geostatistics Environmental Scientists"*. 383-26 posizione Biblio PD.

ZENARI E. (1999) *"Alcune considerazioni sul Controllo Statistico della Qualità"*. Dispensa del Corso di Controllo Statistico della Qualità.

TEXAS NATURAL RESOURCE CONSERVATION COMMISSION
http://www.tnrcc.state.tx.us

AZIENDA REGIONALE PROTEZIONE AMBIENTALE del VENETO
http://www.arpa.veneto.it

ENVIRONMENTAL PROTECTION AGENCY
http://www.epa.gov