# UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA TRIENNALE IN

SCIENZE STATISTICHE E TECNOLOGIE INFORMATICHE

TESI DI LAUREA

## IL CONTROLLO STATISTICO MULTIVARIATO: UN'APPLICAZIONE PRATICA AI DATI DI UN PROCESSO S.E.S.A. S.p.A.

**Relatore: Prof.ssa Giovanna Capizzi**

**Laureando: Feffin Giulio**

**Matricola: 602428 - STI**

Anno Accademico 2010/2011

*Un ringraziamento profondo*

*va ai miei genitori, agli amici e*

*alle persone a me vicine*

*che mi hanno sempre sostenuto*

*nel momento del bisogno*

*A mio cugino Marco,*

*per la sua gioia di vivere la vita*

*tra mille difficoltà;*

*che il tuo insegnamento*

*sia sempre vivo dentro di noi*

# Indice

# Premessa

Alla fine della maturità ho scelto La Facoltà di Scienze Statistiche, più precisamente il corso di laurea in Statistica e Tecnologie Informatiche, perché sono sempre stato attratto dall'informatica e nutrivo un profondo interesse per la matematica e la statistica.

Già dal primo anno non mi sarei mai aspettato che questo percorso di studi mi potesse insegnare così tanti metodi per poter analizzare i dati, aiutare nella risoluzione di problemi e dare un significato ai fenomeni che si verificano quotidianamente.

Tra i vari corsi, il mio interesse era attratto principalmente per "*Il controllo statistico della Qualità"* proprio perché forniva strumenti utili alla risoluzione dei problemi collegati al processo produttivo (studio delle non conformità, Capacità, ecc …). Il mio interesse mi ha portato ad approfondire questo campo, apprendendo alcune delle molte tecniche multivariate presenti nel mondo del Controllo della Qualità, e quindi a svolgere uno stage in un'azienda per poter applicare in prima persona questi metodi d'analisi.

Con la presente tesi voglio presentare le tecniche multivariate più diffuse nell'ambito del controllo Statistico del Processo Produttivo (MSPC) ed applicarle ai dati provenienti da processi S.E.S.A. S.p.A., società dove ha avuto luogo il mio stage.

# CAPITOLO 1

## Presentazione dell'azienda

S.E.S.A. S.p.A. – Società Estense Servizi Ambientali –, opera dal 1995 come partner tecnologico del comune di Este per lo sviluppo di attività ambientali nella raccolta differenziata e nel recupero dei rifiuti.

L'impianto si trova nel comune di Este (in Provincia di Padova), a circa 3 km ad Ovest del centro urbano della località e circa 1,5 km a Nord del nucleo centrale del Comune di Ospedaletto Euganeo.

A partire dalla data di stipulazione della convenzione con l'Amministrazione Comunale, S.E.S.A. S.p.A. ha intrapreso un percorso di crescente e continuo sviluppo, reinvestendo i capitali nella realizzazione dei nuovi impianti di compostaggio, di cogenerazione di energia termica ed elettrica e di selezione dei rifiuti, nonché delle annesse strutture di supporto, dalla viabilità all'officina per la manutenzione dei mezzi, dalle aree di stoccaggio dei rifiuti all'installazione di pannelli fotovoltaici, dalla captazione e successivo trattamento del percolato della discarica.

In particolare S.E.S.A. S.p.A. si impegna a perseguire i seguenti obiettivi specifici:

1. continuo aggiornamento tecnologico/impiantistico per garantire un costante miglioramento per:
   - ✓ LA QUALITÀ: miglioramento e ottimizzazione dei processi aziendali compatibile alle nuove tecnologie scientifiche, strutturali ed impiantistiche;
   - ✓ L'AMBIENTE: riduzione dell'impatto ambientale in termini di:

- intensificazione della raccolta differenziata e relativa diminuzione del rifiuto smaltito in discarica a favore del recupero e della produzione di Materie Prime Secondarie;
- sviluppo della rete di teleriscaldamento con conseguente riduzione delle emissioni domestiche in città e riduzione gas effetto serra per impiego fotovoltaico;
- riutilizzo di acque di processo nelle attività aziendali;
- messa in sicurezza, bonifica e ripristino ambientale;

✓ LA SALUTE E SICUREZZA SUI LUOGHI DI LAVORO mediante:

- intensificazione dei livelli di sicurezza nella prevenzione degli incidenti, degli infortuni e delle malattie professionali;
- costante monitoraggio e analisi degli incidenti, infortuni e situazioni pericolose;
- costante formazione e addestramento nella gestione delle emergenze dal punto di vista della sicurezza.

2. progettazione , sviluppo, costruzione di impianti con le migliori tecnologie disponibili, conformemente alla normativa vigente in materia ambientale e di sicurezza;
3. sensibilizzazione della popolazione locale per un costante aumento della raccolta differenziata e della valorizzazione del rifiuto stesso, puntando al riciclo totale;
4. aumento dell'efficienza del recupero energetico del biogas da rifiuto come fonte energetica alternativa;
5. costante formazione e informazione del proprio personale creando figure qualificate e specializzate e comunicando le modalità definite per la corretta gestione dell'ambiente, dei requisiti relativi al servizio effettuato e dei ruoli e delle responsabilità di ciascuno per la tutela della salute e sicurezza sui luoghi di lavoro;

6. creazione di un rapporto di trasparenza con la comunità locale attraverso la divulgazione della Dichiarazione Ambientale

## Attività svolte presso l'impianto

Presso l'impianto S.E.S.A. di Este sono svolte le seguenti attività soggette ad autorizzazione:

- Smaltimento in discarica di rifiuti urbani e speciali non pericolosi e gestione della stessa: l'impianto, realizzato secondo le disposizioni provinciali e già classificato di prima categoria ai sensi della Delibera C.I. 27/7/84, è gestito direttamente da S.E.S.A. ed è stato riclassificato come discarica per rifiuti non pericolosi, rientrando nelle previsioni del Piano Regionale dei rifiuti solidi urbani ed assimilabili, al servizio dei Comuni del Bacino PD3 (Padova 3).
- Compostaggio: il compost TERRA EUGANEA prodotto da S.E.S.A. è un ammendante di alta qualità ottenuto attraverso un processo di trasformazione e stabilizzazione controllata dei materiali provenienti sia dalla raccolta differenziata effettuata dalle famiglie, sia dalle attività di trasformazione dei prodotti ortofrutticoli.
- Produzione di biogas e recupero energetico con produzione di calore e di energia elettrica: l'eccellenza dell'impianto S.E.S.A. sta anche e soprattutto nella capacità di produrre energia dai rifiuti senza ricorrere a combustione diretta. Infatti, la frazione liquida derivata dalla lavorazione del rifiuto umido per l'impianto di compostaggio, viene inviata ad idonei fermentatori, dove grazie al processo di digestione anaerobica realizzato in appositi digestori, viene prodotto biogas, il quale viene compresso, purificato e destinato ad un gruppo di cogenerazione per la produzione di energia elettrica e termica.
- Selezione e messa in riserva del secco differenziato ed indifferenziato: il rifiuto secco differenziato ed indifferenziato in ingresso viene lavorato in un

moderno impianto che opera una selezione meccanica e in parte manuale, al fine di ottenere materie prime secondarie destinati al recupero presso altri impianti autorizzati.

- Ricevimento e stoccaggio di rifiuti urbani ed assimilati ingombranti.
- Ricevimento e stoccaggio di rifiuti urbani pericolosi – ex RUP.
- Intermediazione e commercio di rifiuti: in alcuni casi S.E.S.A. opera da intermediario di rifiuti, ovvero quando diventa conveniente, sotto il punto di vista ambientale o economico, destinare rifiuti ad altri impianti di trattamento, mantenendone un controllo amministrativo.
- Ripristino e bonifica di siti: come attività marginale, S.E.S.A. svolge, esclusivamente su segnalazione della Pubblica Amministrazione, interventi di messa in sicurezza, ripristino e bonifica di siti contaminati.
- Gestione del laboratorio interno: il laboratorio, accreditato con n. 0688 da ACCREDIA (organismo nazionale di accreditamento dei laboratori di prova riconosciuto a livello internazionale), svolge analisi di tipo chimico, chimico-fisico, microbiologico e merceologico di rifiuti urbani, industriali e agricoli, di ammendanti, di terreni, nonché analisi di controllo dell'inquinamento atmosferico e ambientale.

## Il Sistema di Gestione Integrato

I principali sistemi di gestione oggi riconosciuti hanno molteplici aree completamente sovrapponibili e integrabili tra loro e tale specificità permette a sistemi quali quello ambientale (UNI EN ISO 14001, EMAS), quello della sicurezza (OHSAS 18001, DM 2000) e quello della qualità (UNI EN ISO 9001) di potere progettare un unico **sistema detto Sistema di Gestione Integrato.**

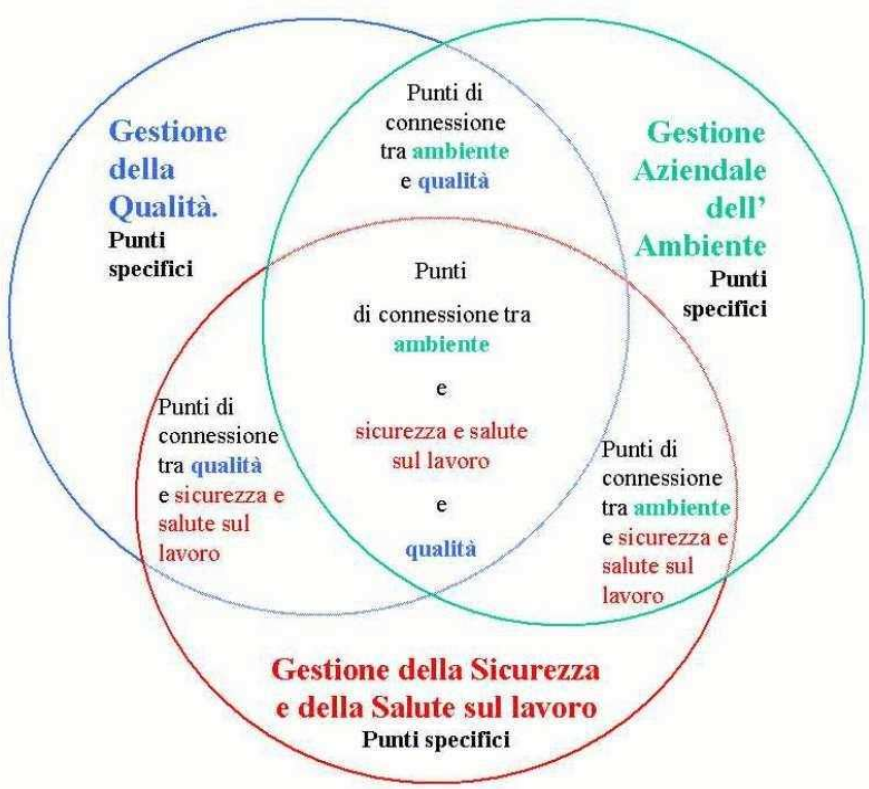


**Figura 1: Sistema di connessione fra le 3 aree di gestione**

L'integrazione dei sistemi apporta benefici organizzativi, semplificando ad esempio la gestione documentale, e consente di sviluppare la logica della gestione unificata dei processi, senza mantenere distinzioni tra ambiente, sicurezza e qualità.

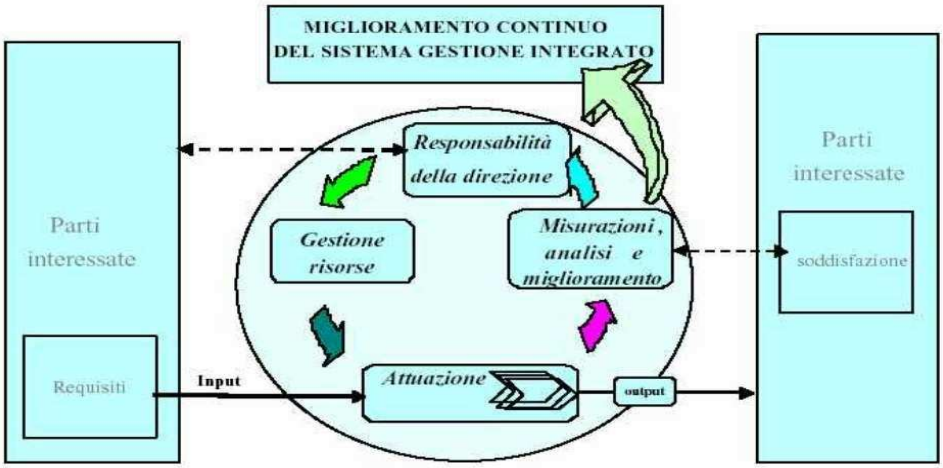


**Figura 2: Schema di miglioramento di una società tramite il Sistema di Gestione Integrato**

I principali vantaggi nell'implementazione di un Sistema di Gestione Integrato sono:

- attuare uno strumento utile per assicurare alle parti interessate (Enti di controllo, azionisti, popolazione, lavoratori) che tutti gli aspetti normativi e cogenti correlati alle attività svolte siano gestiti secondo le modalità prescritte dalle leggi vigenti e secondo le migliori prassi disponibili;
- rendere l'azienda maggiormente competitiva in un mercato sempre più sensibile alle problematiche ambientali;
- contribuisce ad ottenere risultati ottimizzati a beneficio di tutte le parti interessate;
- consentire all'azienda di ottenere risparmi economici e di tempo legati ad una più efficiente gestione di tutti gli aspetti importanti del sistema di gestione aziendale, con una più chiara ed integrata definizione di obiettivi e programmi, con una ottimizzazione di costi e risorse e con una maggiore facilità nel trasferire agli *stakeholders* interni ed esterni gli obbiettivi dell'Organizzazione.

# CAPITOLO 2

## Il compost

Il **compost** è il risultato della decomposizione e dell'umificazione di un misto di materie organiche (come ad esempio residui di potatura, scarti di cucina, letame, liquame o i rifiuti del giardinaggio come foglie ed erba sfalciata) da parte di macro e microrganismi in condizioni particolari: presenza di ossigeno ed equilibrio tra gli elementi chimici della materia coinvolta nella trasformazione.

Il compostaggio tecnicamente è un processo biologico aerobico e controllato dall'uomo che porta alla produzione di una miscela di sostanze umificate (il compost) a partire da residui vegetali sia verdi che legnosi o anche animali mediante l'azione di batteri e funghi.

Il compost può essere utilizzato come fertilizzante su prati o prima dell'aratura. Il suo utilizzo, con l'apporto di sostanza organica migliora la struttura del suolo e la biodisponibilità di elementi nutritivi (composti del fosforo e dell'azoto). Come attivatore biologico aumenta inoltre la biodiversità della microflora nel suolo.

Per avere un buon compost, bisogna ricordarsi che sono gli organismi decompositori del suolo a produrlo. Essi, per vivere, hanno bisogno di tre parametri:

- nutrienti equilibrati, composti da un misto di materie carboniose (brune – dure – secche) e di materie azotate (verdi – molli – umide), di umidità che proviene dalle materie azotate (umide) ed eventualmente dall'acqua piovana o apportata manualmente;
- aria che si infiltra attraverso la porosità prodotta dalla presenza delle sostanze carboniose strutturanti (dure).

Presso l'impianto vengono prodotti :

- Ammendante compostato di qualità;
- Ammendante compostato torboso, costituito da compost vagliato e torba già vagliata in percentuale minima del 50%.

## Il processo di compostaggio

Le fasi principali del processo di produzione del compost sono i seguenti:

1. **Miscelazione delle matrici.**

Le matrici principali, F.O.R.S.U. (Frazione Organica Rifiuti Solidi Urbani) e materiali lignocellulosici, vengono miscelate insieme e rispettando la percentuale prevista dalla normativa, 70% F.O.R.S.U. e 30% materiali lignocellulosici.

2. **Spremitura**

Una volta miscelate le due matrici, la risultante viene spremuta, eliminando gran parte di liquidi presenti. I liquidi verranno poi raccolti nelle apposite cisterne di digestione anaerobica per produrre energia elettrica.

3. **Biossidazione**

Il prodotto derivato dalla miscelazione delle matrici viene introdotto nelle biocelle di ossidazione alla temperatura costante di 70 °C per almeno 1 h.

Il materiale in uscita dalle celle di biossidazione dovrà avere un Indice Respirometrico inferiore a $1300\ \mathrm{mg}\ \mathrm{O_2/Kg}$ di $\mathrm{SV/h}$ come definito dalla D.G.R.V. n. 568/05, Tab. G, ed un'umidità superiore al 40%.

4. **Maturazione**

La maturazione dovrà avere una durata non inferiore a 45 giorni e dovrà avvenire esclusivamente nelle celle di maturazione, celle nelle quali dovrà essere mantenuto un tenore di umidità superiore al 30% ed un flusso d'aria idoneo a garantire una

microaerobiosi della massa, per assicurare la conformità del compost ottenuto alle disposizioni di legge

In uscita dalla fase di maturazione vengono poi presi dei campioni e analizzati per poter verificare che siano rispettati i parametri di conformità previsti dall'Art. 8 Provv. 5149/EC/2007; Art. 2 Provv. 5215/EC/2007 , e quindi è stato compito mio monitorare tali parametri e riferire l'eventuale non conformità.

## L'Ufficio Certificazione S.E.S.A. S.p.A.

Per accertare se il sistema di gestione per la qualità, l'ambiente e la sicurezza risultino conformi ai requisiti della norma di riferimento, la S.E.S.A. conduce verifiche ispettive in forma sistematica e pianificata.

S.E.S.A. si è sempre impegnata per il pieno rispetto delle norme sulla sicurezza, l'ambiente e la qualità, di fatti ha conseguito la Certificazione di Qualità secondo la norma UNI EN ISO 9001, la Certificazione di Sistema di Gestione Ambientale secondo la norma UNI EN ISO 14001 e la Certificazione di Sicurezza secondo la BH OHSAS 18001 e il Certificato di Eccellenza per avere conseguito tutti e tre i Certificati.

Tutto questo attraverso l'implementazione e all'attuazione del Sistema di Gestione Integrato per la sicurezza, l'ambiente e la qualità.

## Risultati dell'analisi del laboratorio

Il laboratorio, parte integrante degli impianti di S.E.S.A. S.p.A., occupa un ruolo importante in tutte le operazioni svolte dall'azienda. In questo capitolo si approfondirà la fase dell'analisi delle matrici in ingresso e in uscita.

In seguito all'arrivo in azienda delle ditte conferitrici, il laboratorio si occupa di campionare e analizzare le matrici, sia in ingresso che in uscita, affinché si rispettino

i limiti di specifica imposti secondo le normative vigenti. In seguito all'analisi i risultati vengono mandati all'Ufficio Certificazione, che si occupa di gestire i risultati del laboratorio.

## Gestione del prodotto non conforme

Di queste operazioni correttive la S.E.S.A. ha stabilito che per i controlli in ingresso si eseguano le seguenti operazioni:

- valutazione esiti analisi chimica: nel caso in cui la concentrazione di uno o più parametri ricercati con le analisi chimiche sui rifiuti in ingresso risulti non conforme ai limiti stabiliti viene messa in riserva la partita di compost in questione, e definite le modalità di trattamento che possono essere la re immissione nel ciclo produttivo o invio a opportuna forma di smaltimento
- verifiche documentali dei rifiuti in ingresso all'impianto: se tali verifiche evidenziano che i rifiuti in ingresso non sono conformi il carico viene respinto.
- verifiche in loco con ispezione visiva prima e dopo lo scarico: se tali controlli evidenziano che il rifiuto non è conforme alle specifiche stabilite (presenza eccessiva di rifiuti non compostabili), il carico viene gestito secondo una delle seguenti alternative: respinto, sottoposto a trattamento di rimozione delle frazioni non compostabili, avviato a smaltimento in discarica.

Nel caso in cui siano rilevate delle non conformità:

- in fase di effettuazione dei controlli sul compost bio ossidato: la partita viene re immessa nel ciclo produttivo, sottoponendola ad un ulteriore ciclo di bio ossidazione;
- in fase di effettuazione dei controlli sul prodotto finito: il cumulo viene riprocessato nella maturazione e/o sottoposto ad ulteriore lavorazione di raffinazione con vagliatura. Nel caso, con ulteriore campionamento e controllo, fosse riconfermata la non conformità, il cumulo di compost viene

riprocessato una seconda volta e/o conferito in discarica come compost non conforme per l'uso consentito dalla normativa.

## Cause di non conformità

Per quanto riguarda il prodotto, possono verificarsi le seguenti situazioni:

- un compost con elevato contenuto di metalli pesanti o di inerti, o ad elevata salinità o dotato di squilibri nutrizionali, deve la sua non conformità alla non adeguatezza delle matrici in ingresso;
- un compost a basso tasso di stabilità biologica (elevato IRD), o contaminato da agenti patogeni o fitotossico, deve la sua non conformità ad un'inadeguata conduzione del processo: fase di bi ossidazione non adeguata, non raggiungimento di temperature adeguate, contaminazioni di vario genere ed origine.

Per quanto riguarda il processo, invece, i problemi possono essere il mancato raggiungimento di temperature adeguate, o loro instabilità nel tempo. Ciò può essere dovuto a:

- rapporto carbonio/azoto non ottimale nella miscela, per cui il processo bio ossidato stenta a partire;
- frequenza troppo bassa di rivoltamento (nel caso di sistemi aperti ), per cui la miscela può risultare disomogenea, con zone in cui il processo bio ossidato si blocca, generando, per l'instaurarsi di condizioni di anaerobiosi, il rilascio di emissioni maleodoranti;
- basso tenore di umidità della miscela (nel caso di sistemi chiusi) causato da un'eccessiva ventilazione dei materiali in trasformazione, a cui non corrisponde un corretto ripristino della stessa umidità;
- frequenza troppo alta di rivoltamento, per cui il calore viene disperso troppo in fretta.

# Azioni correttive

Le azioni correttive necessarie per eliminare le cause della non conformità per il prodotto possono essere le seguenti:

- per il compost ad elevato contenuto di inerti è necessario verificare se è possibile un'ulteriore separazione;
- se il compost non è conforme è necessaria una verifica dell'Indice di Respirazione (indice che dà una misura della qualità di matrici organiche ottenute da processi di biotrasformazione) e un'ottimizzazione dei parametri di processo durante la bio ossidazione;
- un compost che ha rilevato presenza di agenti patogeni o di sostanze fitotossiche deve essere re immesso in testa, nella fase di bio ossidazione, affinché raggiunga nuovamente temperature di 60°C per almeno 5 giorni.

Per il processo, invece:

- se non viene raggiunta la temperatura prefissata, è necessario ottimizzare la frequenza dei rivoltamenti (nel caso di sistemi aperti), garantire il giusto tenore di umidità della miscela iniziale (nel caso di sistemi chiusi) e/o verificare il rapporto C/N (carbonio/azoto) del materiale, rendendolo eventualmente compatibile con l'innesco del processo;
- se sussistono eccessivi odori sgradevoli, è necessario omogeneizzare e riossigenare il materiale il materiale per eliminare le zone di anaerobiosi.

# CAPITOLO 3

# IL CONTROLLO STATISTICO DELLA QUALITÀ

## Introduzione al capitolo

Nei capitoli precedenti si è fatta una panoramica sull'azienda e sulle attività svolte da S.E.S.A. S.p.A., dato il contesto di riferimento, nel seguito si illustrano i principali metodi statistici multivariati impiegati per l'analisi dei dati ricavati dal campionamento del F.O.R.S.U. (Frazione Organica Rifiuti Solidi Urbani) e del compost.

In particolare, nel seguito si assumerà che le osservazioni provenienti dai processi S.E.S.A. S.p.A. si distribuiscono secondo una variabile casuale normale multivariata di grado p e si scriverà $X \sim N_p(\mu, \Sigma)$ con $\mu$, vettore delle medie delle variabili, e $\Sigma$, matrice di varianza e covarianza:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \quad \text{e} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2}^2 & \cdots & \sigma_{1,p}^2 \\ \sigma_{2,1}^2 & \sigma_2^2 & \cdots & \sigma_{2,p}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1}^2 & \sigma_{p,2}^2 & \cdots & \sigma_p^2 \end{pmatrix}$$

La funzione di densità per $X \sim N_p(\mu, \Sigma)$ sarà:

$$f_X(X) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right\}$$

Le curve di livello della funzione di densità di una normale multivariata sono individuate dai punti per cui $(X-\mu)^T\Sigma^{-1}(X-\mu) = costante$, e quest'ultima è l'equazione di un'elissoide con centro in $\mu$.

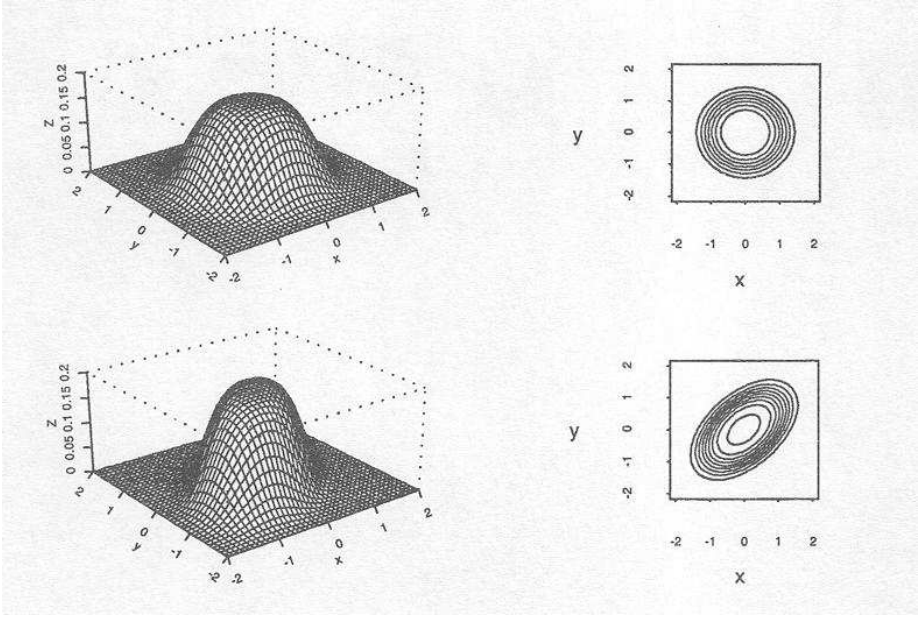


**Figura 3: Esempio di curve di livello per 2 diverse matrici Σ**

## 3.1 Il controllo statistico della qualità del processo produttivo

Il controllo statistico della qualità del processo produttivo (SPC) consiste nell'insieme di strumenti statistici e finalizzati a controllare l'insieme di caratteristiche di un processo produttivo per renderlo sempre più efficiente.

L'SPC in azienda assume sempre più un ruolo importante derivante dalla complessità dei processi produttivi, dalle norme di sicurezza del lavoro, dagli standard qualitativi richiesti da una competitività maggiore.

Qualsiasi impianto, anche il più sofisticato e tecnologicamente avanzato, è caratterizzato da una produzione che non è mai identica: ogni prodotto finito si differenzia da un altro per delle differenze anche minime ma comunque quantificabili e controllabili.

Proprio perché la produzione è un fenomeno caratterizzato da variabilità e da scostamenti dalle specifiche richieste, è necessario studiarla utilizzando metodi statistici appropriati.

Ogni processo produttivo, indipendentemente dalla progettazione e dalle fasi di cui è composto, è sempre soggetto a una certa variabilità provocata da fattori casuali irrilevanti che lo mantengono comunque sotto controllo ma, vi sono anche fattori specifici (macchinari non ben tarati, materiali difettosi, errori degli operai) che influiscono in modo evidente generando delle non conformità.

Di fondamentale importanza è ricordare che i limiti di specifica vengono definiti da procedure standard a cui l'azienda produttrice fa riferimento o da accordi presi col cliente, indipendentemente dal comportamento naturale del processo che invece determina i limiti di tolleranza naturale.

Quando il processo mantiene costante il proprio livello di variabilità si dice in controllo e la maggior parte dei valori della grandezza considerata cadono tra i limiti d'accettabilità.

Al contrario quando vi sono delle alterazioni, la variabilità del processo subisce delle modifiche, un elevato numero di specifiche campionarie cade all'esterno dei limiti portandolo così fuori controllo.

Per il miglioramento della qualità è necessario il controllo statistico del processo produttivo e quindi l'analisi statistica dei risultati dei collaudi.

## 3.2 Controllo statistico del processo produttivo nel caso multivariato

Il controllo statistico della qualità è stato largamente usato per la sorveglianza dei processi chimici. Tecniche SPC molto conosciute sono per esempio: la carta di controllo Shewhart (Shewhart, 1924), la carta CUSUM (Page, 1954) e la carta EWMA (Roberts, 1959). Tali carte sono molto usate per monitorare processi univariati, ma non funzionano bene per processi multivariati. Per esempio, se venissero applicate p statistiche uni variate indipendenti ad un processo multivariato, il numero dei falsi

allarmi aumenterebbe. Se ci sono p variabili indipendenti in un processo e se venisse costruita una carta SPC con P{errore 1° tipo} = α per ciascuna variabile, allora la probabilità di un errore del primo tipo per l'intera procedura di controllo è data da:

$$\alpha' = 1 - (1 - \alpha)^p$$

Pertanto la vera probabilità dell'errore di primo tipo aumenta man mano che aumentano le variabili. Inoltre, se le variabili non fossero indipendenti, sarebbe difficile misurare la distorsione derivante dall'applicazione congiunta di p carte univariate.

Quindi, al fine di estrarre informazioni utili dai dati del processo e utilizzarle per monitorare il processo, è stato sviluppato il controllo statistico del processo multivariato (MSPC).

Una delle prime applicazioni dell'MSPC è quella basata sulla statistica $T^2$ di Hotelling. La carta di controllo $T^2$ di Hotelling (Mason, Tracy, Young , 1992) è considerata come la versione multivariata della carta di controllo Shewhart univariata.

L'analisi delle componenti principali (PCA) è uno strumento statistico di riduzione dimensionale di dati multivariati. La PCA trova le combinazioni lineari delle variabili che descrivono la maggior parte della variabilità presente in un data set. Per sorvegliare un processo la cui dimensione è stata ridotta via PCA, si possono utilizzare due tipi di carte di controllo, la $T^2$ e la Q (Jackson, 1991). La statistica $T^2$ è la somma dei quadrati degli score normalizzati, ed è una misura della variazione entro il modello PCA. Dall'altra parte la statistica Q è la somma dei quadrati dei residui ed è una misura della quantità della variazione non spiegata dal modello PCA. I residui sono definiti come la differenza tra i dati originali e i dati ricostruiti con diverse componenti principali.

Molte applicazioni hanno dimostrato l'utilità delle tecniche di controllo multivariate. Tuttavia i metodi convenzionali descritti prima non funzionano bene sempre, non

potendo individuare i cambiamenti di correlazione tra le variabili di processo ogni volta che le statistiche $T^2$ e Q sono all'interno dei limiti di controllo.

Per migliorare la sorveglianza multivariata del processo, in letteratura è stato introdotto un metodo basato sull'idea che un cambiamento delle condizioni operative può essere individuato monitorando la distribuzione dei dati del processo, poiché la distribuzione riflette le corrispondenti condizioni (Kano, Hasebe, Hashimoto, Ohno, 2002). In particolare, in seguito verrà descritto un indice di dissimilarità in grado di quantificare la differenza tra le distribuzioni dei dati del processo.

## 3.3 Carta di controllo $T^2$

La statistica $T^2$ di Hotelling è un approccio molto comune nelle carte di controllo multivariate. 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 (Mason, Tracy, Young , 1992).

La statistica sulla quale si basa la carta di controllo è la distanza di Mahalanobis, definita come:

$$T^2(X_i) = (X_i - \hat{\mu}\ )^T \hat{\Sigma}^{-1} (X_i - \hat{\mu}\ )$$

dove $X_i$ è un vettore casuale contenente le osservazioni sulle p variabili fatte nell''istante i-esimo, $\mu$ e $\Sigma$ sono le stime dei parametri di luogo e scala rispettivamente. Se $X_i$ si distribuisce come una normale p-variata con vettore delle medie $\mu$ e matrice di varianza e covarianza $\Sigma_{\mathrm{p}\times\mathrm{p}}$, allora $T^2(X_i)$ si distribuisce come un $\chi^2$ con p gradi di libertà (Johnson e Wichern, 2002).

La forma generale di una statistica $T^2$ si ottiene sostituendo $\hat{\mu}$ e $\hat{\Sigma}$ con il vettore medio campionario ($\overline{X}$) e la matrice di varianza e covarianza campionaria ($S$) rispettivamente. La classica $T^2$ di Hotelling è dunque definita allora come:

$$T^2(X_i) = (X_i - \bar{X})^T S^{-1}(X_i - \bar{X}) \qquad (1)$$

dove

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad , \quad S = \frac{1}{n-1}\sum_{i}^{n}(X_i - \overline{X})(X_i - \overline{X})^T$$

Ci sono due fasi distinte nel disegno della carta. La prima fase, definita come Fase I, consiste in un'analisi retrospettiva del processo per verificare che questo sia in controllo. Questa fase di avvio della sorveglianza del processo consiste nell'estrarre campioni di osservazioni in determinati istanti di tempo da un processo produttivo ritenuto in controllo e nello stimare i limiti della carta di controllo da applicare nella Fase II. La seconda fase consiste dunque nell'utilizzare la carta di controllo per individuare ogni deviazione dal processo man mano che le osservazioni future vengono rilevate.

La statistica multivariata $T^2$ è utilizzata spesso come carta statistica per entrambe le fasi di controllo.

La distribuzione di $T^2(X_i)$ viene approssimata (vedi Jackson(1985) o Ryan (1989)) ad una distribuzione Chi-quadro o ad una distribuzione F per ottenere i limiti di controllo. In particolare se si assume che le stime $\overline{X}$ e $S$ coincidono con i veri parametri della popolazione $\mu$ e $\Sigma$ rispettivamente, allora si può dimostrare (Seber (1984)) che la statistica $T^2(X_i)$ è distribuita come un Chi-quadro con p gradi di libertà. In questo caso il limite di controllo inferiore e superiore saranno rispettivamente:

$$LCL = \chi^2(1 - \alpha/2\,; p)$$

e

$$UCL = \chi^2(\alpha/2\,;p)$$

dove $\chi^2(\alpha/2\,;p)$ è il percentile di livello $1-\alpha$ della distribuzione Chi-quadro con p gradi di libertà.

Se si assume che la i-esima osservazione $X_i$ è indipendente da $\overline{X}$ e $S$, allora la statistica $T^2(X_i)$ si distribuisce come una F con p e n-p gradi di libertà. In questo caso i limite di controllo inferiore e inferiore sono pari rispettivamente a:

$$LCL = \frac{p(n-1)(n+1)}{n(n-p)}F(1-\alpha/2\,;\,p,n-p)$$

$$UCL = \frac{p(n-1)(n+1)}{n(n-p)}F(\alpha/2\,;\,p,n-p)$$

dove $F(\alpha;\,p,n-p)$ è il percentile di livello $1-\alpha$ di una distribuzione F con $p$ e $n-p$ gradi di libertà.

Poiché l'ipotesi di indipendenza tra le osservazioni e $\overline{X}$ e $S$ non vale nella fase di avvio descritta, l'approssimazione suggerita per la distribuzione della statistica $T^2(X_i)$ ha qualche svantaggio. Per esempio, finché p è piccolo, è necessario un gran campione affinché la distribuzione Chi-quadro sia adeguata per l'approssimazione.

Fortunatamente questi problemi possono essere evitati in quanto è possibile ricavare l'esatta distribuzione di $T^2(X_i)$. Gnanadesikan e Kettenring (1972) hanno infatti dimostrato, basandosi su di un risultato di Wilks (1962), che la statistica $T^2(X_i)$ (avente tempo costante) ha distribuzione beta. Nello specifico:

$$T^2 \sim \left[\frac{(n-1)^2}{n}\right] B\left(\frac{p}{2}, \frac{n-p-1}{2}\right)$$

Questa distribuzione è corretta solo quando le osservazioni singole $X_i$ raccolte durante la Fase I, vengono controllate per vedere se rientrano nei limiti di controllo.

Quando invece le osservazioni di Fase II verranno raccolte e controllate per vedere se rientrano nei limiti di controllo calcolati nella Fase I, le statistiche ottenute saranno indipendenti da $\overline{X}$ e $S$ e di conseguenza si distribuiranno esattamente come una distribuzione F.

Ora che si è a conoscenza della distribuzione esatta di $T^2(X_i)$, è possibile costruire i limiti di controllo. Il limite di controllo inferiore è dato da:

$$LCL = \left(\frac{(n-1)^2}{n}\right) B\left(1-\alpha/2\ , \frac{p}{2}, \frac{n-p-1}{2}\right) \quad (2)$$

e il limite di controllo superiore è dato da:

$$UCL = \left(\frac{(n-1)^2}{n}\right) B\left(\alpha/2\ , \frac{p}{2}, \frac{n-p-1}{2}\right) \quad (3)$$

dove $B\left(\alpha\ , \frac{p}{2}, \frac{n-p-1}{2}\right)$ è il percentile di livello $1-\alpha$ di una distribuzione beta avente parametri $\frac{p}{2}$ e $\frac{n-p-1}{2}$.

Se le tabelle per la distribuzione beta non sono facilmente disponibili, può essere utilizzata la seguente relazione

$$\frac{(p/(n-p-1))F(\alpha; p, n-p-1)}{1+(p/(n-p-1))F(\alpha; p, n-p-1)} = B(\alpha;\ p/2\ , (n-p-1)/2) \quad (4)$$

tra le variabili casuali beta e F. Applicando la (3) si possono ottenere i limiti

$$LCL = \frac{(n-1)^2}{n} \cdot \frac{(p/(n-p-1))F(1-\alpha/2\,; p, n-p-1)}{1+(p/(n-p-1))F(1-\alpha/2\,; p, n-p-1)} \quad (5)$$

e

$$UCL = \frac{(n-1)^2}{n} \cdot \frac{(p/(n-p-1))F(\alpha/2\,; p, n-p-1)}{1+(p/(n-p-1))F(\alpha/2\,; p, n-p-1)} \quad (6)$$

in termini di percentili della distribuzione F.

Una volta disegnata la carta, i valori che superano l soglie LCL e UCL sono esclusi dalla popolazione di riferimento e i limiti vengono ricalcolati sulla base dei campioni rimanenti.

In molte situazioni il limite di controllo inferiore (LCL) è posto uguale a zero. La ragione di questo è che qualsiasi shift in media porterà ad un incremento della statistica $T^2(X_i)$, e quindi l'LCL può essere ignorato. Comunque, $T^2(X_i)$, è sensibile non solo ai cambiamenti nel vettore medio ma anche ai cambiamenti nella matrice delle covarianze dei dati. Se la matrice delle covarianze cambia, si dovrebbero ottenere valori molto piccoli di $T^2(X_i)$. Da qui la decisione di scegliere un LCL diverso da zero per individuare questi tipi di cambiamenti. Va notato che grandi cambiamenti di $T^2(X_i)$, possono anche essere causati da cambiamenti nella matrice delle covarianze e non solo da cambiamenti nel vettore delle medie (Hawkins (1991)).

Una linea centrale per la carta può essere ottenuta tramite l'equazione (6) con $\alpha = 1$ (Mason, Tracy, Young, 1992).

Dopo la fase di analisi dei dati di Fase I, si può passare all'analisi dei dati di Fase II.

In questa fase si tratta di verificare se il processo corrente risulta in controllo oppure fuori controllo dovuto a cambiamenti nella media o nella varianza provocati da cause determinabili. Prendendo come riferimento il vettore medio $\hat{\mu}_m$ e la matrice di varianza e covarianza $\hat{S}_m$ (dove $m = n - a$ sono le osservazioni del campione di riferimento ripulito dai fuori controllo) calcolati sulla base dei dati in controllo ottenuti nella Fase I, 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

$$T^2{}_f(X_i) = (X_i - \hat{\mu})^T \hat{S}^{-1}(X_i - \hat{\mu}) \quad (7)$$

Va notato che data l'indipendenza delle osservazioni future dal vettore medio e dalla matrice di covarianze, allora la statistica ${T^2}_f(x_i)$ avrà una distribuzione:

$${T^2}_f(X_i) \sim \frac{p(m+1)(m-1)}{m(m-p)} \cdot F(p, m-p)$$

dove $m = n - a$ sono le osservazioni del campione di riferimento ripulito dai fuori controllo (si rimanda all'Appendice).

Perciò i limiti di controllo esatti saranno:

$$LCL = \frac{p(m+1)(m-1)}{m(m-p)} \cdot F(1-\alpha/2\,; p, m-p) \quad (8)$$

$$UCL = \frac{p(m+1)(m-1)}{m(m-p)} \cdot F(\alpha/2\,; p, m-p) \quad (9)$$

Il principale svantaggio della carta $T^2$ è di non poter agevolare l'individuazione di quale variabile o set di variabili sia responsabile di un fuori controllo poiché non fornisce informazioni utili per stabilire quali tra le variabili sono le maggior responsabili dei segnali d'allarme.

Mason, Tracy e Young (1995) proposero il seguente metodo di interpretazione di un segnale fuori controllo. La statistica $T^2$ può essere decomposta in p componenti ortogonali. La forma della decomposizione proposta è data da:

$$T^2 = T_1^2 + T_{2\cdot 1}^2 + T_{3\cdot 1,2}^2 + \cdots + T_{p\cdot 1,2,\ldots,p-1}^2 = T_1^2 + \sum_{j=1}^{p-1} T_{j1\cdot 1,\ldots,j-1}^2$$

Il primo termine della decomposizione, $T_1^2$, è una statistica $T^2$ di Hotelling incondizionata per la prima variabile

$$T_1^2 = \left(\frac{X_1 - \overline{X}_1}{s_1}\right)^2$$

dove $\overline{X}_1$ e $s_1$ sono rispettivamente la media e la deviazione standard della variabile $X_1$.

La forma generale degli altri termini, denominati termini condizionali, è data da:

$$T^2_{j\cdot 1,2,\dots j-1} = \frac{(X_j - \overline{X}_{j\cdot 1,2,\dots,j-1})^2}{s^2_{j\cdot 1,2,\dots,j-1}}, con\ j = 1,2,\dots,p$$

dove

$$\overline{X}_{j\cdot 1,2,\dots,j-1} = \overline{X}_j + b_j^T (X_{i,(j-1)} - \overline{X}_{(p-1)})$$

e $X_{i,(j-1)}$ è il (j-1)-esimo vettore che esclude la j-esima variabile, $\overline{X}_j$ è la media campionaria della variabile j-esima, $b_j = S_{XX}^{-1} \cdot s_{xX}$ è un vettore di dimensioni j-1 che stima i coefficienti di regressione della j-esima variabile sulle precedenti j-1,

$$s^2_{j\cdot 1,2,\dots,j-1} = s_x^2 - s_{xX}^T S_{XX}^{-1} \cdot s_{xX}$$

e

$$S = \begin{bmatrix} S_{XX} & s_{xX} \\ s_{xX} & s_x^2 \end{bmatrix}$$

Conseguentemente, il valore $T^2_{j\cdot 1,2,\dots j-1}$ è il quadrato della j-esima variabile aggiustata dalla stima della media e della deviazione standard delle distribuzioni condizionate di $X_j$ e la sua distribuzione esatta è la seguente:

$$T^2_{j\cdot 1,\dots,j-1} \sim \frac{n+1}{n} F(1, n-1)$$

Tramite questa si possono comparare ciascun termine ad un valore preso dalle tavole di una distribuzione F , per determinare se questa è significativa. Questo processo fornisce un meccanismo per decidere quando un termine sta segnalando un problema.

L'ordinamento delle p componenti non è unico e quello visto prima rappresenta soltanto una delle possibili $p!$ diversi modi di ordinare queste componenti. Ogni ordine genera lo stesso valore complessivo $T^2$ , ma fornisce un partizionamento di $T^2$ in p termini ortogonali. Se escludiamo le ridondanze, ci sono $p \times 2^{p-1}$ componenti distinte tra le $p \times p!$ possibili termini che dovrebbero essere valutati come potenziale aiuto per segnalare.

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. D fatti, uno dei maggiori problemi che si incontrava 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.

## 3.4 L'Analisi delle Componenti Principali

L'obbiettivo dell'analisi delle componenti principali è di ridurre la "dimensionalità", ovvero da un insieme di p variabili correlate si vuole passare ad un insieme di k variabili incorrelate (dove k<p), senza la perdita di informazioni. Le k-p variabili escluse avranno un basso contenuto informativo.

Tale tecnica consiste nel trovare p nuove variabili chiamate componenti principali, combinazioni lineari delle variabili originali centrate

$$y_1 = a_{11}c_1 + a_{12}c_2 + \cdots + a_{1p}c_p$$

$$y_2 = a_{21}c_1 + a_{22}c_2 + \cdots + a_{2p}c_p$$

$$\vdots$$

$$y_p = a_{p1}c_1 + a_{p2}c_2 + \cdots + a_{pp}c_p$$

Proprietà delle componenti principali:

$$E(y_1) = E(y_2) = \cdots = E(y_p) = 0$$

$$V(y_1) = a_{11}^2 s_1^2 + a_{12}^2 s_2^2 + \cdots + a_{1p}^2 s_p^2 + \cdots + 2a_{11}a_{12}rs_1s_2 + \cdots + 2a_{1p}a_{1(p-1)}rs_{(p-1)}s_p$$

$$V(y_2) = a_{21}^2 s_1^2 + a_{22}^2 s_2^2 + \cdots + a_{2p}^2 s_p^2 + \cdots + 2a_{21}a_{22}rs_1s_2 + \cdots + 2a_{2p}a_{2(p-1)}rs_{(p-1)}s_p$$

$$\vdots$$

$$V(y_p) = a_{21}^2 s_1^2 + a_{22}^2 s_2^2 + \cdots + a_{pp}^2 s_p^2 + \cdots + 2a_{p1}a_{p2}rs_1s_2 + \cdots + 2a_{pp}a_{p(p-1)}rs_{(p-1)}s_p$$

dove $s_1^2 = V(X_1)$, $s_2^2 = V(X_2)$,..., $s_p^2 = V(X_p)$

I vettori di coefficienti dovranno rispettare le seguenti condizioni affinché siano quelli ottimali:

a) La varianza della prima combinazione lineare sia la più grande possibile;
b) Tutte le combinazioni lineari siano incorrelate $\underline{a}_i{}'\underline{a}_j = 0$
c) Tutte le combinazioni lineari siano di norma unitaria: $\underline{a}_1{}'\underline{a}_1 = 1$ , $\underline{a}_2{}'\underline{a}_2 = 1$,..., $\underline{a}_p{}'\underline{a}_p = 1$

Per ottenere i coefficienti si deve risolvere un problema di massimo vincolato, risolto da Hotelling(1993).

- Si trovano gli autovalori della matrice

$$S = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{pmatrix} \longrightarrow l_1, l_2, \dots, l_p$$

- Si ordinano gli autovalori $l_{(1)} > l_{(2)} > \cdots > l_{(p)}$

- Si ricavano gli autovettori $l_{(1)} \Rightarrow \underline{a}_1$, $l_{(2)} \Rightarrow \underline{a}_2$, $\vdots$, $l_{(p)} \Rightarrow \underline{a}_p$ } soddisfano i vincoli a), b), c)

La percentuale di variabilità totale spiegata dalla i-esima componente principale è pari a:

$$\frac{l_{(i)}}{l_{(1)} + l_{(2)} + \cdots + l_{(p)}}$$

Se le prime k componenti spiegano una percentuale elevata della variabilità totale esse possono assumersi la rappresentativa delle p variabili.

## 3.5 Carta di controllo $T^2$ costruita col metodo PCA

La carta $T^2$ di Hotelling basata sulle PCA può essere disegnata per tutte le p componenti principali o per le prime k componenti principali.

Facendo uso delle PCA, la forma originale della statistica $T^2$ se vengono usate tutte e p le componenti principali prevede la forma seguente:

$$T^2 = \sum_{i=1}^{p} Y_i^2 \cdot l_i^{-1} \qquad (10)$$

e i valori critici per $T^2$ sono:

$$UCL = \frac{p \cdot (n+1) \cdot (n-1)}{n \cdot (n-p)} \cdot F_{(1-\alpha,p,n-p)} \quad (11)$$

$$LCL = 0 \quad (12)$$

Se invece si considerano solo le prime k componenti principali, la statistica è la seguente:

$$T_k^2 = \sum_{i=1}^{k} Y_i^2 \cdot l_i^{-1} \quad (13)$$

e i valori critici per questa sono dati da:

$$UCL = \frac{k \cdot (n+1) \cdot (n-1)}{n \cdot (n-k)} \cdot F_{(1-\alpha,k,n-k)} \quad (14)$$

$$LCL = 0 \quad (15)$$

Quindi se un valore di $T^2$ è più grande del valore critico UCL il processo produttivo viene definito fuori controllo.

## 3.6 Carta di controllo Q

L'uso esteso 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 verificare la presenza di eventuali valori anomali. A tal 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 delle osservazioni 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 catturare l'ammontare di variazione del processo che non viene colta dal modello basato sulle componenti principali.

Il termine residuo Q può essere controllato mediante la somma dei quadrati dei residui:

$$Q = \left(X - \hat{X}\right)^T \left(X - \hat{X}\right) = \sum_{i=k+1}^{p} l_i t_i^2 = \sum_{i=k+1}^{p} Y_i^2 \qquad (16)$$

dove $\hat{x} = \overline{x} + Uy$ , $U$ è $p \times k$ e $y$ è $k \times 1$.

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

$$Q_\alpha = \vartheta_1 \left[ \frac{c_\alpha \sqrt{2\vartheta_2 h_0^2}}{\vartheta_1} + \frac{\vartheta_2 h_0 (h_0 - 1)}{\vartheta_3^2} + 1 \right]^{\frac{1}{h_0}} \qquad (17)$$

dove $c_\alpha$ è il quantile di livello $(1 - \alpha)$ di una normale $N(0,1)$ e:

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

Un altro test statistico per i residui è stato proposto da Hawkins (1991) utilizzando la somma non ponderata dei quadrati delle p-k componenti principali escluse dal modello

$$T_i^2 = t_{k+1}^2 + \cdots + t_p^2$$

la quale è distribuita secondo una distribuzione F con p-k gradi di libertà al numeratore e n-p+k gradi di libertà al denominatore. I limiti di controllo per questa statistica saranno quindi:

$$UCL = \frac{k \cdot (n-1) \cdot (n+1)}{n \cdot (n-1) - k + 1} \cdot F_{(1-\alpha, p-k, n-p+k)} \qquad (18)$$

$$LCL = 0 \qquad (19)$$

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 a 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 tal proposito nel successivo paragrafo viene descritto un nuovo metodo innovativo capace di tener conto della dipendenza temporale delle osservazioni.

## 3.7 Indice di Dissimilarità

Il concetto di similarità o dissimilarità è spesso usato per classificare un data set. Nella cluster analysis, per esempio, la dissimilarità tra due classi è misurata attraverso la differenza dei baricentri dei dati, e due classi col più basso grado di dissimilarità vengono unite per generare una nuova classe.

Per valutare la differenza tra due diversi data set, in questa tesi lavoro viene usato un metodo di classificazione basatosull'espansione di Karhunen-Loeve (KL) proposta da Fukunaga e Koontz (1970). L'espansione KL è una tecnica ben conosciuta per la riduzione della dimensionalità, ed è matematicamente equivalente alla PCA.

Consideriamo i seguenti due data set:

$$X_i = \begin{bmatrix} x_{11}^{(i)} & x_{12}^{(i)} & \cdots & x_{1P}^{(i)} \\ x_{21}^{(i)} & x_{22}^{(i)} & \cdots & x_{2P}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N_i1}^{(i)} & x_{N_i2}^{(i)} & \cdots & x_{N_iP}^{(i)} \end{bmatrix}, \quad i = 1,2$$

dove $N_i$ è il numero dei campioni dell'i-esimo data set $X_i$ e p è il numero di variabili. Ogni colonna di $X_i$ si assume sia centrata in media. La matrice delle covarianze è data da:

$$R_i = \frac{1}{N_i - 1} X_i^T X_i$$

e la matrice delle covarianze dell'unione di entrambi i data set è dato da:

$$R = \frac{1}{N-1} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}^T \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

$$= \frac{N_1 - 1}{N-1} R_1 + \frac{N_2 - 1}{N-1} R_2$$

Applicando la decomposizione degli autovalori alla matrice R, si ottiene una matrice ortogonale $P_0$ che soddisfa

$$RP_0 = P_0\Lambda$$

dove $\Lambda$ è una matrice dove i cui elementi sulla diagonale sono gli autovalori della matrice R. Con una matrice trasformata P definita come

$$P = P_0 \Lambda^{-1/2}$$

si ottiene la seguente equazione

$$P^T R P = I$$

Quando le matrici di dati $X_i$ sono trasformate come

$$Y_i = \sqrt{\frac{N_i - 1}{N - 1}} X_i P_0 \Lambda^{-1/2}$$

le matrici di covarianze dei dati trasformati

$$S_i = \frac{1}{N_i - 1} Y_i^T Y_i = \frac{N_i - 1}{N - 1} P^T R_i P \quad (20)$$

soddisfano la seguente equazione:

$$S_1 + S_2 = I \quad (21)$$

Applicando la decomposizione degli autovalori alle matrici di covarianza si ottiene

$$S_i w_j^{(i)} = \lambda_j^{(i)} w_j^{(i)} \quad (22)$$

Dalle equazioni (21) e (22) si ricava

$$S_2 w_j^{(1)} = \left(1 - \lambda_j^{(1)}\right) w_j^{(1)} \quad (23)$$

L'equazione (23) implica che gli autovettori di $S_2$ sono gli stessi di $S_1$ e che la seguente relazione è soddisfatta

$$1 - \lambda_j^{(1)} = \lambda_j^{(2)}$$

Come risultato, dato che gli autovettori della matrice di covarianze rappresentano le direzione delle componenti principali e gli autovalori sono equivalenti alle varianze

delle componenti principali, a ciascun data set trasformato corrisponde lo stesso set di componenti principali mentre i corrispondenti autovalori delle matrici di covarianze sono ordinati inversamente.

Così dopo le precedenti trasformazioni, la correlazione più importante per il primo data set è equivalente alla correlazione meno importante del secondo e viceversa. Quando i data set sono quasi simili tra di loro, gli autovalori $\lambda_j^{(i)}$ devono essere pari a 0.5. Dall'altra parte, quando i data set sono diversi tra di loro, il più grande e il più piccolo autovalore dovrebbero essere vicini a uno e a zero rispettivamente. In conclusione, viene definito il seguente indice D per valutare la dissimilarità dei data set:

$$D = \frac{4}{P}\sum_{j=1}^{P}(\lambda_j - 0.5)^2 \qquad (24)$$

L'indice (7) varia tra zero e uno. Quando due data set sono simili tra di loro, D è vicino allo zero;  mentre è uguale a uno quando i data set sono differenti tra di loro.

# CAPITOLO 4

# ANALISI DEI DATI

L'analisi esplorativa dei dati verrà nel seguito svolta facendo uso prevalentemente dei boxplot e dei grafici della correlazione. Verificata l'ipotesi di normalità, si procederà con il disegno della carta $T^2$, prima su di un campione di Fase 1, e in seguito su di un campione di Fase 2. La carta Q e la carta $T^2$ verranno dunque disegnate per la sorveglianza delle componenti principali allo scopo di suggerire agli operatori una possibile riduzione della dimensionalità del processo osservato. In seguito verrà dunque applicato l'Indice di dissimilarità per poter individuare cambiamenti nella struttura delle relazioni tra variabili nel tempo, cambiamenti che, come anticipato, non possono essere individuati dalle carte Q e $T^2$.

**Matrice in ingresso**

La matrice in ingresso è costituita da 8 variabili:

- **Umidit_R** (Umidità residua): variabile che misura la quantità di umidità residua, espressa in percentuale, presente dopo la spremitura delle matrici in ingresso (FASE DI SPREMITURA) ;
- **Cadmio**: variabile che misura la quantità in mg/kg s.s. di cadmio presente nel campione;
- **Cromo_Tot** (Cromo totale): variabile che misura la quantità totale di cromo in mg/kg s.s. presente nel campione;
- **Mercurio**: variabile che misura la quantità in mg/kg s.s. di mercurio presente nel campione;

- **Nichel**: variabile che misura la quantità in mg/kg s.s. di nichel presente nel campione;
- **Piombo**: variabile che misura la quantità in mg/kg s.s. di piombo presente nel campione;
- **Rame**: variabile che misura la quantità in mg/kg s.s. di rame presente nel campione;
- **Zinco**: variabile che misura la quantità in mg/kg s.s. di zinco presente nel campione;

La Tabella 1 mostra i limiti di specifica (LSL=Limite di Specifica Inferiore e USL= Limite di Specifica Superiore) per le variabili in ingresso.

| ELEMENTO CHIMICO | LSL | USL |
|---|---|---|
| Cadmio | 0 | 20 |
| Cromo totale | 0 | 750 |
| Mercurio | 0 | 10 |
| Nichel | 0 | 300 |
| Piombo | 0 | 750 |
| Rame | 0 | 1000 |
| Zinco | 0 | 2500 |

**Tabella 1: Tabella dei limiti di specifica per le 8 variabili in ingresso**

Per l'umidità residua non è previsto un limite di specifica, ma viene considerata nell'analisi per poter verificare la presenza di correlazione tra questa e le altre 7 variabili.

Sono stati raccolti 19 campioni per il F.O.R.S.U., ognuno in un diverso istante temporale.

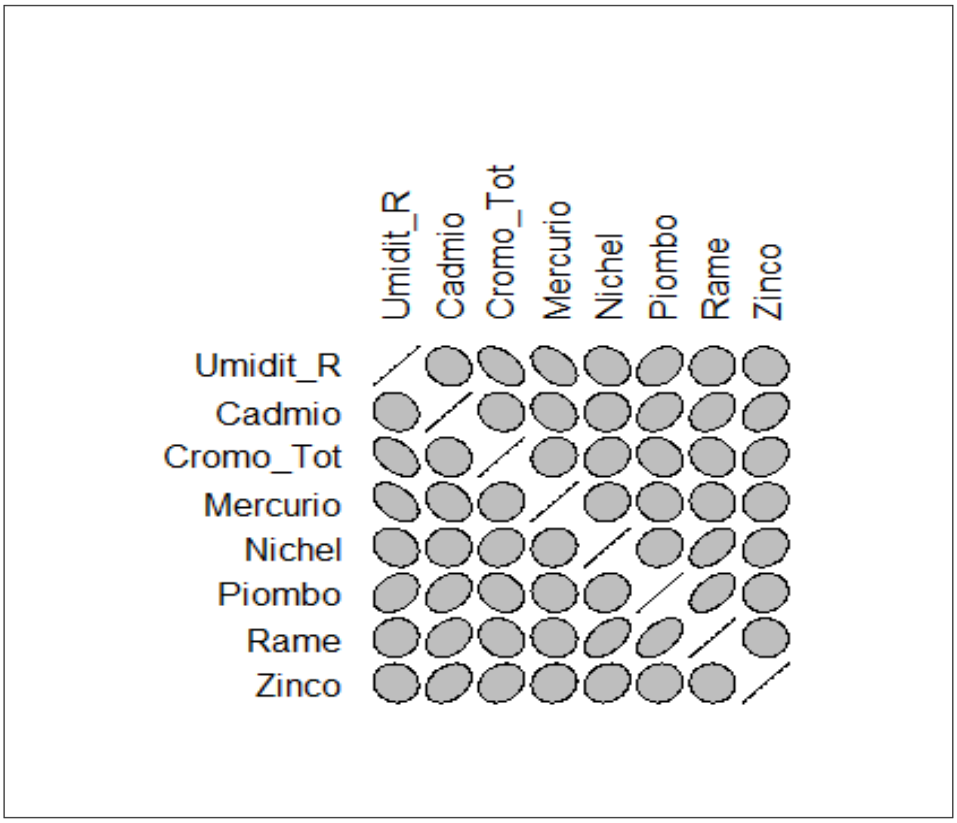


**Figura 4: Grafico delle correlazoni presenti tra le variabili**

Nella Figura 4 è rappresentata graficamente la correlazione esistente tra le variabili: cromo – umidità residua (negativa), mercurio – umidità residua (negativa), piombo – umidità residua, piombo – cadmio, rame – piombo, rame – nichel, rame – cadmio, zinco – cadmio.

Per le prime 2 correlazioni (cromo – umidità residua, mercurio – umidità residua) si hanno i seguenti valori per l'indice di correlazione: $-0.495$ e $-0.462$. Il fatto che sia negativa indica che all'aumentare del cromo o del mercurio l'umidità residua diminuisce. Le restanti correlazioni sono positive e hanno valori per l'indice di correlazione che si aggirano intorno lo $0.318$ e lo 0.529.

La Figura 5 rappresenta graficamente la distribuzione di ciascuna delle variabili per il F.O.R.S.U., standardizzate.

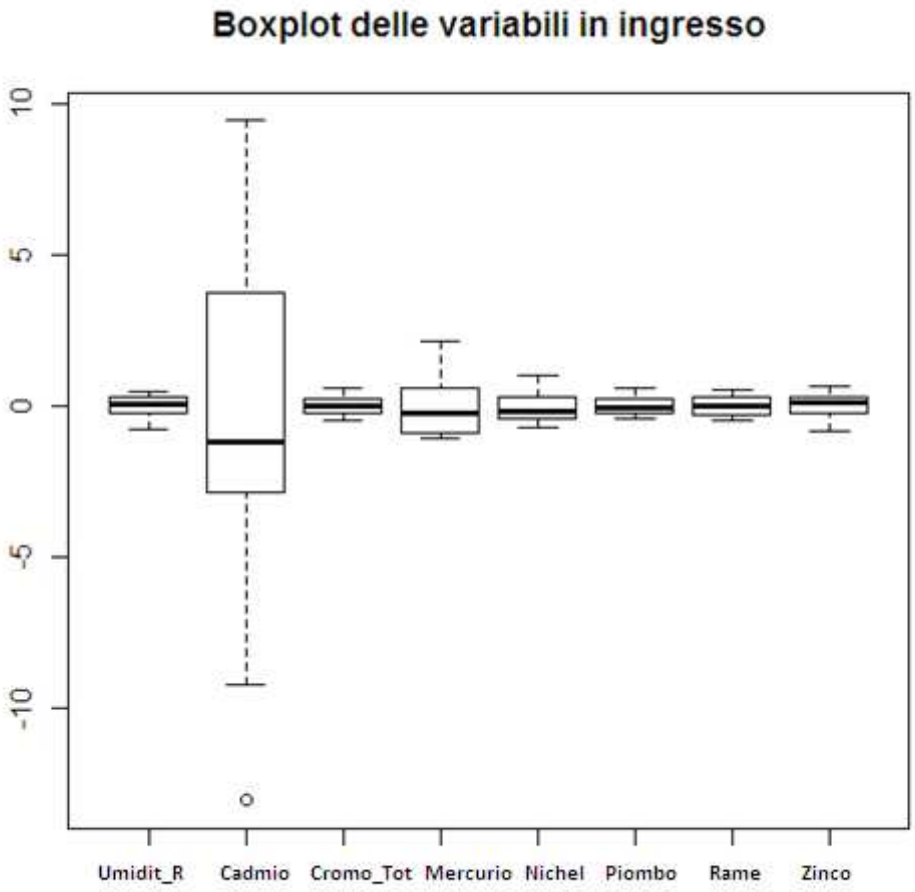


**Figura 5: Boxplot delle variabili in ingresso**

I box-plot mostrano che le distribuzioni sono pressoché simmetriche e con una variabilità contenuta, tranne che per la variabile cadmio che presenta un outliers e un'asimmetria che porterebbero a rifiutare l'ipotesi di normalità.

Il grafico della probabilità normali della variabile cadmio (Figura 6) mostra infatti la presenza di una coda pesante sulla sinistra.

In particolare, il test di Shapiro-Wilk applicato su ciascuna variabile mi fornisce un p-value superiore ad un $\alpha$ fissato pari a 0.05, quindi accetto l'ipotesi di normalità delle variabili in ingresso, compresa la variabile cadmio la cui normalità era stata messa in dubbio dal boxplot e dal QQPlot.

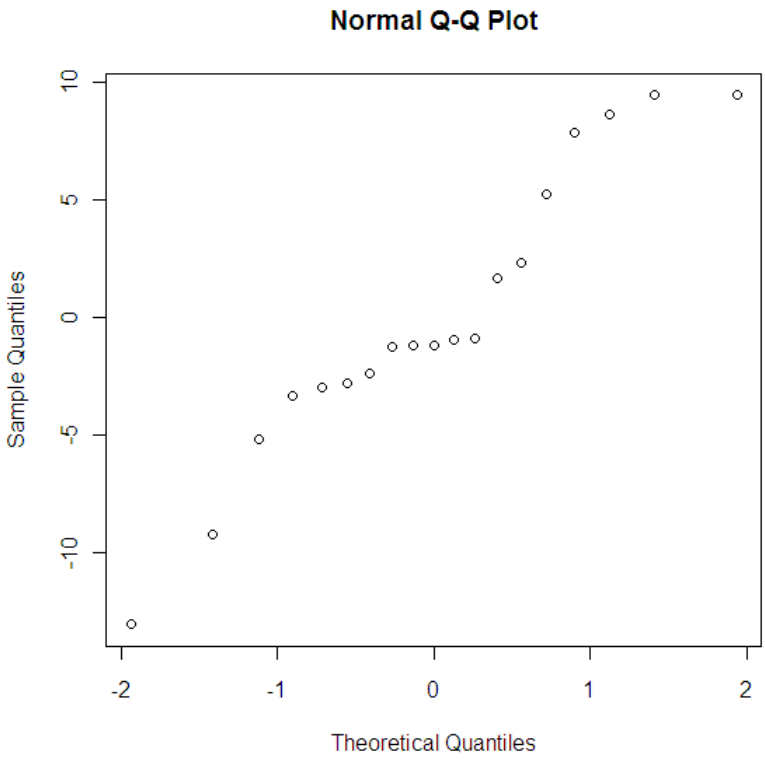


**Figura 6: Grafico QQPlot per la variabile cadmio**

Vista la grande differenza in variabilità tra alcune variabili come zinco e cadmio, l'ipotesi di omoschedasticità viene scartata.

Purtroppo non è possibile analizzare la struttura di autocorrelazione tramite i grafici di autocorrelazione parziale e campionaria poiché non vi è la struttura temporale necessaria perché tale analisi abbia un senso. Questa conclusione è data dal fatto che il F.O.R.S.U. viene campionato ogni 10 mila tonnellate, per cui la cadenza temporale con cui viene analizzato non è regolare.

**Matrici in uscita**

In uscita dal processo di maturazione abbiamo 17 variabili:

- **pH**: variabile che misura il livello del pH del compost in uscita ;
- **Salinità**: variabile che misura il livello di salinità in meq/100 g presente nel campione;

- **Umidità**: variabile che misura la quantità in percentuale di umidità presente nel campione;
- **Carbonio_org** (Carbonio organico): variabile che misura la quantità in percentuale su s.s. di carbonio organico presente nel campione;
- **Acidi_Umici_Fulvici**: variabile che misura la quantità di acidi umici fulvici presenti nel campione;
- **Azoto_org** (Azoto organico): variabile che misura la quantità in percentuale su s.t. di azoto organico presente nel campione;
- **Rapporto_CN** (Rapporto Carbonio Azoto): variabile che misura il rapporto tra carbonio e azoto presente nel campione;
- **Cadmio**: variabile che misura la quantità in mg/kg s.s. di cadmio presente nel campione;
- **Cromo_tot** (Cromo totale): variabile che misura la quantità totale in mg/kg s.s. di cromo presente nel campione;
- **Mercurio**: variabile che misura la quantità in mg/kg s.s. di mercurio presente nel campione;
- **Nichel**: variabile che misura la quantità in mg/kg s.s. di nichel presente nel campione;
- **Piombo**: variabile che misura la quantità in mg/kg s.s. di piombo presente nel campione;
- **Rame**: variabile che misura la quantità in mg/kg s.s. di rame presente nel campione;
- **Zinco**: variabile che misura la quantità in mg/kg s.s. di zinco presente nel campione;
- **Mat_plastico** (Materiale plastico): variabile che misura la quantità in percentuale su s.s. di materiale plastico presente nel campione;
- **Vetro**: variabile che misura la quantità in percentuale su s.s. di vetro presente nel campione;

- **Metalli**: variabile che misura la quantità in percentuale su s.s. di metalli presenti nel campione

La Tabella 2 mostra i limiti di specifica per tali variabili.

| VARIABILE | LSL | USL |
|---|---|---|
| pH | 6 | 8.5 |
| Salinità | 0 | 200 |
| Umidità | 0 | 50 |
| Carbonio_org | 20 | 100 |
| Acidi_Umici_Fulvici | 7 | 100 |
| Azoto_org | 80 | 100 |
| Rapporto_CN | 0 | 25 |
| Cadmio | 0 | 1.5 |
| Cromo_tot | 0 | 150 |
| Mercurio | 0 | 1.5 |
| Nichel | 0 | 100 |
| Piombo | 0 | 140 |
| Rame | 0 | 230 |
| Zinco | 0 | 500 |
| Mat_plastico | 0 | 0.5 |
| Vetro | 0 | 0.5 |
| Metalli | 0 | 0.5 |

**Tabella 2: Tabella dei limiti di specifica per le 17 variabili in uscita**

Per il compost sono stati raccolti 19 campioni, ognuno in un diverso istante temporale. Analizziamo innanzitutto la correlazione esistente tra le variabili.

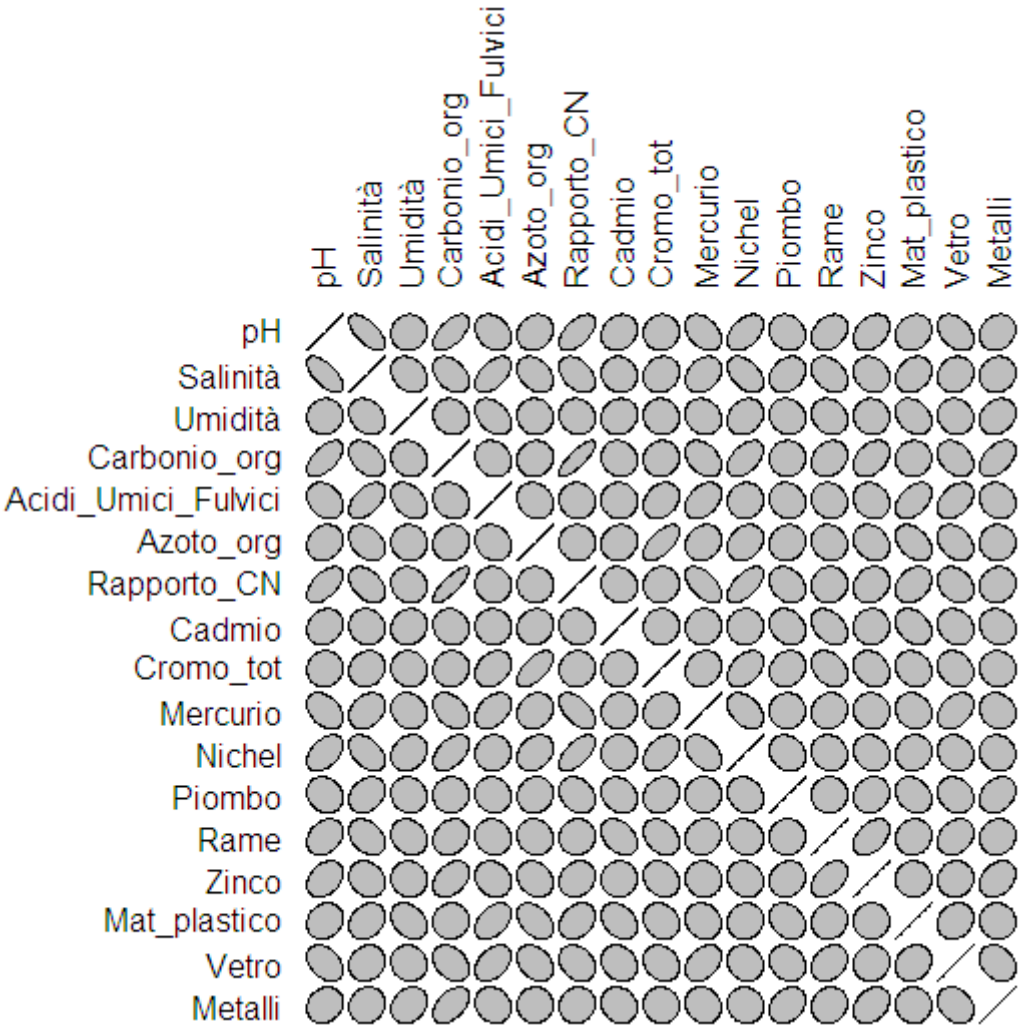


**Figura 7: Grafico delle correlazioni tra tutte le variabili**

La Figura 7 mostra la presenza di correlazione per diverse variabili, in particolare le più evidenti sono: rapporto carbonio/azoto – carbonio organico, salinità – pH (negativa), carbonio organico – pH, mercurio – azoto organico, nichel – rapporto carbonio/azoto, cromo totale – azoto organico, mercurio – rapporto carbonio/azoto (negativa).

Per le correlazioni negative i valori dell'indice di correlazione sono -0.654 (salinità – pH) e -0.560 (mercurio – rapporto carbonio/azoto), mentre per le correlazioni

positive, quella più forte è tra rapporto carbonio/azoto – carbonio organico con un valore dell'indice di correlazione pari a 0.820. Quest'ultima conclusione è ovvia dato che se aumenta il carbonio, aumenterà pure il rapporto tra carbonio e azoto.

Visualizziamo graficamente i box-plot delle variabili standardizzate. Viste le dimensioni del data set, decidiamo di suddividere i grafici boxplot in 3 parti, aventi la medesima scala (Figure 8-10).

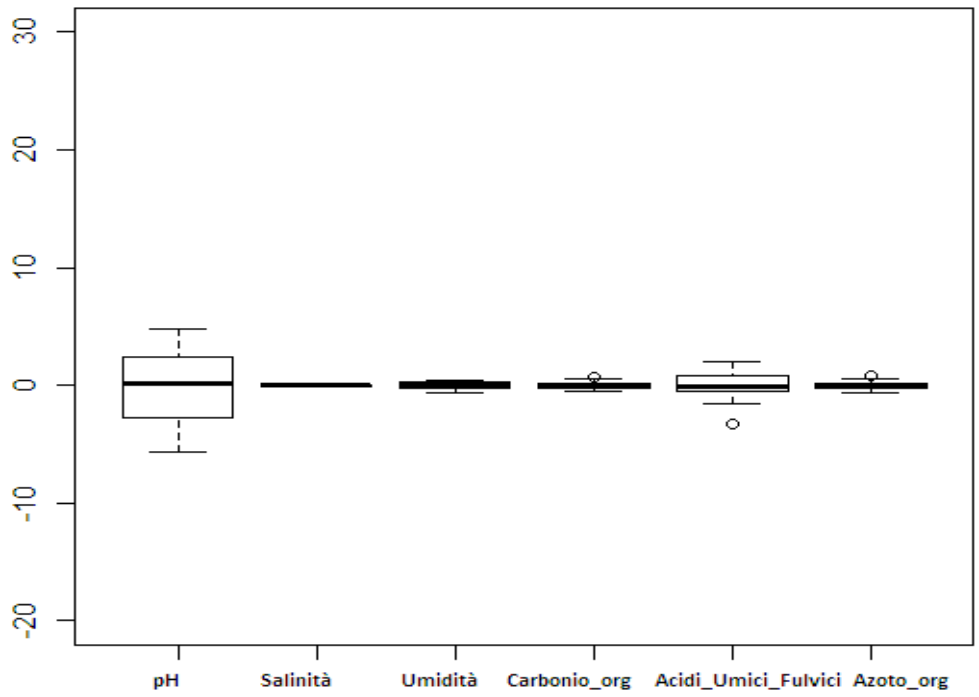


**Figura 8: Boxplot delle variabili pH, salinità, umidità, carbonio organico, acidi umici fulvici e azoto organico**

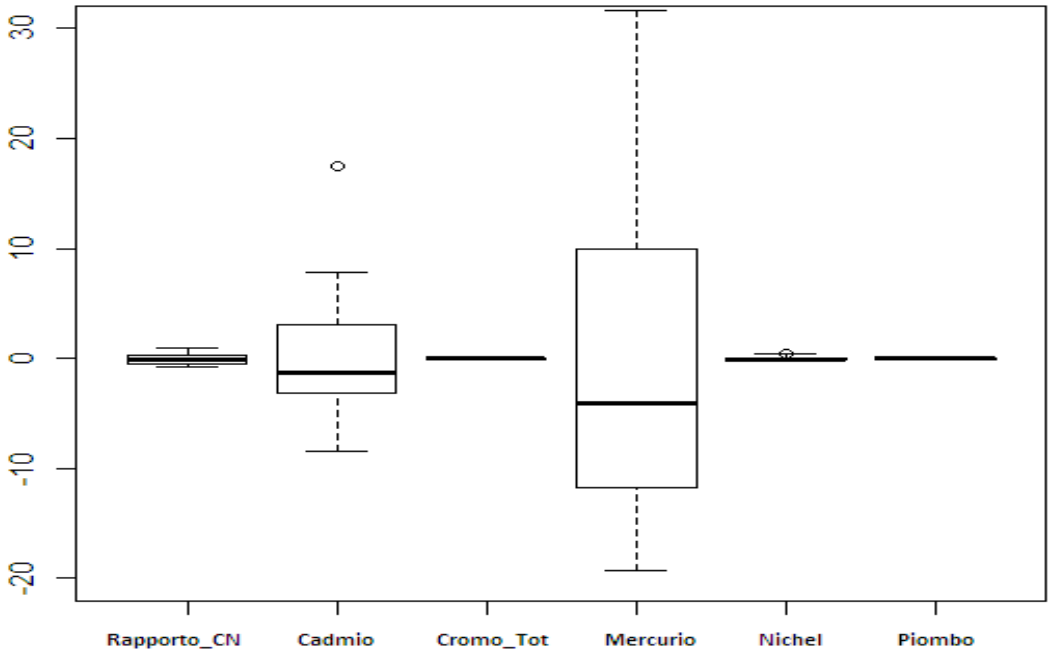


**Figura 9: Boxplot delle variabili rapporto carbonio/azoto, cadmio, cromo totale, mercurio, nichel e piombo**

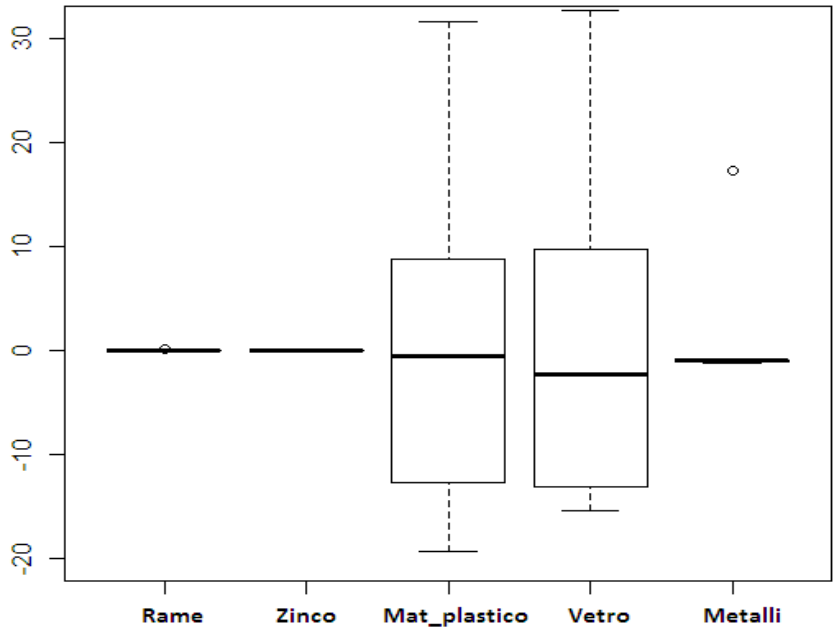


**Figura 10: Boxplot delle variabili rame,zinco, materiale plastico, vetro e metalli**

Possiamo notare la presenza di diversi outliers nelle variabili metalli, rame, nichel, carbonio organico e azoto organico.

Per studiare la normalità posso fare affidamento anche in questo caso al test di Shapiro-Wilk. Tale test porta in particolare al rifiuto dell'ipotesi di normalità per la variabile metalli (vedi Appendice). La media per tale variabile è pari a $7.589 \cdot 10^{-17}$ e il valore massimo è $17.27$ che fa pensare possa essere un outliers. Una conferma di questo la possiamo avere dal box-plot visto in precedenza relativo a tale variabile.

Come era stato detto in precedenza il valore sospetto della causa di non normalità è proprio un outliers rilevato al quinto istante. Un nuovo test di Shapiro-Wilk senza tale valore produce un p-value che rientra nella zona di accettazione (vedi Appendice).

Purtroppo anche per il compost non è possibile analizzare la struttura di autocorrelazione tramite i grafici di autocorrelazione parziale e campionaria poiché non vi è la struttura temporale necessaria perché tale analisi abbia un senso.

Ora che anche quest'analisi è stata conclusa si può passare alla fase principale, ovvero alla fase di monitoraggio del processo produttivo, che comincia con il disegno delle carte di controllo.

# Costruzione delle carte di controllo

Valutando quanto visto nell'analisi descrittiva dei dati, in particolare la validità dell'assunzione di normalità, posso applicare ora le tecniche MSPC introdotte nei precedenti capitoli. Il primo passo da fare per entrambe le matrici però è rendere adimensionali i dati, visto che la matrice è composta da elementi misurati secondo unità di misura diversa le una dalle altre.

### Carte di controllo per la matrice in ingresso

Per l'analisi delle matrice in ingresso, si comincerà con la carta basata sulla statistica $T^2$ di Hotelling. Per costruire tale statistica non sono più necessari il vettore medio della matrice in ingresso $\mu$ e la matrice di varianze covarianze della matrice in ingresso $\Sigma$ secondo la (1), poiché si stanno considerando dati standardizzati, la sola matrice di correlazione R.

Una volta stimata usando le osservazioni di Fase I (vedi Tabella 3) si possono calcolare le seguenti quantità:

$$T^2(X_i) = (X_i)^T R^{-1}(X_i) \quad (25)$$

che, sotto particolari condizioni viste in precedenza (vedi capitolo 3.3), si distribuiscono come una distribuzione **beta** di parametri $\frac{p}{2}$ e $\frac{n-p-1}{2}$, ovvero 4 e 5.

Quindi, posto $\alpha$ pari a 0.05, i valori della (25) dovranno essere confrontati con il seguente limite di controllo superiore:

$$UCL = \left(\frac{(n-1)^2}{n}\right) B\left(\alpha/2\ , \frac{p}{2}, \frac{n-p-1}{2}\right) = 12.87707 \cong 12.88$$

$$LCL = 0$$

Il limite di controllo inferiore stato posto pari a zero poiché qualsiasi shift in media porterà ad un incremento della statistica $T^2$. Si ha quindi la seguente carta di controllo

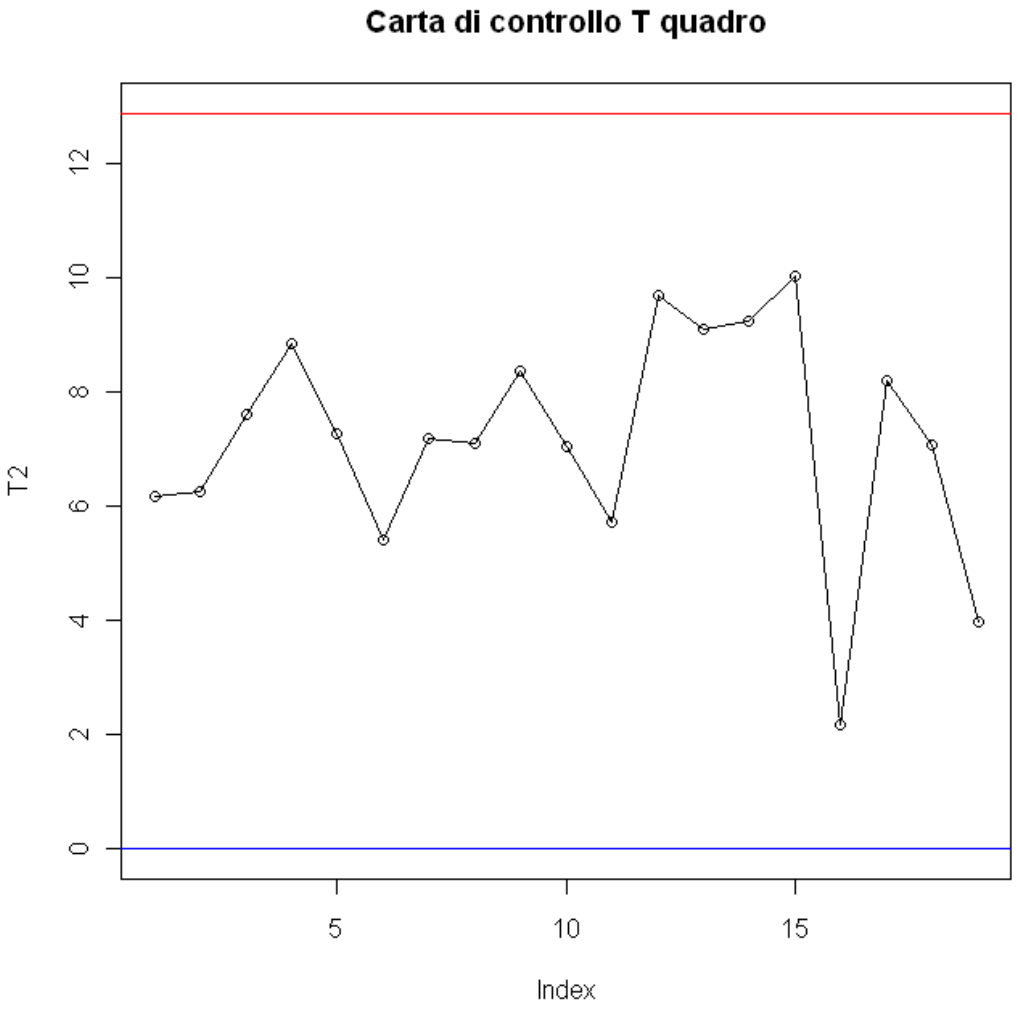


**Figura 11: Carta di controllo $T^2$ per i dati in ingresso di Fase I**

Come è possibile notare non sono presenti fuori controllo. Si può quindi affermare che i campioni di Fase I provengono da un processo in controllo. Completata la Fase I, ovvero la fase di controllo dei dati storici, si passa alla Fase II, durante la quale un le statistiche di controllo calcolate su un nuovo campione di dati, composto da 19 osservazioni riguardanti le 8 variabili d'interesse, vengono messo a confronto con i limiti di controllo calcolati nella Fase I.

La statistica (7) viene calcolata utilizzando la matrice $\hat{R}$ ottenuta durante la prima fase, e tramite questa verrà calcolata la statistica $T^2$ per i nuovi campioni.

I valori ottenuti vengono dunque confrontati con i limiti di controllo introdotti da Mason, Tracy e Young (1992):

$$UCL = \frac{p(n+1)(n-1)}{n(n-p)} \cdot F(\alpha/2\,; p, n-p) = 50.48708 \cong 50.48$$

$$LCL = 0$$

Come si è visto nel paragrafo 3.3 la statistica di controllo si distribuisce sotto l'ipotesi nulla come una F di parametri $p$ ed $n-p$ a causa dell'indipendenza tra le osservazioni e i parametri $\hat{\mu}$ e $\hat{S}$, in questo caso tra le osservazioni e $\hat{R}$. Quindi si usano i percentili della distribuzione in controllo per stabilire se un'osservazione è in controllo o meno.

La carta di controllo per i dati di Fase II è rappresentata nella Figura 12.

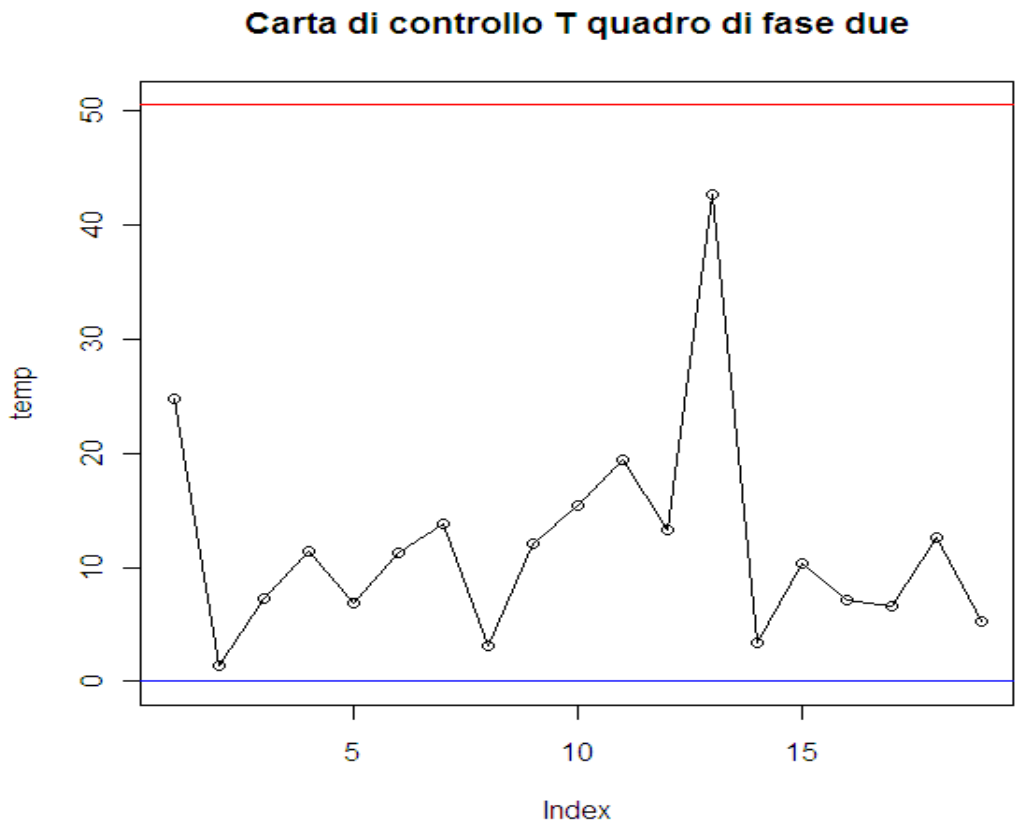


**Figura 12: Carta di controllo $T^2$di per i dati in ingresso di Fase II**

Il processo produttivo sembra dunque in controllo.

Applico ora il metodo delle componenti principali al controllo statistico della qualità, con l'obbiettivo primario di suggerire agli operatori la sorveglianza di un numero ridotto di variabili che si assume possano essere le maggiori responsabili di disturbi e variazioni nel processo. Da quanto visto nel paragrafo 3.5, si riscalano le variabili dividendole per le loro radici caratteristiche.

Dall'analisi delle componenti principali si ottengono gli autovettori, gli autovalori e le percentuali di varianza spiegata riportati nella Tabella 4.

| Autovettori | Comp. 1 | Comp. 2 | Comp. 3 | Comp. 4 | Comp. 5 | Comp. 6 | Comp. 7 | Comp. 8 |
|---|---|---|---|---|---|---|---|---|
| Umidit_R | 0.359 | -0.484 | 0.019 | 0.142 | -0.507 | -0.005 | 0.159 | 0.623 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Cadmio | 0.354 | 0.259 | 0.524 | -0.210 | 0.469 | 0.125 | -0.133 | 0.430 |
| Cromo_Tot | -0.348 | 0.373 | 0.179 | 0.446 | -0.048 | -0.607 | 0.192 | 0.396 |
| Mercurio | -0.295 | 0.264 | -0.395 | -0.691 | -0.177 | -0.038 | 0.087 | 0.457 |
| Nichel | 0.154 | 0.491 | -0.366 | 0.409 | -0.209 | 0.364 | -0.474 | 0.144 |
| Piombo | 0.518 | 0.103 | -0.079 | -0.238 | -0.168 | -0.656 | -0.382 | -0.145 |
| Rame | 0.491 | 0.303 | -0.337 | 0.082 | 0.175 | 0.023 | 0.709 | |
| Zinco | 0.031 | 0.379 | 0.530 | -0.163 | -0.620 | 0.220 | 0.193 | -0.140 |
| | **Comp. 1** | **Comp. 2** | **Comp. 3** | **Comp. 4** | **Comp. 5** | **Comp. 6** | **Comp. 7** | **Comp. 8** |
| **Autovalori** | 2.17 | 1.87 | 1.21 | 0.915 | 0.808 | 0.617 | 0.255 | 0.134 |

**Tabella 3: Tabella degli autovalori e autovettori per i dati del F.O.R.S.U.**

La variabilità totale dei dati è rappresentata dalle 8 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 (di solito si scelgono quelle che spiegano tra il 70 e il 90 per cento della varianza). Si è scelto qui quindi di prendere in esame tre componenti principali per la costruzione della carta di controllo considerando che con queste si ottiene una varianza spiegata pari all'82.4%.

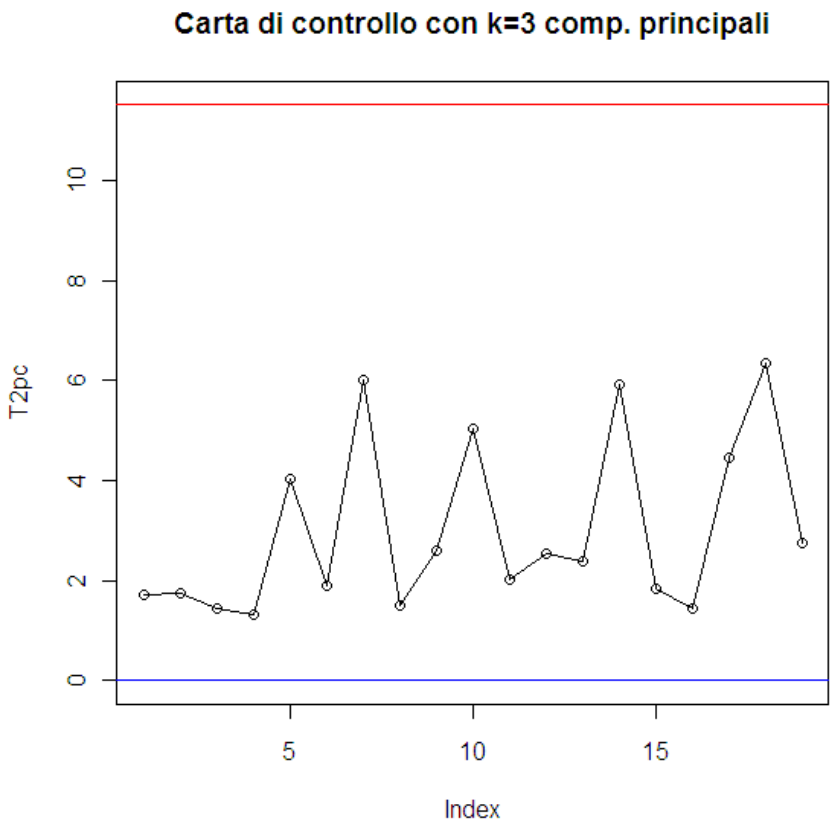


**Figura 13: Carta $T^2$ basata sulle prime 3 componenti per i dati del F.O.R.S.U.**

Dal momento che le prime tre componenti spiegano una buona porzione di variabilità, la carta di controllo, come nel caso in cui si considerano tutte le variabili, non evidenzia alcun fuori controllo (Figura 13). Passiamo alla costruzione della carta Q, applicata alle PCA. Il primo passaggio è calcolare i valori della statistica di controllo tramite l'equazione (16) e quella dei limiti di controllo tramite l'equazione (17) , ovvero:

$$UCL = \vartheta_1 \left[ \frac{c_\alpha \sqrt{2\vartheta_2 h_0^2}}{\vartheta_1} + \frac{\vartheta_2 h_0 (h_0 - 1)}{\vartheta_3^2} + 1 \right]^{\frac{1}{h_0}} = 7.908$$

$$LCL = 0$$

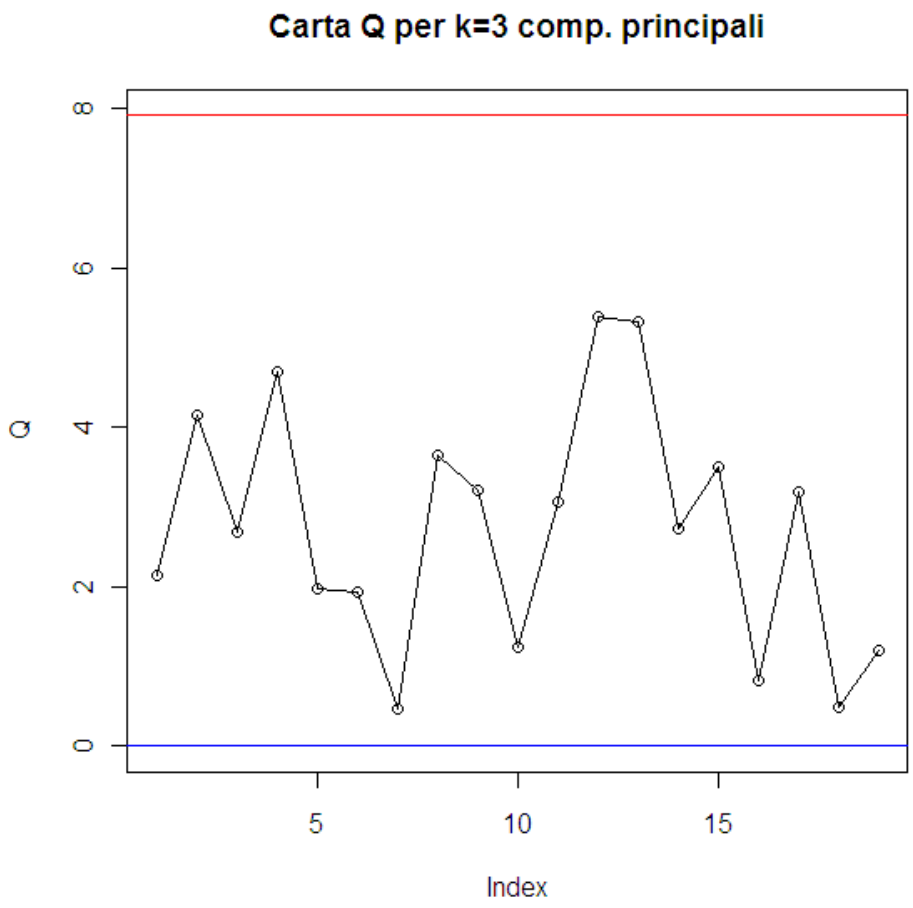


**Figura 14: Carta Q basata sulle prime 3 componenti per i dati del F.O.R.S.U.**

Anche dalla Figura 14 si evidenzia come il processo produttivo, la cui variabilità originale è rappresentata dalle sole prime tre componenti principali, risulta essere in controllo.

È possibile pure usare la variante proposta da Hawkins (1991), la quale considera le componenti escluse dalla statistica $T^2$ (Figura 15). Secondo questa variante la 15^esima osservazione è molto vicina al limite superiore, ma il processo sembra sostanzialmente in controllo.

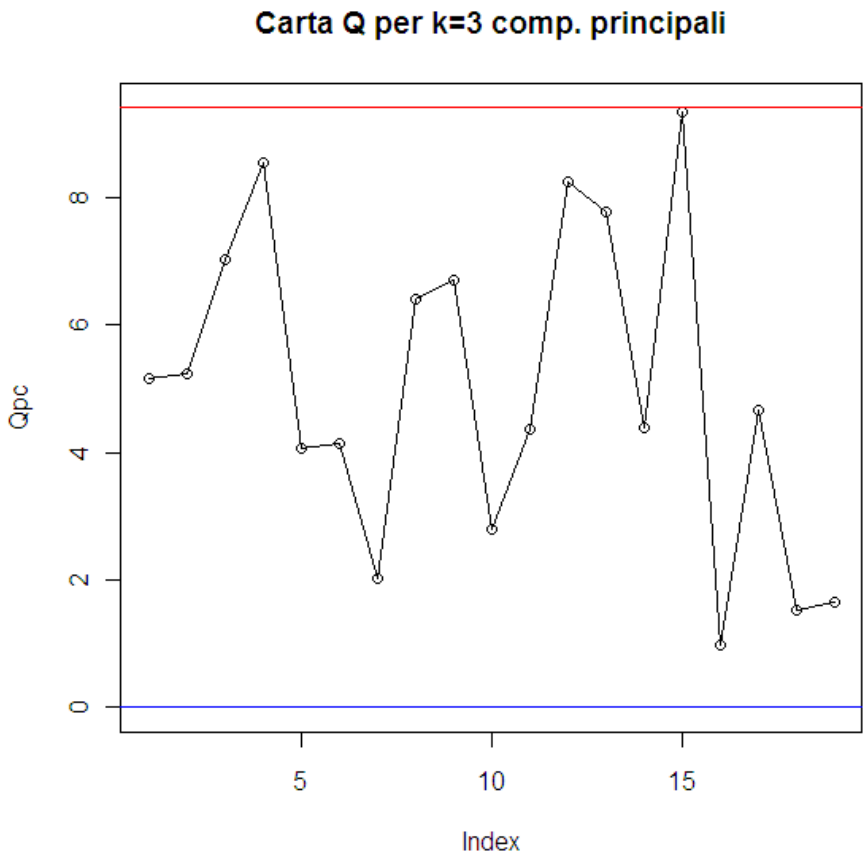


**Figura 15: Carta Q basata sulle prime 3 componenti basata sui dati del F.O.R.S.U. calcolata col metodo di Hawkins (1991)**

Applichiamo ora l'Indice di Dissimilarità, proposto da Kano (2001) per sorvegliare insiemi di dati multivariati e autocorrelati.

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 dell'indice, si è scelto di eseguire l'Indice di Dissimilarità (indice D per abbreviazione) per un valore di w pari a 10.

La procedura è la seguente:

1. Si osserva una matrice $X_{n\times p}$ utilizzando i dati provenienti da un processo operante in condizioni di normalità;
2. Le colonne di tale matrice vengono standardizzate in maniera tale da ottenere osservazioni aventi media nulla e varianza unitaria;

3. Si sceglie una dimensione della finestra temporale w (dove in questo caso w=10);
4. Facendo scorrere tale finestra lungo l'asse temporale si generano sottomatrici di dimensioni (n-w+1)
5. Si calcolano per ogni matrice ottenuta al passo 4 i valori dell'Indice di Dissimilarità

Si è suddiviso il data set del F.O.R.S.U. in due sottomatrici composte rispettivamente da 9 e 10 campioni. Si è calcolata poi la matrice R di varianza e covarianza tra le due matrici e si è applicata a questa l'analisi delle componenti principali ottenendo gli autovettori e autovalori riportati nella Tabella 5.

| **Autovettori** | **Comp. 1** | **Comp. 2** | **Comp. 3** | **Comp. 4** | **Comp. 5** | **Comp. 6** | **Comp. 7** | **Comp. 8** |
|---|---|---|---|---|---|---|---|---|
| Var. 1 | 0.007 | 0.203 | 0.058 | -0.304 | 0.569 | -0.076 | 0.280 | 0.674 |
| Var. 2 | -0.998 | -0.038 | 0.017 | -0.008 | -0.016 | -0.012 | 0.018 | 0.021 |
| Var. 3 | 0.005 | -0.037 | -0.108 | 0.487 | -0.275 | 0.648 | 0.164 | 0.478 |
| Var. 4 | 0.044 | -0.969 | 0.115 | -0.142 | 0.066 | -0.016 | 0.061 | 0.133 |
| Var. 5 | -0.001 | -0.085 | -0.906 | -0.009 | -0.091 | -0.279 | 0.292 | 0.024 |
| Var. 6 | -0.015 | 0.002 | -0.104 | -0.319 | 0.313 | 0.620 | 0.400 | -0.495 |
| Var. 7 | -0.018 | -0.01 | -0.342 | -0.407 | 0.042 | 0.335 | -0.759 | 0.167 |
| Var. 8 | -0.023 | -0.099 | -0.152 | 0.619 | 0.699 | -0.015 | -0.264 | -0.159 |
| | **Comp. 1** | **Comp. 2** | **Comp. 3** | **Comp. 4** | **Comp. 5** | **Comp. 6** | **Comp. 7** | **Comp. 8** |
| **Autovalori** | 37.54 | 0.962 | 0.329 | 0.189 | 0.141 | 0.064 | 0.033 | 0.021 |

**Tabella 4: Tabella degli autovettori e degli autovalori della matrice di varianza e covarianza R**

A questo punto vanno calcolate le osservazioni nelle nuove coordinate per ciascuna sottomatrice secondo la funzione $Y_i = \sqrt{\frac{N_i}{N_1+N_2}} \cdot X_i \cdot P_0 \cdot \Lambda^{-1/2}$, dove $P_0$ è la matrice degli autovettori e $\Lambda$ è la matrice degli autovalori. Per ciascuna $Y_i$ viene calcolata la matrice di varianza e covarianza e per una di queste viene applicata l'analisi delle

componenti principali. Si calcola dunque l'indice D a partire dagli autovalori ottenuti dall'analisi effettuata su una delle due matrici di varianza e covarianza, ottenendo $\mathrm{D} = 1$. Questo significa che le due sottomatrici del F.O.R.S.U. differiscono tra di loro, ovvero c'è stato un cambiamento nella struttura di correlazione che non è stato possibile individuare tramite le carte di controllo Q e $T^2$.

Purtroppo non è stato possibile eseguire ulteriori suddivisioni a causa della bassa numerosità del campione.

**Carte di controllo per la matrice in uscita**

Anche per l'analisi della matrice in uscita, si partirà dalla carta basata sulla statistica $T^2$ di Hotelling. Si calcola quindi la matrice di correlazione R dai dati standardizzati di Fase I.

Si possono ora ricavare le statistiche $T^2$ per l'output del processo secondo la (25). Posto quindi $\alpha$ pari a 0.05, i valori della (25) vanno confrontati con il limite di controllo:

$$UCL = \left(\frac{(n-1)^2}{n}\right) B\left(\alpha/2\ , \frac{p}{2}, \frac{n-p-1}{2}\right) = 17.05162 \cong 17.05$$

Il limite di controllo inferiore è stato posto pari a zero poiché qualsiasi shift in media porterà ad un incremento della statistica $T^2$. Si ha quindi la seguente carta di controllo.

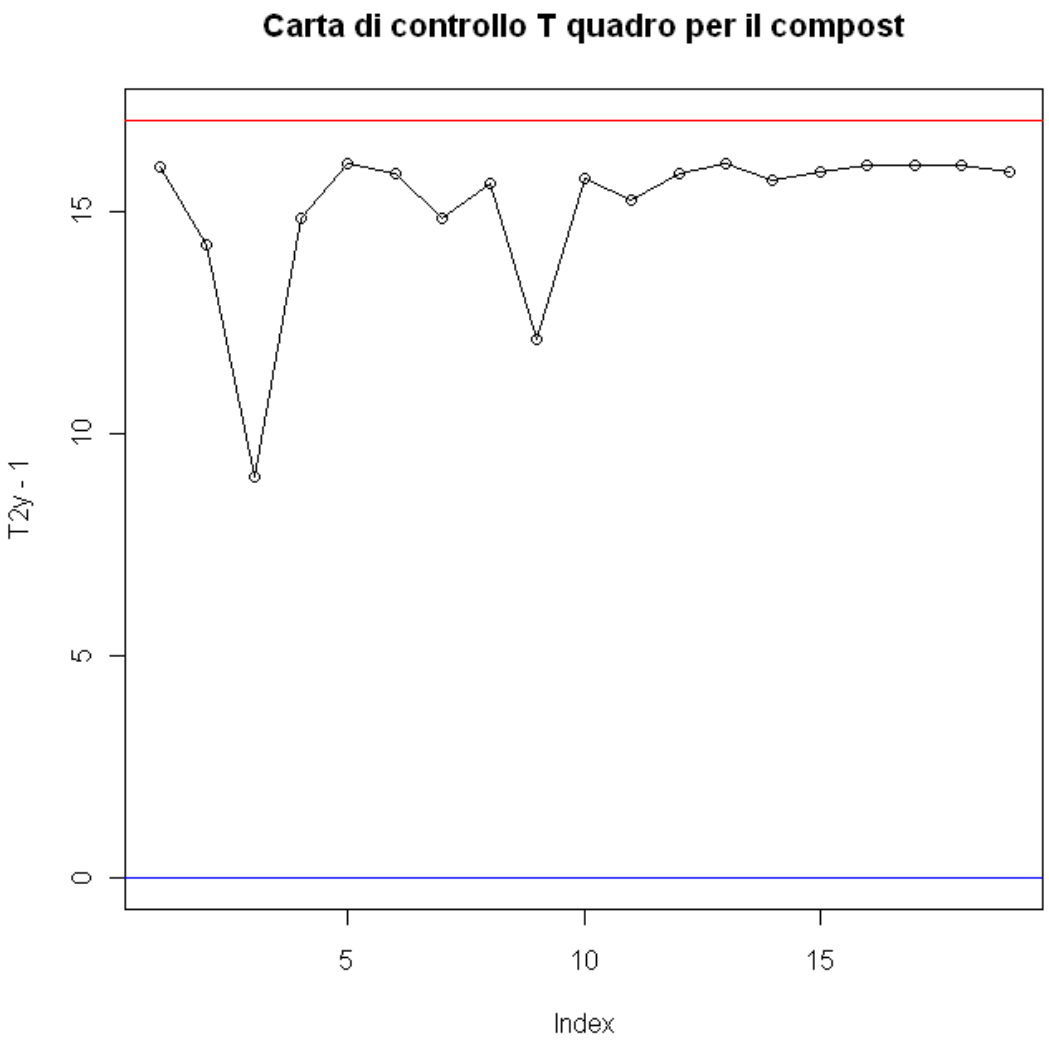


**Figura 16: Carta di controllo $T^2$ per i dati in uscita di Fase I**

Come è possibile notare non sono presenti fuori controllo. Si può quindi affermare che i campioni di Fase I provengono da un processo in controllo. Completata la Fase I, ovvero la fase di controllo dei dati storici, si passa alla Fase II, durante la quale un le statistiche di controllo calcolate su un nuovo campione di dati, composto da 19 osservazioni riguardanti le 17 variabili d'interesse, vengono messo a confronto con i limiti di controllo calcolati nella Fase I.

La statistica (7) viene calcolata utilizzando la matrice $\hat{R}$ ottenuta durante la prima fase, e tramite questa verrà calcolata la statistica $T^2$ per i nuovi campioni.

I valori ottenuti vengono dunque confrontati con i limiti di controllo introdotti da Mason, Tracy e Young (1992):

$$UCL = \frac{p(n+1)(n-1)}{n(n-p)} \cdot F(\alpha/2\,; p, n-p) = 6351.77$$

$$LCL = 0$$

La carta di controllo per i dati di Fase II è rappresentata nella Figura 17.

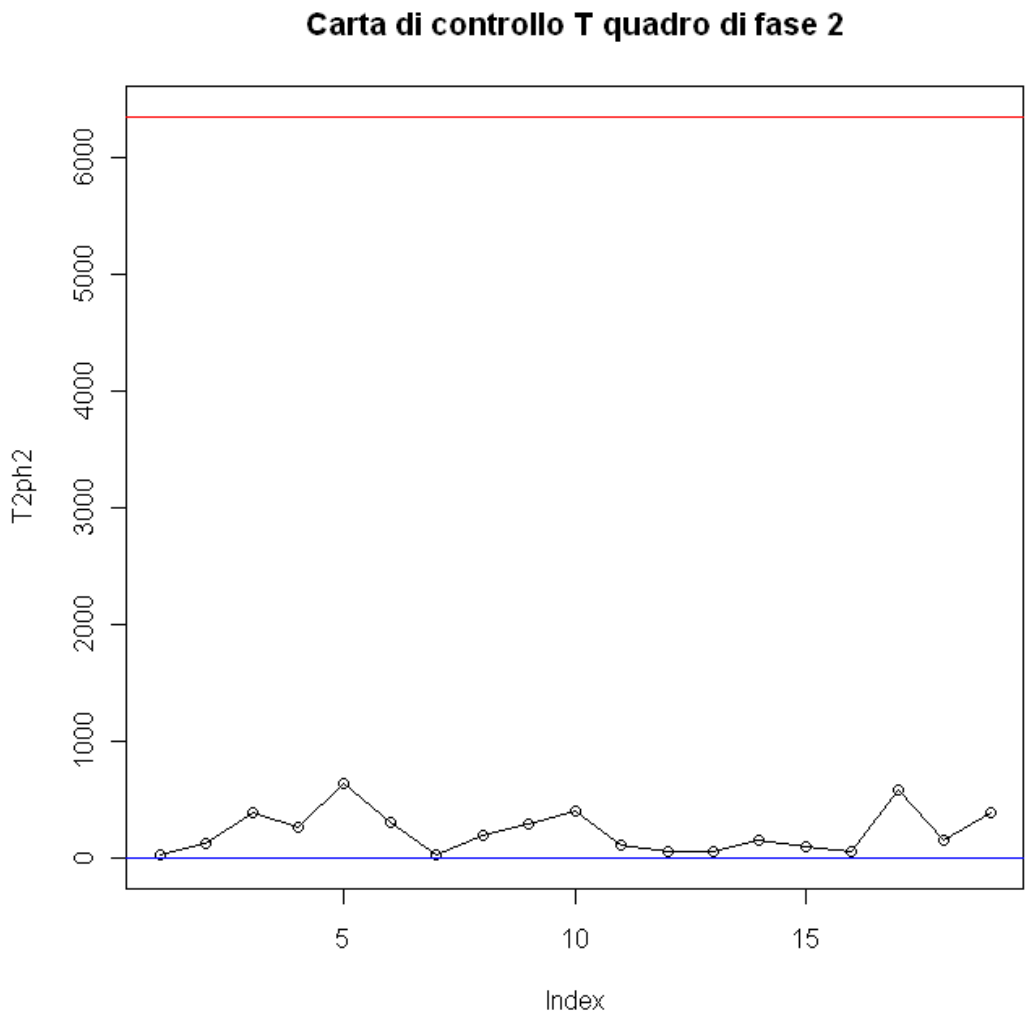


**Figura 17: Carta di controllo $T^2$ basata sui dati in uscita di Fase II**

Il processo produttivo sembra dunque in controllo.

Applico ora il metodo delle componenti principali al controllo statistico della qualità, con l'obbiettivo primario di suggerire agli operatori la sorveglianza di un numero

ridotto di variabili che si assume possano essere le maggiori responsabili di disturbi e variazioni nel processo. Da quanto visto nel paragrafo 3.5, si riscalano le variabili dividendole per le loro radici caratteristiche.

Dall'analisi delle componenti principali si ottengono gli autovettori, gli autovalori e le percentuali di varianza spiegata riportati nella Tabella 7.

| Autovettori | Comp. 1 | Comp. 2 | Comp. 3 | Comp. 4 | Comp. 5 | Comp. 6 | Comp. 7 | Comp. 8 | Comp. 9 |
|---|---|---|---|---|---|---|---|---|---|
| pH | 0.411 | 0.059 | -0.023 | 0.048 | 0.135 | -0.229 | 0.016 | 0.191 | 0.464 |
| Salinità | -0.359 | 0.051 | -0.123 | -0.306 | -0.347 | -0.131 | -0.068 | 0.007 | 0.021 |
| Umidità | 0.086 | -0.199 | 0.244 | 0.287 | -0.217 | 0.293 | -0.582 | 0.043 | 0.261 |
| Carbonioorg | 0.407 | 0.121 | -0.056 | -0.263 | 0.005 | -0.092 | -0.118 | -0.054 | -0.406 |
| Acid_Umici_Fulvici | -0.222 | 0.137 | -0.468 | -0.150 | 0.066 | -0.182 | -0.269 | -0.056 | 0.045 |
| Azoto_org | 0.110 | -0.446 | -0.062 | -0.180 | 0.379 | 0.049 | 0.093 | 0.256 | -0.037 |
| Rapporto_CN | 0.395 | 0.150 | -0.279 | -0.025 | -0.089 | -0.008 | 0.041 | -0.116 | -0.304 |
| Cadmio | 0.012 | -0.257 | 0.032 | 0.291 | -0.048 | -0.718 | -0.087 | -0.138 | 0.050 |
| Cromo_Tot | 0.028 | -0.384 | -0.319 | -0.325 | 0.163 | 0.023 | -0.162 | -0.093 | 0.205 |
| Mercurio | -0.309 | -0.059 | 0.010 | -0.094 | 0.444 | -0.157 | -0.248 | 0.261 | -0.093 |
| Nichel | 0.316 | -0.130 | -0.358 | 0.075 | 0.030 | 0.228 | -0.235 | -0.277 | 0.123 |
| Piombo | -0.058 | -0.092 | 0.244 | -0.520 | -0.077 | 0.118 | 0.167 | -0.452 | 0.335 |
| Rame | 0.087 | 0.362 | 0.206 | -0.065 | 0.477 | 0.154 | 0.053 | 0.100 | 0.150 |
| Zinco | 0.149 | 0.287 | 0.285 | -0.147 | 0.157 | -0.370 | -0.198 | -0.330 | 0.152 |
| Mat_plastico | -0.006 | 0.400 | -0.362 | -0.013 | -0.177 | -0.002 | 0.055 | 0.338 | 0.433 |
| Vetro | -0.241 | 0.276 | -0.085 | 0.178 | 0.292 | 0.174 | -0.352 | -0.335 | -0.118 |
| Metalli | 0.144 | 0.062 | 0.256 | -0.402 | -0.231 | -0.037 | -0.457 | 0.388 | -0.154 |
| Autovalori | Comp. 1 | Comp. 2 | Comp. 3 | Comp. 4 | Comp. 5 | Comp. 6 | Comp. 7 | Comp. 8 | Comp. 9 |
| | 4.059 | 2.568 | 2.189 | 1.805 | 1.552 | 1.137 | 0.978 | 0.696 | 0.593 |

| Comp. 10 | Comp. 11 | Comp. 12 | Comp. 13 | Comp. 14 | Comp. 15 | Comp. 16 | Comp. 17 |
|---|---|---|---|---|---|---|---|
| 0.157 | 0.003 | -0.058 | 0.351 | -0.359 | 0.177 | 0.220 | -0.373 |
| -0.060 | -0.059 | 0.248 | 0.290 | 0.313 | -0.218 | 0.160 | -0.539 |
| -0.098 | -0.269 | -0.013 | -0.091 | -0.111 | -0.337 | -0.175 | -0.136 |
| 0.108 | -0.174 | -0.137 | 0.177 | 0.048 | 0.036 | -0.641 | -0.217 |
| 0.263 | -0.108 | 0.148 | -0.542 | -0.397 | 0.043 | -0.055 | -0.116 |
| -0.098 | 0.390 | -0.167 | -0.301 | 0.104 | -0.345 | -0.019 | -0.343 |
| -0.140 | -0.190 | -0.062 | 0.029 | -0.162 | 0.547 | 0.462 | 0.157 |
| 0.319 | 0.113 | 0.040 | 0.059 | 0.197 | -0.289 | -0.107 | 0.213 |
| -0.332 | 0.017 | 0.357 | 0.346 | -0.133 | 0.024 | -0.162 | 0.368 |
| -0.020 | -0.509 | -0.420 | 0.159 | 0.141 | 0.051 | 0.187 | 0.102 |
| 0.221 | -0.030 | -0.041 | -0.105 | 0.580 | 0.301 | 0.233 | -0.042 |
| 0.324 | -0.079 | -0.354 | -0.020 | -0.087 | -0.188 | 0.008 | 0.102 |
| 0.280 | -0.135 | 0.526 | -0.015 | 0.227 | -0.301 | -0.057 | 0.075 |
| -0.581 | 0.056 | -0.010 | -0.279 | 0.133 | 0.102 | 0.061 | -0.080 |
| -0.135 | 0.135 | -0.336 | 0.018 | 0.233 | -0.195 | -0.270 | 0.232 |
| 0.073 | 0.501 | -0.188 | 0.354 | -0.129 | -0.137 | 0.0001 | -0.070 |
| 0.208 | 0.347 | 0.043 | -0.030 | 0.008 | 0.111 | 0.238 | 0.268 |
| Comp. 10 | Comp. 11 | Comp. 12 | Comp. 13 | Comp. 14 | Comp. 15 | Comp. 16 | Comp. 17 |
| 0.411 | 0.323 | 0.258 | 0.170 | 0.138 | 0.0939 | 0.015 | 0.007 |

**Tabella 5: Tabella degli autovalori e degli autovettori per i dati del compost**

La variabilità totale dei dati è rappresentata dalle 17 componenti principali ciascuna delle quali spiega a sua volta una proporzione decrescente di varianza. Seguendo il precedente criterio per la scelta del numero di componenti da includere (scegliere quelle che spiegano tra il 70 e il 90 per cento della varianza) si è scelto qui quindi di prendere in esame cinque componenti principali per la costruzione della carta di controllo considerando che con queste si ottiene una varianza spiegata pari al 71%.

Si possono calcolare i punti dalla (13) e disegnare la carta insieme ai limiti dati dalla (14) secondo quanto proposto da Jackson (1991).

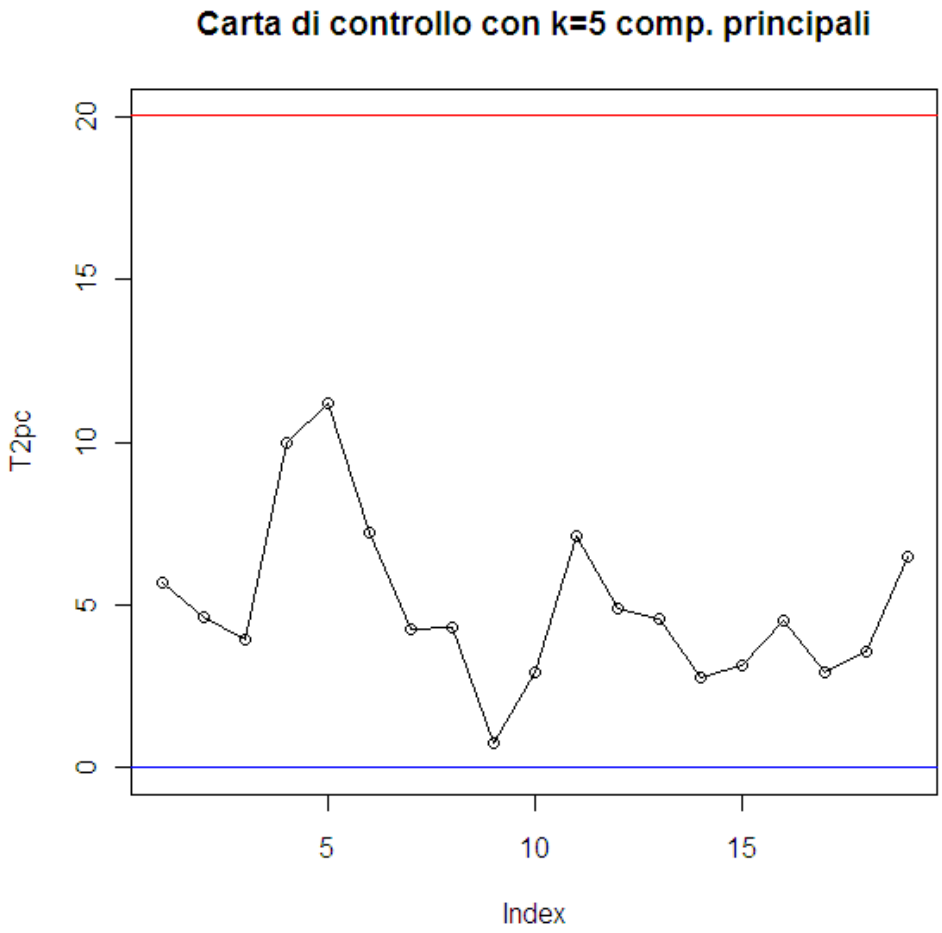


**Figura 18: Carta $T^2$ basata sulle prime 5 componenti per i dati del compost**

Dal momento che le prime cinque componenti spiegano una buona porzione di variabilità, la carta di controllo, come nel caso in cui si considerano tutte le variabili, non evidenzia alcun fuori controllo (Figura 19). Passiamo alla costruzione della carta Q, applicata alle PCA. Il primo passaggio è calcolare i valori della statistica di controllo tramite l'equazione (16) e quella del limite di controllo tramite l'equazione (17) , ovvero:

$$UCL = \vartheta_1 \left[ \frac{c_\alpha \sqrt{2\vartheta_2 h_0^2}}{\vartheta_1} + \frac{\vartheta_2 h_0 (h_0 - 1)}{\vartheta_3^2} + 1 \right]^{\frac{1}{h_0}} = 11.585$$

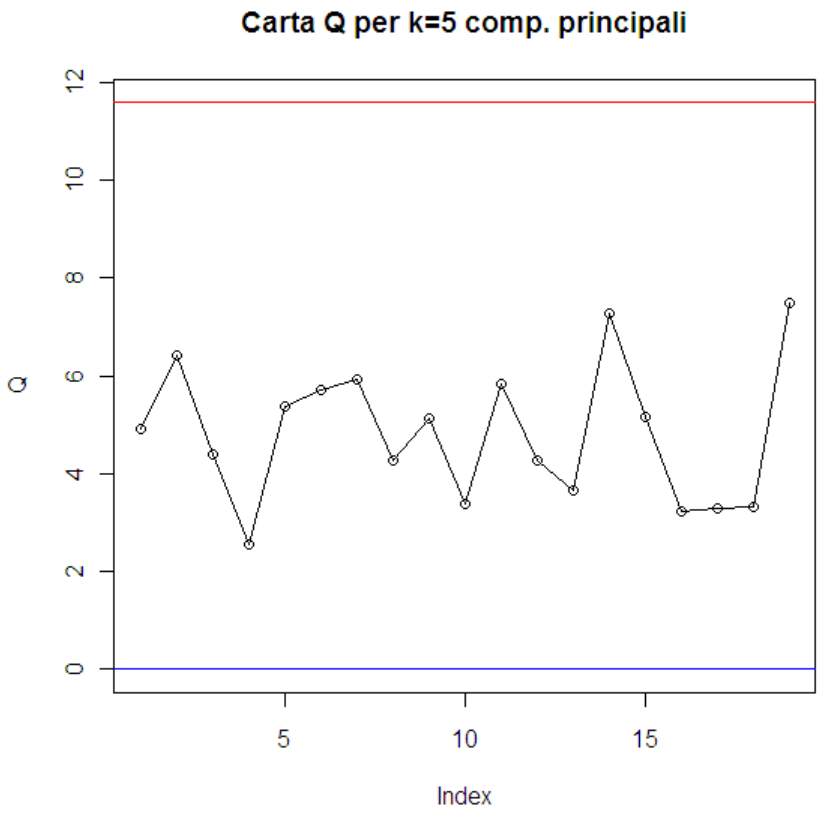


**Figura 19: Carta Q basata sulle prime 5 componenti per i dati del compost**

Anche dalla Figura 20 si evidenzia come il processo produttivo, la cui variabilità originale è rappresentata dalle sole prime cinque componenti principali, risulta essere in controllo.

È possibile pure usare la variante proposta da Hawkins (1991), la quale considera le componenti escluse dalla statistica $T^2$ (Figura 21).

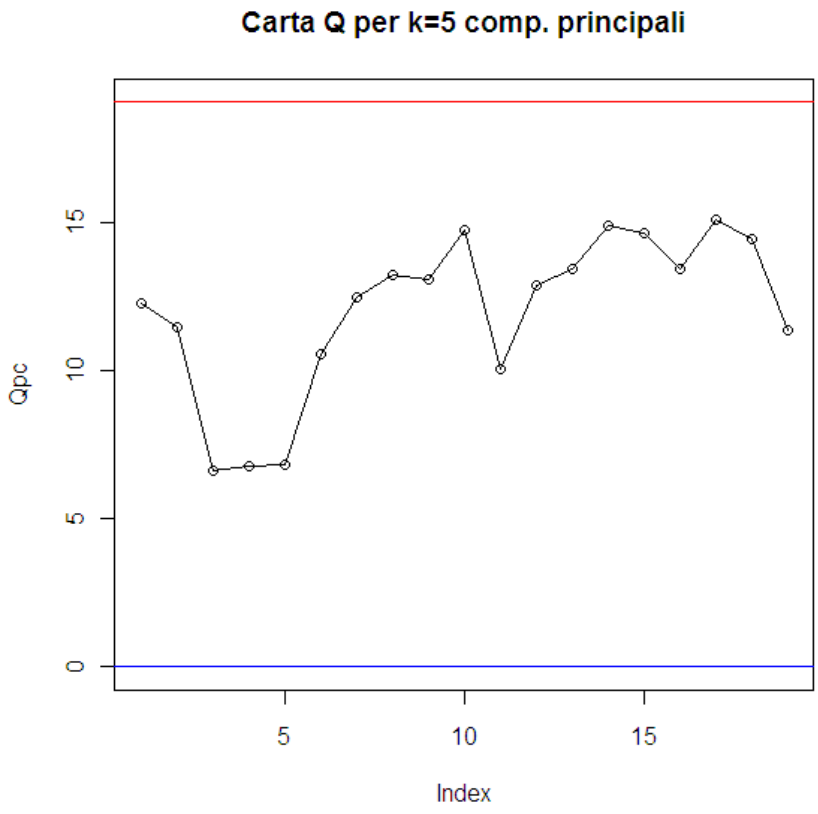


**Figura 20: Carta Q basata sulle prime 5 componenti per i dati del compost calcolata col metodo di Hawkins**

Ugualmente con il secondo metodo proposto il processo risulta essere in controllo. superiore.

Applichiamo ora l'Indice di Dissimilarità, proposto da Kano (2001) per sorvegliare insiemi di dati multivariati e autocorrelati.

Si è suddiviso il data set del compost in due sottomatrici composte rispettivamente da 9 e 10 campioni. Si è calcolata poi la matrice R di varianza e covarianza tra le due matrici e si è applicata a questa l'analisi delle componenti principali ottenendo gli autovettori e autovalori riportati nella Tabella 8.

| | Autovalori | Metalli | Vetro | Frat_plastico | Zinco | Rame | Piombo | Nichel | Mercurio | Cromo_Tot | Cadmio | Rapporto_CN | Azoto_org | Acidi_Umici_Fulvici | Carbonioorg | Umidità | Salinità | pH | Autovettori |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Comp. 1 | 0.056 | -0.629 | 0.004 | 0.00006 | -0.001 | 0.0004 | 0.004 | -0.768 | -0.00005 | 0.028 | 0.014 | 0.001 | -0.037 | 0.009 | 0.003 | -0.001 | 0.089 | Comp. 1 |
| | 313.44 | | | | | | | | | | | | | | | | | | |

| Comp. 2 | Comp. 3 | Comp. 4 | Comp. 5 | Comp. 6 | Comp. 7 | Comp. 8 | Comp. 9 | Comp. 10 | Comp. 11 | Comp. 12 | Comp. 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.027 | -0.041 | -0.123 | 0.222 | 0.949 | 0.093 | 0.105 | -0.034 | -0.027 | 0.013 | -0.001 | 0.022 |
| 0.001 | -0.001 | 0.0006 | 0.0009 | -0.022 | 0.010 | 0.025 | -0.029 | -0.045 | -0.07 | -0.229 | 0.373 |
| -0.005 | 0.005 | 0.0003 | 0.016 | 0.0006 | -0.083 | 0.221 | -0.182 | 0.879 | 0.148 | -0.179 | -0.185 |
| 0.001 | -0.0007 | 0.009 | 0.036 | 0.050 | 0.091 | -0.534 | -0.081 | -0.048 | -0.370 | -0.060 | -0.320 |
| 0.034 | -0.022 | -0.0351 | -0.0071 | -0.118 | 0.969 | 0.179 | -0.066 | 0.011 | 0.051 | 0.004 | -0.045 |
| -0.011 | -0.008 | 0.004 | -0.0049 | 0.036 | 0.053 | -0.028 | 0.945 | 0.097 | 0.177 | -0.077 | -0.182 |
| 0.009 | 0.002 | 0.0007 | 0.0048 | 0.063 | 0.140 | -0.776 | -0.073 | 0.217 | 0.387 | -0.099 | 0.168 |
| -0.153 | -0.037 | -0.957 | 0.1759 | -0.158 | -0.041 | -0.024 | 0.011 | 0.0002 | -0.003 | -0.003 | -0.003 |
| -0.001 | -0.002 | 0.001 | -0.0022 | 0.003 | 0.044 | 0.010 | 0.134 | 0.062 | -0.093 | -0.315 | 0.528 |
| -0.229 | -0.592 | 0.032 | 0.0206 | 0.055 | -0.024 | -0.014 | -0.007 | 0.006 | -0.001 | 0.0004 | 0.002 |
| 0.002 | 0.003 | -0.002 | 0.0032 | 0.029 | 0.083 | -0.114 | 0.169 | 0.372 | -0.739 | 0.290 | 0.099 |
| -0.001 | -0.0004 | 0.005 | 0.002 | -0.005 | 0.007 | 0.055 | 0.0005 | -0.100 | -0.307 | -0.752 | -0.146 |
| 0.001 | 0.0008 | 0.003 | 0.003 | 0.015 | -0.006 | 0.011 | -0.045 | -0.082 | 0.016 | 0.158 | -0.445 |
| 0.0002 | 0.0006 | -0.0003 | 0.007 | 0.007 | -0.009 | -0.027 | -0.065 | -0.065 | 0.025 | -0.345 | -0.387 |
| 0.918 | -0.369 | -0.118 | 0.013 | -0.055 | -0.049 | -0.006 | 0.011 | 0.005 | -0.003 | -0.002 | -0.003 |
| 0.280 | 0.711 | -0.08 | 0.100 | 0.049 | -0.012 | -0.003 | 0.009 | -0.001 | 0.001 | -0.002 | 0.002 |
| -0.016 | -0.04 | 0.214 | 0.952 | -0.201 | -0.006 | 0.002 | 0.017 | -0.005 | 0.006 | 0.011 | 0.014 |

| Comp. 2 | Comp. 3 | Comp. 4 | Comp. 5 | Comp. 6 | Comp. 7 | Comp. 8 | Comp. 9 | Comp. 10 | Comp. 11 | Comp. 12 | Comp. 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 280.68 | 118.63 | 40.23 | 15.52 | 7.38 | 0.987 | 0.143 | 0.09 | 0.081 | 0.013 | 0.005 | 0.003 |

| Comp. 14 | 0.008 | -0.106 | 0.027 | -0.012 | 0.008 | 0.030 | 0.079 | 0.001 | -0.450 | 0.001 | -0.094 | 0.467 | -0.152 | -0.725 | -0.0002 | 0.001 | 0.001 | Comp. 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Comp. 15 | -0.002 | 0.064 | -0.160 | -0.611 | -0.024 | -0.074 | 0.344 | 0.003 | -0.093 | 0.003 | 0.346 | 0.262 | 0.512 | 0.104 | -0.002 | -0.001 | 0.021 | Comp. 15 |
| Comp. 16 | 0.007 | -0.355 | -0.140 | -0.273 | 0.003 | -0.049 | 0.085 | -0.004 | -0.317 | 0.002 | 0.197 | 0.040 | -0.683 | 0.400 | -0.002 | -0.0006 | 0.010 | Comp. 16 |
| Comp. 17 | -0.013 | -0.810 | -0.039 | 0.036 | -0.007 | -0.106 | 0.017 | 0.003 | 0.524 | 0.0005 | -0.0495 | 0.121 | 0.115 | -0.149 | 0.002 | -0.0005 | 0.006 | Comp. 17 |
| Comp. 14 | 0.008 | -0.106 | 0.027 | -0.012 | 0.008 | 0.030 | 0.079 | 0.001 | -0.450 | 0.001 | -0.094 | 0.467 | -0.152 | -0.725 | -0.0002 | 0.001 | 0.001 | Comp. 14 |
| Comp. 15 | -0.002 | 0.064 | -0.160 | -0.611 | -0.024 | -0.074 | 0.344 | 0.003 | -0.093 | 0.003 | 0.346 | 0.262 | 0.512 | 0.104 | -0.002 | -0.001 | 0.021 | Comp. 15 |
| Comp. 16 | 0.007 | -0.355 | -0.140 | -0.273 | 0.003 | -0.049 | 0.085 | -0.004 | -0.317 | 0.002 | 0.197 | 0.040 | -0.683 | 0.400 | -0.002 | -0.0006 | 0.010 | Comp. 16 |
| Comp. 17 | -0.013 | -0.810 | -0.039 | 0.036 | -0.007 | -0.106 | 0.017 | 0.003 | 0.524 | 0.0005 | -0.0495 | 0.121 | 0.115 | -0.149 | 0.002 | -0.0005 | 0.006 | Comp. 17 |

**Tabella 6: Tabella degli autovalori e autovettori per la matrice di varianza e covarianza R applicata ai dati del compost**

A questo punto vanno calcolate le osservazioni nelle nuove coordinate per ciascuna sottomatrice secondo la funzione $Y_i = \sqrt{\frac{N_i}{N_1+N_2}} \cdot X_i \cdot P_0 \cdot \Lambda^{-1/2}$, dove $P_0$ è la matrice degli autovettori e $\Lambda$ è la matrice degli autovalori. Per ciascuna $Y_i$ viene calcolata la matrice di varianza e covarianza e per una di queste viene applicata l'analisi delle componenti principali. Si calcola dunque l'indice D a partire dagli autovalori ottenuti dall'analisi effettuata su una delle due matrici di varianza e covarianza, ottenendo $\mathrm{D} = 1$. Questo significa che le due sottomatrici del compost differiscono tra di loro, ovvero c'è stato un cambiamento nella struttura di correlazione che non è stato possibile individuare tramite le carte di controllo Q e $T^2$.

Purtroppo non è stato possibile eseguire ulteriori suddivisioni a causa della bassa numerosità del campione.

## Carte di controllo univariate: il confronto col caso S.E.S.A. S.p.A.

Si effettua ora l'analisi univariata per le 8 variabili in ingresso. Lo scopo di questo è dovuto al fatto che le carte univariate spesso non segnalano (o viceversa) laddove quelle multivariate lo fanno visto che sono in grado di tenere in considerazione la

struttura di correlazione esistente tra le variabili. In pratica quanto emergerebbe da un'analisi univariata non sarebbe coerente con le conclusioni a livello multivariato.

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. Contrariamente una carta univariata non tiene conto della correlazione esistente nella popolazione, ma della sola deviazione standard che è presente. Per dimostrare ciò si farà uso della carta univariata Shewhart per la media e per la deviazione standard poiché la carta $T^2$ è il corrispettivo della carta Shewhart nel caso multivariato.

Per studiare quindi gli shift in media si procede alla costruzione della carta per la media: la carta Shewhart. Per questa sono necessari i limiti LCL, UCL così ottenuti:

$$UCL = \mu + L \cdot \sigma$$

$$LCL = \mu - L \cdot \sigma$$

Infine si riportano questi limiti, insieme ai punti della variabile presa in esame e al limite centrale, nella carta di controllo.

Per studiare invece la variabilità delle osservazioni si procede alla costruzione di una serie dette "escursioni mobili" date dalla differenza tra ogni valore e quello precedente. Si passa poi al calcolo di due costanti $d_2$ e $d_3$, le quali dipendono dalla numerosità campionaria e hanno il ruolo di correggere l'inconsistenza di $E[R_t]$. Queste due costanti verranno poi usate per il calcolo dei limiti LCL e UCL:

$$UCL = d_2\sigma + L \cdot d_3 \cdot \sigma$$

$$LCL = d_2\sigma - L \cdot d_3 \cdot \sigma$$

In conclusione si riportano questi limiti, insieme alle escursioni mobili, nella carta di controllo.

In tutte e due le carte si considererà $L = 3$.

Vediamo quindi entrambe le carte di controllo per ciascuna variabile:

- Variabile `Umidit_R`

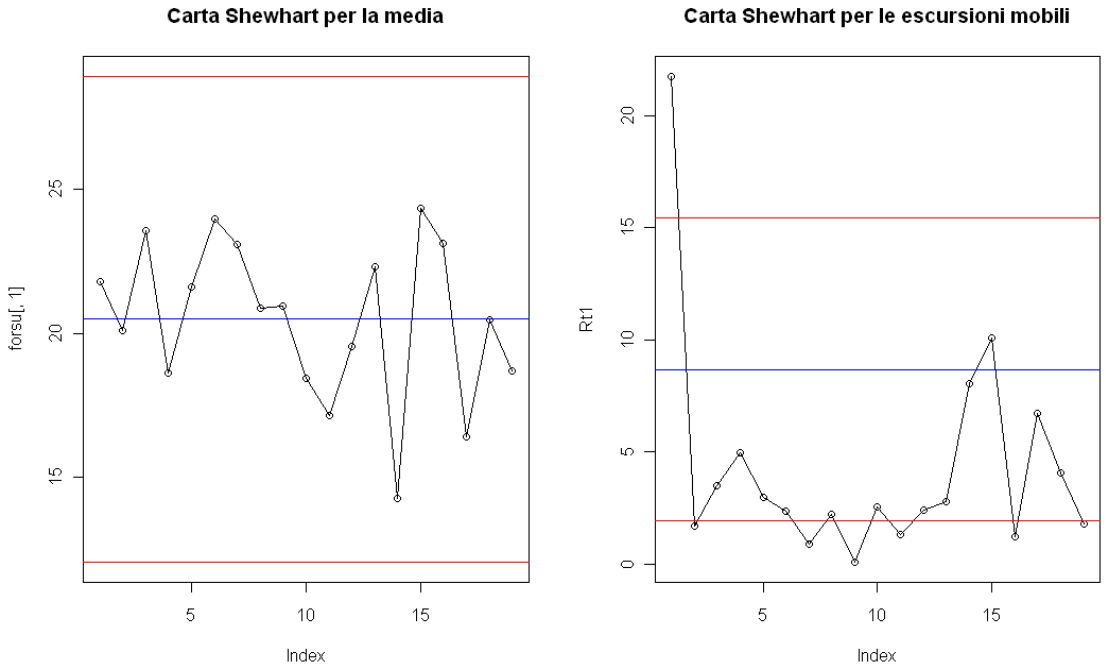


**Figura 21: Carte di controllo per la media e le escursioni mobili per la variabile Umidit_R**

Le carte di controllo univariate riferite alla variabile umidità residua mostrano un processo in controllo nella carta per la media, e un processo fuori controllo nella carta per le escursioni mobili. La varianza risulta quindi instabile. Quanto emerso risulta coerente con la carta $T^2$, nel caso della media, e con l'Indice di Dissimilarità, nel caso ella varianza.

Prendendo in esame le altre variabili si arriva alla stessa conclusione.

- Variabile `Cadmio`

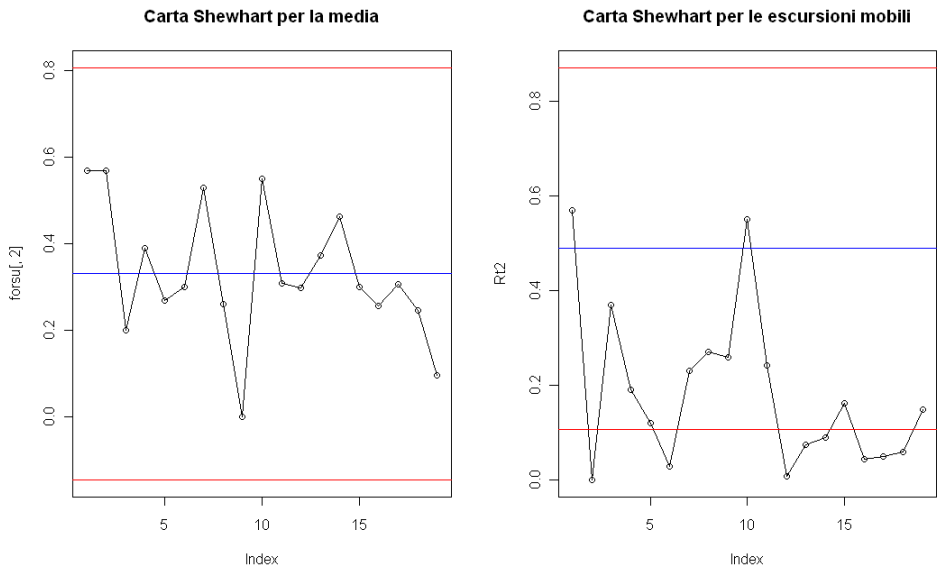


**Figura 22: Carte di controllo per la media e le escursioni mobili per la variabile Cadmio**

- Variabile `Cromo_Tot`

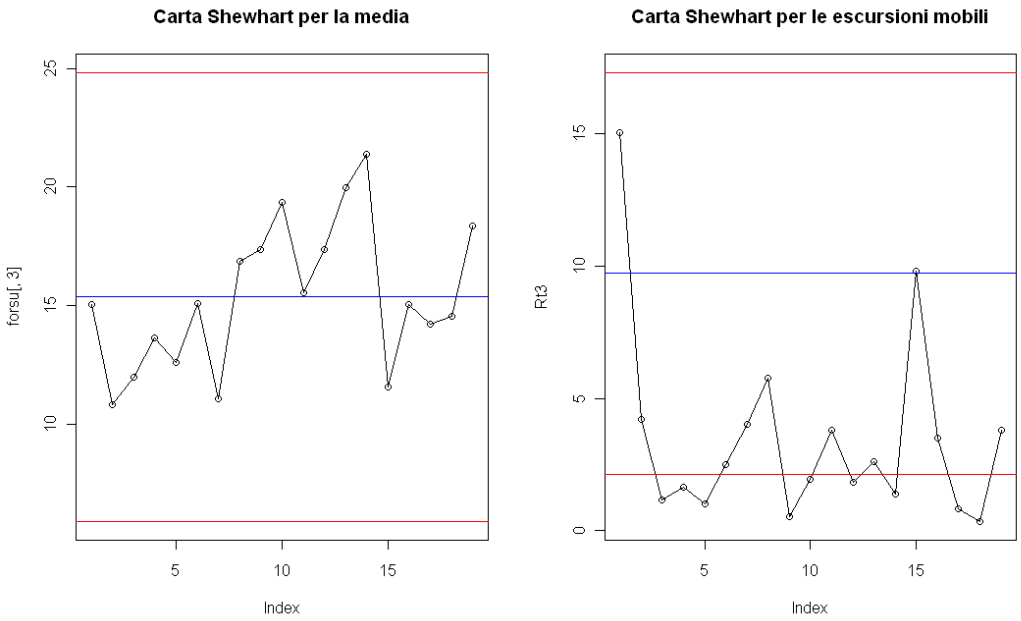


**Figura 23: Carte di controllo per la media e le escursioni mobili per la variabile Cromo_Tot**

- Variabile `Mercurio`

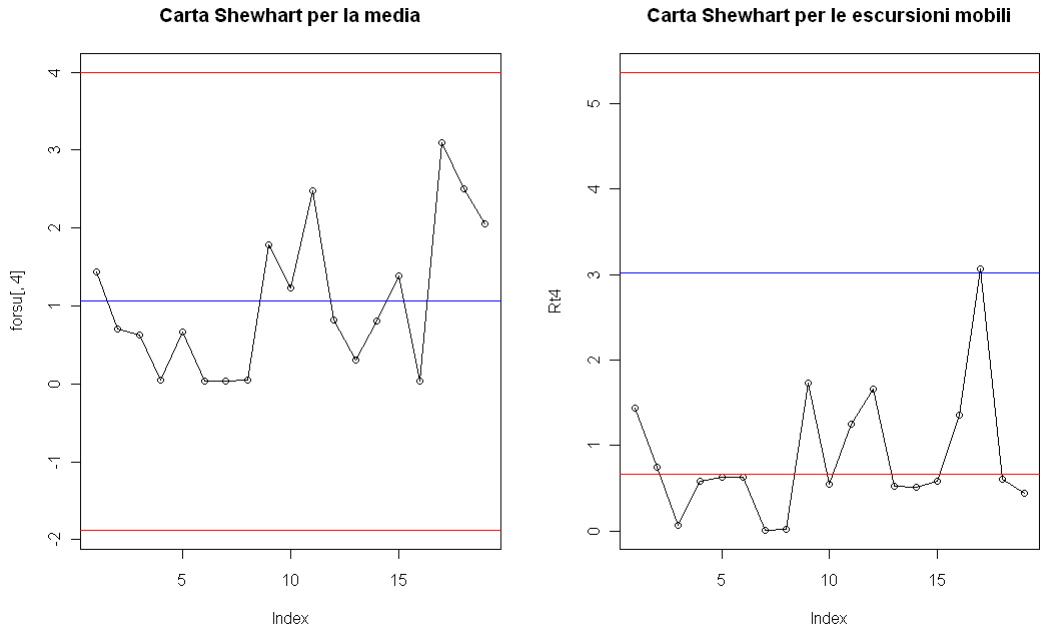


**Figura 24: Carte di controllo per la media e le escursioni mobili per la variabile Mercurio**

- Variabile `Nichel`

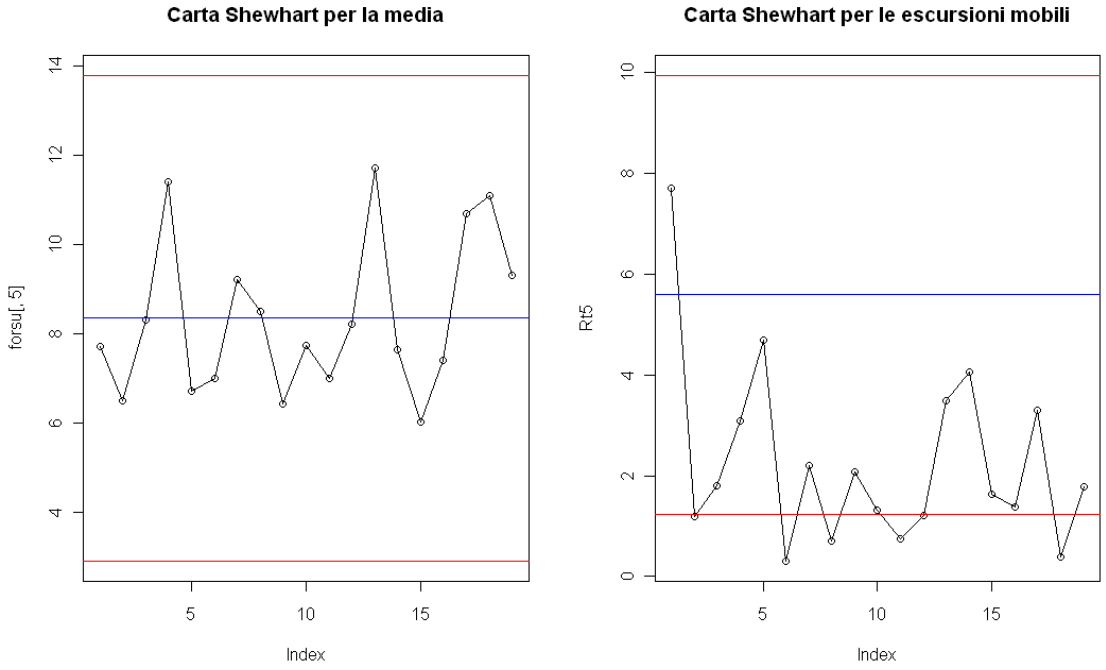


**Figura 25: Carte di controllo per la media e le escursioni mobili per la variabile Nichel**

- Variabile `Piombo`

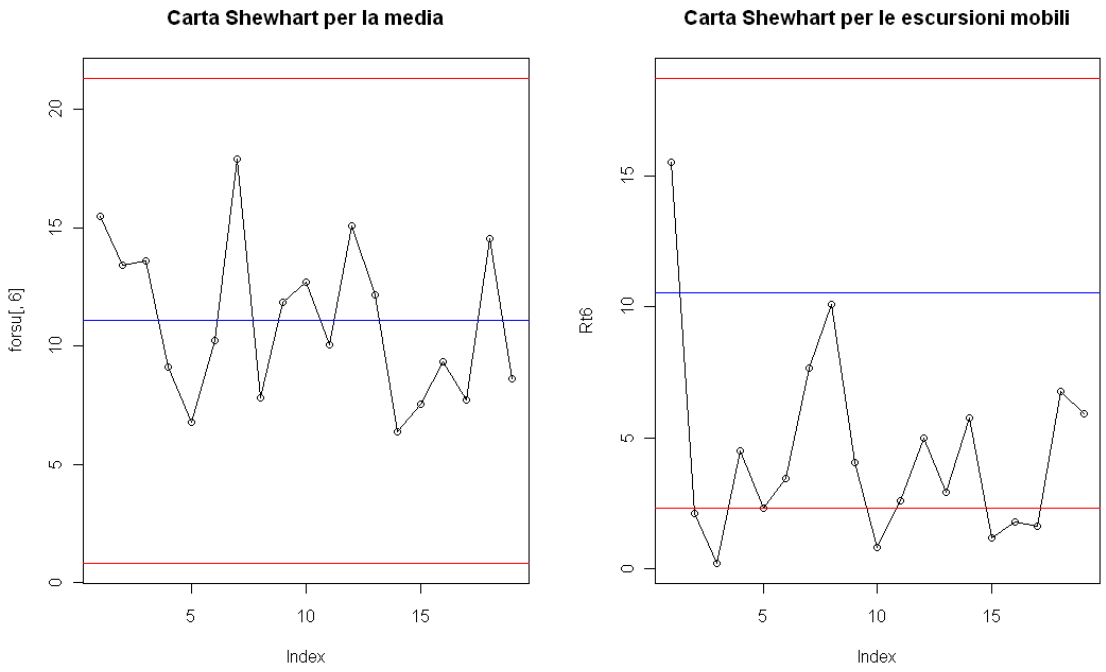


**Figura 26: Carte di controllo per la media e le escursioni mobili per la variabile Piombo**

- Variabile `Rame`

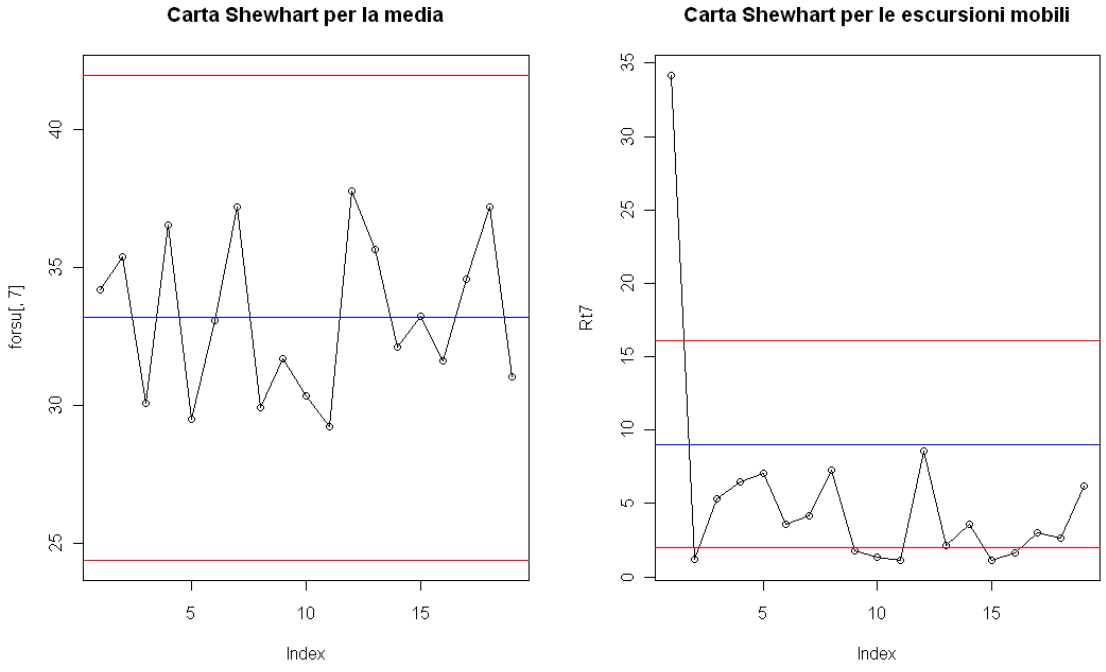


**Figura 27: Carte di controllo per la media e le escursioni mobili per la variabile Rame**

- Variabile `Zinco`

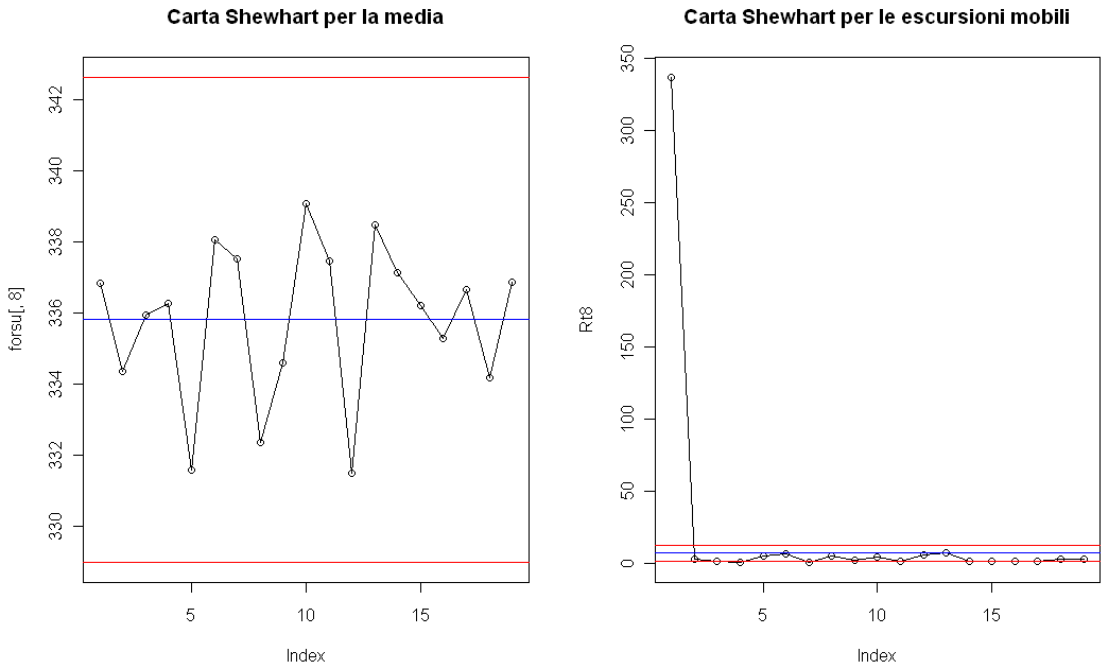


**Figura 28: Carte di controllo per la media e le escursioni mobili per la variabile Zinco**

Come si può notare, tutte le carte presentano fuori controllo solo nella carta per le escursioni mobili. Diventa quindi evidente pensare che è avvenuto uno shift nella varianza o eventualmente nella struttura di correlazione. Purtroppo però le sole carte univariate non sono sufficienti per affermare quanto detto. È proprio grazie all'Indice di Dissimilarità, che tale ipotesi risulta corretta. Infatti sia la carta per le escursioni mobili, che l'indice D (Indice di Dissimilarità) sono giunti alla stessa conclusione implicando che è avvenuto effettivamente un cambiamento nella varianza o nella struttura di correlazione del processo.

In conclusione, considerando quanto visto da quest'analisi, si può affermare che il contributo dato dalle tecniche multivariate è di importanza vitale.

# Conclusioni

Lo scopo di questa tesi è stato quello di analizzare il controllo del processo di analisi del F.O.R.S.U. e di produzione del compost da parte della società S.E.S.A. S.p.A.,

società dove si è svolto lo stage e la raccolta dei dati. In particolare per poter effettuare un'analisi corretta e attendibile, è stato necessario porre l'attenzione solo su materiali in ingresso e in uscita simili per caratteristiche (escludendo quindi il materiale verde, il S.O.A. in input e il compost torboso in output). Al variare dei materiali di fatti variano le caratteristiche, i limiti di specifica e la loro composizione. Quindi dai test report si sono ricavate le variabili d'interesse per l'analisi dei conferimenti e del processo di produzione, entrambi processi aziendali conclusivi di vitale importanza in quanto il risultato positivo ne determina l'approvazione e quindi la vendita del prodotto, se si tratta del compost, o l'accettazione negli impianti, se si tratta del F.O.R.S.U.. Per poter fare ciò è stato necessario ricorrere a metodologie di controllo del processo che riguardano l'ambito multivariato. Partendo dalla costruzione delle carte per la statistica $T^2$, si è passati all'Analisi delle Componenti Principali che ha permesso di costruire le carte $T^2$ per PCA e quella per i residui Q. È stato infine calcolato un indice, denominato Indice di Dissimilarità, il quale tiene conto della struttura di correlazione dei dati, per poter riuscire a identificare shift nella matrice di correlazione che le carte di controllo non possono rilevare.

È stata quindi eseguita un'analisi esplorativa sia sulla matrice in ingresso del F.O.R.S.U., che in quella in uscita del compost, facendo notare una certa correlazione tra le diverse variabili quantitative di ciascun data set, l'accettazione dell'ipotesi di normalità, necessaria per poter applicare le tecniche MSPC e il rifiuto dell'ipotesi di omoschedasticità.

Per i dati in ingresso la carta di controllo $T^2$ di Hotelling, ha dimostrato come il processo produttivo fosse in controllo. È stata quindi costruita una carta di Fase II dove i dati attuali sono apparsi anch'essi in controllo. In seguito all'analisi delle componenti principali, la carta Q e la carta $T^2$ per PCA, hanno ridotto la "dimensionalità" del problema, considerando tre componenti principali, ritenute quelle ad alto contenuto esplicativo tra i dati del F.O.R.S.U.. Il processo è risultato in

controllo sia per le carte $T^2$ applicata alle PCA e Q. Successivamente l'indice di Dissimilarità applicato ai dati è risultato pari a 1, indicando uno shift nella struttura di correlazione e portando a considerare fuori controllo il processo.

Per i dati in uscita la carta di controllo $T^2$ ha sempre indicato il processo in controllo, sia sui dati di Fase I che su quelli di Fase II. L'analisi delle componenti principali ha sottolineato la presenza di cinque componenti che riassumevano il 71% della variabilità totale (tralasciandone quindi ben dodici), e che quindi potevano essere considerate per la costruzione delle carte Q e $T^2$ per PCA. Quest'ultime hanno dimostrato che seppur si considerassero solo 5 componenti il processo risultava in controllo. L'Indice di Dissimilarità applicato ai dati ha dimostrato però in questi vi è stato uno shift nella struttura di correlazione che non è stato rilevato e che quindi porta a considerare fuori controllo i dati.

Grazie a tecniche multivariate come l'Indice di Dissimilarità è stato possibile applicare il controllo statistico della qualità a processi multivariati laddove le tecniche univariate da sole non avrebbero portato ad una conclusione certa. Questo lo si è potuto dimostrare nell'ultimo paragrafo, in quanto si è visto come la carta per le escursioni mobili, segnalasse uno shift nella varianza, proprio come l'indice D ha indicato ai dati per il F.O.R.S.U..

Il metodo DISSIM, inoltre, si è dimostrato efficace nel determinare variazioni nella relazione temporale tra variabili. Tale tecnica infatti ha permesso di individuare, non con grande esattezza a causa della bassa numerosità, il periodo di tempo in cui è più evidente la differenza tra i due insiemi.

L'utilizzo quindi di queste tecniche risulta appropriato e i risultati ottenuti da queste tecniche dimostra seppur il processo sia in controllo in media, vi è un probabilità che la correlazione tra le osservazioni pregiudichi tali risultati oscurando di conseguenza eventuali fuori controllo.

Il materiale non andato a buon fine sono un'ulteriore spesa di manodopera, risulta quindi fondamentale per il miglioramento della qualità, in termini di efficienza e di risparmio economico, cercare di capire i motivi che hanno determinato tali perdite: solitamente le cause sono dovute da ditte esterne che non effettuano controlli precisi sui materiali che vengono a conferire presso S.E.S.A. S.p.A. se si tratta del F.O.R.S.U., oppure ad una fase di biossidazione che non è stata eseguita alla temperatura giusta e costante. L'intervento con azioni correttive opportune quali l'ulteriore biossidazione del compost a 60 C° per cinque giorni oppure una separazione di questo in diverse parti per la presenza di inerti portano a una visibile rientranza nei limiti di controllo e quindi risultati positivi.

# Appendice

- *I 19 campioni osservati provenienti dai controlli in ingresso:*

| Umidit_R | Cadmio | Cromo_Tot | Mercurio | Nichel | Piombo | Rame | Zinco |
|---|---|---|---|---|---|---|---|
| 21.78 | 0.570 | 15.03 | 1.440 | 7.70 | 15.50 | 34.2 | 336.82 |
| 20.08 | 0.570 | 10.81 | 0.700 | 6.50 | 13.40 | 35.4 | 334.35 |
| 23.56 | 0.200 | 11.99 | 0.630 | 8.30 | 13.60 | 30.08 | 335.92 |
| 18.62 | 0.390 | 13.64 | 0.046 | 11.40 | 9.10 | 36.54 | 336.25 |
| 21.61 | 0.270 | 12.61 | 0.670 | 6.70 | 6.78 | 29.5 | 331.57 |
| 23.98 | 0.300 | 15.11 | 0.040 | 7.00 | 10.25 | 33.07 | 338.05 |
| 23.09 | 0.530 | 11.08 | 0.037 | 9.20 | 17.90 | 37.2 | 337.50 |
| 20.88 | 0.260 | 16.86 | 0.052 | 8.50 | 7.80 | 29.9 | 332.36 |
| 20.94 | 0 | 17.38 | 1.780 | 6.43 | 11.87 | 31.7 | 334.58 |
| 18.42 | 0.550 | 19.34 | 1.230 | 7.74 | 12.69 | 30.33 | 339.07 |
| 17.13 | 0.308 | 15.55 | 2.480 | 6.99 | 10.07 | 29.22 | 337.46 |
| 19.53 | 0.299 | 17.38 | 0.820 | 8.21 | 15.07 | 37.790 | 331.49 |
| 22.31 | 0.373 | 19.98 | 0.300 | 11.70 | 12.15 | 35.66 | 338.48 |
| 14.27 | 0.463 | 21.38 | 0.810 | 7.64 | 6.39 | 32.100 | 337.11 |
| 24.36 | 0.301 | 11.57 | 1.390 | 6.01 | 7.56 | 33.240 | 336.21 |
| 23.13 | 0.256 | 15.06 | 0.039 | 7.39 | 9.35 | 31.6 | 335.26 |
| 16.41 | 0.306 | 14.23 | 3.100 | 10.69 | 7.74 | 34.575 | 336.66 |
| 20.47 | 0.246 | 14.57 | 2.500 | 11.08 | 14.53 | 37.21 | 334.18 |
| 18.70 | 0.097 | 18.35 | 2.060 | 9.30 | 8.62 | 31.02 | 336.87 |

- *Analisi esplorativa dei dati in ingresso:*

```
#Grafico della correlazione tra le variabili
library(ellipse)
plotcorr(cor(forsu))

#Statistiche riassuntive dei i dati per il forsu
summary(forsu)
#    Umidit_R         Cadmio          Cromo_Tot        Mercurio
#Min.   :14.27   Min.   :0.0000   Min.   :10.81   Min.   :0.037
#1st Qu.:18.66   1st Qu.:0.2580   1st Qu.:13.12   1st Qu.:0.176
#Median :20.88   Median :0.3010   Median :15.06   Median :0.810
#Mean   :20.49   Mean   :0.3310   Mean   :15.36   Mean   :1.059
#3rd Qu.:22.70   3rd Qu.:0.4265   3rd Qu.:17.38   3rd Qu.:1.610
#Max.   :24.36   Max.   :0.5700   Max.   :21.38   Max.   :3.100
#     Nichel           Piombo           Rame           Zinco
#Min.   : 6.010   Min.   : 6.39   Min.   :29.22   Min.   :331.5
#1st Qu.: 6.995   1st Qu.: 8.21   1st Qu.:30.68   1st Qu.:334.5
```

```
#Median : 7.740   Median :10.25   Median :33.07   Median :336.2
#Mean   : 8.341   Mean   :11.07   Mean   :33.18   Mean   :335.8
#3rd Qu.: 9.250   3rd Qu.:13.50   3rd Qu.:35.53   3rd Qu.:337.3
#Max.   :11.700   Max.   :17.90   Max.   :37.79   Max.   :339.1
```

```
#Analisi dei grafici boxplot
#Standardizzo la matrice dei dati in ingresso
n=19
p=8
media=function(p,x){
      temp=double(p)
      for(j in 1:p){
            temp[j]=mean(x[,j])
      }
      temp
}
xbar=media(p,forsu)
S=var(forsu)*(n)/(n-1)
x=matrix(0,19,8)
for(j in 1:8){
      for(i in 1:19){
            x[i,j]=(forsu[i,j]-xbar[j])/S[j,j]
      }
}
boxplot(x)
```

```
#Verifico assunti normalità
#Test sulla normalità della variabile Umidità_R
>shapiro.test(forsu[,1])
#W = 0.9624, p-value = 0.6209
#Test sulla normalità della variabile Cadmio
>shapiro.test(forsu[,2])
#W = 0.9367, p-value = 0.2299
```

```
#Test sulla normalità della variabile Cromo_Tot
>shapiro.test(forsu[,3])
#W = 0.9658, p-value = 0.6903
#Test sulla normalità della variabile Mercurio
>shapiro.test(forsu[,4])
#W = 0.9017, p-value = 0.0522
#Test sulla normalità della variabile Nichel
>shapiro.test(forsu[,5])
#W = 0.9086, p-value = 0.0698
#Test sulla normalità della variabile Piombo
>shapiro.test(forsu[,6])
#W = 0.954, p-value = 0.4602
#Test sulla normalità della variabile Rame
>shapiro.test(forsu[,7])
#W = 0.9307, p-value = 0.1787
#Test sulla normalità della variabile Zinco
>shapiro.test(forsu[,8])
#W = 0.9301, p-value = 0.1743
```

- *I 19 campioni osservati provenienti dai controlli in uscita:*

| pH | Salinità | Umidità | Carbonio_org | Umici_Fulvic. | Azoto_org | Rapporto_CN |
|---|---|---|---|---|---|---|
| 8.3 | 42.9 | 41.6 | 25.2 | 9.6 | 83 | 13 |
| 8.3 | 35.4 | 37 | 25.2 | 8 | 88 | 12.6 |
| 8 | 44.5 | 33.3 | 29.4 | 9.4 | 87 | 14.6 |
| 8.3 | 35.6 | 35.5 | 26.4 | 10.6 | 88 | 12.1 |
| 8.2 | 53.8 | 41 | 30.4 | 9.2 | 88 | 12.4 |
| 8.4 | 37.6 | 39.3 | 29.8 | 9.8 | 90 | 15.4 |
| 8.1 | 43.3 | 42.4 | 23.4 | 9.7 | 87 | 11.1 |
| 7.6 | 59.6 | 39.2 | 23.7 | 10.3 | 86 | 12.4 |
| 7.6 | 58.6 | 35.8 | 25.4 | 9.5 | 88 | 11.2 |
| 8.2 | 35.5 | 40.1 | 27.8 | 9.5 | 90 | 14.9 |
| 8.2 | 35 | 41.6 | 24.1 | 8.9 | 94 | 11.3 |
| 7.9 | 56.3 | 35.7 | 21.8 | 9.9 | 92 | 9.3 |
| 8 | 54.9 | 41.2 | 22.5 | 8.9 | 86 | 8.5 |
| 7.7 | 55.5 | 39.5 | 22 | 9.7 | 85 | 10.8 |
| 7.6 | 52.38 | 41.52 | 23 | 9.41 | 85.78 | 9.66 |
| 7.8 | 64.38 | 34.96 | 23.7 | 10.7 | 88.74 | 9.91 |
| 7.91 | 33.25 | 37.21 | 23.5 | 9.37 | 89.11 | 9.67 |
| 7.5 | 68.1 | 38.4 | 21.09 | 10.5 | 87.81 | 9.8 |
| 8.2 | 66 | 32.5 | 25 | 10.67 | 85 | 12.7 |

| Cadmio | Cromo_tot | Mercurio | Nichel | Piombo | Rame | Zinco | Mat_plastico | Vetro | Metalli |
|---|---|---|---|---|---|---|---|---|---|
| 0.62 | 24.2 | 0.04 | 24.9 | 44.4 | 114 | 223 | 0.21 | 0.09 | 0.017 |
| 0.63 | 24.5 | 0.06 | 12.8 | 38.2 | 115 | 222 | 0.09 | 0.02 | 0.009 |
| 0.48 | 33.6 | 0.08 | 19.3 | 53.2 | 121 | 217 | 0.09 | 0.07 | 0.008 |
| 0.48 | 37.3 | 0.18 | 21 | 44.8 | 139 | 207 | 0.11 | 0.19 | 0.016 |
| 0.57 | 36.08 | 0.08 | 17.6 | 57.8 | 108 | 221 | 0.08 | 0.01 | 1 |
| 0.82 | 60.5 | 0.07 | 28.6 | 47.2 | 82.3 | 216 | 0.03 | 0.01 | 0.001 |
| 1.03 | 27.1 | 0.12 | 17.3 | 35.8 | 105 | 211 | 0.04 | 0.12 | 0.02 |
| 0.49 | 32.4 | 0.05 | 18.1 | 45.6 | 91.8 | 186 | 0.12 | 0.01 | 0.008 |
| 0.75 | 37 | 0.1 | 20.6 | 57.7 | 101 | 197 | 0.03 | 0.11 | 0.02 |
| 0.63 | 40 | 0.12 | 23 | 36.1 | 96.3 | 188 | 0.14 | 0.07 | 0.02 |
| 0.63 | 57.8 | 0.07 | 27.9 | 39.6 | 101.8 | 180 | 0.05 | 0.01 | 0.013 |
| 0.68 | 45.2 | 0.19 | 15.5 | 69.2 | 97.3 | 200 | 0.09 | 0.03 | 0.011 |
| 0.7 | 39.12 | 0.13 | 13.3 | 64.7 | 103 | 200 | 0.02 | 0.02 | 0.015 |
| 0.49 | 32.5 | 0.05 | 18.2 | 49.7 | 92.1 | 197 | 0.13 | 0.23 | 0.014 |
| 0.615 | 25.84 | 0.24 | 16.77 | 37.15 | 98.91 | 194.74 | 0.05 | 0.13 | 0.012 |
| 0.628 | 46.76 | 0.22 | 15.71 | 43.29 | 102.2 | 216.3 | 0.17 | 0.15 | 0.009 |
| 0.817 | 25.99 | 0.12 | 16.2 | 39.96 | 99.5 | 206.6 | 0.02 | 0.083 | 0.011 |
| 0.63 | 54.33 | 0.19 | 14.7 | 41.9 | 101 | 204.4 | 0.07 | 0.13 | 0.012 |
| 0.824 | 36.25 | 0.09 | 18.54 | 34.78 | 91.46 | 186.62 | 0.21 | 0.04 | 0.02 |

- *Analisi esplorativa dei dati in uscita:*

```
#Analisi della correlazione
library(ellipse)
plotcorr(cor(compost))

#Statistiche riassuntive dei dati per il compost
summary(compost)
#       pH          Salinità        Umidità       Carbonio_org
#Min.   :7.50   Min.   :33.25   Min.   :32.50   Min.   :21.09
#1st Qu.:7.75   1st Qu.:36.60   1st Qu.:35.75   1st Qu.:23.20
#Median :8.00   Median :52.38   Median :39.20   Median :24.10
#Mean   :7.99   Mean   :49.08   Mean   :38.30   Mean   :24.92
#3rd Qu.:8.20   3rd Qu.:57.45   3rd Qu.:41.10   3rd Qu.:25.90
#Max.   :8.40   Max.   :68.10   Max.   :42.40   Max.   :30.40
```

```
#Acidi_Umici_Fulvici   Azoto_org      Rapporto_CN         Cadmio
#Min.   :8.000  Min.   :83.00   Min.   : 8.500   Min. :0.4800
#1st Qu.:9.385  1st Qu.:86.00   1st Qu.: 9.855   1st Qu.:0.5925
#Median :9.600  Median :88.00   Median :11.300   Median :0.6300
#Mean   :9.666  Mean   :87.81   Mean   :11.649   Mean   :0.6586
#3rd Qu.:10.100 3rd Qu.:88.92   3rd Qu.:12.650   3rd Qu.:0.7250
#Max.   :10.700 Max.   :94.00   Max.   :15.400   Max.   :1.0300
#  Cromo_tot        Mercurio          Nichel          Piombo
#Min.   :24.20   Min.   :0.0400   Min.   :12.80   Min.   :34.78
#1st Qu.:29.75   1st Qu.:0.0700   1st Qu.:15.96   1st Qu.:38.90
#Median :36.25   Median :0.1000   Median :18.10   Median :44.40
#Mean   :37.71   Mean   :0.1158   Mean   :18.95   Mean   :46.37
#3rd Qu.:42.60   3rd Qu.:0.1550   3rd Qu.:20.80   3rd Qu.:51.45
#Max.   :60.50   Max.   :0.2400   Max.   :28.60   Max.   :69.20
#     Rame            Zinco        Mat_plastico        Vetro
#Min.   : 82.3   Min.   :180.0   Min.   :0.0200   Min.   :0.01000
#1st Qu.: 96.8   1st Qu.:195.9   1st Qu.:0.0450   1st Qu.:0.02000
#Median :101.0   Median :204.4   Median :0.0900   Median :0.07000
#Mean   :103.2   Mean   :203.9   Mean   :0.0921   Mean   :0.08016
#3rd Qu.:106.5   3rd Qu.:216.2   3rd Qu.:0.1250   3rd Qu.:0.12500
#Max.   :139.0   Max.   :223.0   Max.   :0.2100   Max.   :0.23000
#   Metalli
#Min.   :0.00100
#1st Qu.:0.01000
#Median :0.01300
#Mean   :0.06505
#3rd Qu.:0.01850
#Max.   :1.00000

#Analisi dei grafici boxplot
#Standardizzo la matrice dei dati in uscita
n=19
p=17
media=function(p,x){
      temp=double(p)
      for(j in 1:p){
```

```
            temp[j]=mean(x[,j])
      }
      temp
}
xbar=media(p,compost)
S=var(compost)*(n)/(n-1)
x=matrix(0,19,17)
for(j in 1:17){
      for(i in 1:19){
            x[i,j]=(compost[i,j]-xbar[j])/S[j,j]
      }
}
boxplot(x[,1:6],ylim=c(-20,31))
boxplot(x[,7:12],ylim=c(-20,31))
boxplot(x[,13:17],ylim=c(-20,31))

#Verifico assunti normalità
#Test sulla normalità della variabile pH
shapiro.test(compost[,1])
#W = 0.9176, p-value = 0.1022
#Test sulla normalità della variabile Salinità
shapiro.test(compost[,2])
#W = 0.8889, p-value = 0.03078
#Test sulla normalità della variabile Umidità
shapiro.test(compost[,3])
#W = 0.934, p-value = 0.2055
#Test sulla normalità della variabile Carbonio_org
shapiro.test(compost[,4])
#W = 0.919, p-value = 0.1085
#Test sulla normalità della variabile Acidi_Umici_Fulvici
shapiro.test(compost[,5])
#W = 0.9434, p-value = 0.3028
#Test sulla normalità della variabile Azoto_org
shapiro.test(compost[,6])
#W = 0.9649, p-value = 0.6721
#Test sulla normalità della variabile Rapporto_CN
```

```
shapiro.test(compost[,7])
#W = 0.9546, p-value = 0.4708
#Test sulla normalità della variabile Cadmio
shapiro.test(compost[,8])
#W = 0.9056, p-value = 0.06154
#Test sulla normalità della variabile Cromo_tot
shapiro.test(compost[,9])
#W = 0.9163, p-value = 0.09667
#Test sulla normalità della variabile Mercurio
shapiro.test(compost[,10])
#W = 0.9086, p-value = 0.06964
#Test sulla normalità della variabile Nichel
shapiro.test(compost[,11])
#W = 0.9215, p-value = 0.1207
#Test sulla normalità della variabile Piombo
shapiro.test(compost[,12])
#W = 0.9047, p-value = 0.05932
#Test sulla normalità della variabile Rame
shapiro.test(compost[,13])
#W = 0.9073, p-value = 0.06615
#Test sulla normalità della variabile Zinco
shapiro.test(compost[,14])
#W = 0.9511, p-value = 0.4117
#Test sulla normalità della variabile Mat_plastico
shapiro.test(compost[,15])
#W = 0.9195, p-value = 0.1109
#Test sulla normalità della variabile Vetro
shapiro.test(compost[,16])
#W = 0.906, p-value = 0.06254
#Test sulla normalità della variabile Metalli
shapiro.test(compost[,17])
#W = 0.263, p-value = 7.133e-09
#Nuovo test di Shapiro Wilk per la variabile Metalli
compost[,17]=compost[-5,17]
shapiro.test(compost[,17])
#W = 0.9365, p-value = 0.2522
```

- *Costruzione delle carte multivariate per i dati in ingresso con R*

```
##Carico i dati, e li trasformo in matrice per poter effettuare i calcoli
forsu=as.matrix(read.table("E:/stage/dati/forsu.txt",header=T))
#Stimo la media e la varianza per i dati
n=19
p=8
media=function(p,x){
      temp=double(p)
      for(j in 1:p){
            temp[j]=mean(x[,j])
      }
      temp
}
xbar=media(8,forsu)
S=var(forsu)*(n)/(n-1)
#Standardizzo i dati
x=matrix(0,n,p)
for(j in 1:8){
      for(i in 1:19){
            x[i,j]=(forsu[i,j]-xbar[j])/sqrt(S[j,j])
      }
}
#Stimo la matrice di correlazione R
R=var(x)*n/(n-1)
#Calcolo i punti della statistica T di Hotelling
cartaT2=function(n,medie,S,x){
      T2=double(n)
      for(i in 1:n){
            T2[i]=t(x[i,]-medie)%*%solve(S)%*%(x[i,]-medie)
      }
      T2
}
T2=cartaT2(19,0,R,x)
T2
```

```
#6.171489   6.266713   7.602180   8.842353   7.255779   5.412582   7.194656
#7.093030   8.362971   7.040245   5.727870   9.691219   9.102240   9.234261
#10.021305  2.166818  8.192350  7.077591  3.965400

#Calcolo ora i limiti di controllo

alpha=0.05

UCL=(((n-1)^2)/n)*qbeta(1-alpha/2,(p/2),(n-p-1)/2)

LCL=0

#Disegno la carta

plot(T2,ylim=c(min(LCL,T2),max(UCL,T2)),type="l",main="Carta di controllo
T quadro")

points(T2)

abline(h=LCL,col="blue")

abline(h=UCL,col="red")

#Nessun fuori controllo, il processo è in controllo.

#Carico i dati di Fase II

forsuph2=as.matrix(read.table("E:/stage/dati/forsu_ph2.txt",header=T))

xbar2=media(p,forsuph2)

S2=var(forsuph2)*(n)/(n-1)

#Standardizzo i dati

x2=matrix(0,n,p)

for(j in 1:8){
     for(i in 1:19){
          x2[i,j]=(forsuph2[i,j]-xbar2[j])/sqrt(S2[j,j])
     }
}

#Calcolo i punti

T2ph2=cartaT2(19,0,R,x2)

T2ph2

#29.218466  1.700167 10.061975 14.057960  8.310258 15.703400 18.182499

#3.275655 13.366326 18.556601 21.999283 19.797220 50.081846  3.931679

#14.223575  9.312014  8.331782 13.861508  7.431986

#Limiti di Fase II

alpha=0.05

UCLph2=((p*(n-1)*(n+1))/(n*(n-p)))*qf(1-alpha/2,p,n-p)

LCLph2=0

#Disegno la carta
```

```
plot(T2ph2,type="l",ylim=c(min(T2ph2,LCLph2),max(T2ph2,UCLph2)),main="Car
ta di controllo T quadro di Fase II")

points(T2ph2)

abline(h=UCLph2,col="red")

abline(h=LCLph2,col="blue")

#Il proceso risulta essere in controllo

#Analisi delle componenti principali applicata alla matrice dei dati del
#F.O.R.S.U. standardizzata

prforsu=princomp(x,cor=T)

#Carico gli scores, i punteggi delle nuove coordinate

Y=prforsu$scores

#Carico gli autovalori

lambda=diag(prforsu$sd^2)

#Carico gli autovettori

gamma=prforsu$loadings

(prforsu$sd^2)/sum((prforsu$sd^2))

#      Comp.1       Comp.2       Comp.3       Comp.4       Comp.5       Comp.6

#4.660804e-01 2.001567e-01 1.580395e-01 8.893798e-02 5.044961e-02 2.970191e-02

      Comp.7       Comp.8

#6.633892e-03 1.315207e-17

#Le prime 3 componenti principali spiegano insieme 46.6+20+16=82.6%

#Costruisco quindi la carta T di Hotelling con le prime 3 comp. pr.

n=19

T2pc=double(n)

for(i in 1:n){

     T2pc[i]=Y[i,c(1:3)]%*%solve(lambda[c(1:3),c(1:3)])%*%(as.matrix(Y[i
,c(1:3)]))

}

T2pc

#1.708282 1.743962 1.442345 1.311799 4.021962 1.888178 5.994592 1.498850
#2.612679 5.037156 2.032588 2.551404 2.380819 5.910313 1.837209 1.442414
#4.469841 6.356329 2.759279

k=3

alpha=0.05

LI=0

LS=((k*(n+1)*(n-1))/((n-k)*n))*qf(1-alpha,k,(n-k))

plot(T2pc,ylim=c(LI,max(T2pc,LS)),type="l",main="Carta di controllo con
k=3 comp. principali")
```

```
points(T2pc)
abline(h=LI,col="blue")
abline(h=LS,col="red")
#Il processo è in controllo
#Applico la carta Q ai dati con le prime 3 comp. principali
Q=double(n)
for(i in 1:n){
     Q[i]=Y[i,c(4:8)]%*%as.matrix(Y[i,c(4:8)])
}
Q
#2.1354310  4.1591916  2.6837895  4.6852873  1.9823617  1.9347900  0.4768852
#3.6557177  3.2056663  1.2366746  3.0725791  5.3885891  5.3257874  2.7346280
#3.5103000 0.8345476 3.1916379 0.4972835 1.1984570
#Limiti per la carta Q
delta1=sum(lambda[4:8,4:8])
delta2=sum((lambda[4:8,4:8])^2)
delta3=sum((lambda[4:8,4:8])^3)
h0=1-((2*delta1*delta3)/(3*(delta2)^2))
alpha=0.05
LS=delta1*(((qnorm(1-alpha/2)*sqrt(2*delta2*(h0^2)))/(delta1))+(delta2*h0*(h0-
1))/(delta1^2)+1)^(1/h0)
LI=0
plot(Q,type="l",ylim=c(LI,max(LS,Q)),main="Carta   Q   per   k=3   comp.
principali")
points(Q)
abline(h=LS,col="red")
abline(h=LI,col="blue")
#Oppure Secondo Hawkins
n=19
Qpc=double(n)
for(i in 1:n){
     Qpc[i]=Y[i,c(4:8)]%*%solve(lambda[c(4:8),c(4:8)])%*%(as.matrix(Y[i,
c(4:8)]))
}
#Limiti per la carta Q di Hawkins
k=3
p=8
```

```
LI=0
LS=(k*(n+1)*(n-1)/(n*(n-1)-k+1))*qf(1-alpha,p-k,n-p+k)
plot(Qpc,type="l",ylim=c(LI,max(LS,Qpc)),main="Carta Q per k=3 comp. principali")
points(Qpc)
abline(h=LS,col="red")
abline(h=LI,col="blue")
#Il processo risulta in controllo
```

- *Costruzione delle carte multivariate per i dati in uscita con R*

```
##Carico i dati, e li trasformo in matrice per poter effettuare i calcoli
compost=as.matrix(read.table("E:/stage/dati/compost.txt",header=T))
#Stimo la media e la varianza per i dati
n=19
p=17
media=function(p,x){
      temp=double(p)
      for(j in 1:p){
            temp[j]=mean(x[,j])
      }
      temp
}
xbar=media(17,compost)
S=var(compost)*(n)/(n-1)
#Standardizzo i dati
x=matrix(0,n,p)
for(j in 1:p){
      for(i in 1:n){
            x[i,j]=(compost[i,j]-xbar[j])/sqrt(S[j,j])
      }
}
#Stimo la matrice di correlazione
R=var(x)*n/(n-1)
#Calcolo i punti della statistica T di Hotelling
cartaT2=function(n,medie,S,x){
      T2=double(n)
```

```
      for(i in 1:n){
            T2[i]=t(x[i,]-medie)%*%solve(S)%*%(x[i,]-medie)
      }
      T2
}
T2y=cartaT2(19,0,R,x)
T2y
#16.102742 14.444378  9.490723 15.013247 16.155009 15.948640 15.009275
#15.730362 12.419975 15.853752 15.394571 15.966937 16.152921 15.807408
#15.974264 16.137668 16.145102 16.144813 16.002948
#Calcolo ora i limiti di controllo
alpha=0.05
UCL=(((n-1)^2)/n)*qbeta(1-alpha/2,(p/2),(n-p-1)/2)
LCL=0
#Disegno la carta
plot(T2y,ylim=c(min(LCL,T2y),max(UCL,T2y)),type="l",main="Carta        di
controllo T quadro")
points(T2y)
abline(h=LCL,col="blue")
abline(h=UCL,col="red")
#Nessun fuori controllo, il processo è in controllo.
#Carico i dati di Fase II e li rendo adimensionali
compostph2=as.matrix(read.table("E:/stage/dati/compost_ph2.txt",header=T))
xbar2=media(p,compostph2)
S2=var(compostph2)*(n)/(n-1)
#Rendo adimensionali i dati
x2=matrix(0,n,p)
for(j in 1:p){
      for(i in 1:n){
            x2[i,j]=(compostph2[i,j]-xbar2[j])/sqrt(S2[j,j])
      }
}
#Calcolo i punti
T2ph2=cartaT2(19,0,R,x2)
T2ph2
#26.74337 116.61688 367.35008 257.74581 609.17105 287.49016  27.04251
```

```
#192.36774 276.23581 380.87562 102.70170  50.42253  61.44510 142.16059
#95.97107  55.70206 561.12485 148.40957 377.75428
#Limiti di Fase II
alpha=0.05
UCLph2=((p*(n-1)*(n+1))/(n*(n-p)))*qf(1-alpha/2,p,n-p)
LCLph2=0
#Disegno la carta
plot(T2ph2,type="l",ylim=c(min(T2ph2,LCLph2),max(T2ph2,UCLph2)),main="Car
ta di controllo T quadro di Fase II")
points(T2ph2)
abline(h=UCLph2,col="red")
abline(h=LCLph2,col="blue")
#Il proceso risulta essere in controllo
#Analisi delle componenti principali applicata alla matrice
#standardizzata del compost
prcompost=princomp(x,cor=T)
#Carico gli scores, i punteggi delle nuove coordinate
Y=prcompost$scores
#Carico gli autovalori
lambda=diag(prcompost$sd^2)
#Carico gli autovettori
gamma=prcompost$loadings
(prcompost$sd^2)/sum((prcompost$sd^2))
Comp.1       Comp.2       Comp.3       Comp.4       Comp.5       Comp.6
0.2387950209 0.1510652578 0.1287784323 0.1062099215 0.0913468426 0.0668864642
      Comp.7       Comp.8       Comp.9      Comp.10      Comp.11      Comp.12
0.0575685489 0.0409884100 0.0349248456 0.0242249626 0.0190098287 0.0151891268
     Comp.13      Comp.14      Comp.15      Comp.16      Comp.17
0.0100075057 0.0081426283 0.0055245767 0.0008865345 0.0004510929
#Le prime 5 componenti principali spiegano insieme 24+15+13+10+9=71%
#Costruisco quindi la carta T di Hotelling con le prime 5 comp. pr.
n=19
T2pc=double(n)
for(i in 1:n){
     T2pc[i]=Y[i,c(1:5)]%*%solve(lambda[c(1:5),c(1:5)])%*%(as.matrix(Y[i
,c(1:5)]))
}
```

```
T2pc

#5.676955   4.632012   3.963800   9.973786 11.194507   7.216289   4.281359
#4.327124   0.781459   2.959553   7.092543   4.918679   4.558621   2.755813
#3.150198  4.549942  2.910156  3.578909  6.478291

k=5

alpha=0.05

LI=0

LS=((k*(n+1)*(n-1))/((n-k)*n))*qf(1-alpha,k,(n-k))

plot(T2pc,ylim=c(LI,max(T2pc,LS)),type="l",main="Carta  di  controllo  con
k=5 comp. principali")

points(T2pc)

abline(h=LI,col="blue")

abline(h=LS,col="red")

#Il processo è in controllo

#Applico la carta Q ai dati con le prime 5 comp. principali

Q=double(n)

for(i in 1:n){

     Q[i]=Y[i,c(4:8)]%*%as.matrix(Y[i,c(4:8)])

}

Q

#4.919278 6.419589 4.393051 2.565615 5.385339 5.725118 5.932666 4.278519
#5.137804 3.379849 5.831900 4.281023 3.668516 7.279211 5.156297 3.224786
#3.293172 3.320814 7.476314

#Limiti per la carta Q

delta1=sum(lambda[6:17,6:17])

delta2=sum((lambda[6:17,6:17])^2)

delta3=sum((lambda[6:17,6:17])^3)

h0=1-((2*delta1*delta3)/(3*(delta2)^2))

alpha=0.05

LS=delta1*(((qnorm(1-alpha/2)*sqrt(2*delta2*(h0^2)))/(delta1))+(delta2*h0*(h0-
1))/(delta1^2)+1)^(1/h0)

LI=0

plot(Q,type="l",ylim=c(LI,max(LS,Q)),main="Carta   Q   per   k=5   comp.
principali")

points(Q)

abline(h=LS,col="red")

abline(h=LI,col="blue")

#Oppure Secondo Hawkins
```

```
n=19
Qpc=double(n)
for(i in 1:n){
      Qpc[i]=Y[i,c(6:17)]%*%solve(lambda[c(6:17),c(6:17)])%*%(as.matrix(Y
      [i,c(6:17)]))
}
#Limiti per la carta Q di Hawkins
k=5
p=17
LI=0
LS=(k*(n+1)*(n-1)/(n*(n-1)-k+1))*qf(1-alpha,p-k,n-p+k)
plot(Qpc,type="l",ylim=c(LI,max(LS,Qpc)),main="Carta  Q  per  k=5  comp.
principali")
points(Qpc)
abline(h=LS,col="red")
abline(h=LI,col="blue")
#Il processo risulta in controllo
```

- *Applicazione dell'Indice di Dissimilarità ai dati in ingresso con R*

```
#Applico L'Indice di Dissimilarità in ingresso
#Carico i dati
forsu=as.matrix(read.table("i:/stage/dati/forsu.txt",header=T))
#Standardizzo i dati
n=19
p=8
media=function(p,x){
      temp=double(p)
      for(j in 1:p){
            temp[j]=mean(x[,j])
      }
      temp
}
xbar=media(p,forsu)
S=var(forsu)*n/(n-1)
x=matrix(0,n,p)
for(i in 1:p){
```

```
      x[,i]=(forsu[,i]-xbar[i])/sqrt(S[i,i])
}
#Creo le sottomatrici
x2=x[c(10:19),]
x=x[c(1:9),]
#Creo la matrice R di covarianza
R=((9-1)/(10+9-1))*(1/8)*t(x)%*%x+((10-1)/(10+9-1))*(1/9)*t(x2)%*%x2
#Applico l'analisi delle component principali alla matrice R
prR=princomp(covmat=R,cor=F)
#Ottengo gli autovalori
lambda=diag(prR$sdev^2)
#Ottengo gli autovettori
gamma=prR$loadings
#Creo le osservazioni nelle nuove coordinate per la prima matrice
Y1=sqrt(9/19)*x%*%gamma%*%sqrt(solve(lambda))
#Creo le osservazioni nelle nuove coordinate per la seconda matrice
Y2=sqrt(10/19)*x2%*%gamma%*%sqrt(solve(lambda))
S1=var(Y1)
S2=var(Y2)
round(S1+S2)
#Soddisfa l'equazione S1+S2=I
#Applico quindi l'analisi delle componenti principali ad una delle due
#matrici di varianza
pcS=princomp(covmat=S1,cor=F)
#Ottengo lgi autovalori per questa
lambdaS=diag(pcS$sdev^2)
#Calcolo l'indice di dissimilarità
(4/p)*sum((lambdaS[c(1:8)]-0.5)^2)
#1
#il valore dell'indice è vicino ad 1 dunque i due dataset sono differenti
#tra loro,e quindi c'è stato uno shift nella struttura di correlazione
#che non è stato possibile individuare tramite le tecniche precedenti.
```

- *Applicazione dell'Indice di Dissimilarità ai dati in uscita con R*

```
#Applico L'Indice di Dissimilarità in uscita
```

```
#Carico i dati
compost=as.matrix(read.table("i:/stage/dati/compost.txt",header=T))
#Standardizzo i dati
n=19
p=p
media=function(p,x){
      temp=double(p)
      for(j in 1:p){
            temp[j]=mean(x[,j])
      }
      temp
}
xbar=media(p,compost)
S=var(compost)*n/(n-1)
x=matrix(0,n,p)
for(i in 1:p){

      x[,i]=(forsu[,i]-xbar[i])/sqrt(S[i,i])
}
#Creo le sottomatrici
x2=x[c(10:19),]
x=x[c(1:9),]
#Creo la matrice R di covarianza
R=((9-1)/(10+9-1))*(1/8)*t(x)%*%x+((10-1)/(10+9-1))*(1/9)*t(x2)%*%x2
#Applico l'analisi delle component principali alla matrice R
prR=princomp(covmat=R,cor=F)
#Ottengo gli autovalori
lambda=diag(prR$sdev^2)
#Ottengo gli autovettori
gamma=prR$loadings
#Creo le osservazioni nelle nuove coordinate per la prima matrice
Y1=sqrt(9/19)*x%*%gamma%*%sqrt(solve(lambda))
#Creo le osservazioni nelle nuove coordinate per la seconda matrice
Y2=sqrt(10/19)*x2%*%gamma%*%sqrt(solve(lambda))
S1=var(Y1)
S2=var(Y2)
```

```
round(S1+S2)
#Soddisfa l'equazione S1+S2=I
#Applico quindi l'analisi delle componenti principali ad una delle due
#matrici di varianza
pcS=princomp(covmat=S1,cor=F)
#Ottengo lgi autovalori per questa
lambdaS=diag(pcS$sdev^2)
#Calcolo l'indice di dissimilarità
(4/p)*sum((lambdaS[c(1:17)]-0.5)^2)
#1
#Il valore dell'indice è vicino ad 1 dunque i due dataset sono differenti
#tra loro,e quindi c'è stato uno shift nella struttura di correlazione
#che non è stato possibile individuare tramite le tecniche precedenti.
```

- *Teorema (Seber (1984, pp. 30-31)):*

Sia $T^2 = my^T W^{-1} y$, dove $y \sim N_d(0, \Sigma)$, $W \sim W_d(m, \Sigma)$, e $y$ e $W$ sono statisticamente indipendenti. $N_d$ e $W_d$ rappresentano rispettivamente una normale d-variata e una distribuzione Wishart.

Allora

$$\frac{m-d+1}{d} \cdot \frac{T^2}{m} \sim F(d, m-d+1)$$

Data questa notazione, consideriamo un set di osservazioni iniziali multivariate $X_1, X_2, \ldots, X_m$ e una futura osservazione $X_f$, dove ogni $X_i$ è un vettore di osservazioni su p variabili. Se

$$X_i \sim N_p(\mu, \Sigma)$$

allora

$$\overline{X}_m \sim N_p(\mu, \Sigma/m)$$

$$(m-1)S_m \sim W_p(m-1, \Sigma)$$

Supponiamo ora che $X_f$, $\overline{X}_m$, e $S_m$ siano indipendenti.

Allora

$$X_f - \overline{X}_m \sim N_p(0, (\frac{m+1}{m})\Sigma)$$

e

$$\sqrt{(\frac{m}{m+1})}(X_f - \overline{X}_m) \sim N_p(0, \Sigma)$$

Se si definisce la statistica

$$T^2 = \left(\frac{m}{m+1}\right)(X_f - \overline{X}_m)^T {S_m}^{-1}(X_f - \overline{X}_m)$$

allora

$$\frac{(m-p)}{p} \cdot \frac{T^2}{(m-1)} \sim F(p, m-1-p+1)$$

che porta

$$\frac{(m-p)}{p(m-1)} \cdot \frac{m}{(m+1)}(X_f - \overline{X}_m)^T {S_m}^{-1}(X_f - \overline{X}_m) \sim F(p, m-p)$$

e di conseguenza

$$(X_f - \overline{X}_m)^T {S_m}^{-1}(X_f - \overline{X}_m) \sim \frac{p(m-1)(m+1)}{m(m-p)} F(p, m-p)$$

# Bibliografia

Hawkins D. M. (1991), *Multivariate quality control based on regression-adjusted variables*, Technometrics, Vol. 33, 61-75

Iacus, S.M., Masarotto G. (2003), *Laboratorio di Statistica con R*, McGraw-Hill, Milano

Jackson, Mudholkar (1979), *Control Procedures for residuals associated with principal components analysis*, Technometrics, Vol. 21, No. 3, 341-349

Jackson J. E., *A User Guide to Principal Components*, Wiley, New York, 1991

Kano M., Ohno H., Hasebe (2001), *A new multivariate statistical process monitoring method using principal components analysis,* Computers & Chemical Engineering, Vol. 25, Agosto 2001, 1103-1113

Montgomery D. C., *Controllo Statistico della Qualità*, Mc Graw-Hill, Milano, 2000

Manabu Kano, Shinji Hasebe, Iori Hashimoto, Hiromu Ohno (2002), *Statistical Process Monitoring Based on Dissimilarity of Process Data*, AIChe Journal, Giugno 2002

Mardia, Kent, Bibby (1997), *Multivariate Analysis, Academic Press*, 1979

Piccolo, D. (1998), *Statistica*, Il Mulino, Bologna.

Robert L. Mason, Nola D. Tracy, John C. Young (1996) Monitoring a *Multivariate Step Process*, Journal of Quality Technology, Vol. 28, No. 1, 39-50

Robert L. Mason, Nola D. Tracy, John C. Young (1995) *Decomposition of $T^2$ for Multivariate Control Chart Interpretation*, Journal of Quality Technology, Vol. 27, No. 2, 99-108

Robert L. Mason, Nola D. Tracy, John C. Young (1992) *Multivariate Control Charts for Individual Observations*, Journal of Quality Technology, Vol. 24, No. 2, 88-95

S. Bersimis, S. Psarakis, J. Panaretos (2006), *Multivariate Statistical Process Control Charts: An Overview*, Quality and Reliability Engineering International, Vol. 23, 517-543