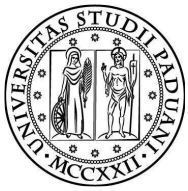


UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE

TESI DI LAUREA MAGISTRALE IN
INGEGNERIA CHIMICA E DEI PROCESSI INDUSTRIALI

TRASFERIMENTO DI MODELLI PER IL MONITORAGGIO DI PROCESSI CONTINUI E BATCH MEDIANTE TECNICHE STATISTICHE MULTIVARIATE

Relatore: Prof. Massimiliano Barolo
Correlatore: Ing. Pierantonio Facco

Laureanda: MARTINA LARGONI

ANNO ACCADEMICO 2011 – 2012

Riassunto

In questa Tesi è stato affrontato il problema del trasferimento di modelli per il monitoraggio di processo tra impianti che, pur essendo simili, presentano differenze, per esempio per scala, collocazione geografica, strumentazione o posizionamento dei sensori.
Due sono gli obiettivi perseguiti nella Tesi: il primo consiste nell'uniformare i metodi esistenti per il trasferimento e confrontarne le prestazioni, applicandoli ad un processo continuo di *spray-drying*. Il secondo è lo sviluppo di metodi per il trasferimento nel caso di processi batch, in particolare per un processo di produzione di penicillina.
L'ipotesi principale sui cui si basa il trasferimento è che il processo nei diversi impianti sia guidato dalle stesse forze motrici. Si può assumere pertanto che queste determinino una simile struttura di correlazione tra le variabili che vengono misurate in ciascun impianto. In particolare, la struttura di correlazione tra variabili misurate comuni ai diversi impianti è sfruttata per effettuare il trasferimento mediante il metodo PCA (analisi delle componenti principali) o il metodo JY-PLS (proiezione su strutture latenti a Y congiunta).
I risultati mostrano che il trasferimento è vantaggioso, in quanto permette di costruire modelli per il monitoraggio più efficienti rispetto a un modello costruito sui soli dati dell'impianto oggetto del trasferimento qualora siano disponibili pochi dati da esso. Il metodo PCA dà buone prestazioni nel trasferimento con una minore richiesta di dati rispetto al metodo JY-PLS. Solo quest'ultimo è in grado tuttavia di rilevare anomalie che si sviluppano esclusivamente su variabili non comuni tra i due impianti. È stato verificato che queste considerazioni sono valide sia per il processo continuo che per quello batch.
Infine, per il caso batch sono stati analizzati gli effetti della scelta delle variabili comuni, evidenziando come l'inserimento delle variabili controllate in questo insieme peggiori le prestazioni del modello.

Indice

Nomenclatura

a	=	indicatore generico per il numero di componenti principali (-)
A	=	numero di componenti principali (-)
A_n	=	allarme segnato al campione n (-)
A_{tot}	=	numero totale di allarmi generati (-)
c_p	=	calore specifico
d	=	valore generico della variabile indicatrice (%)
D_0	=	valore iniziale della variabile indicatrice (g/L)
D_f	=	valore finale della variabile indicatrice (g/L)
$\mathbf{e}_i$	=	vettore riga della matrice dei residui **E** (-)
$\mathbf{e}_k$	=	vettore riga contenente i residui del campione $\mathbf{x}_k$ (-)
$\mathbf{e}_{new}$	=	vettore riga contenente i residui del campione $\mathbf{x}_{new}$ (-)
$\mathbf{E}$	=	matrice degli errori nei metodi statistici multivariati per la matrice **X** (-)
$\mathbf{E}^J$	=	matrice congiunta dei residui per il modello JY-PLS (-)
$F_{A,(k-A),\alpha}$	=	distribuzione statistica F (-)
$\mathbf{F}$	=	matrice degli errori nei metodi statistici multivariati per la matrice **Y** (-)
h_0	=	coefficiente numerico della formula di Jackson-Mudholkar (-)
i	=	indicatore generico di un'osservazione o pedice generico (-)
I	=	numero totale di osservazioni (campioni o batch) (-)
j	=	numero campioni in normali condizioni operative (-)
k	=	indicatore per l'istante a cui si acquisisce un nuovo campione o indicatore generico (-)
K	=	istanti temporali di campionamento totali (-)
$\mathbf{L}$	=	matrice di correlazione degli *score* delle variabili latenti (-)
$m_{SPE,k}$	=	media degli $SPE_{Lim,k}$ (-)
$\dot{M}_{gas}$	=	portata di gas che entra nella camera di essiccamento
$\dot{M}_{sol}$	=	portata di soluzione
$\mathbf{M}_r$	=	generica matrice delle variabili di processo di rango r (-)
n	=	indicatore generico per i campioni proiettati (-)
N	=	campioni totali di un impianto (-)
$\mathbf{p}_i$	=	generico vettore colonna della matrice dei *loading* **P** (-)
$\mathbf{p}_r$	=	*loading* della generica matrice $\mathbf{M}_r$ (-)
$\mathbf{P}$	=	matrice dei *loading* (-)
$PRESS_j$	=	errore di predizione sulla somma dei quadrati dei residui (-)

$\mathbf{Q}$ = matrice dei *loading* generica per la matrice $\mathbf{Y}$ (-)

$\mathbf{Q}^{\mathrm{J}}$ = matrice dei *loading* congiunta nel metodo JY-PLS (-)

$\mathbf{Q}_k^{\mathrm{J\,T}}$ = trasposta della matrice dei *loading* congiunta all'istante (k-1) nel metodo JY-PLS (-)

r = indicatore generico per il rango di una matrice (-)

R = rango di una generica matrice (-)

s_a = generico semiasse dell'ellissoide di confidenza nel diagramma degli *score* (-)

S = campioni proiettati nel modello di monitoraggio (-)

$\mathrm{SPE}_{\mathrm{Lim},k}$ = limite dell'errore di predizione al quadrato all'istante k (-)

SPE_k = errore di predizione al quadrato per il nuovo campione $\mathbf{x}_k$ all'istante k (-)

SPE_i = errore di predizione al quadrato per il generico campione i (-)

SPE_n = errore di predizione al quadrato per il campione n proiettato (-)

$\mathrm{SPE}_{\alpha,Lim}$ = limite dell'errore di predizione al quadrato (-)

$\mathbf{t}_i$ = generico vettore colonna della matrice degli *score* $\mathbf{T}$ (-)

$\mathbf{t}_r$ = *score* della generica matrice $\mathbf{M}_r$ (-)

$\hat{\mathbf{t}}_k^{\mathrm{B}}$ = predizione del vettore degli *score* per $\mathbf{x}_k^{\mathrm{B}}$ all'istante k (-)

$\hat{\mathbf{t}}_{new}$ = predizione del vettore degli *score* per un nuovo campione $\mathbf{x}_{new}$ (-)

$\mathbf{T}$ = matrice degli *score* sulle variabili di processo (-)

T^{in} = temperatura in ingresso (°C)

T^{out} = temperatura in uscita (°C)

T^2 = statistica di Hotelling (-)

$T^2_{\mathrm{A},k,\alpha}$ = limite di confidenza per il diagramma degli *score* e T^2 (-)

T_i^2 = generica distanza dall'origine del diagramma degli *score* nel loro piano (-)

T^2_k = statistica di Hotelling per il nuovo campione all'istante k (-)

T^2_{Lim} = limite della statistica T^2 di Hotelling

T^2_n = statistica di Hotelling del campione n (-)

$\mathbf{U}$ = matrice degli *score* per $\mathbf{Y}$ (-)

v = indicatore generico per le variabili (-)

$\mathbf{v}'$ = variabili comuni di un impianto per il metodo PCA (-)

$\mathbf{v}''$ = variabili comuni di un impianto per il metodo JY-PLS nel caso continuo o variabili non comuni nel caso batch (-)

$\mathbf{v}'''$ = variabili non comuni di un impianto per il metodo JY-PLS (-)

V = numero totale delle variabili di processo misurate (-)

wtd = variabile indipendente dell'impianto (-)

$\mathbf{wtd}_j$ = vettore delle variabili indipendenti dell'impianto all'istante k (-)

W = ampiezza iniziale della finestra di campioni dell'impianto B (campioni)

$\mathbf{W}_k^{*\mathrm{B}}$ = matrice dei vettori dei pesi di $\mathbf{X}''^{\mathrm{B}}$ all'istante k (-)

$\mathbf{x}^{\mathrm{BF}}$ = vettore di dati di un'anomalia (-)

$\mathbf{x}_i$ = vettore riga di $\mathbf{X}$ (-)

$x_{i,v}$ = elemento della matrice $\mathbf{X}$ (-)

$\hat{x}_{i,v}$ = stima del vettore $x_{i,v}$ (-)

$\mathbf{x}_{new}$ = generico vettore di nuovi dati (-)

$\hat{\mathbf{x}}_{new}$ = predizione del vettore $\mathbf{x}_{new}$ (-)

$\mathbf{x}_v$ = vettore colonna della matrice $\mathbf{X}$ (-)

$\bar{\mathbf{x}}_v$ = vettore dei valori medi per ogni colonna della matrice $\mathbf{X}$ (-)

x_{solid} = frazione di massa di solido in soluzione (-)

$\mathbf{x}_k^{\text{B}}$ = nuovo campione dell'impianto B all'istante k (-)

$\mathbf{x}_k'^{\text{B}}$ = nuovo campione dell'impianto B all'istante k in cui le variabili sono $\mathbf{v}'^{\text{B}}$ (-)

$\mathbf{x}_k''^{\text{B}}$ = nuovo campione dell'impianto B all'istante k in cui le variabili sono $\mathbf{v}'''^{\text{B}}$ (-)

$\mathbf{X}$ = matrice bidimensionale delle variabili di processo misurate (-)

$\underline{\mathbf{X}}$ = matrice tridimensionale delle variabili di processo misurate (-)

$\mathbf{X}_k'$ = matrice derivante dal concatenamento verticale di $\mathbf{X}'^{\text{A}}$ e $\mathbf{X}_j'^{\text{B}}$ all'istante k (-)

$\mathbf{X}'$ = matrice dei dati delle variabili v' (-)

$\mathbf{X}''$ = matrice dei dati delle variabili $\mathbf{v}'''$ per il metodo JY-PLS (-)

$\mathbf{X}_{\text{SUB}}^{\text{A}}$ = matrice ottenuta da $\mathbf{X}^{\text{A}}$ per selezione di campioni simili a quelli di $\mathbf{X}_j^{\text{B}}$ (-)

$\mathbf{X}_k^{\text{AB}}$ = matrice che deriva dal concatenamento verticale di $\mathbf{X}_{\text{SUB}}^{\text{A}}$ e $\mathbf{X}_j^{\text{B}}$ (-)

$\mathbf{X}_j^{\text{B}}$ = matrice dei dati dell'impianto B disponibile all'istante k (-)

$\mathbf{X}_j'^{\text{B}}$ = matrice dei dati delle variabili $\mathbf{v}'^{\text{B}}$ dell'impianto B disponibile all'istante k (-)

$\mathbf{X}_j''^{\text{B}}$ = matrice dei dati delle variabili $\mathbf{v}'''^{\text{B}}$ dell'impianto B disponibile all'istante k (-)

$y_{k,v}''^{B}$ = valore misurato della variabile comune v all'istante k per l'impianto B (-)

$\hat{y}_{k,v}''^{B}$ = valore della predizione di $y_{k,v}^{B}$ (-)

$\mathbf{y}_k''^{\text{B}}$ = nuovo campione dell'impianto B all'istante k in cui le variabili sono $\mathbf{v}''^{\text{B}}$ (-)

$\mathbf{Y}$ = matrice di variabili comuni tra impianti (-)

$\mathbf{Y}''$ = matrice dei dati delle variabili $\mathbf{v}''$ per il metodo JY-PLS (-)

$\mathbf{Y}_j^{\text{B}}$ = vettore delle variabili indipendenti dell'impianto B all'istante k (-)

$\mathbf{Y}_j''^{\text{B}}$ = matrice dei dati delle variabili $\mathbf{v}''^{\text{B}}$ dell'impianto B disponibile all'istante k (-)

$\mathbf{Y}^{\text{J}}$ = matrice $\mathbf{Y}$ congiunta per il metodo JY-PLS (-)

z_α = deviazione normale standard (-)

Apici

A = relativo all'impianto A

B = relativo all'impianto B

T = trasposto

x = relativo allo spazio delle variabili non comuni

y = relativo allo spazio delle variabili comuni

$^{-1}$ = inversa di una matrice

Lettere greche

α = limite di fiducia (-)
$\mathbf{\Lambda}$ = matrice diagonale degli autovalori (-)
$\boldsymbol{\lambda}$ = vettore delle varianze degli *score* delle variabili latenti (-)
λ_a = autovalore della matrice $\mathbf{\Lambda}$ associato alla *a*-esima componente principale (-)
Δ = numero di campioni consecutivi (-)
ΔH^{vap} = calore di vaporizzazione
θ_i = coefficienti della formula di Jackson-Mudholkar (-)
σ = varianza (-)
σ^2 = deviazione standard (-)
$\chi^2_{2m^2_{SPE}/\sigma_{SPE},\alpha}$ = funzione di distribuzione χ^2 con $\frac{2m^2_{SPE}}{\sigma^2_{SPE}}$ gradi di libertà e probabilità α (-)
ψ_n = variabile che indica se un campione *k* è all'interno o all'esterno del limite di confidenza delle carte di monitoraggio SPE e T^2 (-)

Acronimi

AR = frequenza degli allarmi
DAE = equazioni differenziali e algebriche
JY-PLS = proiezione *joint*-Y su strutture latenti
NOC = normali condizioni operative
PCA = metodo dell'analisi delle componenti principali
PC = componenti principali
PID = proporzionale integrale differenziale
PLS = metodo della proiezione su strutture latenti
RMSECV$_j$ = *root-mean square error of cross validation*
SPE = errore di predizione al quadrato
TD = ritardo di rilevazione
Wtd = *weighted temperature difference*

Introduzione

Il monitoraggio di processo è indispensabile per salvaguardare l'impianto da problemi di malfunzionamento delle apparecchiature e dei sensori, in modo che vengano assicurate le condizioni normali dell'esercizio e il prodotto finale abbia le qualità desiderate. Una possibilità è il controllo di processo statistico multivariato (MSPC) (Nomikos e MacGregor, 1994). Durante le fasi iniziali di funzionamento di un impianto, si hanno a disposizione pochi dati delle variabili di processo misurate e il monitoraggio può non essere efficiente. Se il modello per il monitoraggio non è adeguato e non è in grado di rilevare anomalie quando si manifestano, queste potrebbero essere segnalate con elevato ritardo, cioè quando il prodotto è ormai fuori specifica. Il trasferimento di modelli per il monitoraggio di processo è proposto per risolvere questo problema. Si considerano due impianti che hanno gli stessi meccanismi fisici alla base, i quali realizzano lo stesso processo. Gli impianti possono tuttavia presentare delle differenze di scala, configurazione, condizioni operative, sistema di misurazione, posizionamento geografico. Spesso, accade che in uno solo dei due impianti sia stata condotta una sperimentazione adeguata e siano disponibili dati sul processo per la costruzione di un modello di monitoraggio per l'operazione nell'impianto stesso. Da qui nasce l'idea di sviluppare dei metodi per trasferire il modello di monitoraggio dall'impianto in cui i dati sono disponibili all'impianto, simile al primo, che marcia da poco tempo. Lo scopo di questi metodi è di utilizzare i dati dell'impianto su cui sono disponibili per costruire un modello con cui monitorare il processo del nuovo impianto, fino a che non è acquisito da quest'ultimo un *set* di dati sufficiente per poter definire solo con essi un efficiente modello di monitoraggio.

Il problema del trasferimento di modello è già stato affrontato per altre applicazioni (Feudale *et al.*, 2002; Lu *et al.*, 2008, 2009). Metodi per il trasferimento di modelli di monitoraggio sono stati recentemente proposti da Facco *et al.* (2012) e Tomba *et al.* (2012). In particolare, Facco *et al.* (2012) hanno sviluppato dei metodi sulla base delle informazioni di processo disponibili (dati delle variabili di processo), mentre Tomba *et al.* (2012) hanno fatto riferimento anche a leggi fisiche (bilanci di conservazione) sul processo. In entrambi i casi, il trasferimento si basa su un approccio di modellazione a variabili latenti, in cui sono utilizzati i dati degli impianti disponibili in condizioni operative normali (NOC), assumendo che la struttura di correlazione delle variabili di processo misurate su entrambi gli impianti (variabili comuni) sia simile. Il modello di monitoraggio viene poi aggiornato in modo adattativo con i nuovi dati fino a che non si acquistano sufficienti dati dall'impianto oggetto del trasferimento.

Sulla base di questi lavori, in questa Tesi si applicano metodi statistici multivariati per effettuare il trasferimento di modelli di monitoraggio nel caso di un reale processo continuo di

spray-drying per l'industria farmaceutica e nel caso simulato di un processo batch di produzione di penicillina.

Nel caso del processo continuo si utilizzano i metodi per il trasferimento proposti da Facco *et al.* (2012) e Tomba *et al.* (2012). Questi metodi presentano diverse procedure di aggiornamento del modello ai nuovi dati e diversi criteri di generazione degli allarmi quando rilevano un'anomalia. Nella Tesi i metodi vengono uniformati in modo da operare un confronto tra essi e stabilire il metodo che permette di ottenere le prestazioni migliori per il monitoraggio. È necessario stabilire il metodo più veloce nell'adattamento ai nuovi dati, poiché un processo industriale deve essere monitorato in modo efficace in tempi brevi. Inoltre, nella Tesi si stabilisce la metodologia più robusta che possa rilevare diversi tipi di anomalie.

Nel caso del processo batch, invece, oltre a risolvere problemi di sincronizzazione e gestione della dinamica, la Tesi sviluppa nuovi metodi per il trasferimento. In particolare, è dimostrato il vantaggio del trasferimento, indipendentemente dalla natura del processo, e viene studiata la sensitività dei modelli di monitoraggio alla scelta delle variabili con cui sono costruiti.

La Tesi è organizzata in quattro capitoli. Il Capitolo 1 spiega in dettaglio il monitoraggio statistico di processo mediante l'analisi delle componenti principali (PCA) con l'introduzione delle carte di monitoraggio e dei limiti di controllo. Inoltre, vengono definiti i metodi per il trasferimento dei modelli, con riferimento per entrambi i casi ai processi continui e a quelli batch. Nel Capitolo 2 sono descritti i processi studiati e sono presentati i dati disponibili per ogni caso di studio. Il Capitolo 3 presenta i risultati del trasferimento per il processo continuo, operando il confronto tra i metodi proposti da Facco *et al.* (2012) e da Tomba *et al.* (2012), precedentemente modificati in modo che i risultati del trasferimento siano confrontabili. Infine il Capitolo 4 riporta l'analisi del processo batch, analizzando le prestazioni del monitoraggio nel caso si utilizzassero solo i dati dell'impianto oggetto del trasferimento, e il miglioramento apportato allo stesso caso nel momento in cui si introduce il trasferimento. Inoltre, viene testato l'effetto delle variabili con i cui dati si costruisce il modello di monitoraggio.

Capitolo 1

Metodi statistici multivariati per il monitoraggio di processo

Questo Capitolo illustra le tecniche statistiche che si utilizzano in questa Tesi per il monitoraggio di processo. Le applicazioni di queste tecniche sono relative sia a processi continui che a processi discontinui. Inoltre vengono presentate alcune tecniche per il trasferimento di modelli per il monitoraggio.

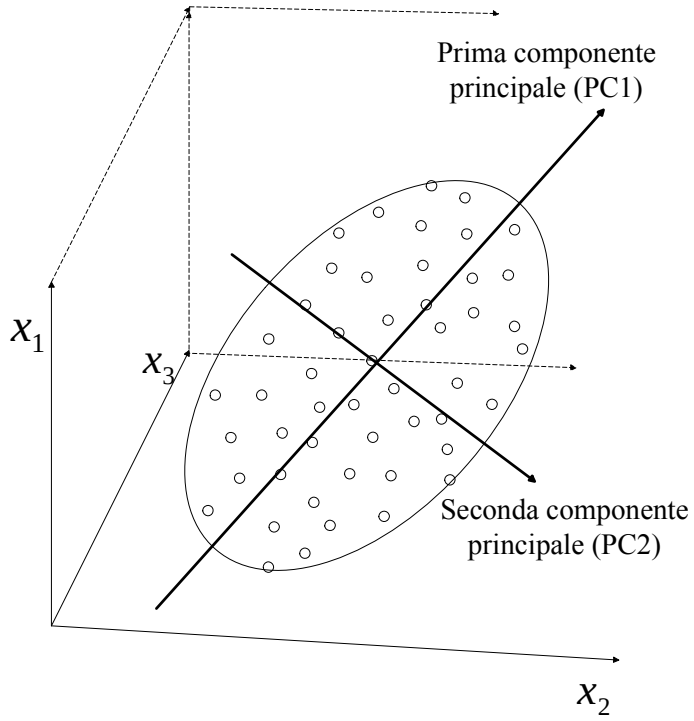
1.1 Monitoraggio statistico di processo

I processi industriali vengono monitorati al fine di garantire la qualità desiderata per i prodotti finali, capire se un processo sta marciando correttamente e rilevare eventuali problemi all'impianto, i quali possono compromettere le qualità del prodotto o addirittura le apparecchiature e la sicurezza delle persone. Usualmente, i sensori presenti nell'impianto raccolgono dati di processo in tempo reale, i quali vengono contestualmente registrati. Spesso, la quantità di informazioni raccolta dalla strumentazione è ridondante e il controllo diventa molto complesso, specie nel caso di processi batch. Al fine di monitorare in modo efficiente un processo durante la sua marcia, sono state proposte metodologie di controllo statistico multivariato (Nomikos e MacGregor, 1994). Il controllo statistico di processo si basa su metodi statistici multivariati, mediante i quali si monitorano i dati acquisiti in un impianto rispetto alle condizioni operative normali. Inoltre, i metodi statistici multivariati sono in grado di eliminare la ridondanza dei dati ed estrarre le informazioni utili per il monitoraggio in forma sintetica.
La tecnica principalmente utilizzata per il monitoraggio statistico multivariato è l'analisi delle componenti principali e può essere applicata a processi continui o a processi batch con lo sviluppo di apposite carte di monitoraggio. Con esse, fissati dei limiti di confidenza da dati "normali" di processo, si osserva, con una predeterminata soglia di probabilità, se il processo marcia in condizioni normali o anomale.

1.1.1 Analisi delle componenti principali PCA

L'analisi delle componenti principali (PCA, *principal component analysis*) è un metodo statistico multivariato molto efficace nel comprimere una serie di dati correlati ed estrarne le

informazioni più rilevanti che descrivono la variabilità sistematica dei dati (Jackson, 1991). Spesso, infatti, nelle misure di processo ci sono informazioni molto correlate e per questo ridondanti. Con PCA si possono estrarre le informazioni di covarianza e correlazione, trovando la combinazione delle variabili che descrive la variabilità dei dati (Wise e Gallagher, 1996). A livello geometrico, l'obiettivo della PCA è di individuare le direzioni di massima variabilità dei dati, dette anche componenti principali (PC). Le componenti principali sono combinazioni lineari delle variabili originarie e sono tra loro ortogonali. In Figura 1.1 è illustrato un esempio di riduzione di uno spazio da tridimensionale a bidimensionale. Nello spazio originario, la variabilità dei dati si sviluppa su tre variabili di processo, genericamente x_1, x_2, x_3. Essa può essere rappresentata in uno spazio fittizio bidimensionale di componenti principali, definite da PCA, che rappresentano le informazioni contenute nello spazio originario, senza perdite rilevanti di informazioni. Infatti, lo spazio delle variabili originarie è intrinsecamente bidimensionale per la correlazione esistente tra x_1, x_2, x_3. Senza perdita di generalità, questo discorso è stato affrontato per un sistema tridimensionale, ma può essere esteso ad un generico spazio *V*-dimensionale, con qualsivoglia *V*.



PrinComp.vsd

Figura 1.1. *Riduzione da uno spazio tridimensionale ad un piano definito dalle componenti principali (PC1 e PC2), che descrivono la massima variabilità dei dati.*

In altri termini, PCA realizza una decomposizione delle variabili di processo in autovettori della matrice di covarianza di $\mathbf{X}(I \times V)$, matrice di *I* campioni (osservazioni) delle *V* variabili di processo:

$$\mathrm{cov}(\mathbf{X}) = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{I-1}. \tag{1.1}$$

Il metodo suddivide una matrice **X** di rango *R* in una somma di *R* matrici $\mathbf{M}_r$ di rango 1, con $r = 1, ..., R$:

$$\mathbf{X} = \mathbf{M}_1 + \mathbf{M}_2 + \mathbf{M}_3 + ... + \mathbf{M}_r + ... + \mathbf{M}_R. \tag{1.2}$$

La generica matrice $\mathbf{M}_r$ può essere rappresentata con il prodotto esterno di due vettori $\mathbf{t}_r$ e $\mathbf{p}_r$, rispettivamente *score* e *loading*. Riscrivendo l'Equazione (1.2) si ottiene:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \mathbf{t}_3\mathbf{p}_3^{\mathrm{T}} + ... + \mathbf{t}_r\mathbf{p}_r^{\mathrm{T}} + ... + \mathbf{t}_R\mathbf{p}_R^{\mathrm{T}}, \tag{1.3}$$

dove l'apice T indica la trasposizione del vettore.

PCA esegue l'operazione algebrica di approssimazione:

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{t}_a\mathbf{p}_a^{\mathrm{T}} + \mathbf{E} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}, \tag{1.4}$$

dove **E** è la matrice dei residui, **T** la matrice degli *score*, **P** la matrice dei *loading* e *A* è il numero di componenti principali ($A \le \min(I,V)$), le quali descrivono la parte rilevante della variabilità dei dati. *A* viene scelto sulla base di analisi di convalida incrociata, come sarà descritto al Paragrafo 1.1.1.2.

In dettaglio, gli *score* sono combinazioni dei dati originari secondo:

$$\mathbf{t}_i = \mathbf{X}\mathbf{p}_i. \tag{1.5}$$

La matrice degli *score* **T**, che ha per righe i vettori $\mathbf{t}_i$, ha dimensioni (*I*×*A*) e rappresenta le coordinate dei dati sullo spazio individuato dalle componenti principali. Gli *score* contengono le informazioni su come i campioni si relazionano tra loro.

La matrice **P** dei *loading* (*A*×*V*) ha per righe i vettori $\mathbf{p}_i$, gli autovettori della matrice di covarianza. Contiene le informazioni di come le variabili si relazionano tra loro e i suoi elementi sono i coseni direttori di ciascuna componente principale. Poiché gli *score* sono tra loro ortogonali e i *loading* ortonormali, le componenti principali sono tra loro non correlate.

Le coppie $\mathbf{t}_i$ e $\mathbf{p}_i$ possono essere disposte in ordine decrescente dei rispettivi autovalori, i quali sono misure della varianza spiegata dalla *a*-esima componente principale. Tale varianza può essere intesa come una quantità di informazioni del *set* originario di dati nello spazio definito dalle componenti principali, se esiste grande correlazione tra le variabili originarie. Il dato viene dunque rappresentato da un numero di variabili inferiore a quello originario (usualmente, $A << \min(I,V)$) , senza perdere informazioni rilevanti e sistematiche, qualora *A* sia scelto opportunamente. I residui raccolti in **E** corrispondono alle informazioni non rappresentate dal modello (per esempio: rumore). È necessario definire delle statistiche che quantificano questa capacità di rappresentare i dati da parte del modello PCA. In questo modo, per i dati disponibili e per eventuali nuovi dati che vengono proiettati sul modello, è possibile definire la loro normalità sulla base dei valori delle statistiche, rispetto ad un limite definito nella fase di calibrazione del modello.

La mancanza di accuratezza statistica nel regredire i dati è rappresentata dall'errore quadratico medio SPE_i (SPE, *squared prediction error*). SPE_i è la somma dei quadrati di ciascun campione (riga) di **E**, ovvero per l'*i*-esimo campione:

$$SPE_i = \mathbf{e}_i\mathbf{e}_i^T = \mathbf{x}_i\left(\mathbf{I} - \mathbf{P}\mathbf{P}^T\right)\mathbf{x}_i^T, \tag{1.6}$$

dove $\mathbf{e}_i$ è un vettore riga della matrice dei residui, $\mathbf{x}_i$ il campione *i*-esimo e **I** la matrice identità. La statistica SPE indica quanto bene ogni campione viene rappresentato dal modello e, in termini geometrici, il valore $\sqrt{SPE_i}$ rappresenta la distanza euclidea dell'*i*-esimo punto dall'iperpiano di dimensioni ridotte costituito dalle variabili latenti.

Per quantificare quanto un'osservazione è lontana dalla media, cioè quanto un punto è lontano dall'origine del sistema delle componenti principali, si introduce la statistica T^2 di Hotelling. Essa è la somma al quadrato degli *score* normalizzati secondo la varianza spiegata ed è definita come:

$$T_i^2 = \mathbf{t}_i\mathbf{\Lambda}^{-1}\mathbf{t}_i^T, \tag{1.7}$$

dove $\mathbf{\Lambda}^{-1}$ è l'inversa della matrice diagonale degli autovalori λ_i (Wise e Gallagher, 1996).

Le due statistiche vengono utilizzate per il monitoraggio di processo. Il modello per il monitoraggio viene costruito con un *set* di dati di calibrazione, costituito da campioni del processo in condizioni operative normali (NOC, *normal operating conditions*). Quando si ha un nuovo campione $\mathbf{x}_{new}$ lo si proietta all'interno del modello, predicendo lo *score* secondo:

$$\hat{\mathbf{t}}_{new} = \mathbf{x}_{new}\mathbf{P}. \tag{1.8}$$

Anche in questo caso ne deriva un vettore dei residui $\mathbf{e}_{new}$:

$$\mathbf{e}_{new} = \mathbf{x}_{new} - \hat{\mathbf{x}}_{new}, \tag{1.9}$$

dove:

$$\hat{\mathbf{x}}_{new} = \hat{\mathbf{t}}_{new}\mathbf{P}^T. \tag{1.10}$$

Il modello costruito con PCA permette di fare il monitoraggio di un processo, sia continuo che discontinuo mediante carte di monitoraggio. Esse sono luoghi di punti in cui sono definiti dei limiti di controllo statistici. Normalmente si costruiscono carte di monitoraggio relative agli *score* e alle statistiche SPE e T^2. Quando il nuovo campione $\mathbf{x}_{new}$ è disponibile si confrontano *score*, SPE e T^2 con i rispettivi limiti stabiliti dal modello sulle NOC. Se le statistiche sono all'interno dei limiti con predeterminata probabilità, il processo sta marciando in condizioni normali; quando le statistiche eccedono i limiti si segnala la presenza di anomalie e le variabili responsabili possono essere individuate in modo da diagnosticare le cause dell'anomalia.

1.1.1.1 Trattamento preliminare dei dati

Al fine di estrarre le caratteristiche di correlazione e non semplicemente di covarianza, la matrice dei dati **X** deve essere pre-trattata. Eseguendo un *autoscaling*, la matrice di covarianza delle misure corrisponde alla matrice di correlazione. L'*autoscaling* consiste in un

centramento al valor medio (*mean centering*) e una riduzione a varianza unitaria (*scaling*). Il *mean centering* consiste nel sottrarre la media per ogni variabile (Kourti, 2003):

$$\bar{\mathbf{x}}_v = \frac{\sum_{i=1}^{I} x_{i,v}}{I}, \tag{1.11}$$

in cui $x_{i,v}$ è l'elemento della matrice $\mathbf{X}(I \times V)$ situato nella riga i e nella colonna v.

Lo *scaling* compensa le differenze di unità di misura diverse tra variabili, in modo da dare a tutte lo stesso peso. Si effettua dividendo tutte le misure di una variabile per la deviazione standard della variabile stessa, in modo che la varianza per tutte le variabili risulti unitaria:

$$\mathrm{var}(\mathbf{x}_v) = \frac{\sum_{i=1}^{I} \left(x_{i,v} - \bar{\mathbf{x}}_v\right)^2}{I} \tag{1.12}$$

e

$$\sigma = \sqrt{\mathrm{var}(\mathbf{x}_v)}\,. \tag{1.13}$$

Tutti i metodi statistici multivariati che vengono in seguito analizzati trattano matrici di dati che hanno subito come trattamento preliminare l'*autoscaling*.

1.1.1.2 Selezione del numero di componenti principali

Quando si esegue l'operazione di approssimazione definita nella (1.4), il residuo **E** deve essere minimizzato per aumentare la rappresentatività del modello. La scelta del numero di componenti principali con cui costruire il modello è quindi un passo chiave. Le metodologie di scelta seguite per questa Tesi sono principalmente due.

La prima è la convalida incrociata (*cross validation*), dovuta a Mosteller e Wallace (1963) e Stone (1974). In essa, la matrice dei dati **X** viene suddivisa in segmenti, costituiti da una o più righe, e viene costruito un modello con PCA sulla matrice a meno di un segmento; con questo segmento viene verificato il modello in convalida. La procedura si applica per più segmenti e ad ogni iterazione si valuta l'errore in termini di errore medio quadratico di convalida incrociata (*RMSECV, root-mean squared error of cross validation*):

$$RMSECV_v = \sqrt{\frac{PRESS_v}{I}}, \tag{1.14}$$

in cui

$$PRESS_v = \sum_{i=1}^{I} \left(x_{i,v} - \hat{x}_{i,v}\right)^2. \tag{1.15}$$

Abitualmente l'aggiunta di componenti principali al modello fa decrescere il valore dell'errore nel *set* di calibrazione. Quando però il numero di componenti principali è eccessivo si descrive una varianza poco rilevante del *set* di calibrazione, dovuta unicamente al rumore, per cui l'errore sul *set* di convalida cresce. Il punto di minimo del *RMSECV* al variare del numero di componenti principali individua il numero ottimale per costruire il modello, in relazione allo specifico *set* di dati.

Un secondo criterio adottato per la maggior parte delle analisi riportate in questa Tesi è la regola dell'autovalore λ_a (Martens e Naes, 1989). Si tratta di una regola pratica che risulta più immediata e di semplice applicazione anche dal punto di vista computazionale. Essa si basa sul fatto che l'autovalore λ_a è una stima indiretta del numero delle *J* variabili originarie rappresentate dalla *a*-esima componente principale. Per questo motivo, gli autovalori λ_a con valore superiore all'unità rappresentano più di una variabile. Dunque, secondo la regola, il numero di componenti principali da adottare per la costruzione del modello deve corrispondere al numero di autovalori che soddisfano il criterio $\lambda_a \geq 1$. Questa regola viene utilizzata principalmente per lo studio dei processi continui, mentre nel caso dello studio del processo batch è necessaria un'analisi più dettagliata con la convalida incrociata.

1.1.2 Carte di monitoraggio per processi continui e limiti di controllo

Le carte di monitoraggio servono a capire se un processo sta marciando in condizioni normali o se si verificano della anomalie. Per fare questo, si definiscono dei limiti di fiducia del modello, cioè i confini che stabiliscono, con un predeterminato margine di fiducia (1-α), le condizioni operative normali su cui il modello è stato costruito. Il monitoraggio deve essere fatto sullo spazio del modello e su quello al di fuori del modello, caratterizzato dai residui.
Per il monitoraggio all'interno del modello può essere utilizzato un diagramma degli *score*. Secondo Jackson (1991) i limiti per il controllo nello spazio degli *score* sono definiti da un'ellissoide di fiducia che ha come semiassi:

$$s_a = \sqrt{\lambda_a T^2_{A,I,\alpha}}\ ,\ \forall\, a = 1, 2, ..., A\,, \tag{1.16}$$

dove $T^2_{A,I,\alpha}$ è:

$$T^2_{A,I,\alpha} = \frac{A(I-1)}{I-A} F_{A,(I-A),\alpha}\,, \tag{1.17}$$

nella quale compare la distribuzione *F*, il cui valore dipende dal numero di componenti principali *A*, dal numero di campioni *I* e dal limite di confidenza (1-α).
Per lo stesso scopo si fa riferimento alla statistica di Hotelling T^2, che è una rappresentazione cumulativa e pesata degli *score*. In questo caso il limite è definito dalla (1.17): $T^2_{\text{Lim}} = T^2_{A,I,\alpha}$.
Lo spazio esterno al modello è caratterizzato dalla statistica SPE^x_i, calcolata sullo spazio delle *x* per ogni campione *i*:

$$\text{SPE}^x_i = \sum_{v=1}^{V} \left(x_{new,v} - \hat{x}_{new,v}\right)^2\,, \tag{1.18}$$

dove $\hat{x}_{new,v}$ è la proiezione calcolata dal modello PCA (MacGregor e Kourti, 1995) secondo l'Equazione (1.10).
Per questa statistica il limite $\text{SPE}_{\alpha,\text{Lim}}$ è definito da Jackson e Mudholkar (Jackson, 1991):

$$\mathrm{SPE}_{\alpha,\mathrm{Lim}} = \theta_1 \left(\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right)^{\frac{1}{h_0}}, \tag{1.19}$$

in cui:

$$\theta_n = \sum_{r=A+1}^{R} \lambda_r^n \text{ per } n = 1, 2, 3 \tag{1.20}$$

e

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2}, \tag{1.21}$$

e infine z_α la deviazione normale standard per la percentuale di confidenza (1-α).

1.1.3 Monitoraggio di processi batch

Molti processi dell'industria chimica sono condotti in modo discontinuo o semicontinuo. Per il monitoraggio di processo è possibile estendere l'impiego di carte di monitoraggio su SPE e T^2, in modo da tenere sorvegliato il comportamento di nuovi batch.

Nel caso batch, però, si deve considerare che i dati relativi alle variabili evolvono nel tempo e i dati sono disponibili in forma tridimensionale $\underline{\mathbf{X}}$ ($I \times V \times K$), dove I è il numero di batch, V le variabili di processo e K gli istanti temporali in cui avviene il campionamento delle variabili. Ai fini del monitoraggio di questi processi, deve essere considerata la dinamica del batch e, per fare questo, Nomikos e MacGregor (1994) hanno proposto di applicare le tecniche statistiche multivariate su una matrice bidimensionale **X**, ricavata dalla matrice tridimensionale $\underline{\mathbf{X}}$ mediante srotolamento o *unfolding* (Nomikos e MacGregor, 1994).

Le possibilità per l'*unfolding* sono due:

- *unfolding* nel senso delle variabili (*variable-wise unfolding*): ogni sezione orizzontale ($V \times K$) viene disposta sotto a quella precedente e si ottiene una matrice $\mathbf{X}(KI \times V)$ che corrisponde a trattare i dati di ogni variabile (nelle colonne) in tutti i batch e in tutti gli istanti temporali. Il metodo è rappresentato in Figura 1.2.

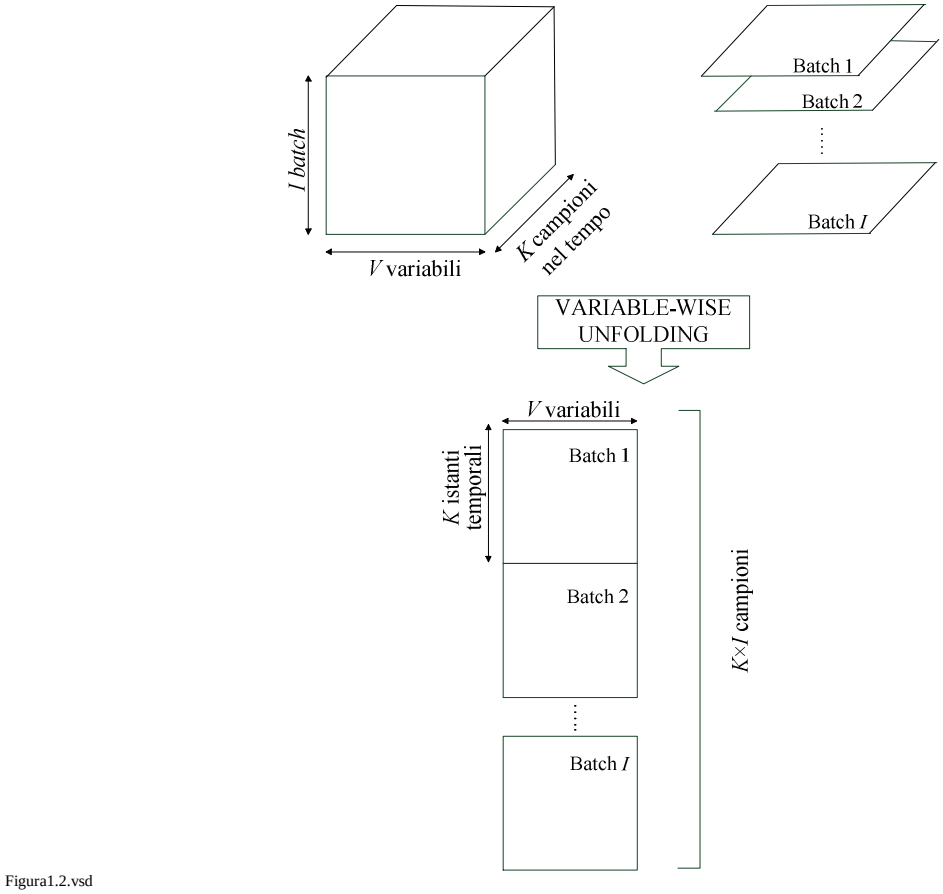


Figura1.2.vsd

Figura 1.2 *Rappresentazione dell'unfolding nel senso delle variabili per la matrice* $\underline{\mathbf{X}}$.

Applicando PCA a questa matrice, si analizzano le traiettorie delle variabili nel tempo rispetto alla media globale per ciascuna variabile e in tutti gli istanti. Ciò specifica che il *variable-wise unfolding* ha l'inconveniente di non considerare la dinamica del batch;

- *unfolding* nel senso dei batch (*batch-wise unfolding*): si dispongono le *K* sezioni verticali (*I*×*V*) affiancate le une alle altre e si ottiene una matrice **X**(*I*×*VK*) in cui ciascuna riga contiene i dati di tutte le variabili di un batch per tutti gli istanti temporali, come rappresentato in Figura 1.3.

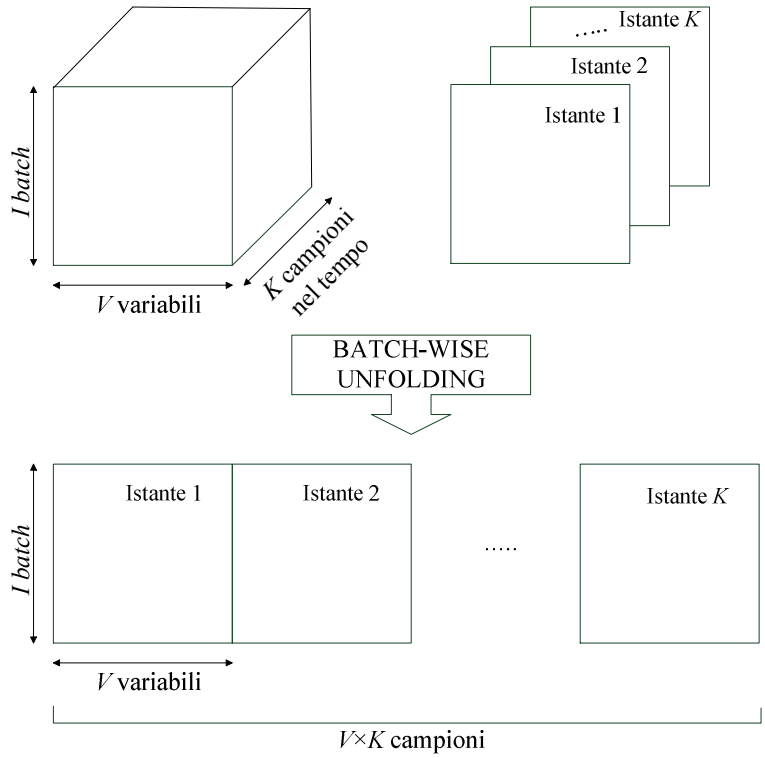


Figura1.3.vsd

Figura 1.3 *Rappresentazione dell'unfolding nel senso dei batch per la matrice* $\underline{\mathbf{X}}$.

In questo caso, applicando PCA si considera la variazione nel tempo delle traiettorie delle variabili in tutti i batch rispetto alla traiettoria media della variabile nel batch stesso, e quindi si considera la dinamica. Il *batch-wise unfolding* presenta, comunque, degli inconvenienti per l'applicazione in linea; infatti, la matrice srotolata presenta dati mancanti se il processo non è concluso, e il numero di colonne delle matrici srotolate potrebbe non essere lo stesso se i batch hanno diversa durata.

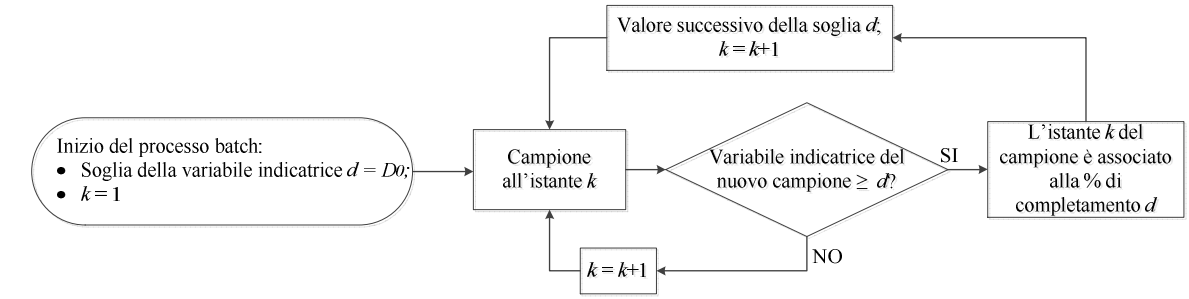
In questa Tesi si fa riferimento al metodo *batch-wise unfolding*. Per il monitoraggio in linea nel caso di *batch-wise unfolding* sono stati proposti dei metodi per la risoluzione degli inconvenienti di batch incompleto e batch non sincronizzato.

Nel monitoraggio in linea di un nuovo batch $\mathbf{x}_{new}$, si deve poter monitorare il processo ad ogni istante di campionamento disponibile. L'applicazione dei metodi statistici multivariati si riferisce al vettore $\mathbf{x}_{new}$, il quale deve contenere i dati dell'intero batch, cioè per tutte le VK variabili. Però, durante il processo, all'istante k il vettore $\mathbf{x}_{new}$ contiene solo le informazioni disponibili fino a k; da (k+1) a K $\mathbf{x}_{new}$ non è completo, poiché mancano le osservazioni future. Nomikos e MacGregor (1994) hanno proposto diversi metodi di riempimento. Una possibilità di riempimento è di utilizzare il valore medio di ciascuna variabile nel tempo per la parte dei dati mancanti; ciò equivale a mettere a zero la restante parte del vettore autoscalato. Un'altra possibilità è quella di assumere che i dati mancanti abbiano future deviazioni dal valore medio

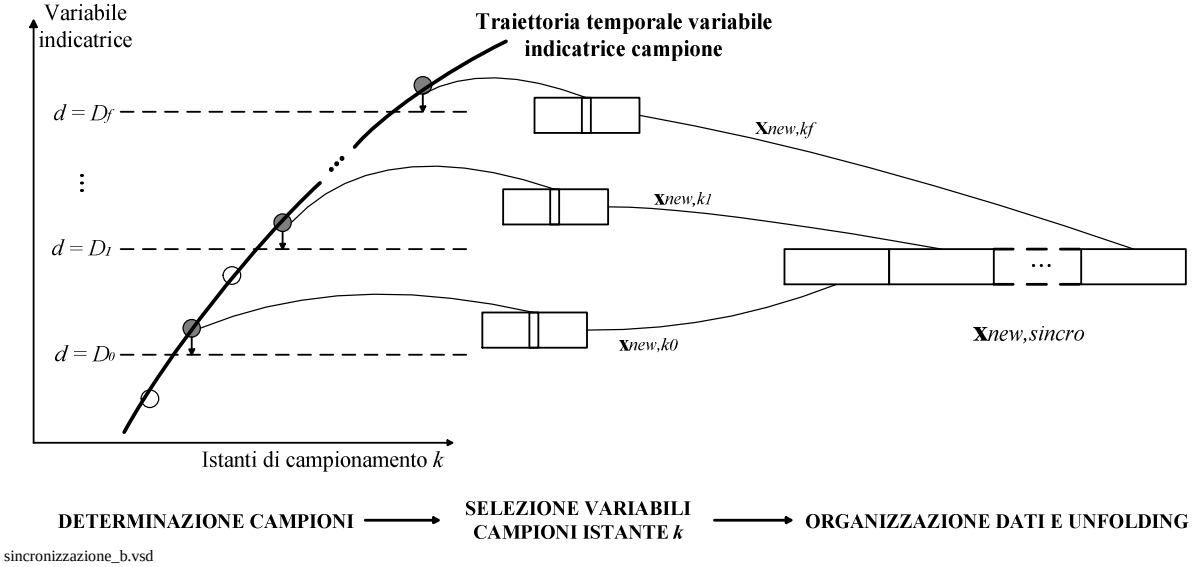
uguali a quelle dell'ultimo istante di campionamento. Quest'ultimo metodo è quello utilizzato nella Tesi.

Il secondo problema è relativo alla diversa durata dei batch. In questo caso è necessario allineare o sincronizzare i diversi batch. Secondo Garcia-Muñoz *et al.* (2003), allineare il *set* di dati di una traiettoria batch significa compiere una trasformazione sulle traiettorie dei batch, in modo che alla fine dell'allineamento ciascuna evolva in modo simile alle altre e ad ognuna corrisponda lo stesso numero di campioni.

Diverse tecniche sono disponibili in letteratura, ad esempio la sincronizzazione basata su *dynamic time warping* (Kassidas et al., 1998). In questa Tesi, la tecnica adottata per la sincronizzazione è simile a quella basata sull'uso della variabile indicatrice (*indicator variable*), proposta da Nomikos e MacGregor (1995) e ripresa da Garcia-Muñoz *et al.* (2003). La variabile indicatrice deve essere una variabile di processo che si sviluppa nel tempo in modo monotono e deve assumere gli stessi valori iniziale e finale per tutti i batch. La procedura di sincronizzazione in questa Tesi è modificata per l'applicazione in linea, secondo lo schema proposto in Figura 1.4.



(a)



(b)

Figura 1.4. *Schema della (a) procedura di sincronizzazione e (b) costruzione della matrice sincronizzata e srotolata in cui d è il valore di soglia della variabile indicatrice ad una certa percentuale di completamento del batch e k l'istante temporale.*

Per la sincronizzazione, si stabiliscono i valori iniziale (D_0) e finale (D_f) della variabile indicatrice, ai quali corrispondono rispettivamente le percentuali 0% e 100% di completamento del batch. L'intervallo è suddiviso in un numero stabilito di sottointervalli, a ciascuno dei quali corrisponde una percentuale di completamento, per la quale la soglia della variabile indicatrice assume un valore *d*. Seguendo lo schema di Figura 1.4a, quando è disponibile un campione $\mathbf{x}_{new}$:

1. $\forall d$, con $d = D_0, \ldots, D$, si determina l'istante *k* del campione $\mathbf{x}_{new,k}$ per cui la sua variabile indicatrice assume un valore $\geq d$;
2. le variabili del campione $\mathbf{x}_{new,k}$ all'istante *k*, individuato al punto 1 (punti in grigio di Figura 1.4b), vengono associate alla percentuale di completamento del batch *d* (Figura 1.4b);
3. le variabili del punto 2 sono utilizzate per la costruzione del vettore srotolato $\mathbf{x}_{new,sincro}$ alla percentuale di completamento del batch *d* (Figura 14.b), affiancando le variabili a quelle determinate alle precedenti *d*;
4. si riprende dal punto 1 per ogni percentuale di completamento *d*.

La procedura è applicata per i dati delle proiezioni di nuovi campioni. I dati di calibrazione sono trattati analogamente, applicando la procedura fuori linea.

La procedura si differenzia dalla proposta di Garcia-Muñoz *et al.* (2003) in quanto quando il campione non assume perfettamente il valore della variabile indicatrice alla percentuale di completamento, non si procede con l'interpolazione con il valore precedente per determinarne il valore preciso. Questo perché non si vogliono inserire valori fittizi delle variabili, ma solo reali.

In questa Tesi, per l'analisi del processo batch si farà riferimento all'indice *k* per indicare gli istanti di campionamento sincronizzati, anziché quelli realmente disponibili per ogni batch.

1.1.3.1 Inserimento della dinamica nei limiti di controllo per il monitoraggio

Se lo scopo principale è il monitoraggio del processo batch, è importante poter rilevare eventuali anomalie quando il processo è in corso, piuttosto che alla fine del batch. Da qui la necessità di costruire carte di monitoraggio per i residui e per gli *score* in cui i limiti di controllo statistico riguardino la dinamica nel tempo del batch. Ciò è possibile solo se i limiti si sviluppano temporalmente.

Nomikos e MacGregor (1994) hanno proposto la metodologia per lo sviluppo temporale dei limiti per il metodo PCA. In questa Tesi, per il monitoraggio si utilizzano le carte T^2 ed SPE.

Il limite temporale per la statistica T^2 di Hotelling è costante nel tempo ed è dato dalla (1.17), applicata sui dati disponibili alla conclusione del batch.

I limiti temporali di SPE sono definiti usando l'idea della distribuzione esterna di Box (1978), assumendo che la distribuzione degli errori quadratici è del tipo χ^2:

$$\mathrm{SPE}_{\mathrm{Lim},k} = \frac{\sigma_{\mathrm{SPE},k}}{2m_{\mathrm{SPE},k}} \chi^2_{2m^2_{\mathrm{SPE},k}/\sigma_{\mathrm{SPE},k},\alpha}, \tag{1.22}$$

con $\sigma_{\mathrm{SPE},k}$ e $m_{\mathrm{SPE},k}$ rispettivamente varianza media degli $\mathrm{SPE}_{\mathrm{Lim},k}$ sugli I batch, mentre α è il limite di fiducia. La procedura per la determinazione di $\sigma_{\mathrm{SPE},k}$ e $m_{\mathrm{SPE},k}$ è definita da Nomikos e MacGregor (1994).

1.1.4 Diagnostiche del modello di monitoraggio

Quando si realizzano delle situazioni anomale di funzionamento dell'impianto, per cui il processo non è più in condizioni operative normali, si parla di accadimento di un'anomalia (*fault*). Grazie alla costruzione delle carte di controllo e dei rispettivi limiti, è possibile il monitoraggio del processo, con prestazioni che dipendono dal modello adottato. Proiettando i nuovi campioni nel modello costruito con il *set* di dati di calibrazione è possibile stabilire, in base ai valori di SPE e T^2 della proiezione, se i campioni siano normali o, nel caso eccedano i limiti di controllo, siano anomali. Per segnalare un allarme per l'accadimento di un'anomalia, non basta che un unico campione sia al di fuori dei limiti, perché questo comportamento potrebbe essere dovuto semplicemente al rumore della misura. Per questo si devono stabilire dei criteri più robusti.

Il criterio adottato in questa Tesi consiste nel segnalare un allarme quando almeno Δ-1 su Δ campioni consecutivi sono al di fuori dei limiti di controllo statistico, con Δ = 5. Le prestazioni del modello di monitoraggio sono valutate in termini di frequenza degli allarmi (AR, *alarm rate*) e ritardo di rilevazione dell'anomalia (TD, *time delay*).

La frequenza degli allarmi è il rapporto percentuale tra il numero totale di allarmi A_{tot} segnalati su S campioni proiettati nel modello di monitoraggio e il numero S di campioni proiettati (Facco *et al.*, 2012), secondo:

$$\mathrm{AR} = 100 \times \frac{A_{\mathrm{tot}}}{S}, \tag{1.23}$$

dove:

$$A_{\mathrm{tot}} = \sum_{n=1}^{S} A_n \,. \tag{1.24}$$

L'allarme A_n segnato al campione n può assumere due valori:

$$A_n = \begin{cases} 1, \text{se} \sum_{n=k-\Delta+1}^{k} \psi_n \geq \Delta - 1 \\ 0, \text{se} \sum_{n=k-\Delta+1}^{k} \psi_n < \Delta - 1 \end{cases}, \tag{1.25}$$

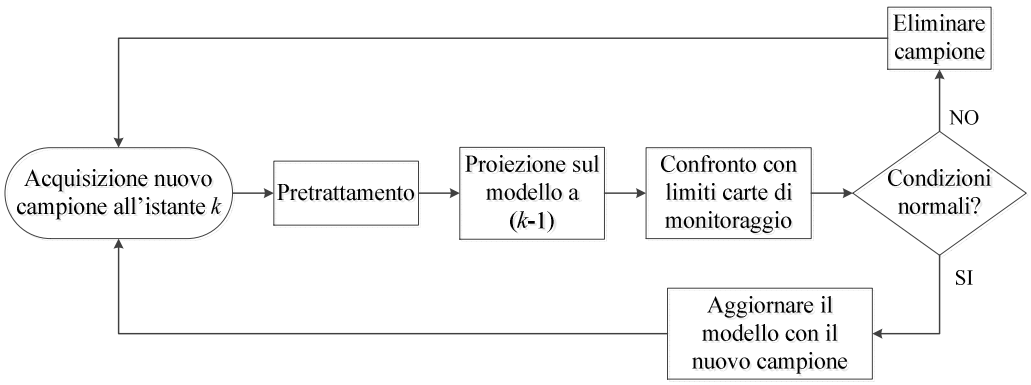
dove $k = \Delta, \Delta+1, ..., S$, e:

$$\psi_n = \begin{cases} 1, \text{se } T_n^2 > T^2_{\mathrm{Lim}} \text{ o } \mathrm{SPE}_n > \mathrm{SPE}_{\alpha,\mathrm{Lim}} \\ 0, \text{se } T_n^2 \leq T^2_{\mathrm{Lim}} \text{ e } \mathrm{SPE}_n \leq \mathrm{SPE}_{\alpha,\mathrm{Lim}} \end{cases}. \tag{1.26}$$

Il ritardo di rilevazione, invece, corrisponde al numero di campioni necessari a generare un allarme dopo l'inizio dell'anomalia.
Le prestazioni del modello di monitoraggio sono migliori quanto più AR è vicino allo zero per campioni NOC e vicino al 100% per i campioni dell'anomalia (i primi appartenenti alla cosiddetta fase 1, i secondi alla fase 2). Inoltre, si deve considerare anche il ritardo di rilevazione TD, che deve essere quanto più prossimo possibile a Δ-1.

1.2 Trasferimento di modelli per il monitoraggio

Quando si deve monitorare un impianto B nelle fasi iniziali di esercizio, si dispone di un quantitativo di dati che spesso è insufficiente per costruire modelli di monitoraggio autosufficienti. Tuttavia, si può pensare di sfruttare le informazioni derivanti da impianti A che, pur essendo simili a B, presentano delle differenze dovute per esempio alla diversa scala, alla collocazione geografica, alla strumentazione, ecc. Per gli impianti A possono essere disponibili molti dati e un sistema di monitoraggio efficiente. Quindi, con il trasferimento si vuole monitorare il nuovo impianto B fin dagli istanti iniziali di funzionamento a partire dalle informazioni di A. Inizialmente, il modello di monitoraggio viene costruito sui dati disponibili dall'impianto A; quindi viene aggiornato sui dati derivanti da B, secondo lo schema di Figura 1.5, fino a che non è possibile avere un modello di monitoraggio solamente su dati di B che sia autosufficiente.



trasferimento.vsd

Figura 1.5. *Schema a blocchi generico per il trasferimento.*

La procedura riportata in Figura 1.5 è generica per il trasferimento, secondo la quale il nuovo campione acquisito all'istante *k* viene proiettato sul modello definito all'istante precedente. L'aggiornamento del modello avviene solo se il campione è in condizioni operative normali.
Le tecniche presentate per il trasferimento si possono basare su PCA o JY-PLS, descritte in seguito.

1.2.1 Trasferimento mediante PCA

La tecnica PCA non è un vero e proprio metodo per il trasferimento, ma può essere ugualmente utilizzata per tal scopo. Il monitoraggio dell'impianto B avviene nello spazio delle variabili misurate comuni tra i due impianti, supponendo che la struttura di correlazione tra le variabili venga mantenuta tra gli impianti A e B. Il modello PCA per il monitoraggio è costruito inizialmente sui soli dati di A e viene aggiornato ai nuovi dati provenienti dall'impianto B che si trovano all'interno dei limiti di controllo statistico, definiti dal modello PCA. Quando i dati di B sono sufficienti per costruire un modello PCA indipendente, che abbia buone prestazioni, si passa ad un modello di monitoraggio solo con dati di B.
Seguendo lo schema a blocchi di Figura 1.5, applicando il metodo PCA, quando è disponibile un nuovo campione all'istante *k*:

- il campione viene trattato all'interno del *set* di dati a cui appartiene;
- il campione trattato è proiettato nel modello PCA costruito sui dati normali disponibili a (*k*-1);
- dalla proiezione si ricavano i valori delle statistiche SPE e T^2, e le si confrontano con i rispettivi limiti;
- se il campione è in condizioni normali, è possibile aggiornare il modello; altrimenti il campione viene scartato.

Il modello PCA permette di definire i luoghi delle carte di monitoraggio per stabilire le condizioni del nuovo campione, ma utilizza solo lo spazio definito dalle variabili comuni, quindi descrive la correlazione dei dati di questo insieme.

1.2.2 Trasferimento mediante JY-PLS

Un altro tipo di metodo per il trasferimento è joint-Y PLS (JY-PLS), sempre basato sulle variabili latenti (García-Muñoz *et al.*, 2005). Questo metodo è stato sviluppato nell'ambito del problema del trasferimento del prodotto. In questa Tesi il metodo è utilizzato nel problema del trasferimento dei modelli per il monitoraggio (Facco *et al.*, 2012; Tomba *et al.*, 2012).
I dati provenienti dai due impianti per cui si vuole fare il trasferimento, classificati A e B, appartengono a 4 matrici principali, $\mathbf{X}''^{A}$, $\mathbf{Y}''^{A}$, $\mathbf{X}''^{B}$, $\mathbf{Y}''^{B}$. Le matrici $\mathbf{X}''$ contengono variabili esclusivamente misurate in un impianto, mentre le matrici $\mathbf{Y}''$ presentano solo variabili comuni. Verificato che $\mathbf{Y}''^{A}$ e $\mathbf{Y}''^{B}$ abbiano la stessa struttura di correlazione, JY-PLS modella lo spazio delle variabili comuni unitamente allo spazio delle variabili non comuni, specifiche di ogni impianto. In questo modo si trasferiscono informazioni tra $\mathbf{X}''^{A}$ e $\mathbf{X}''^{B}$ attraverso lo spazio latente comune.
Per ottenere il modello JY-PLS si applicano due separati modelli PLS (*partial least-squares regression*)[1] alle coppie di matrici $\mathbf{X}''$ e $\mathbf{Y}''$, ottenendo per entrambe la scomposizione:

[1] Per un'analisi dettagliata del metodo PLS si rimanda a Nomikos e MacGregor (1995) e a Wise e Gallagher (1996).

$$\begin{aligned}\mathbf{X}'' &= \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \\ \mathbf{Y}'' &= \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F}\end{aligned}, \tag{1.27}$$

dove **T** e **U** sono le matrici degli *score*, **P** e **Q** le matrici dei *loading*, **E** ed **F** i residui che vengono determinati per entrambi gli impianti A e B.

Ma, assumendo che le matrici $\mathbf{Y}''^{\mathrm{A}}$ e $\mathbf{Y}''^{\mathrm{B}}$ giacciano nel piano di variabili latenti comuni, i *loading* $\mathbf{Q}^{\mathrm{A}}$ e $\mathbf{Q}^{\mathrm{B}}$ per i due modelli sono una rotazione l'uno dell'altro, per cui definiscono una singola matrice dei *loading* congiunta $\mathbf{Q}^{\mathrm{J}}$. La tecnica JY-PLS di regressione alle variabili latenti modella la matrice $\mathbf{Y}^{\mathrm{J}}$ congiunta definendo lo spazio comune attraverso $\mathbf{Q}^{\mathrm{J}}$:

$$\mathbf{Y}^{\mathrm{J}} = \begin{bmatrix}\mathbf{Y}''^{\mathrm{A}} \\ \mathbf{Y}''^{\mathrm{B}}\end{bmatrix} = \begin{bmatrix}\mathbf{T}^{\mathrm{A}} \\ \mathbf{T}^{\mathrm{B}}\end{bmatrix}\mathbf{Q}^{\mathrm{J}\mathrm{T}} + \begin{bmatrix}\mathbf{E}^{\mathrm{JA}} \\ \mathbf{E}^{\mathrm{JB}}\end{bmatrix}. \tag{1.28}$$

Gli *score* sono definiti attraverso:

$$\begin{aligned}\mathbf{T}^{\mathrm{A}} &= \mathbf{X}''^{\mathrm{A}}\mathbf{W}^{*\mathrm{A}} \\ \mathbf{T}^{\mathrm{B}} &= \mathbf{X}''^{\mathrm{B}}\mathbf{W}^{*\mathrm{B}}\end{aligned}, \tag{1.29}$$

dove $\mathbf{W}^{*\mathrm{A}}$ e $\mathbf{W}^{*\mathrm{B}}$ sono i pesi del modello JY-PLS.

Per la struttura dell'algoritmo del modello e la determinazione dei suoi parametri ci si riferisce a García-Muñoz *et al.* (2005).

Lo spazio delle direzioni di massima variabilità congiunta tra $\mathbf{Y}''^{\mathrm{A}}$ e $\mathbf{Y}''^{\mathrm{B}}$ è utilizzato per monitorare il processo dell'impianto B attraverso le carte di controllo che contengono informazioni dovute sia ai dati in $\mathbf{X}''$ che a quelli in $\mathbf{Y}''$. In Facco *et al.* (2012) è definito il metodo per la proiezione di un nuovo campione nel modello JY-PLS. Si predice lo *score* del nuovo campione $\mathbf{x}_k''^{\mathrm{B}}$, $\mathbf{y}_k''^{\mathrm{B}}$ e questo lo si usa per ottenere la predizione della variabile di risposta comune:

$$\hat{\mathbf{t}}_k^{\mathrm{B}} = \mathbf{x}_k''^{\mathrm{B}}\mathbf{W}_k^{*\mathrm{B}}, \tag{1.30}$$

$$\hat{\mathbf{y}}_k''^{\mathrm{B}} = \hat{\mathbf{t}}_k^{\mathrm{B}}\mathbf{Q}_k^{\mathrm{J}\mathrm{T}}, \tag{1.31}$$

dove $\mathbf{W}_k^{*\mathrm{B}}$ e $\mathbf{Q}_k^{\mathrm{J}\mathrm{T}}$ sono i pesi e la matrice dei *loading* congiunta definiti dal modello di calibrazione disponibile all'istante *k*. Ottenuta la proiezione si calcolano le statistiche, che vengono utilizzate per il monitoraggio. Nel caso JY-PLS ci sono tre spazi da monitorare:

- SPE_k^y, errore quadratico medio del nuovo campione $\mathbf{y}_k''^{\mathrm{B}}$ su *y*:

$$\mathrm{SPE}_k^y = \sum_{v=1}^{V}\left(y_{k,v}''^{B} - \hat{y}_{k,v}''^{B}\right)^2, \tag{1.32}$$

  in cui *V* sono le variabili dello spazio comune;

- SPE_k^x, errore quadratico medio del nuovo campione $\mathbf{x}_k''^{\mathrm{B}}$ su *x*, per il quale si fa riferimento alla Equazione (1.18), in cui il pedice *i* corrisponde all'istante *k*;

- statistica di Hotelling T_k^2:

$$T_k^2 = \hat{\mathbf{t}}_k^{\mathrm{B}}\mathbf{L}\hat{\mathbf{t}}_k^{\mathrm{B}\mathrm{T}}, \tag{1.33}$$

  con

$$\mathbf{L} = \mathrm{diag}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, ..., \frac{1}{\lambda_A}\right), \tag{1.34}$$

dove λ_a con a = 1,2,…, A sono gli autovalori, ovvero gli elementi della matrice:

$$\mathbf{\Lambda} = \frac{\mathrm{diag}\left(\mathbf{T}^{\mathrm{B}^{\mathrm{T}}}\mathbf{T}^{\mathrm{B}}\right)}{I-1}. \tag{1.35}$$

Tali statistiche dipendono non solo dal nuovo campione ma anche dall'intero *set* di dati A attraverso $\mathbf{Q}^{\mathrm{J}}$.

Anche in questo caso, per il monitoraggio vengono elaborate delle carte di controllo, in cui i limiti per le statistiche sono definiti secondo il formulario dei Paragrafi 1.1.2, per il caso continuo, e 1.1.3.1, per il caso batch.

Capitolo 2

Processi considerati e dati disponibili

In questo Capitolo sono presentati i due casi di studio affrontati in questa Tesi. Il primo è un impianto industriale di *spray-drying* continuo. Il secondo è un impianto batch di produzione di penicillina simulato. Per entrambi sono presentati i dati disponibili su cui si applicano i metodi di trasferimento dei modelli per il monitoraggio. Entrambi i *set* di dati sono relativi a due impianti di diversa scala con dati sia di condizioni operative normali che di anomalie di funzionamento.

2.1 Processo industriale continuo di *spray-drying*

Lo *spray-drying* è un processo molto utilizzato nell'industria farmaceutica; tra le varie applicazioni, viene adottato per la preparazione di dispersioni di solidi amorfi, ma anche per la preparazione di eccipienti, realizzazione di particelle bioterapeutiche, disidratazione dei principi attivi cristallini e incapsulamento (Dobry *et al.*, 2009). Il processo consiste nell'inviare alla camera di essiccamento una soluzione, costituita dal solido da essiccare disciolto in un solvente. La soluzione viene atomizzata in gocce e posta in contatto con un gas inerte caldo (spesso azoto), che fa evaporare il solvente e disidrata le particelle, trascinate fuori dalla camera dal gas stesso. In uscita dalla camera si ha la separazione tra le particelle e il gas di processo in un ciclone; poiché le particelle fini possono rimanere disciolte, queste vengono trattate in una serie di filtri (a maniche, di tipo HEPA, …). Il gas passa in un condensatore per separare il solvente evaporato e viene riciclato all'essiccatore alla temperatura di processo. La Figura 2.1 rappresenta schematicamente il processo descritto.

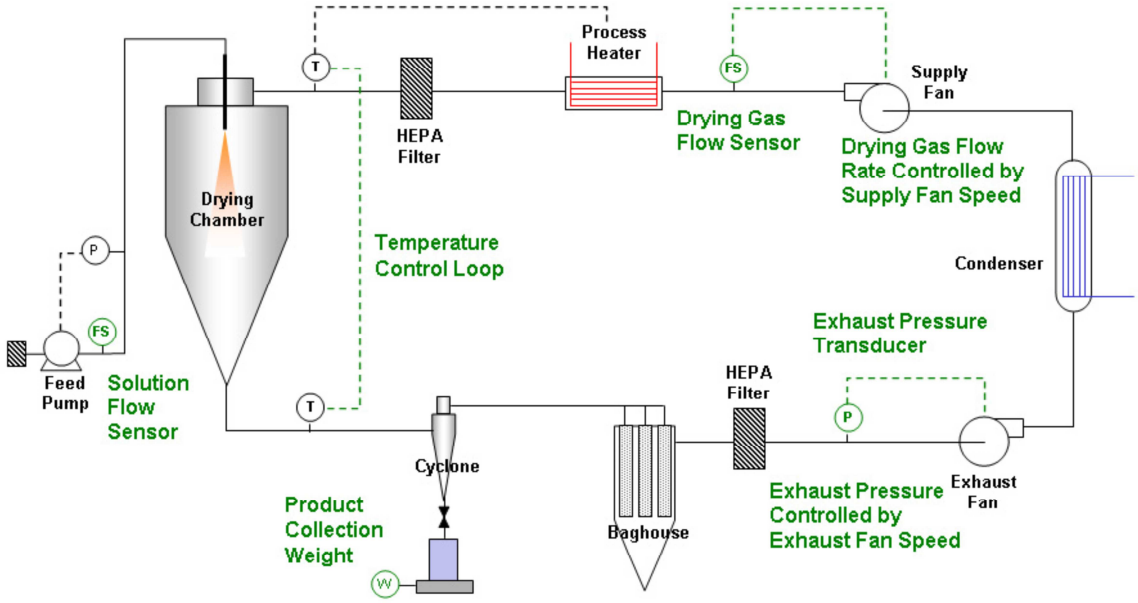


Figura2.1

Figura 2.1. *Schema del processo industriale di spray-drying (Garcia-Muñoz e Settell, 2009).*

Nello studio di questa Tesi si considerano due impianti di *spray-drying* geometricamente simili, strumentati in modo simile, ma di diverse dimensioni: un impianto pilota (A) e uno su scala industriale (B).

2.1.1 Dati disponibili

Sia per l'impianto su scala pilota (A) che su scala industriale (B), si hanno a disposizione dati relativi a variabili caratteristiche del processo. Queste variabili sono riportate in Tabella 2.1. Alcune risultano comuni tra i due impianti (nella Tabella sono rappresentate dal carattere corsivo), altre sono esclusivamente misurate in una delle due scale. Inoltre, si hanno dati ottenuti in condizioni operative normali e dati di anomalie.

Per l'impianto A i dati NOC sono raccolti in una matrice $\mathbf{X}^{A}[N^{A}\times V^{A}]$, dove N^{A} comprende 15031 campioni e V^{A} 16 variabili. Per l'impianto B sono raccolti in $\mathbf{X}^{B}[N^{B}\times V^{B}]$, dove N^{B} sono 4224 campioni e V^{B} 10 variabili.

Tabella 2.1. *Variabili di processo per gli impianti su scala pilota A e su scala industriale B di spray-drying. Le variabili caratterizzate dal carattere corsivo appartengono all'insieme delle variabili comuni.*

Impianto pilota (A)		Impianto scala industriale (B)	
Variabile	**Tipo di variabile**	**Variabile**	**Tipo di variabile**
1	*Pressione atomizzatore* (psig)	*1*	*Pressione atomizzatore* (psig)
2	Temperatura gas ingresso 1 (°C)	*2*	*Temperatura ingresso* (°C)
3	*Temperatura gas ingresso 2* (°C)	*3*	*Temperatura in uscita* (°C)
4	*Temperatura in uscita 1* (°C)	*4*	*Portata gas* (kg/h)
5	*Portata gas* (kg/h)	*5*	*Portata soluzione* (kg/h)
6	Temperatura gas dopo cond (°C)	*6*	*Pressione camera* (mbar)
7	*Portata soluzione* (kg/h)	*7*	*Differenza pressione ciclone* (mbar)
8	*Pressione camera* (mm_{H2O})	*8*	*Differenza pressione filtro* (mbar)
9	*Differenza pressione ciclone* (mm_{H2O})	9	Split Range Pressure (mbar)
10	*Differenza pressione filtro* (mm_{H2O})	*10*	*Pressione uscita* (mbar)
11	Pressione sistema (mm_{H2O})		
12	*Pressione di scarico* (mm_{H2O})		
13	Temperatura in uscita 2 (°C)		
14	Velocità ventola scarico 1 (%)		
15	Velocità ventola scarico 2 (%)		
16	Velocità pompa alimentazione (%)		

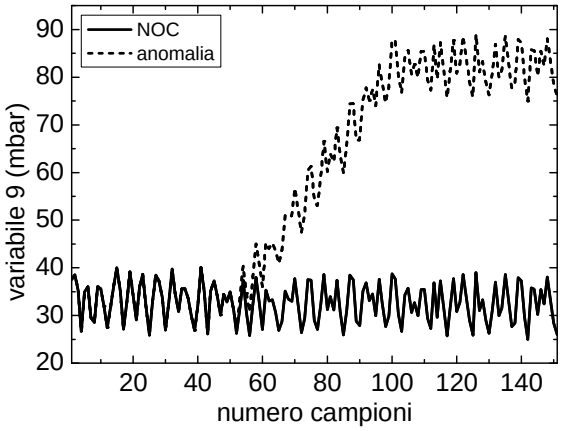
Per le variabili si definiscono i sottoinsiemi delle variabili comuni $\mathbf{v}'^{A} = \{1, 3, 4, 5, 7, 8, 9, 10, 12\}$, appartenenti all'impianto A, e $\mathbf{v}'^{B} = \{1, 2, 3, 4, 5, 6, 7, 8, 10\}$ dell'impianto B, in quanto assumono lo stesso significato fisico per entrambi.

Per quanto riguarda i dati relativi alle anomalie, per l'impianto B si ha a disposizione un *set* di dati relativo ad un'anomalia. Si tratta di un'ostruzione alla valvola dell'atomizzatore. Per questo è disponibile un *set* $\mathbf{x}^{BF}$ di 81 campioni che viene suddiviso in:

- fase 1: i primi 24 campioni in normali condizioni operative;
- fase 2: i rimanenti 57 campioni che rappresentano l'anomalia.

Su questi dati, poi, sono state create diverse realizzazioni dell'anomalia, aggiungendo rumore random di tipo Gaussiano per ciascuna variabile del *set* originario. Queste realizzazioni differiscono dai dati reali solamente per il rumore. Le realizzazioni sono state create per verificare che i risultati siano indipendenti dal rumore delle misure.

Inoltre, sono stati creati dati relativi a un'anomalia artificiale su una variabile dell'impianto B (Figura 2.2). La variabile interessata è la 9 dell'impianto B, che risulta una variabile non comune tra gli impianti.



variabile9_fault.opj

Figura 2.2. *Rappresentazione dell'anomalia artificiale sulla variabile 9 non comune. Inizio della fase 2 dell'anomalia al campione 51.*

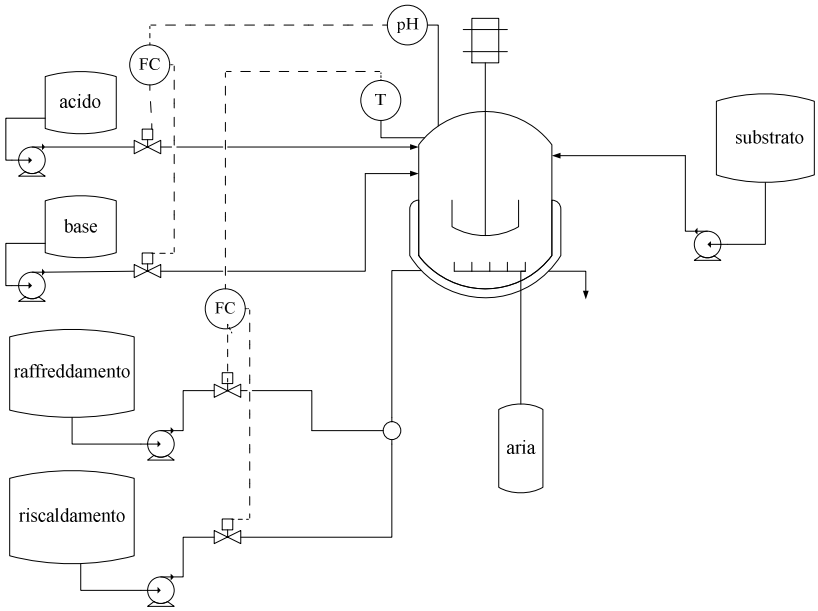
La Figura 2.2 mostra le caratteristiche dell'anomalia, una rampa di pendenza unitaria e pari a 50 mbar di altezza. L'anomalia artificiale è costituita da 151 campioni divisibili in due fasi:

- fase 1: i primi 50 campioni in condizioni normali operative;
- fase 2: i restanti campioni dal 51 al 151, che portano il valore medio della variabile da 32 a 82 mbar.

Analogamente al caso precedente si definiscono 30 realizzazioni con l'aggiunta del rumore random Gaussiano. La scelta di lavorare con le realizzazioni è stata fatta anche in questo caso per mediare i risultati, rendendo l'analisi indipendente dal rumore e più robusta.

2.2 Simulazione di un processo batch per la produzione di penicillina

Le penicilline sono metaboliti secondari prodotti da microrganismi filiformi. Poiché la loro formazione di solito non è associata alla crescita dei microrganismi, comunemente si fanno crescere questi ultimi in una coltura batch, seguita da un'operazione di tipo fed-batch per favorire la sintesi dei metaboliti (Birol *et al.*, 2001). La caratteristica di questa operazione è di poter avere cicli produttivi molto lunghi, mantenendo costanti alcune variabili operative. Dunque, il processo è suddiviso in due fasi: una prima fase di crescita in cui la biomassa viene prodotta in presenza di ossigeno e di un substrato (zuccheri) e una seconda fase in cui si produce penicillina ancora in presenza di ossigeno, ma in assenza di substrato.
La Figura 2.3 rappresenta il diagramma di flusso del processo.



processo.vsd

Figura 2.3. *Diagramma di flusso del processo di produzione di penicillina.*

Il diagramma di flusso di Figura 2.3 mostra il sistema di alimentazioni al reattore, il quale è di tipo agitato. Esso viene alimentato da una portata di substrato, durante la fase batch, e da una portata d'aria per tutta la durata del processo. Ulteriori ingressi al processo sono dovuti alle variabili manipolate dal sistema di controllo. Questo agisce sul pH e sulla temperatura, che devono rimanere costanti per tutto il processo. Infatti, con il procedere della reazione il sistema tende a diventare più acido e le reazioni esotermiche portano ad un aumento della temperatura. Questi creano delle condizioni sfavorevoli per la crescita dei microrganismi, con conseguente diminuzione della produzione di penicillina. Ciò rende necessario un sistema di regolazione.

In questa Tesi è stato utilizzato il simulatore PenSim (Birol *et al.*, 2002), descritto al Paragrafo 2.2.1, per la simulazione del processo.

2.2.1 Il simulatore PenSim

Mediante il simulatore PenSim, proposto da Birol *et al.* (2002), sono stati costruiti i *set* di dati che rappresentano il processo batch di produzione di penicillina. Esso risolve il sistema di equazioni differenziali e algebriche (DAE) che rappresenta il modello del processo.

Il simulatore è implementato in Matlab ed è definito dal codice *pensim2*, che richiede come dati in ingresso i valori delle variabili di processo iniziali, gli *input* dei parametri progettuali, le caratteristiche del regolatore e alcune informazioni sulla durata del batch e sul tempo di campionamento. Queste informazioni sono necessarie per la risoluzione definita del sistema di DAE. Il simulatore dà come *output* l'evoluzione temporale delle grandezze caratteristiche misurate nel processo.

Le variabili iniziali da definire sono:

- concentrazione di substrato;
- concentrazione di ossigeno disciolto;
- concentrazione di biomassa;
- concentrazione di penicillina;
- volume di coltura;
- concentrazione di anidride carbonica;
- pH nel reattore;
- temperatura nel reattore;
- calore generato.

Ulteriori variabili da definire sono i *set point* dei sistemi di regolazione, ovvero:

- velocità di aerazione;
- potenza di agitazione;
- portata di alimentazione di substrato;
- temperatura del substrato in alimentazione;
- *set point* del pH nel reattore;
- *set point* della temperatura nel reattore.

Per ogni parametro è stabilito un *range* per il quale, assegnato il valore del parametro in questo intervallo, la soluzione offerta dal simulatore ha significato fisico (Çinar *et al.*, 2003).

La definizione del sistema di controllo dipende dal tipo di variabile controllata. Mentre per la temperatura è assegnato un regolatore PID, per il pH si può selezionare il tipo di regolazione, on-off o PID. Le grandezze caratteristiche del regolatore, tra cui il guadagno o le costanti integrali e derivative costituiscono ulteriori ingressi per il simulatore.

Gli *output* di *pensim2* sono i profili delle variabili di processo nel tempo, raccolti in una matrice bidimensionale **X** per ogni batch. Le variabili sono definite in Tabella 2.2.

Tabella 2.2. *Numerazione e tipo di misure degli output nel processo simulato per la produzione di penicillina.*

Numero variabile	Tipo variabile
1	Velocità di aerazione
2	Potenza di agitazione
3	Velocità di alimentazione substrato
4	Temperatura substr. in alimentazione
5	Profilo concentrazione substrato
6	Conc. Ossigeno disciolto
7	Conc. Biomassa
8	Conc. penicillina
9	Volume di coltura
10	Conc. CO2
11	pH nel reattore
12	Temperatura nel reattore
13	Calore generato
14	Portata acido
15	Portata base
16	Portata acqua riscaldamento/raffreddamento

Nel simulatore si possono anche simulare delle anomalie sulle variabili di processo e in particolare su velocità di aerazione (1), potenza di agitazione (2) e portata di substrato in alimentazione (3). Per l'anomalia si devono definire, come ulteriore ingresso per il simulatore, il tipo di anomalia (step o rampa), i tempi di inizio e conclusione dell'anomalia, l'entità e la variabile coinvolta.

2.2.2 Struttura dei dati

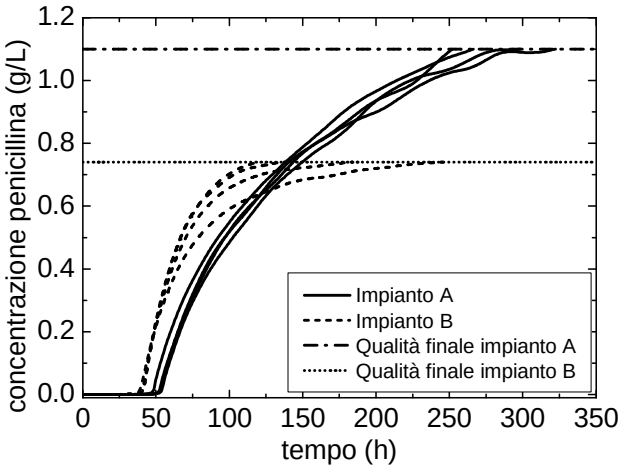
Sono stati simulati due impianti simili di diversa scala, A e B. L'impianto A è un impianto più piccolo, caratterizzato da un volume medio di coltura di 105 L e una potenza di agitazione di 20 W. L'impianto B ha dimensioni maggiori con un volume di coltura medio di 195 L e una potenza di agitazione di 40 W, derivanti da uno scale-up che tiene costante il rapporto potenza su volume. Per creare diversi batch per ogni scala è stato simulato un rumore al valore di alcune variabili di *input*, come riportato in Tabella 2.3. In questo modo attraverso il simulatore si ottengono gli andamenti nel tempo delle variabili di processo per 100 diversi batch di piccole dimensioni e 100 batch di dimensioni superiori. Le variabili, il loro valore medio e la massima variazione data dal rumore per ciascuna sono definiti in Tabella 2.3.

Tabella 2.3. *Simulazione del processo fed-batch: valore medio stabilito per i parametri di processo e loro variabilità per gli impianti A e B.*

	Impianto A		**Impianto B**	
Variabili	**Valore medio**	**Massima variazione data dal rumore**	**Valore medio**	**Massima variazione data dal rumore**
Concentrazione substrato (g/L)	15	±2.5	15	±2.5
Conc. O_2 disciolto (mmol/L)	1.16	0	1.16	0
Concentrazione biomassa (g/L)	0.05	±0.025	0.15	±0.025
Conc. penicillina iniziale (g/L)	0	0	0	0
Volume di coltura (L)	105	±2.5	195	±2.5
Concentrazione CO_2 (mmol/L)	0.65	±0.0625	0.85	±0.0625
pH	5	±0.25	5	±0.25
Temperatura (K)	299	±0.25	299	±0.25
Calore generato iniziale (cal)	0	0	0	0
Velocità aerazione (L/h)	4	±0.25	8	±0.25
Potenza agitazione (W)	20	0	40	0
Portata substr. alimentata (L/h)	0.037	±0.00125	0.042	±0.00125
Temperatura substr. ingresso (K)	296.5	±0.125	297.5	±0.125
Set point pH	5	0	5	0
Set point temperatura (K)	298	0	298	0

Un'ulteriore distinzione tra gli impianti riguarda il sistema di controllo del pH, in cui per l'impianto A si è definito un regolatore di tipo on-off, mentre è di tipo PID per B. Questo porta ad avere diversi andamenti delle portate di acido e base inviate al processo, che corrispondono alle variabili manipolabili dal sistema di controllo.

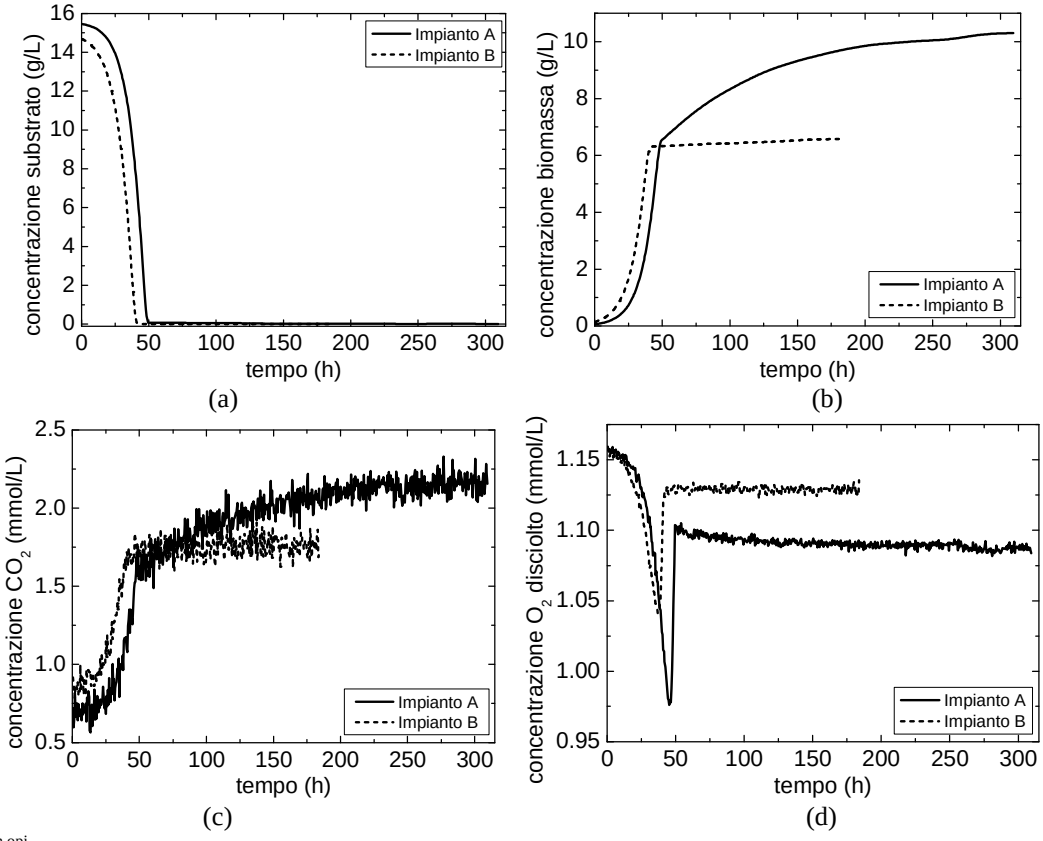
I dati di processo per ogni batch, ottenuti dalla simulazione, sono raccolti in matrici tridimensionali. Perché i dati siano maggiormente rappresentativi di un processo reale, i batch simulati hanno una durata che dipende dalla concentrazione finale di penicillina desiderata, nel senso che i batch di A vengono arrestati quando la concentrazione assume il valore 1.1 g/L, mentre i batch di B raggiungono 0.74 g/L. I tempi per cui ogni batch raggiunge la specifica sono differenti, come si vede dalla Figura 2.4, nella quale è rappresentata la concentrazione di penicillina per 4 batch caratteristici di ciascun impianto.



penicillina.opj

Figura 2.4. *Profilo temporale della concentrazione di penicillina per 4 batch dell'impianto A e 4 batch dell'impianto B con i rispettivi limiti legati alla qualità finale desiderata.*

Dalla Figura 2.4 si può vedere come la concentrazione di penicillina raggiunga valori finali diversi nei batch dell'impianto A e dell'impianto B, a seconda della specifica. Inoltre, in ogni impianto i batch raggiungono la qualità finale specificata a tempi diversi.
Nelle Figure 2.5 e 2.6, si riportano esempi di profili di alcune variabili di ciascun impianto.



variabilibatch.opj

Figura 2.5. *Profilo temporale delle variabili concentrazione di (a) substrato, (b) biomassa, (c) anidride carbonica e (d) ossigeno disciolto per il primo batch dell'impianto A e il primo batch dell'impianto B.*

Dalla Figura 2.5 si nota come le variabili di processo abbiano profili di forma simile tra gli impianti, nonostante il processo termini a tempi diversi. Per ogni variabile è visibile la discontinuità che segna il passaggio dalla fase batch a quella fed-batch. Queste variabili, avendo profili simili, hanno la stessa correlazione tra gli impianti A e B. Analizzando la relazione tra esse, si può dire che:

- la concentrazione di substrato (Figura 2.5a) è anticorrelata con la concentrazione di biomassa (Figura 2.5b) e di CO_2 (Figura 2.5c); inoltre, lo è anche con la concentrazione di penicillina (Figura 2.4);
- la concentrazione di O_2 disciolto (Figura 2.5d) ha un profilo simile alla concentrazione di substrato per la fase batch, mentre nella fase fed-batch è maggiormente assimilabile alle altre concentrazioni.

Vi sono però differenze sostanziali che risiedono nel fatto che:

- i batch non sono sincronizzati;
- la durata relativa delle fasi batch e fed-batch è diversa;
- il tempo di innesco della fase fed-batch è diverso.

La concentrazione di substrato sarà utilizzata per caratterizzare la fase batch, mentre quella di penicillina descriverà la fase fed-batch del processo.

Per quanto riguarda le variabili di controllo, si pone l'attenzione sulle variabili legate al pH. L'effetto della diversa tipologia di regolazione sul pH è visibile in Figura 2.6.

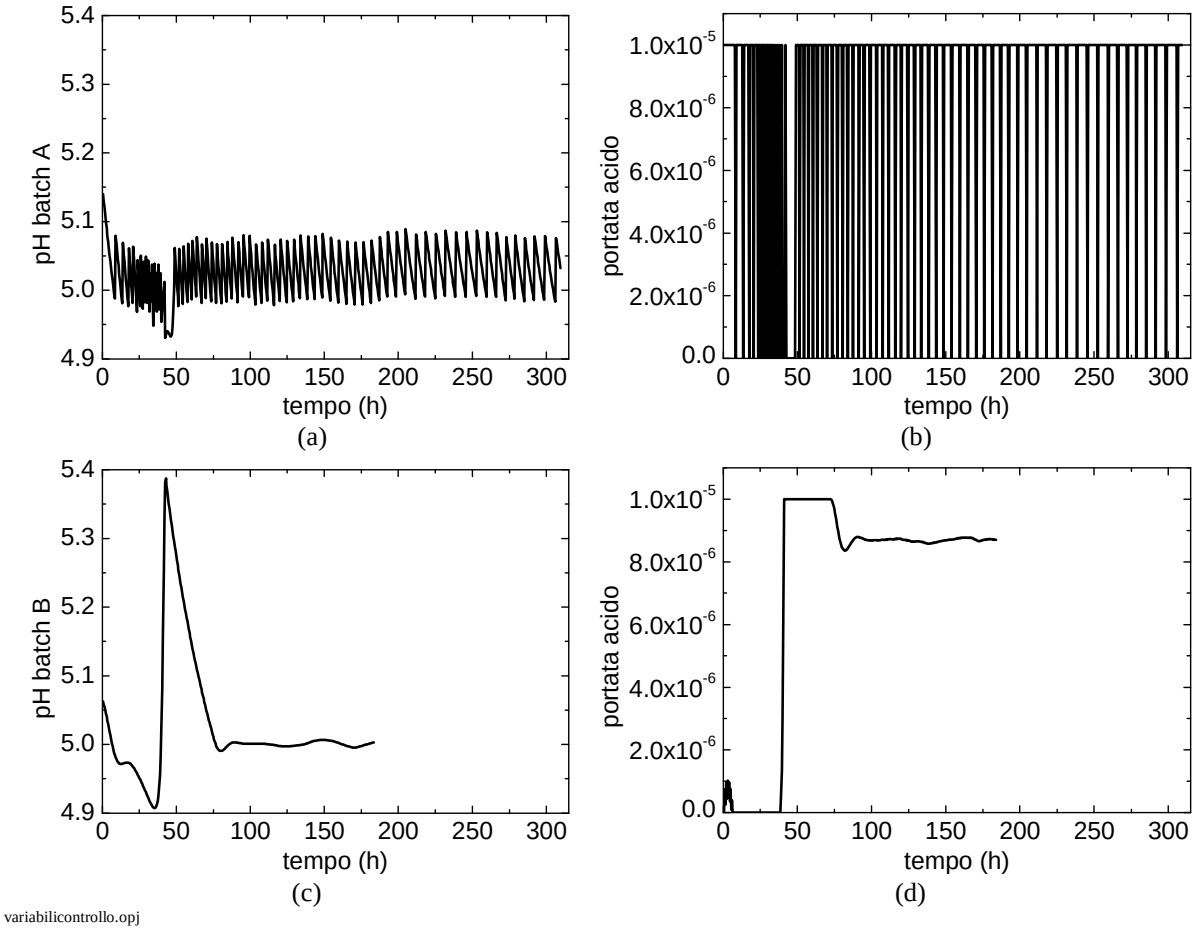
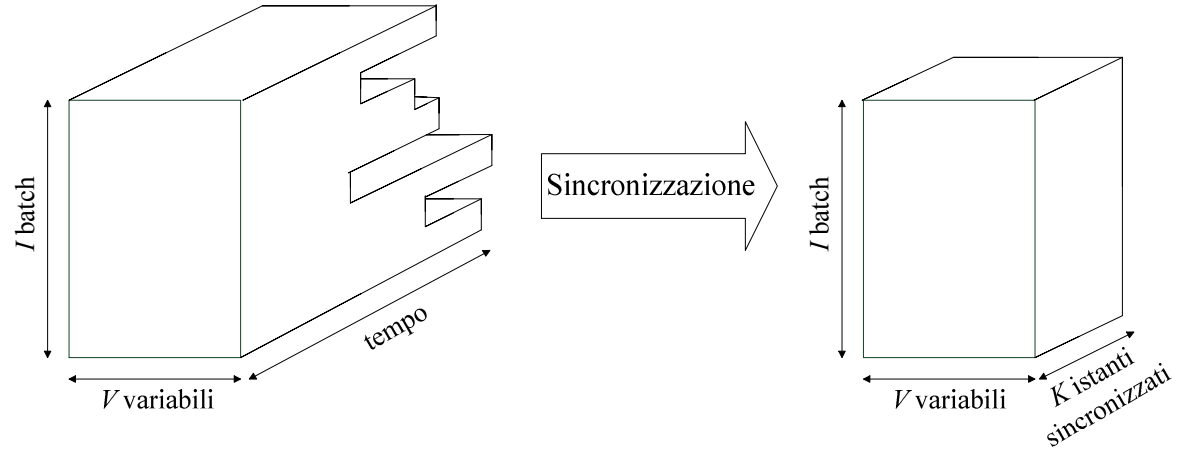


Figura 2.6. *Profilo temporale delle variabili di (a) pH e (b) portata di acido per il primo batch dell'impianto A, (c) pH e (d) portata di acido per il primo batch dell'impianto B.*

Dalla Figura 2.6 è evidente come la regolazione influisca sulle variabili; il pH è correlato alla portata di acido dello stesso impianto. Infatti, il profilo del pH di Figura 2.6a è dovuto alla variazione della portata di acido nel tempo (Figura 2.6b). Analoga considerazione si può fare per l'impianto B (Figura 2.6c e Figura 2.6d). Poiché il pH è una variabile importante del processo di produzione di penicillina, si dovrà considerare la diversa correlazione tra le variabili per i modelli sviluppati con esse.

Nelle Figure 2.4, 2.5 e 2.6, è visibile la diversa durata dei batch tra gli impianti e all'interno del singolo impianto (Figura 2.4). I problemi delle differenze di durata dei batch sono superati sincronizzando i batch, applicando la procedura definita al Paragrafo 1.1.3. La Figura 2.7 rappresenta una schematizzazione delle matrici che si ottengono dalla simulazione e successiva sincronizzazione.



sincronizzazione.opj

Figura 2.7. *Effetto della sincronizzazione sulla matrice di dati del processo.*

In particolare, per la sincronizzazione si utilizzano due variabili indicatrici, in modo da rappresentare la fase batch e quella fed-batch. Le due variabili sono la concentrazione di substrato, per la fase batch, e la concentrazione di penicillina, per la fase fed-batch. La prima varia nell'intervallo 12 - 0.56 g/L e allinea 25 campioni; la seconda va da 0.18 g/L alla concentrazione di specifica desiderata nell'impianto e allinea 175 campioni. Alla fine della sincronizzazione si dispone di una matrice $\underline{\mathbf{X}}$ (100×16×200) per entrambi gli impianti.
Infine, sono state costruite 5 anomalie per l'impianto B, definite dalla Tabella 2.4. Le anomalie si differenziano per le variabili coinvolte dall'anomalia, per la forma dell'anomalia, che può essere di tipo step o a rampa, e infine per l'entità dell'anomalia stessa. Tutte le anomalie iniziano al tempo 100 h e persistono per tutta la durata del batch. Anche nel caso batch si parlerà di fase 1 per i campioni normali e fase 2 per quelli anomali.

Tabella 2.4. *Caratteristiche delle anomalie dell'impianto B; ampiezza della rampa espressa in pendenza, ampiezza dello step espressa in percentuale di variazione dal set point.*

N° anomalia	Variabile	Tipo anomalia	Entità anomalia per batch B
1	1	Step	-25%
2	1	Rampa	-0.5
3	1	Rampa	-1
4	2	Step	-15%
5	3	Rampa	-0.001

Dalla Tabella 2.4 si vede che alcune anomalie, ad esempio 1 e 4, sono più evidenti delle altre. In questo modo, applicando i metodi statistici per il monitoraggio del sistema batch e il trasferimento dei modelli di monitoraggio, si valuta la loro efficienza per diversi casi a seconda dell'anomalia considerata. L'analisi che ne deriva è più generale.

Capitolo 3

Trasferimento di modelli per il monitoraggio di un processo di spray-drying: confronto tra due metodi

In questo Capitolo vengono illustrati diversi metodi per trasferire modelli di monitoraggio da un impianto pilota ad un impianto industriale per un processo continuo di *spray-drying*. Le tecniche sviluppate si basano su PCA e JY-PLS adattativi. Le prestazioni dei diversi metodi sono confrontate.

3.1 Metodi per il trasferimento

Viene considerato un processo industriale continuo di *spray-drying* e il trasferimento avviene dall'impianto A, un impianto pilota su cui è stata fatta un'ampia sperimentazione, all'impianto B di scala industriale, che inizia ad essere esercito. Il modello per il monitoraggio viene costruito sui dati di A e su quelli di B disponibili inizialmente, e viene aggiornato ogni volta che un nuovo campione di B viene acquisito, fino a che non ci siano sufficienti informazioni per utilizzare esclusivamente i dati di B per costruire un modello di monitoraggio autonomo.

Il trasferimento di modelli per il monitoraggio può essere fatto perché gli impianti sono geometricamente simili e alla loro base ci sono le stesse forze motrici e leggi fisiche. Sull'impianto A sono disponibili molti dati delle variabili di processo, dato che su di esso esiste una sperimentazione. L'impianto B si trova negli istanti iniziali di funzionamento, per cui sono note poche informazioni. In queste condizioni, un modello per il monitoraggio costruito sui soli dati dell'impianto B non sarebbe efficiente. Il trasferimento ha l'obiettivo di poter rilevare le anomalie che si sviluppano nell'impianto B anche durante la fase iniziale del processo utilizzando le informazioni dell'impianto A, nonostante i due impianti non siano identici (per esempio: diversa scala, diversa strumentazione, diversa collocazione dei sensori, diverse unità di misura delle variabili, …).

A seconda delle informazioni dell'impianto disponibili, Facco *et al.* (2012) e Tomba *et al.* (2012) hanno sviluppato differenti scenari, come mostrato in Figura 3.1. I metodi si differenziano a seconda che si utilizzino i soli dati di processo o si prendano in considerazione

anche informazioni aggiuntive. Inoltre, i metodi di trasferimento dipendono dal tipo di variabili di processo che sono impiegate per costruire il modello di monitoraggio.

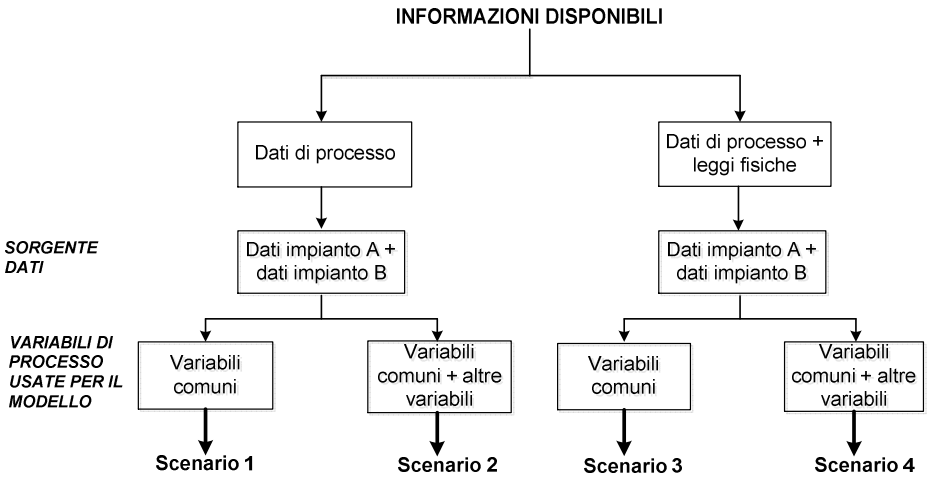


Figura3.1.vsd

Figura 3.1. *Struttura dello sviluppo degli approcci alle variabili latenti per il trasferimento di modelli di monitoraggio tra impianti differenti A e B.*

I quattro scenari proposti (Figura 3.1) derivano dalle informazioni disponibili:

- scenario 1: il modello di monitoraggio per l'impianto B è costruito a partire da dati di processo, considerando le variabili comuni tra gli impianti;
- scenario 2: il modello di monitoraggio per l'impianto B è costruito a partire da dati di processo, considerando le variabili comuni e non comuni tra gli impianti;
- scenario 3: per lo sviluppo del modello di monitoraggio per l'impianto B si utilizzano i dati di processo e le informazioni fisiche caratteristiche del processo stesso, considerando le variabili comuni tra gli impianti;
- scenario 4: per lo sviluppo del modello di monitoraggio per l'impianto B si utilizzano i dati di processo e le informazioni fisiche caratteristiche del processo stesso, considerando le variabili comuni e non comuni tra gli impianti.

Gli scenari 1 e 3 utilizzano un metodo di trasferimento basato su PCA, mentre gli scenari 2 e 4 impiegano un metodo che si basa su JY-PLS.

In tutti i casi, per la costruzione del modello si dispongono dei dati dell'impianto A e di quelli NOC dell'impianto B che vengono aggiornati in modo adattativo per l'uso in linea (Rännar *et al.*, 1998; Qin, 1998; Li *et al.*, 2000). I dati devono essere pretrattati mediante *autoscaling* secondo media e varianza del proprio *set* di origine. Questo compensa alcune delle differenze che sussistono tra gli impianti (per esempio le differenze di unità di misura).

Poiché i metodi trattati in letteratura si basano su diversi criteri di aggiornamento del modello e diversi criteri di valutazione delle prestazioni del modello per il monitoraggio, non è possibile fare delle valutazioni di confronto. Il lavoro svolto in questa Tesi consiste

nell'uniformare i metodi in modo da fare un confronto tra i quattro scenari e stabilire i vantaggi e gli svantaggi di ognuno, in riferimento all'applicazione al caso industriale.

3.2 Adattamento del modello a nuovi dati

I metodi su cui si basano i quattro scenari definiti al Paragrafo 3.1 sviluppano modelli di monitoraggio a partire dai dati di A e dai campioni NOC di B disponibili. Tali modelli vengono aggiornati in modo adattativo ai nuovi dati che vengono acquisiti dall'impianto B in condizioni NOC. Si assume che all'istante *k*, a cui è disponibile un nuovo campione $\mathbf{x}_k^B$, si abbia un modello di monitoraggio costruito sui dati dell'impianto A e su *j* dati NOC dell'impianto B ottenuti fino all'istante (*k*-1).

Per il nuovo campione $\mathbf{x}_k^B$ si segue la seguente procedura:

1. pre-processare il campione con i valori di media e varianza dei *j* campioni NOC dell'impianto B disponibili all'istante (*k*-1), dove *j* è un sotto *set* dei *k* campioni precedentemente giudicati appartenere alle NOC;
2. proiettare $\mathbf{x}_k^B$ nello spazio del modello di monitoraggio definito all'istante (*k*-1);
3. calcolare le statistiche T^2 e SPE e confrontarle con il limite di confidenza delle rispettive carte di controllo; se entrambe sono al di sotto del limite di confidenza, si aggiorna il modello usando $\mathbf{x}_k^B$ e si ritorna al punto 1; altrimenti si passa a 4;
4. se almeno Δ-1 degli ultimi Δ campioni, con $\Delta = 5$, sono al di fuori dei limiti per T^2 o per SPE si procede con il punto 5 altrimenti si torna al punto 1;
5. si valuta la "normalità" di $\mathbf{x}_k^B$ rispetto un modello costruito localmente sugli ultimi *W* campioni dell'impianto B, dove *W* rappresenta il numero di campioni NOC di B disponibili all'inizio del trasferimento;
6. se almeno Δ-1 degli ultimi Δ campioni sono al di fuori dei limiti o per T^2 o per SPE del modello locale si genera un allarme per segnalare l'anomalia, si scarta $\mathbf{x}_k^B$ e si riprende la procedura dal punto 1, altrimenti il modello locale diventa il nuovo modello di monitoraggio e si riprende ancora dal punto 1.

Si noti che *W* è l'ampiezza iniziale di una finestra di campioni, ovvero corrisponde al numero di campioni disponibili inizialmente dall'impianto B. Al generico istante *k*, i campioni NOC possono essere in numero superiore, rappresentato da *j*. Quando si costruisce il modello locale, l'ampiezza del *dataset* è fissata da *W*, indipendentemente da *j*.

Con questa procedura quando un nuovo campione $\mathbf{x}_k^B$ è disponibile dall'impianto B, l'adattamento del modello avviene se $\mathbf{x}_k^B$ è valutato:

- normale rispetto al modello di monitoraggio definito all'istante (*k*-1);
- non normale rispetto al modello (*k*-1)-esimo, ma è rappresentativo delle condizioni definite da una finestra locale costruita sugli ultimi *W* campioni disponibili dall'impianto B all'istante (*k*-1) assieme a quelli dell'impianto A.

La logica della procedura di adattamento è rappresentata in Figura 3.2, utilizzata per tutti gli scenari.

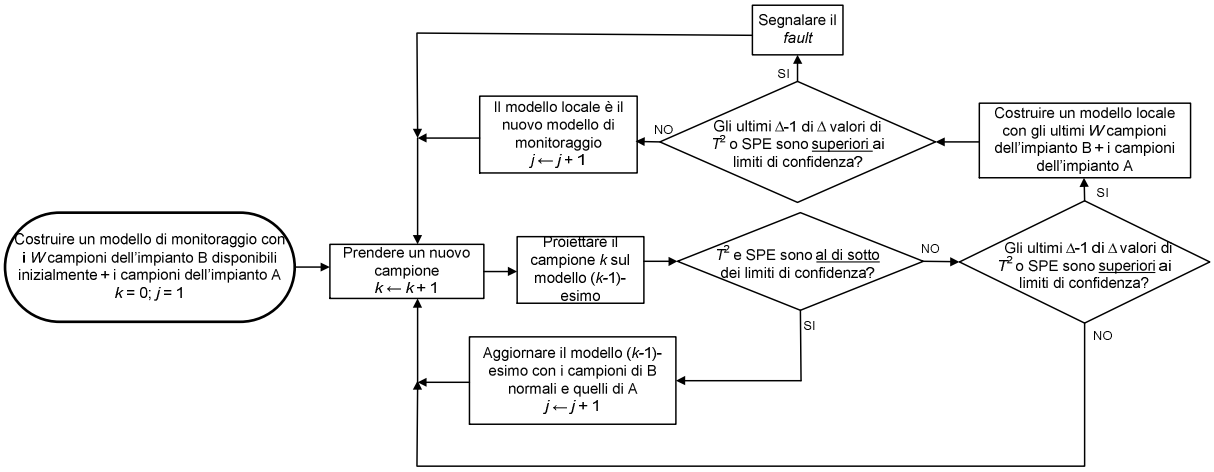


Figura3.2.vsd

Figura 3.2. *Schema a blocchi del meccanismo di adattamento del modello per il monitoraggio on-line. L'indice k rappresenta il nuovo campione mentre j è il numero di campioni NOC che costituiscono il set di calibrazione del modello.*

Il modello di monitoraggio iniziale viene costruito a partire dai dati disponibili per gli impianti A e B, i quali sono pre-processati secondo media e varianza del *set* di dati dell'impianto da cui derivano per compensare le differenze relative alle stesse variabili nei due impianti, come definito al Paragrafo 3.1. Il pretrattamento del nuovo campione consiste nell'*autoscaling* per il metodo PCA e nel metodo riportato da Garcia-Muñoz *et al.* (2005) per il metodo JY-PLS.
Definito il criterio di adattamento, si analizzano in seguito i principi su cui si basano i singoli scenari e si evidenziano le modifiche operate per l'uniformazione.

3.3 Trasferimento basato su dati di processo

Spesso le variabili di processo misurate in un impianto sono correlate fra loro e la struttura di correlazione è legata ai meccanismi fondamentali del processo. In impianti di diversa scala che producono lo stesso prodotto, nel caso in esame gli impianti A e B, ci si aspetta che le strutture di correlazione tra le variabili all'interno di ciascuna scala siano le stesse perché i meccanismi fondamentali che governano i processi non cambiano. D'altra parte, anche le variabili non comuni possono contenere una serie di informazioni relative alle forze motrici del processo che possono essere utili nel monitoraggio; se queste non vengono considerate, il loro apporto di informazioni al processo viene perso. Sulla base di queste considerazioni sono stati proposti da Facco *et al.* (2012) due metodi per il trasferimento di modelli per il monitoraggio, basati su soli dati di processo, che sono descritti in seguito.

3.3.1 Scenario 1: utilizzo di variabili comuni

Per costruire il modello di monitoraggio si utilizza il metodo statistico PCA, basato sui dati dell'impianto A; quando dall'impianto B è disponibile un nuovo campione *k*, il modello di monitoraggio viene eventualmente ricostruito utilizzando le nuove informazioni, per cui si realizza l'aggiornamento del modello. Ad ogni istante *k* la matrice sulla quale si costruisce il modello è definita da:

$$\mathbf{X}'_k = \begin{bmatrix} \mathbf{X}'^{\mathrm{A}} \\ \mathbf{X}'^{\mathrm{B}}_j \end{bmatrix}. \tag{3.1}$$

Le matrici $\mathbf{X}'^{\mathrm{A}}$ e $\mathbf{X}'^{\mathrm{B}}_j$ sono le matrici derivanti da $\mathbf{X}^{\mathrm{A}}$ e $\mathbf{X}^{\mathrm{B}}$, definite al Paragrafo 2.1.1, in cui si considerano le sole variabili comuni, ovvero $\mathbf{v}'^{\mathrm{A}} = \{1, 3, 4, 5, 7, 8, 9, 10, 12\}$ e $\mathbf{v}'^{\mathrm{B}} = \{1, 2, 3, 4, 5, 6, 7, 8, 10\}$. In particolare nella fase iniziale del processo si immagina di avere a disposizione un numero limitato *j* di campioni NOC di B. La matrice $\mathbf{X}'_k$ è il concatenamento verticale dei dati dell'impianto A con i *j* dell'impianto B in condizioni NOC disponibili fino all'istante (*k*-1), considerando le sole variabili comuni.

All'istante *k*, il monitoraggio del nuovo campione procede secondo il seguente algoritmo (Facco *et al.*, 2012):

1. costruire un modello PCA sulla matrice $\mathbf{X}'_k$;
2. costruire le carte di monitoraggio basate su T^2 e su SPE con un limite di confidenza del 95%;
3. autoscalare il nuovo campione $\mathbf{x}'^{\mathrm{B}}_k$ proveniente dall'impianto B secondo media e varianza calcolate sul *set j* di dati di B disponibile fino all'istante precedente (*k*-1);
4. proiettare il nuovo campione sullo spazio delle componenti principali definito dal modello PCA sviluppato al punto 1;
5. calcolare la statistica di Hotelling T^2 e la statistica SPE del nuovo campione e confrontarla con i limiti di confidenza nelle carte di controllo.

Il modello viene aggiornato in modo adattativo solo se le statistiche calcolate al punto 5 sono all'interno dei limiti; se tale condizione è verificata il nuovo campione viene incluso in $\mathbf{X}'^{\mathrm{B}}_j$, altrimenti $\mathbf{x}'^{\mathrm{B}}_k$ viene scartato.

Rispetto al lavoro di Facco *et al.* (2012), l'algoritmo è stato uniformato con gli altri scenari. In particolare, le modifiche riguardano:

- l'introduzione dello schema di adattamento descritto al Paragrafo 3.2;
- la costruzione del modello per il monitoraggio, per cui si è fatto in modo che il numero di componenti principali sia definito automaticamente ad ogni iterazione come il numero di autovalori del modello PCA costruito sulla matrice $\mathbf{X}'_k$ che ha valore superiore all'unità.

3.3.2 Scenario 2: utilizzo di variabili comuni e non comuni

Utilizzando solo le variabili comuni non si considerano le informazioni contenute in quelle che sono esclusivamente misurate in uno degli impianti (variabili non comuni). Tuttavia le correlazioni tra queste ultime e le variabili comuni all'interno di una scala potrebbero racchiudere informazioni molto importanti ai fini del monitoraggio. Per questo motivo si fa riferimento al metodo JY-PLS, proposto da Garcia-Muñoz *et al.*(2005), che le considera entrambe. Infatti, il modello JY-PLS considera lo spazio delle variabili comuni unitamente allo spazio delle variabili non comuni; ciò permette di monitorare il processo in uno spazio di dimensioni ridotte costruito da variabili latenti che tengono conto sia della correlazione tra le variabili comuni in diversi impianti, che della correlazione tra tutte le variabili all'interno di ciascun impianto. Come nel caso PCA, il modello e le carte di controllo vengono aggiornati quando è disponibile un nuovo campione dall'impianto B che soddisfi determinate condizioni per cui si parla di JY-PLS adattativo.

I dati corrispondenti alle variabili comuni sono raccolti nelle matrici $\mathbf{Y}''^{\mathrm{A}}$ e $\mathbf{Y}''^{\mathrm{B}}$, dove per questo metodo sono scelte come variabili comuni per l'impianto A $\mathbf{v}''^{\mathrm{A}} = \{1, 4, 8, 9, 10, 12\}$ e per l'impianto B $\mathbf{v}''^{\mathrm{B}} = \{1, 3, 6, 7, 8, 10\}$. I dati corrispondenti alle variabili non comuni, invece, costituiscono le matrici $\mathbf{X}''^{\mathrm{A}}$ e $\mathbf{X}''^{\mathrm{B}}$ in cui le variabili sono $\mathbf{v}'''^{\mathrm{A}} = \{2, 3, 5, 6, 7, 11, 13, 14, 15, 16\}$ per l'impianto A e $\mathbf{v}'''^{\mathrm{B}} = \{2, 4, 5, 9\}$ per l'impianto B. La Figura 3.3 rappresenta schematicamente la suddivisione delle variabili delle matrici.

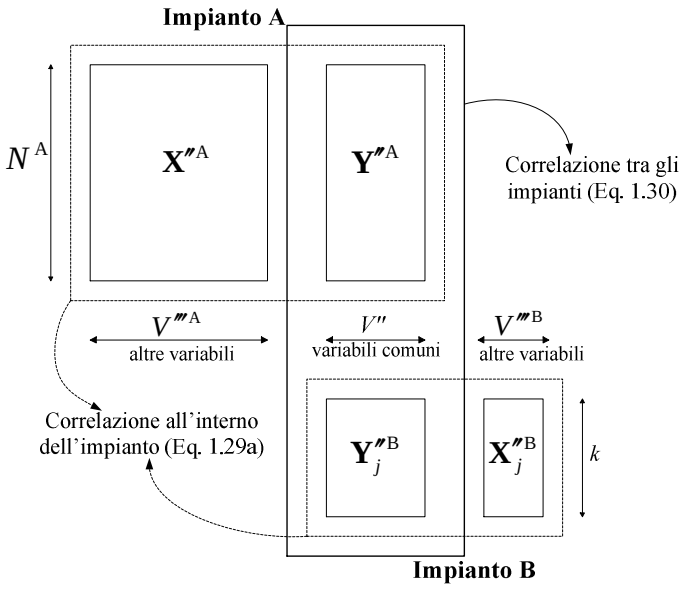


Figura3.3.vsd

Figura 3.3. *Schema delle variabili per l'applicazione del metodo JY-PLS all'istante k.*

La costruzione del modello di monitoraggio si basa sui campioni dell'intero *set* per l'impianto A e sui *j* campioni NOC provenienti dall'impianto B fino all'istante (*k*-1).

Il monitoraggio si svolge secondo l'algoritmo (Facco *et al.*, 2012):

1. costruire il modello JY-PLS su tutti i dati dell'impianto A, $\mathbf{X}''^{\mathrm{A}}$ e $\mathbf{Y}''^{\mathrm{A}}$, più i *j* dati NOC disponibili dall'impianto B fino all'istante (*k*-1), $\mathbf{X}''^{\mathrm{B}}_j$ e $\mathbf{Y}''^{\mathrm{B}}_j$;
2. costruire le carte di monitoraggio basate su T^2 e su SPE con un limite di confidenza del 95%;
3. autoscalare le nuove osservazioni $\mathbf{x}''^{\mathrm{B}}_k$, $\mathbf{y}''^{\mathrm{B}}_k$ provenienti dall'impianto B secondo media e varianza calcolate sul *set j* di dati NOC disponibili in B fino all'istante precedente (*k*-1);
4. proiettare il nuovo campione sullo spazio del modello JY-PLS definito al punto 1 secondo $\hat{\mathbf{t}}^{\mathrm{B}}_k = \mathbf{x}''^{\mathrm{B}}_k \mathbf{W}^{*\mathrm{B}}_k$, $\hat{\mathbf{y}}''^{\mathrm{B}}_k = \hat{\mathbf{t}}^{\mathrm{B}}_k \mathbf{Q}^{\mathrm{J}\,\mathrm{T}}_k$;
5. calcolare la statistica di Hotelling T^2 e la statistica SPE della nuova osservazione e confrontarla con i limiti di confidenza nelle carte di controllo.

Il modello viene aggiornato solo se le statistiche calcolate al punto 5 sono all'interno dei limiti.

Le modifiche apportate al metodo sono le stesse dello scenario 1:

- si applica lo schema di adattamento definito al Paragrafo 3.2;
- il numero di componenti principali per il modello viene fatto selezionare automaticamente come il numero di autovalori del modello PCA costruito sulla matrice $\mathbf{Y}''^{\mathrm{B}}_j$ che ha valore superiore all'unità.

3.4 Trasferimento basato su dati di processo e conoscenza fisica sul processo

Poiché gli impianti A e B sono simili non solo dal punto di vista geometrico, ma anche perché le leggi fondamentali che descrivono i fenomeni fisici nei due sistemi sono le stesse, particolare attenzione viene rivolta alle variabili fisiche che sono determinate dai fenomeni fisici e dovrebbero essere indipendenti dalle dimensioni dell'impianto (Tomba *et al.*, 2012). Il loro valore dipende esclusivamente dalle condizioni a cui lavora il sistema, per cui possono essere considerate indipendenti dall'impianto. Nel processo di *spray-drying* si può individuare una variabile di questo tipo a partire dal bilancio globale di energia alla camera di *spray-drying*. Considerando che l'energia richiesta per la vaporizzazione del solvente è uguale a quella che il gas deve cedere al sistema, si ottiene:

$$\frac{\dot{M}_{\mathrm{gas}}}{\dot{M}_{\mathrm{sol}}}\left(T^{\mathrm{in}} - T^{\mathrm{out}}\right) = \frac{\Delta H^{\mathrm{vap}}}{c_p}\left(1 - x_{\mathrm{solid}}\right) = \mathrm{cost}\,, \tag{3.2}$$

dove $\dot{M}_{\mathrm{gas}}$ è la portata di gas che entra nella camera di essiccamento, $\dot{M}_{\mathrm{sol}}$ è la portata di soluzione, T^{in} e T^{out} sono le temperature del gas in ingresso e in uscita, ΔH^{vap} è il calore di vaporizzazione, c_p è il calore specifico del gas e x_{solid} è la frazione di massa del solido in soluzione. Il primo termine dell'uguaglianza contiene le proprietà fisiche e la variabile x_{solid}, quest'ultima legata alla soluzione in ingresso, che è la medesima in entrambi gli impianti. Ne deriva che il secondo termine della (3.2) assume valori simili tra gli impianti. Questo termine

è definito differenza di temperatura pesata *wtd* (*weighted temperature difference*) e si può calcolare dai dati disponibili per entrambi i processi. Esso costituisce un'ulteriore variabile e può essere utilizzata per rappresentare stati simili raggiunti dagli impianti. In particolare, *wtd* identifica lo spazio termodinamico del processo (Tomba *et al.*, 2012).

Tomba *et al.* (2012) propone due strategie per trasferire il modello di monitoraggio dall'impianto A all'impianto B. Entrambe sono basate su *wtd* per trasferire informazioni relative alla conoscenza del processo, ma differiscono per i metodi adottati per il trasferimento. In questa Tesi vengono modificati il criterio per la generazione degli allarmi e le diagnostiche del modello di monitoraggio.

3.4.1 Scenario 3: utilizzo di variabili comuni

I meccanismi fondamentali che caratterizzano un processo possono essere considerati gli stessi nei due impianti, pertanto si suppone che la struttura di correlazione tra variabili nell'impianto A sia la stessa di quella nell'impianto B per variabili comuni nei due impianti. Il modello di monitoraggio per l'impianto B, dunque, viene costruito applicando il metodo PCA alla matrice $\mathbf{X}_k^{\mathrm{AB}}$ ottenuta concatenando verticalmente i campioni disponibili dall'impianto B $\mathbf{X}_j^{\mathrm{B}}$ e i campioni dell'impianto A $\mathbf{X}_{\mathrm{SUB}}^{\mathrm{A}}$, che risultano più simili a quest'ultimi:

$$\mathbf{X}_k^{\mathrm{AB}} = \begin{bmatrix} \mathbf{X}_{\mathrm{SUB}}^{\mathrm{A}} \\ \mathbf{X}_j^{\mathrm{B}} \end{bmatrix}. \tag{3.3}$$

La similarità tra i campioni dei due impianti è determinata attraverso la variabile indipendente *wtd*: questa viene calcolata per ciascun campione in $\mathbf{X}^{\mathrm{A}}$ e per i *j* campioni NOC disponibili inizialmente per l'impianto B ($\mathbf{X}_j^{\mathrm{B}}$). Per ciascun valore di *wtd* calcolato dai campioni dell'impianto B (riportato nel vettore colonna $\mathbf{wtd}_j^{\mathrm{B}}$), tutti i campioni di $\mathbf{X}^{\mathrm{A}}$ che hanno valori di *wtd* (riportato nel vettore colonna $\mathbf{wtd}_j^{\mathrm{A}}$) più simili a quelli dell'impianto B sono selezionati per formare la matrice $\mathbf{X}_{\mathrm{SUB}}^{\mathrm{A}}$[2]. Dunque, il modello PCA all'istante *k* viene costruito sulla matrice $\mathbf{X}_k^{\mathrm{AB}}$ dell'Equazione (3.3), cioè sulla matrice data dal concatenamento verticale dei campioni dell'impianto A selezionati nel *set* totale $\mathbf{X}^{\mathrm{A}}$, in base alla similitudine dei *wtd* di B, e i *j* campioni NOC di B disponibili, considerando le sole variabili comuni tra gli impianti.

In questo caso le variabili comuni sono quelle definite in $\mathbf{v}'^{\mathrm{A}}$ e $\mathbf{v}'^{\mathrm{B}}$, come nello scenario 1, con l'aggiunta di *wtd* come variabile comune. Il numero di componenti principali necessario per la costruzione del modello corrisponde al numero di autovalori con valore superiore all'unità e ad ogni passo è determinato automaticamente.

In questa Tesi il metodo è stato uniformato ai precedenti modificando:

- il criterio per il passaggio all'analisi locale e il criterio di chiamata degli allarmi; si introduce il criterio 4 *outlier* su 5 consecutivi, che rende l'analisi più robusta in quanto

[2] La procedura per la selezione dei dati di A è definita in Tomba *et al.* (2012).

contribuisce a ridurre il numero di falsi allarmi rispetto al criterio precedente (3 *outlier* consecutivi);

- le diagnostiche per valutare le prestazioni del modello.

3.4.2 Scenario 4: utilizzo di variabili comuni e non comuni

Negli impianti A e B ci sono alcune variabili che non sono comuni, ma che potrebbero essere utili per il monitoraggio del processo. Il metodo JY-PLS permette di tenere conto di tutte le variabili, comuni e non comuni. Le variabili indipendenti degli impianti, come *wtd*, condividono la stessa struttura di correlazione e vengono utilizzate per generare lo spazio comune su cui andare ad analizzare i campioni provenienti dagli impianti. JY-PLS modella le informazioni interne a ciascun impianto in modo congiunto a quelle comuni tra gli impianti, caratterizzate da *wtd*. Per l'applicazione del metodo JY-PLS, le variabili indipendenti sono calcolate dai dati disponibili per ciascun campione dell'impianto A in $\mathbf{X}^{A}$ e vengono raccolte nella matrice $\mathbf{Y}^{A}$; la stessa cosa vale anche per i *j* campioni NOC disponibili dall'impianto B in $\mathbf{X}_j^{B}$, costruendo il vettore $\mathbf{Y}_j^{B}$.

Verificato che $\mathbf{Y}^{A}$ e $\mathbf{Y}_j^{B}$ condividano la stessa struttura di correlazione, il metodo JY-PLS determina lo spazio latente comune tra i dati in modo da massimizzare allo stesso tempo la covarianza tra $\mathbf{X}^{A}$ e $\mathbf{Y}^{A}$ e tra $\mathbf{X}_j^{B}$ e $\mathbf{Y}_j^{B}$ (informazioni interne all'impianto), assieme alla covarianza congiunta di $\mathbf{Y}^{A}$ e $\mathbf{Y}_j^{B}$ (informazioni comuni tra impianti). Ciò significa che lo spazio latente di $\mathbf{X}^{A}$ e $\mathbf{X}_j^{B}$ è opportunamente ruotato per allinearsi alla direzione di massima variabilità congiunta delle variabili in $\mathbf{Y}^{A}$ e $\mathbf{Y}_j^{B}$. In questo modo lo spazio congiunto è usato per mettere in relazione dati da impianti differenti, trascurando le differenze nel tipo di variabili misurate (Tomba *et al.*, 2012).

Il modello JY-PLS viene costruito con un numero di componenti principali scelto automaticamente considerando il numero di autovalori con valore superiore all'unità che vengono calcolati con il metodo PCA applicato alla matrice $\mathbf{X}_j^{B}$. Il monitoraggio viene fatto sullo spazio delle variabili non comuni, ovvero le carte SPE sono costruite sullo spazio delle variabili non comuni anziché su quello delle variabili comuni.

Come per lo scenario 3, l'intervento di uniformazione è stato compiuto:

- sul criterio per lo sviluppo della finestra locale e della generazione di allarmi;
- sul tipo di diagnostiche con cui valutare il modello.

3.5 Confronto dei metodi: risultati e discussione

Nei seguenti Paragrafi si valuteranno le prestazioni di monitoraggio di un modello costruito sui dati dell'impianto B. Poi si confronteranno i risultati del trasferimento per i vari metodi degli scenari definiti.

3.5.1 Monitoraggio di condizioni operative normali

Supponendo di avere a disposizione tutti i dati dell'impianto B si vogliono valutare le prestazioni del metodo statistico per il controllo di processo. Questa è una condizione limite per il caso di studio. Il modello è costruito mediante PCA in cui il numero di componenti principali è posto pari a 4 secondo la regola dell'autovalore. I dati utilizzati sono $\mathbf{X}^{B}$, ovvero tutti i dati disponibili dall'impianto industriale, mentre l'anomalia proiettata è quella rappresentata da $\mathbf{x}^{BF}$ (definita al Paragrafo 2.1.1). Le statistiche SPE e T^2 per ogni campione proiettato sono rappresentate in Figura 3.4.

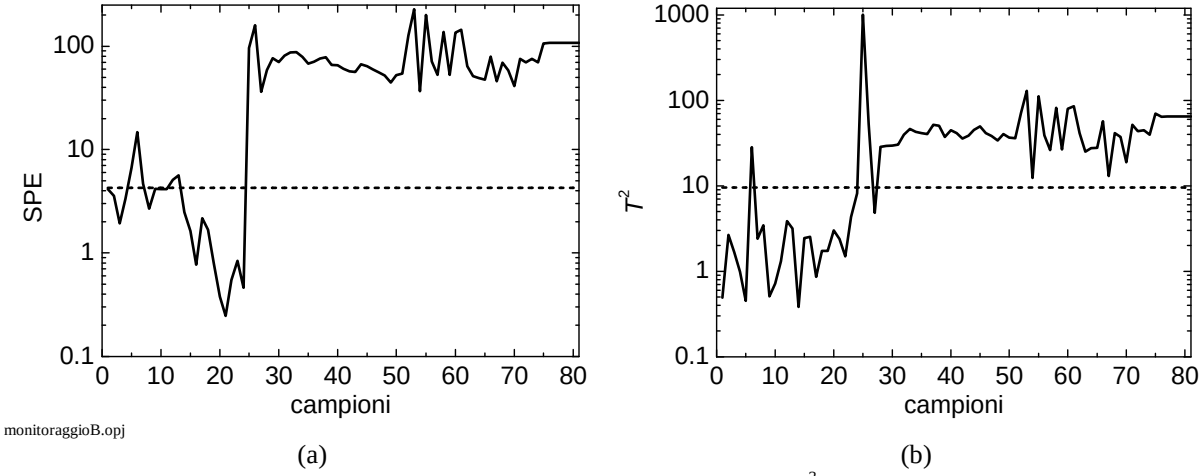


Figura 3.4. *Risultati del monitoraggio: carte di controllo (a) T^2 e (b) SPE per l'anomalia dell'impianto B proiettata nel modello PCA a 4 componenti principali, costruito utilizzando l'intero set di dati di B e nessun dato dell'impianto A.*

Dalla Figura 3.4, si osserva che l'anomalia viene rilevata dal campione 25 e si distingue facilmente la fase 1 (NOC) dalla fase 2 (anomalia). Quando molti dati sono disponibili il monitoraggio è molto efficiente e l'anomalia è rilevata; lo dimostrano i valori delle diagnostiche in Tabella 3.1.

Tabella 3.1. *Diagnostiche per il modello di monitoraggio costruito con soli dati $\mathbf{X}^{B}$ quando viene proiettata l'anomalia $\mathbf{x}^{BF}$.*

Diagnostica	**Valore calcolato**
Frequenza degli allarmi in fase 1	0%
Frequenza degli allarmi in fase 2	94.7%
Ritardo di rilevazione	4 campioni

Si noti che la frequenza degli allarmi in fase 2 non è esattamente il 100%, come risulterebbe dalle carte di monitoraggio, a causa del ritardo di rilevazione dovuto al criterio adottato per la segnalazione dell'allarme. Infatti, nei primi istanti dell'anomalia non si verifica la condizione di 4 *outlier* sulle ultime 5 osservazioni.

Le prestazioni sono eccellenti. Tuttavia ciò non accade quando, nelle fasi iniziali, si hanno solo pochi dati di B a disposizione. Dunque, quando negli istanti iniziali si hanno pochi campioni non è possibile monitorare il processo solo con i dati dell'impianto B.

3.5.2 Rilevamento di anomalie su variabili comuni

Le strategie per il trasferimento proposte vengono applicate ai dati NOC degli impianti A e B e a 30 realizzazioni dell'anomalia dell'impianto B. L'analisi delle diagnostiche dà una valutazione quantitativa delle prestazioni del modello nella rilevazione dell'anomalia per ogni scenario. I grafici risultanti rappresentano il valore della diagnostica di riferimento in funzione dell'ampiezza della finestra *W*, parametrizzato nel numero di campioni NOC *j* disponibili prima dell'anomalia vera e propria (75, 100, 125, 150, 175 campioni). Ciascun punto dell'asse delle ascisse indica il numero di campioni *W* disponibili inizialmente dall'impianto B, che corrisponde anche al numero di campioni consecutivi usati per il monitoraggio locale. Le curve terminano all'ampiezza della finestra pari al numero di campioni NOC *j* della parametrizzazione poiché *W* non può contenere campioni dell'anomalia. I risultati del trasferimento vengono analizzati per tutti gli scenari confrontando i metodi basati su PCA (scenari 1 e 3) e per quelli basati su JY-PLS (scenari 2 e 4).

3.5.2.1 Confronto tra i risultati dello scenario 1 e dello scenario 3

I risultati del trasferimento per lo scenario 1 sono presentati in Figura 3.5, in termini di frequenza degli allarmi in fase 1 e fase 2, e in Figura 3.6, in termini di ritardo di rilevazione.

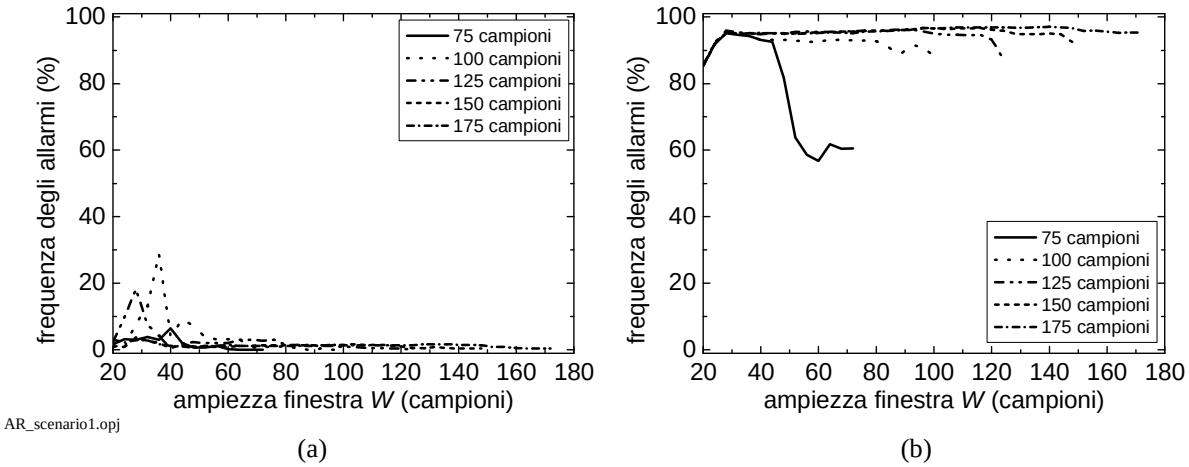
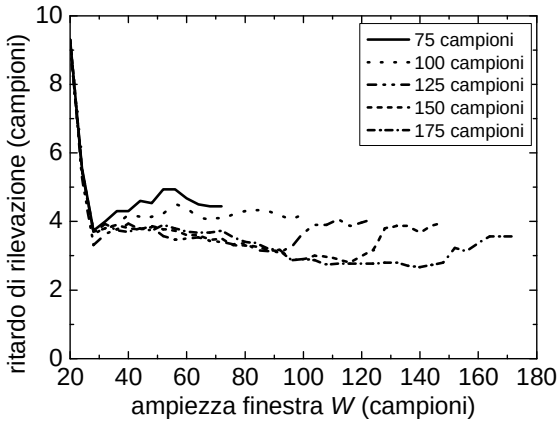


Figura 3.5. *Trasferimento di modelli con PCA adattativa. Effetto del numero di campioni NOC j disponibili nello scenario 1 sulla frequenza degli allarmi in (a) fase 1 e (b) fase 2 dell'anomalia.*



TD_scenario1.opj

Figura 3.6. *Trasferimento di modelli con PCA adattativa. Effetto del numero di campioni NOC j disponibili nello scenario 1 sul ritardo di rilevazione.*

Le Figure 3.5 e 3.6 indicano che le prestazioni del monitoraggio sono molto soddisfacenti. In riferimento alla Figura 3.5a, indipendentemente dal numero di campioni NOC *j* prima dell'anomalia, la frequenza degli allarmi in fase 1 è bassa se sono disponibili finestre di campioni dall'impianto B con ampiezza $W \geq$ a 40 campioni NOC dell'impianto B. D'altra parte, perché anche la fase 2 sia monitorata adeguatamente, dalla Figura 3.5b si nota che sono necessari più di 75 campioni NOC *j*. Per quanto riguarda il ritardo di rilevazione dell'anomalia, dalla Figura 3.6 si può dire che per finestre *W* superiori a 40 campioni il ritardo tende a 4 campioni in accordo con il criterio per la generazione di un allarme, secondo il quale devono verificarsi almeno 4 campioni al di fuori dei limiti delle carte di controllo su 5 campioni consecutivi. Quando le finestre hanno ampiezze inferiori, il ritardo aumenta perché i limiti di controllo delle carte di monitoraggio si adattano ai dati dell'anomalia piuttosto che ai dati dei campioni NOC. Se i dati NOC aumentano con *W*, l'adattamento dell'anomalia è minore e il modello di monitoraggio più robusto.

In sintesi, per lo scenario 1, il monitoraggio è buono per finestre $W \geq 40$ campioni e un numero di campioni NOC $j > 75$.

Lo scenario 3 considera le informazioni fisiche del processo attraverso *wtd* e i risultati delle proiezioni di nuovi campioni di B si modificano dallo scenario 1, come si nota dalle Figure 3.7 e 3.8.

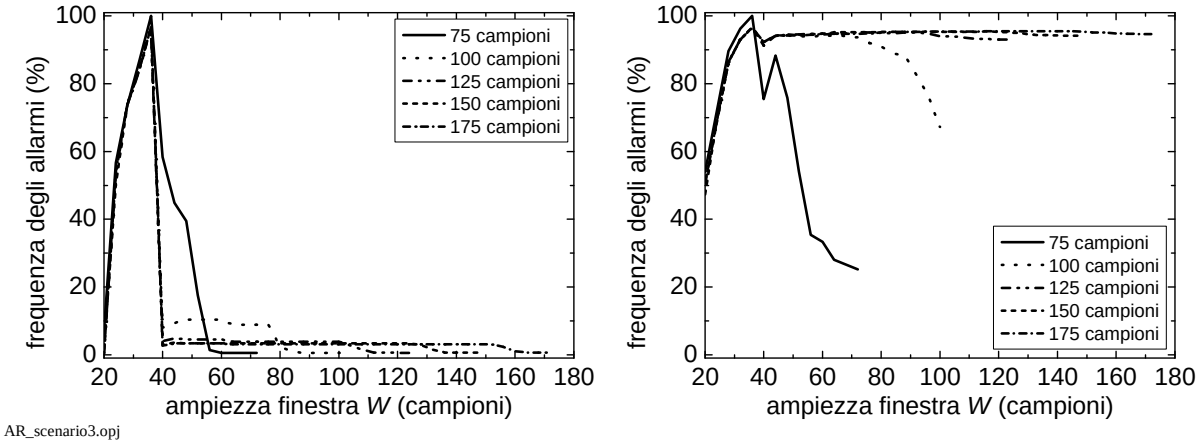


Figura 3.7. *Trasferimento di modelli con PCA adattativa. Effetto del numero di campioni NOC j disponibili nello scenario 3 sulla frequenza degli allarmi nella (a) fase 1 e (b) fase 2 dell'anomalia.*

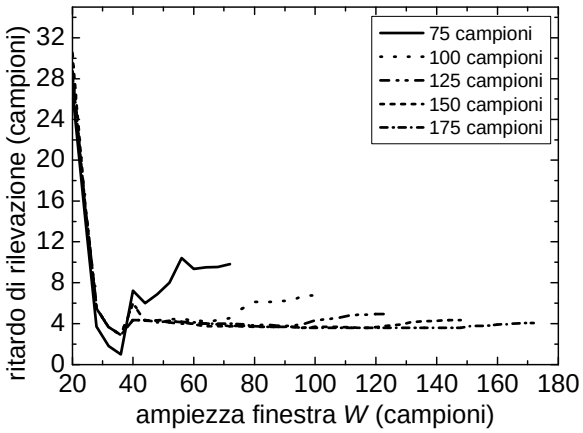


Figura 3.8. *Trasferimento di modelli con PCA adattativa. Effetto del numero di campioni NOC j disponibili nello scenario 3 sul ritardo di rilevazione.*

Anche per lo scenario 3 le prestazioni del monitoraggio sono buone, ma per condizioni differenti dallo scenario 1. Infatti, la Figura 3.7a evidenzia che la frequenza degli allarmi in fase 1 è bassa per finestre di ampiezze *W* superiori a 60 campioni indipendentemente dal numero di campioni NOC disponibili (*j*). Se tale numero, però, risulta superiore a 75 campioni, allora sono sufficienti finestre *W* con 40 campioni NOC di B. Le elevate frequenze di segnalazione di allarmi nella fase 1, quando si hanno pochi campioni di B, sono dovute al fatto che i dati NOC del *set* dell'anomalia si discostano dai dati del *set* di calibrazione, che nelle fasi iniziali hanno una prevalenza di campioni di A. Pertanto, i campioni sono erroneamente segnalati come anomali, in particolare ciò avviene per gli ultimi campioni della fase 1. Di conseguenza la frequenza degli allarmi della fase 2 si avvicina al 100% e il ritardo di rilevazione presenta degli anticipi. In dettaglio, la Figura 3.7b mostra che l'anomalia viene segnalata correttamente per finestre superiori a $W = 40$ campioni e per un numero di campioni NOC *j* maggiore di 100. Analogamente la Figura 3.8 rappresenta il ritardo di rilevazione che,

sulle basi di quanto esposto per la fase 1, presenta degli anticipi per finestre inferiori a 40 campioni. Inoltre, con meno di 75 NOC il ritardo aumenta a causa di un adattamento dei limiti ai campioni dell'anomalia, infatti la frequenza degli allarmi in fase 2 decresce notevolmente. Pertanto, per lo scenario 3 si deve disporre di finestre con ampiezza $W \geq 40$ campioni e un numero di campioni NOC $j > 75$.

Confrontando i risultati dei due scenari si può fare un confronto:

- considerando le variabili indipendenti del processo *wtd*, che introducono le informazioni fisiche, per il metodo PCA non si riscontrano notevoli miglioramenti;
- quando si dispone di pochi campioni NOC il metodo sviluppato nello scenario 1 è più efficiente di quello dello scenario 3;
- i risultati dei due scenari non differiscono, specialmente con molti NOC *j* e per ampiezze *W* della finestra elevate;
- perché il monitoraggio sia efficiente, in entrambi i casi si deve disporre di finestre *W* sufficientemente ampie in modo da non adattarsi ai dati dell'anomalia.

La variabile *wtd*, inserita nel *set* di calibrazione, dà in questo caso un contributo poco rilevante, infatti le variabili di processo utilizzate, $\mathbf{v}'^{A}$ e $\mathbf{v}'^{B}$, sono sufficienti per rappresentare il processo.

3.5.2.2 Confronto tra i risultati dello scenario 2 e dello scenario 4

L'analisi dello scenario 2 porta ad ottenere i risultati delle Figure 3.9, per la frequenza degli allarmi in fase 1 e in fase 2, e 3.10, per il ritardo di rilevazione dell'anomalia.

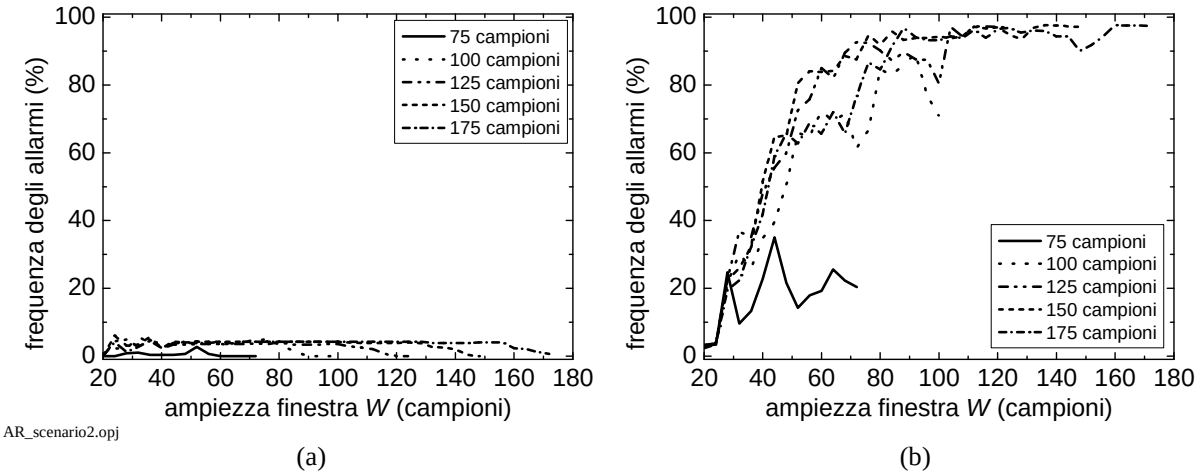
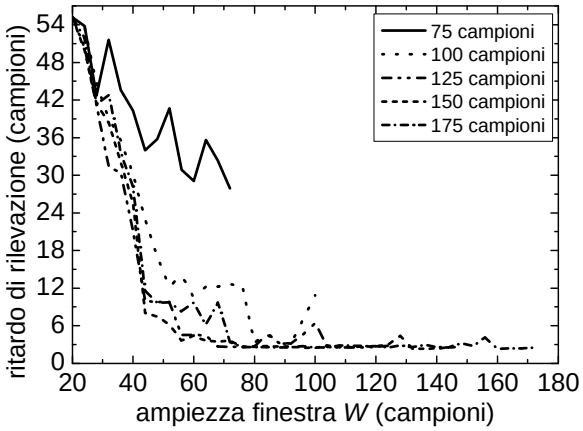


Figura 3.9. *Trasferimento di modelli con JY-PLS adattativo. Effetto del numero di campioni NOC j disponibili nello scenario 2 sulla frequenza degli allarmi nella (a) fase 1 e (b) fase 2 dell'anomalia.*

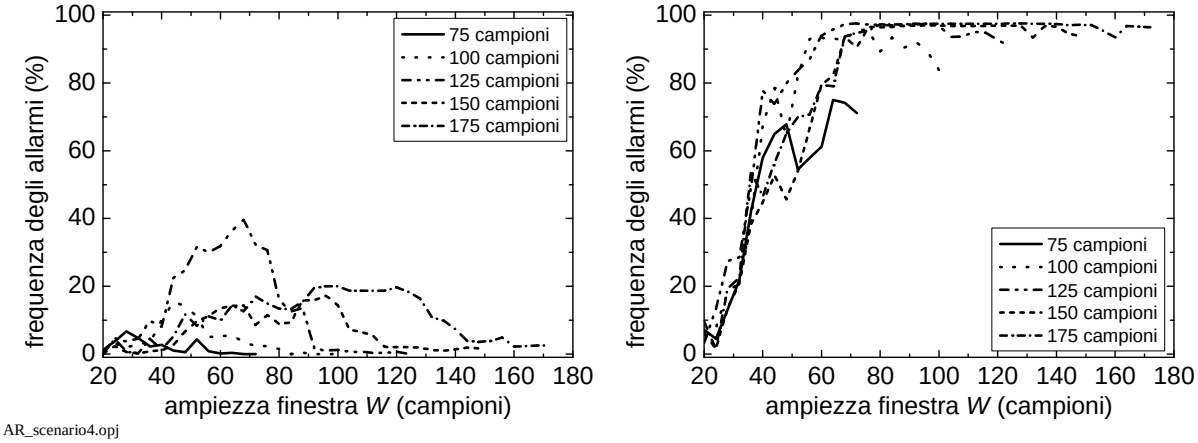


TD_scenario2.opj

Figura 3.10. *Trasferimento di modelli con JY-PLS adattativo. Effetto del numero di campioni NOC j disponibili nello scenario 2 sul ritardo di rilevazione.*

Le prestazioni del monitoraggio della fase 1 dell'anomalia sono molto buone, infatti, in riferimento alla Figura 3.9a, la frequenza degli allarmi è prossima allo 0% per ogni finestra e con qualsiasi disponibilità di campioni NOC. Le stesse considerazioni non valgono per la fase 2, infatti dalla Figura 3.9b si vede che la frequenza degli allarmi cresce con l'aumentare dell'ampiezza della finestra *W*. Il valore della diagnostica è accettabile per finestre con ampiezza di $W \geq 80$ campioni e un numero di NOC prima dell'anomalia $j > 100$. Quando si dispone di pochi dati dall'impianto B i campioni dell'anomalia vengono proiettati al di sotto dei limiti di controllo e non si generano allarmi. Dalla Figura 3.10 si vede che il ritardo di rilevazione è elevato per finestre piccole e poi decresce all'aumentare di *W* fino al valore atteso di 4 campioni quando sono disponibili più di 100 campioni NOC *j* di B, ma già ad 80 campioni il ritardo è solo di 2 campioni superiore.

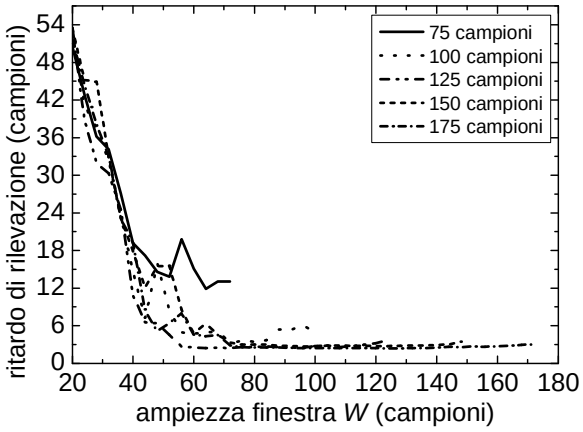
Infine, i risultati per lo scenario 4 sono mostrati nelle Figure 3.11 e 3.12.



AR_scenario4.opj

(a) (b)

Figura 3.11. *Trasferimento di modelli con JY-PLS adattativo. Effetto del numero di campioni NOC j disponibili nello scenario 4 sulla frequenza degli allarmi nella (a) fase 1 e (b) fase 2 dell'anomalia.*



TD_scenario4.opj

Figura 3.12. *Trasferimento di modelli con JY-PLS adattativo. Effetto del numero di campioni NOC j disponibili nello scenario 4 sul ritardo di rilevazione.*

La Figura 3.11a riporta la frequenza degli allarmi nella fase 1, la quale non è eccessivamente elevata, ma presenta oscillazioni attorno al 20% e peggiora se sono disponibili molti campioni NOC per poi diminuire ad ampiezze della finestra W elevate. Con pochi campioni NOC ($j \leq 75$) si ha un adattamento maggiore del modello ai dati disponibili, per cui anche la frequenza con cui gli allarmi vengono generati diminuisce. Nella Figura 3.11b, che rappresenta la frequenza degli allarmi nella fase 2, si vedono condizioni migliori per il monitoraggio con finestre di ampiezza $W \geq 60$ campioni e con disponibilità di $j > 75$ campioni NOC. L'analisi è avvalorata dall'andamento del ritardo di rilevazione, riportato in Figura 3.12. Poiché con pochi campioni il modello si adatta ai dati dell'anomalia, la frequenza di allarme generati in fase 2 è bassa e il ritardo con cui l'anomalia è rilevata risulta piuttosto elevato. In definitiva, sono necessarie finestre di ampiezza $W \geq 60$ campioni e un numero di campioni NOC $j > 75$.
Dal confronto dei risultati ottenuti per i due scenari deriva che:

- nel metodo JY-PLS, le informazioni fisiche del processo portano un miglioramento nelle prestazioni per il monitoraggio specialmente per quanto riguarda la fase 2 dell'anomalia, ovvero richiedono ampiezze di finestre minori e un minor numero di campioni NOC per rilevare con sufficiente esattezza l'anomalia;
- nella fase 1, lo scenario 2 risulta migliore in quanto si generano meno falsi allarmi, ma il ritardo con cui si rileva l'anomalia potrebbe essere eccessivo rispetto al caso dello scenario 4 a parità di ampiezza di finestra.

Non è possibile fare un confronto diretto dei risultati ottenuti per tutti gli scenari in quanto i metodi PCA modellano uno spazio intrinsecamente diverso da quello dei modelli JY-PLS. Si può comunque affermare che entrambe le tecniche sono efficaci nel trasferimento dei modelli per il monitoraggio.

3.5.3 Rilevamento di anomalie su variabili non comuni

In questo Paragrafo si presentano i risultati del monitoraggio di anomalie che compaiono su variabili che non sono comprese nei *set* comuni dei modelli PCA e JY-PLS. I dati utilizzati riguardano un'anomalia artificiale su una variabile non comune in modo da valutare l'efficienza dei metodi di trasferimento basati su JY-PLS rispetto a quelli basati su PCA per il monitoraggio del processo.

3.5.3.1 Scenari 1 e 2

I risultati dello scenario 1, applicato all'anomalia artificiale, sono riportati in Figura 3.13, dove sono rappresentati gli andamenti della frequenza degli allarmi nelle fasi 1 e 2, parametrizzati sul numero di campioni NOC *j* disponibili prima dell'anomalia.

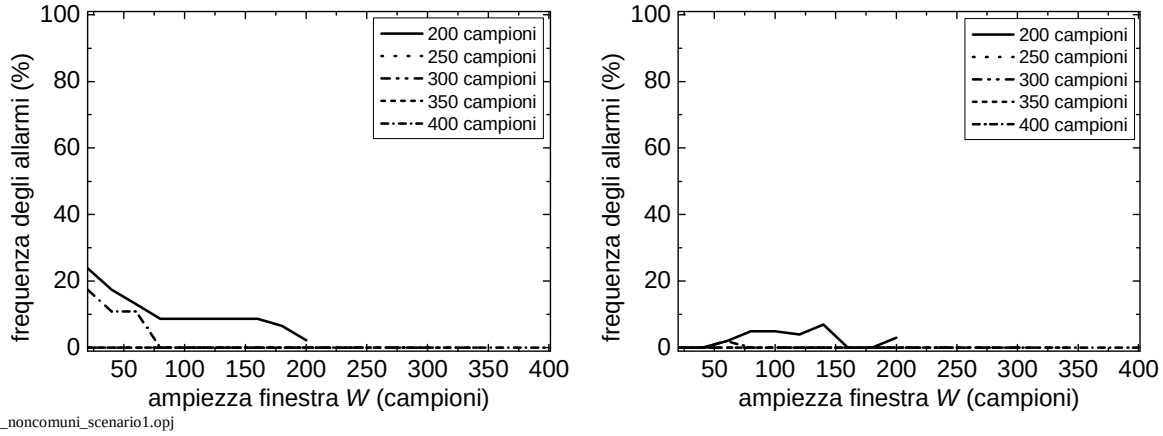


Figura 3.13. *Effetto del numero di campioni NOC j disponibili sulla frequenza degli allarmi nella (a) fase 1 e (b) fase 2 per lo scenario 1 con PCA adattativa per l'anomalia sulle variabili non comuni.*

Dalla Figura 3.13a la frequenza degli allarmi nella fase 1 è nulla se sono disponibili più di 200 campioni NOC *j*. I falsi allarmi vengono generati soprattutto per oscillazioni del SPE dei campioni iniziali proiettati al di sopra del limite, che indica una scarsa rappresentatività iniziale del modello. La Figura 3.13b mostra la frequenza degli allarmi per la fase 2, che è nulla per qualsiasi ampiezza della finestra quando dovrebbe essere prossimo al 100%. In questo caso l'anomalia non viene rilevata.

È evidente come un metodo basato su PCA adattativa, dove il modello non è costruito con le variabili coinvolte dall'anomalia, non è in grado di svolgere il monitoraggio, a meno che questo non si ripercuota sulle variabili incluse nel modello di monitoraggio mediante la loro correlazione con la variabile responsabile dell'anomalia.

Considerazioni differenti si possono fare se il metodo è basato su JY-PLS (scenario 2) in cui l'anomalia sulla variabile non comune è rilevata, come si dimostra nelle Figure 3.14 e 3.15.

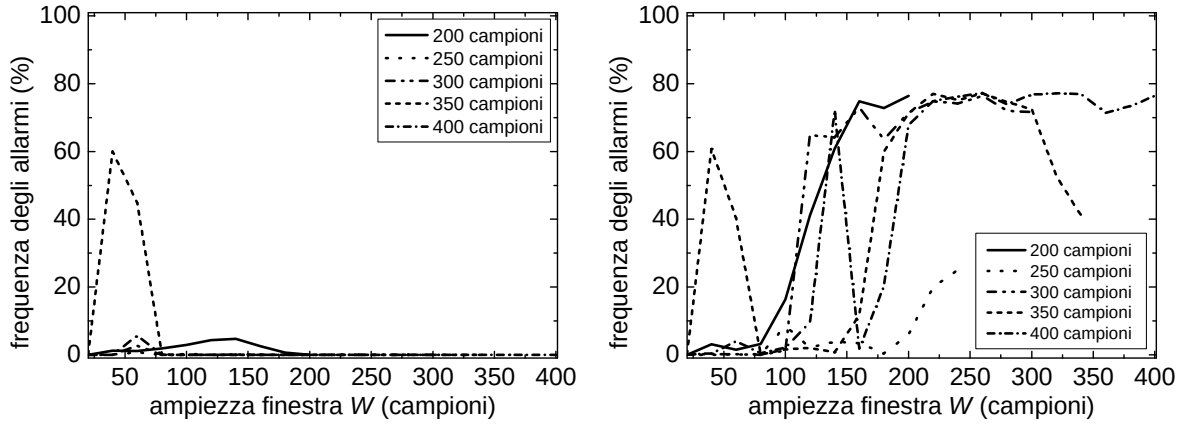


Figura 3.14. *Effetto del numero di campioni NOC j disponibili sulla frequenza degli allarmi nella (a) fase 1 e (b) fase 2 per lo scenario 2 con JY-PLS adattativo per l'anomalia sulle variabili non comuni.*

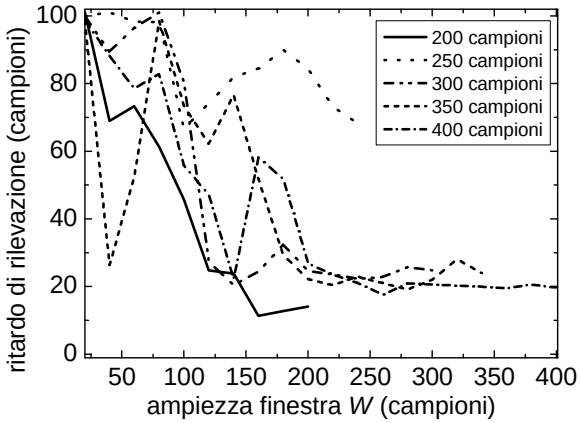


Figura 3.15. *Effetto del numero di campioni NOC j disponibili sul ritardo di rilevazione per lo scenario 2 con JY-PLS adattativo per l'anomalia sulle variabili non comuni.*

La Figura 3.14a rappresenta la frequenza degli allarmi in fase 1, che è vicino allo 0% per ampiezze della finestra $W \geq 75$ campioni, indipendentemente dal numero di campioni NOC. L'interesse maggiore è per la fase 2 dell'anomalia; dalla Figura 3.14b si può vedere che vengono segnalati allarmi senza continue oscillazioni per finestre di ampiezza $W > 200$ campioni e un numero di campioni NOC $j \geq 300$. Tuttavia la frequenza degli allarmi non è elevata. Ciò è dovuto per due motivi principali: il primo riguarda il monitoraggio per SPE, che viene svolto nel dominio delle variabili comuni (Equazioni 1.18 e 1.19), mentre l'anomalia è rappresentata sullo spazio delle variabili non comuni; il secondo è relativo alla forma a rampa dell'anomalia, che comporta un certo ritardo di rilevamento che penalizza anche la frequenza degli allarmi. Infatti, l'anomalia viene rilevata con ritardo, come si nota dalla Figura 3.15, e di conseguenza gli allarmi generati sono inferiori. Il ritardo di rilevazione diminuisce all'aumentare dell'ampiezza della finestra W e all'aumentare del numero j di

campioni NOC di B perché l'adattamento è minore e la frequenza degli allarmi della fase 2 cresce con esso. Però, in ogni caso, non raggiunge il valore stabilito dal criterio (4 campioni).
In sintesi JY-PLS è efficace nel rilevamento di anomalie che si ripercuotono su variabili non comuni, mentre PCA non è in grado di rilevare anomalie a meno che non si riflettano sulle variabili comuni.

3.5.3.2 Scenari 3 e 4

Con lo scenario 3 si considerano anche le variabili fisiche, ovvero la variabile *wtd*. Il metodo si basa su PCA per cui, nonostante l'introduzione della variabile indipendente *wtd* nel *set* comune, se essa non è definita dalle variabili coinvolte dall'anomalia, allora quest'ultima non viene rilevata. I risultati dello scenario 2, applicato all'anomalia artificiale, sono riportati in Figura 3.16 in termini di frequenza degli allarmi in fase 1 e 2.

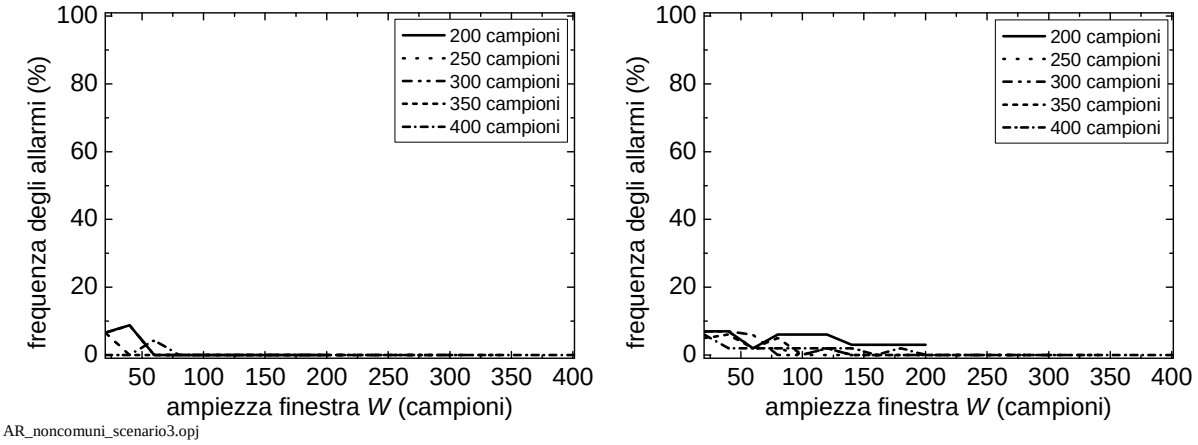


Figura 3.16. *Effetto del numero di campioni NOC j disponibili sulla frequenza degli allarmi nella (a) fase 1 e (b) fase 2 per lo scenario 3 con PCA adattativa per l'anomalia sulle variabili non comuni.*

La Figura 3.16a mostra che la frequenza degli allarmi nella fase 1 è nulla per finestre di ampiezza W > 50 campioni indipendentemente dal numero di campioni NOC *j*. Nella fase 2, in riferimento alla Figura 3.16b, la frequenza degli allarmi è prossima allo 0% in ogni condizione. È chiaro che la variabile indipendente non ha portato miglioramenti nel modello PCA rispetto al caso dello scenario 1 e l'anomalia sulle variabili non comuni continua a non venire rilevata. Questo perché la variabile responsabile dell'anomalia non influenza la *wtd*.
I risultati dell'analisi per lo scenario 4, invece, basato su JY-PLS applicato all'anomalia artificiale, sono riportati nelle Figure 3.17 e 3.18.

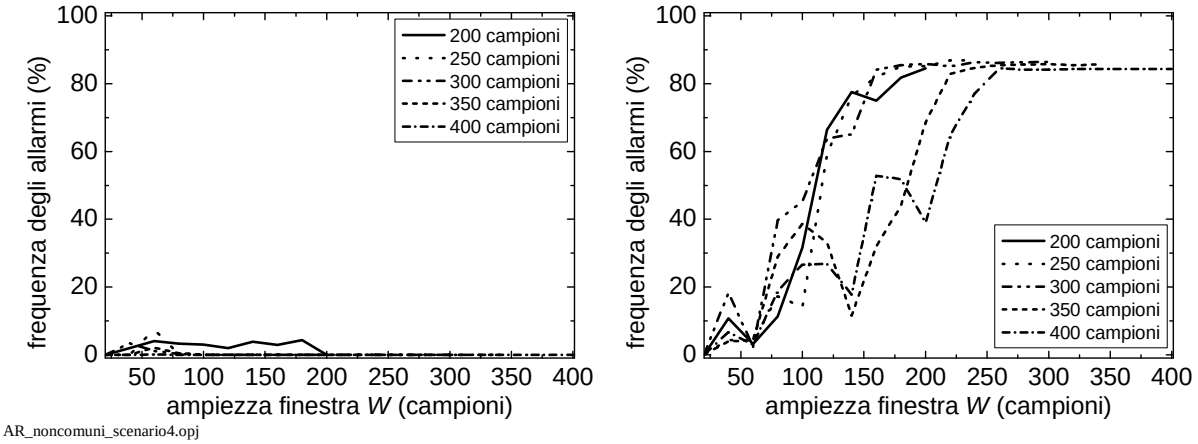


Figura 3.17. *Effetto del numero di campioni NOC j disponibili sulla frequenza degli allarmi nella (a) fase 1 e (b) fase 2 per lo scenario 4 con JY-PLS adattativo per l'anomalia sulle variabili non comuni.*

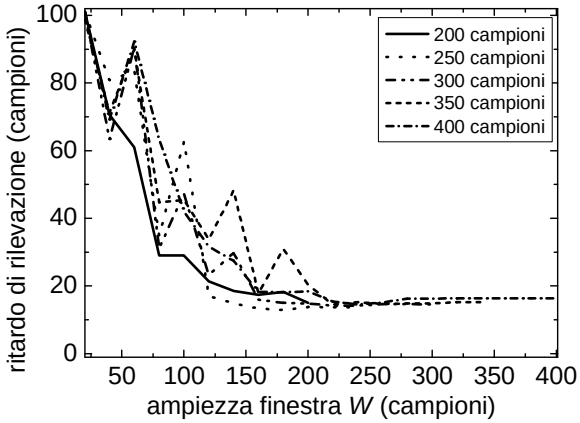


Figura 3.18. *Effetto del numero di campioni NOC j disponibili sul ritardo di rilevazione per lo scenario 4 con JY-PLS adattativo per l'anomalia sulle variabili non comuni.*

Riguardo la Figura 3.17a la frequenza degli allarmi in fase 1 è molto bassa per qualsiasi ampiezza *W* della finestra e numero di campioni NOC *j*. Per quanto riguarda la fase 2 invece, dalla Figura 3.17b si osserva che l'anomalia viene rilevata per finestre di ampiezza $W > 200$ campioni nonostante la frequenza degli allarmi sia inferiore allo scenario 2. Analogamente la Figura 3.18 rappresenta il ritardo di rilevazione che diminuisce all'aumentare dell'ampiezza *W* della finestra similmente per tutti i possibili campioni NOC disponibili, non raggiungendo il valore atteso di 4 campioni. Anche in questo caso, come per lo scenario 2, il maggior ritardo è dovuto alla natura dell'anomalia a rampa, che permette un maggiore adattamento. Rispetto al caso dello scenario 2 il ritardo è minore per il fatto che il monitoraggio avviene sul dominio delle variabili non comuni dove si sviluppa l'anomalia. Per questo motivo non è possibile avere un confronto diretto tra il metodo dello scenario 2 e quello dello scenario 4. Entrambi, rispetto ai metodi basati su PCA, sono in grado di rilevare l'anomalia seppur con un ritardo superiore al valore atteso.

3.6 Conclusioni sul trasferimento per processi continui

Dato che con pochi campioni per l'impianto B non è possibile un buon monitoraggio, con il trasferimento di modelli per il monitoraggio la richiesta di dati di B per un modello efficiente decresce notevolmente se sono disponibili dati di un impianto A con cui costruire un modello di monitoraggio, che viene trasferito all'impianto B.

In sintesi, le valutazioni conclusive generali sono riportate in Tabella 3.2 che mostra le condizioni per cui si possono avere buone prestazioni per ogni scenario nel caso in cui l'anomalia si sviluppi su variabili comuni.

Tabella 3.2. *Riassunto delle condizioni che si devono verificare per ogni scenario perché le prestazioni del modello di monitoraggio siano buone; caso di rilevamento di anomalie su variabili comuni.*

Scenario	Ampiezza *W* finestra (campioni)	Campioni NOC *j* prima dell'anomalia
1	≥ 40	> 75
2	≥ 80	> 100
3	≥ 40-60	> 75
4	≥ 60	> 75

Dalla Tabella 3.2 e dall'analisi precedente emerge che:

- il metodo PCA è più veloce ad adattarsi; infatti, sono sufficienti ampiezze della finestra *W* e un numero di campioni NOC *j* inferiori rispetto ai metodi JY-PLS. Però, in questo caso, si perdono le informazioni delle variabili non comuni, e un'anomalia che si sviluppa su queste viene rilevata solo se il suo effetto si ripercuote anche sulle variabili comuni;
- il metodo JY-PLS è più lento, ma efficace nel rilevare le anomalie sulle variabili non comuni;
- introdurre l'informazione di *wtd* sembra non portare miglioramenti nel metodo PCA, mentre rende più veloce l'adattamento per il metodo JY-PLS. Gli scenari 2 e 4 presentano condizioni simili, anche se lo scenario 4 risulta più veloce ad adattarsi rispetto allo scenario 2. Infatti, in esso diminuiscono sia l'ampiezza minima della finestra di campioni *W* che il numero di dati NOC *j* di B.

Capitolo 4

Trasferimento di un modello per il monitoraggio del processo batch per la produzione di penicillina

In questo Capitolo si riportano i risultati del trasferimento di un modello per il monitoraggio di un processo batch simulato per la produzione di penicillina. Il trasferimento viene fatto da un impianto di dimensioni inferiori e strumentato in modo diverso (A) verso un nuovo impianto di dimensioni maggiori e con una strumentazione più sofisticata (B). Le prestazioni del modello sono confrontate con quelle di un modello adattativo costruito sui soli dati del nuovo impianto, evidenziando i vantaggi apportati dal trasferimento. Infine, si analizzano i fattori che influenzano il trasferimento.

4.1 Monitoraggio in linea del processo

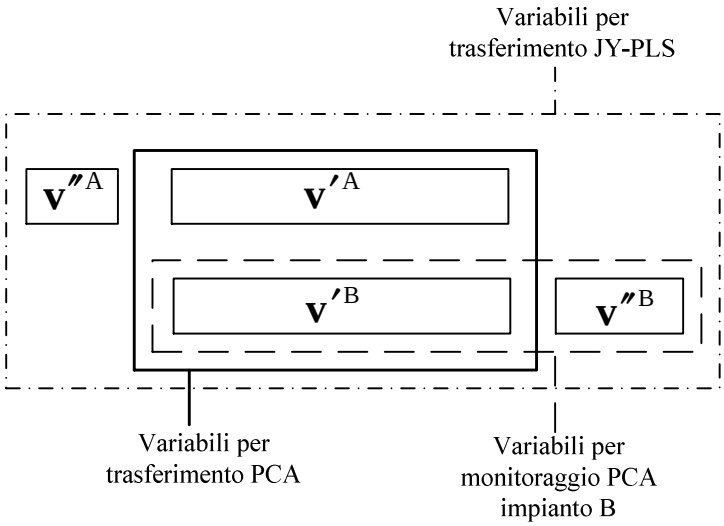
Il problema del trasferimento è molto sentito nelle produzioni batch, in cui servono molti dati di batch completati per costruire un modello di monitoraggio autonomo. Questo può comportare settimane o mesi di produzione per la loro acquisizione. Inoltre, poiché la differenziazione del prodotto è frequente, spesso non è possibile ottenere un numero di dati sufficiente per la costruzione di un modello di monitoraggio adeguato.

Il minimo numero di batch condotti nell'impianto che consenta di costruire un modello statistico che rilevi le anomalie del sistema è abitualmente molto alto. Tuttavia questo si scontra con l'esigenza di monitorare il processo su un nuovo impianto fin dagli istanti iniziali di funzionamento dell'impianto stesso. Per questo, quando un impianto è nelle fasi iniziali di funzionamento e si è nella condizione di insufficienza di dati, una possibile soluzione per monitorare l'impianto è quella di ricorrere al trasferimento dei modelli per il monitoraggio.

In questo Capitolo vengono analizzati diversi metodi per il trasferimento e ne vengono confrontate le prestazioni con quelle di modelli di monitoraggio costruiti con dati di un singolo impianto. Per la costruzione dei modelli, si considera la suddivisione delle variabili di Figura 4.1, che individua gli insiemi delle variabili comuni ($\mathbf{v}'$) e non comuni ($\mathbf{v}''$) di ogni impianto. In particolare, $\mathbf{v}'^{A}$ e $\mathbf{v}''^{A}$ sono le variabili di processo misurate nell'impianto A e $\mathbf{v}'^{B}$ e $\mathbf{v}''^{B}$ sono quelle dell'impianto B. Le variabili dell'impianto B, che è assunto essere di più

recente costruzione, vengono misurate da una strumentazione più sofisticata (per esempio si misurano molte concentrazioni).

L'insieme delle variabili comuni è definito da $\mathbf{v}'^{A} = \mathbf{v}'^{B} = \{1, 2, 3, 4, 8, 9, 10, 14, 15\}$, rispettivamente velocità di aerazione (1), potenza di agitazione (2), portata di alimentazione di substrato (3), temperatura substrato (4), concentrazione di penicillina (8), volume di coltura (9), concentrazione di CO_2 (10), portata di acido (14), portata di base (15). Le variabili non comuni, invece, sono $\mathbf{v}''^{A} = \{11, 12, 16\}$ nell'impianto A, rispettivamente pH (11), temperatura (12) e portata di acqua di raffreddamento (16), e $\mathbf{v}''^{B} = \{5, 6, 7, 13\}$ nell'impianto B, rispettivamente concentrazioni di substrato (5), di ossigeno disciolto (6), di biomassa (7) e calore generato (13)[3]. La scelta delle variabili viene fatta considerando che le variabili comuni tra impianti hanno una correlazione simile. Inoltre, l'impianto B che è più recente ha una strumentazione più complessa ed è possibile la disponibilità di dati su variabili comunemente più difficili da misurare. In seguito sarà condotta un'analisi per valutare l'effetto della scelta delle variabili.



variabili.vsd

Figura 4.1. *Suddivisione degli insiemi delle variabili di processo per il monitoraggio e il trasferimento.*

Ai fini del monitoraggio dell'impianto B, quindi, si aprono diverse alternative, secondo i seguenti scenari che utilizzano i dati menzionati sopra secondo lo schema di Figura 4.1:

- monitoraggio dell'impianto B con un modello PCA costruito sulle sole variabili $\mathbf{v}'^{B}$ e $\mathbf{v}''^{B}$, che sono misurate in questo impianto. Questo permette di stabilire il numero di batch B per costruire un modello autosufficiente e un termine di confronto con i risultati del trasferimento, al fine di valutare se esso sia opportuno;

[3] La numerazione delle variabili è riportata in Tabella 2.2 del Paragrafo 2.2.1.

- trasferimento dei modelli per il monitoraggio dall'impianto A all'impianto B mediante tecnica PCA sulle variabili comuni $\mathbf{v}'^{A}$ e $\mathbf{v}'^{B}$, per determinare l'effetto che le variabili comuni hanno sull'adattamento del modello di monitoraggio ai dati dei batch dell'impianto B;
- trasferimento dei modelli per il monitoraggio dall'impianto A all'impianto B mediante tecnica JY-PLS sulle variabili comuni $\mathbf{v}'^{A}$ e $\mathbf{v}'^{B}$ e sulle non comuni $\mathbf{v}''^{A}$ e $\mathbf{v}''^{B}$. Con questo studio si analizza il contributo che le variabili non comuni hanno nel trasferimento.

Perché il trasferimento sia possibile il processo nei due impianti deve essere guidato dalle stesse forze motrici, da cui deriva che la correlazione tra le variabili comuni deve essere la stessa tra gli impianti; questa è l'ipotesi fondamentale per il trasferimento.

Inizialmente si analizzano le prestazioni dei modelli costruiti con i soli dati di B e si definisce il numero di batch necessari per sviluppare un modello di monitoraggio autonomo. Successivamente si applica il trasferimento per dimostrare il miglioramento. Le prestazioni sono valutate in termini di frequenza degli allarmi e ritardo di rilevazione (Paragrafo 1.1.4), come per il caso continuo.

4.1.1 Monitoraggio di condizioni operative normali

L'impianto B, di costruzione più recente rispetto all'impianto A, ha pochi batch conclusi quindi anche i dati a disposizione sono pochi. Infatti, il numero di batch conclusi è molto basso. Si ritiene di avere a disposizione i dati di 4 batch dell'impianto B, che una volta sincronizzati costituiscono la matrice $\mathbf{X}^{B}\left(4\times\mathbf{v}^{B}\times 200\right)$, dove $\mathbf{v}^{B}$ sono tutte le variabili misurate nell'impianto B.

Tali variabili sono $\mathbf{v}^{B}=\mathbf{v}'^{B}\cup\mathbf{v}''^{B}=\{1,2,3,4,5,6,7,8,9,10,13,14,15\}$.

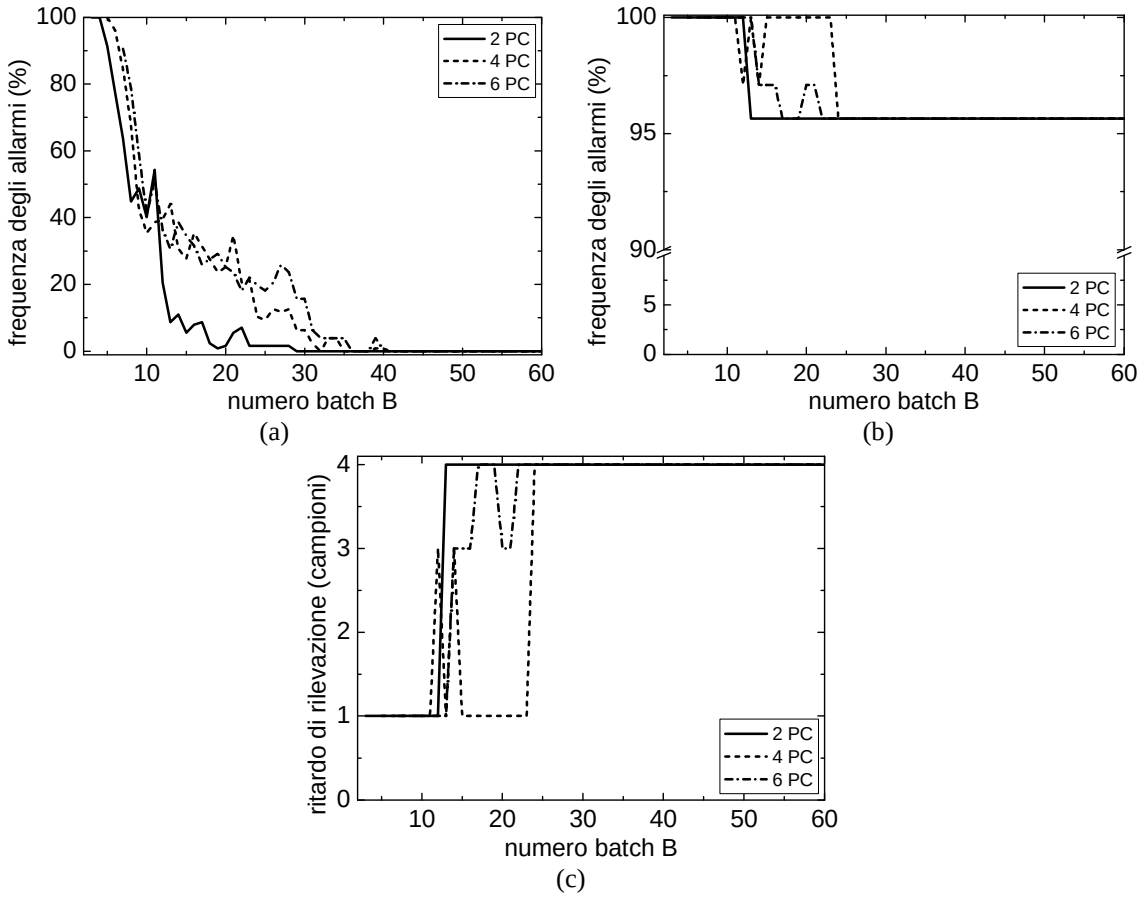
Le prestazioni del modello vengono definite testando le diverse anomalie dell'impianto B, definite al Paragrafo 2.2.2. La Tabella 4.1 riporta i valori delle diagnostiche, ricavate dal monitoraggio.

Tabella 4.1. *Diagnostiche del modello costruito su 4 batch di B con due componenti principali per le anomalie dell'impianto B.*

N° anomalia	Frequenza allarmi fase 1 (%)	Frequenza allarmi fase 2 (%)	Ritardo di rilevazione (campioni)
1	74.8	100	1
2	88.2	100	1
3	100	100	1
4	100	100	1
5	100	100	1

Qualunque sia la natura dell'anomalia, il modello non è in grado di tenere sorvegliato il batch. Infatti, la fase 1 è erroneamente considerata anomala con frequenza degli allarmi prossima a 100%. Di conseguenza, la frequenza degli allarmi della fase 2 è 100% e il ritardo di rilevazione di 1 campione indica un anticipo nella segnalazione dell'allarme perché già la fase 1 è vista come anomala. Questo succede per la scarsa rappresentatività del modello.

Quindi, si intende trovare il numero minimo di batch dell'impianto B necessari per costruire un modello di monitoraggio efficiente. Si prende come riferimento per l'analisi l'anomalia 4, relativa ad una diminuzione istantanea della potenza di agitazione del 15% a 100 h di lavoro. Considerando le variabili $\mathbf{v}^B$, si costruisce un modello PCA con i batch di B via via disponibili. Per ogni nuovo batch che viene incluso nel modello di monitoraggio si proiettano i dati dell'anomalia, valutando le diagnostiche del modello, in termini di frequenza degli allarmi e ritardo di rilevazione. La Figura 4.2 è il risultato del monitoraggio con soli dati dell'impianto B. Le curve ottenute, riportate in Figura 4.2, sono parametriche sul numero di componenti principali per il modello PCA.



PCAbatch_variePC_caso1.opj

Figura 4.2. *Risultati del monitoraggio con modello PCA costruito con i dati delle variabili* $\mathbf{v}^B$ *dell'impianto B: effetto del numero di componenti principali sulla frequenza degli allarmi in (a) fase 1, (b) fase 2 e sul (c) ritardo di rilevazione dell'anomalia.*

In particolare dalla Figura 4.2a si nota come, all'aumentare del numero di componenti principali, le prestazioni del modello peggiorino. Infatti, la frequenza degli allarmi in fase 1 si avvicina meno rapidamente allo 0%, specialmente per i primi 40 batch B disponibili. Analogamente la frequenza degli allarmi nella fase 2, riportata nella Figura 4.2b, è migliore con sole 2 componenti principali. Infatti, ciò è confermato dalla Figura 4.2c, nella quale si vede che il ritardo di rilevazione assume il valore di 4 campioni, stabilito dal criterio, con un adattamento più veloce. Da un'analisi di convalida incrociata, *RMSECV* presenta un minimo per 2 componenti principali, per cui un modello costruito con esse sarà più efficiente. Ciò conferma i risultati ottenuti.
Il modello di monitoraggio PCA, costruito solo sui dati di B e con due componenti principali, presenta buone prestazioni se si dispone di circa 20 batch di B. Si ottengono ottime prestazioni con più di 30 batch di B. Sebbene i risultati del monitoraggio sembrino buoni anche senza trasferimento, per ogni batch si devono considerare in media 175 h di funzionamento per raggiungere la concentrazione di penicillina specificata. Ciò comporta tempi lunghi per l'acquisizione dei dati. Per questo, per avere un modello di monitoraggio efficiente con un ridotto numero di batch di B, si ricorre al trasferimento.

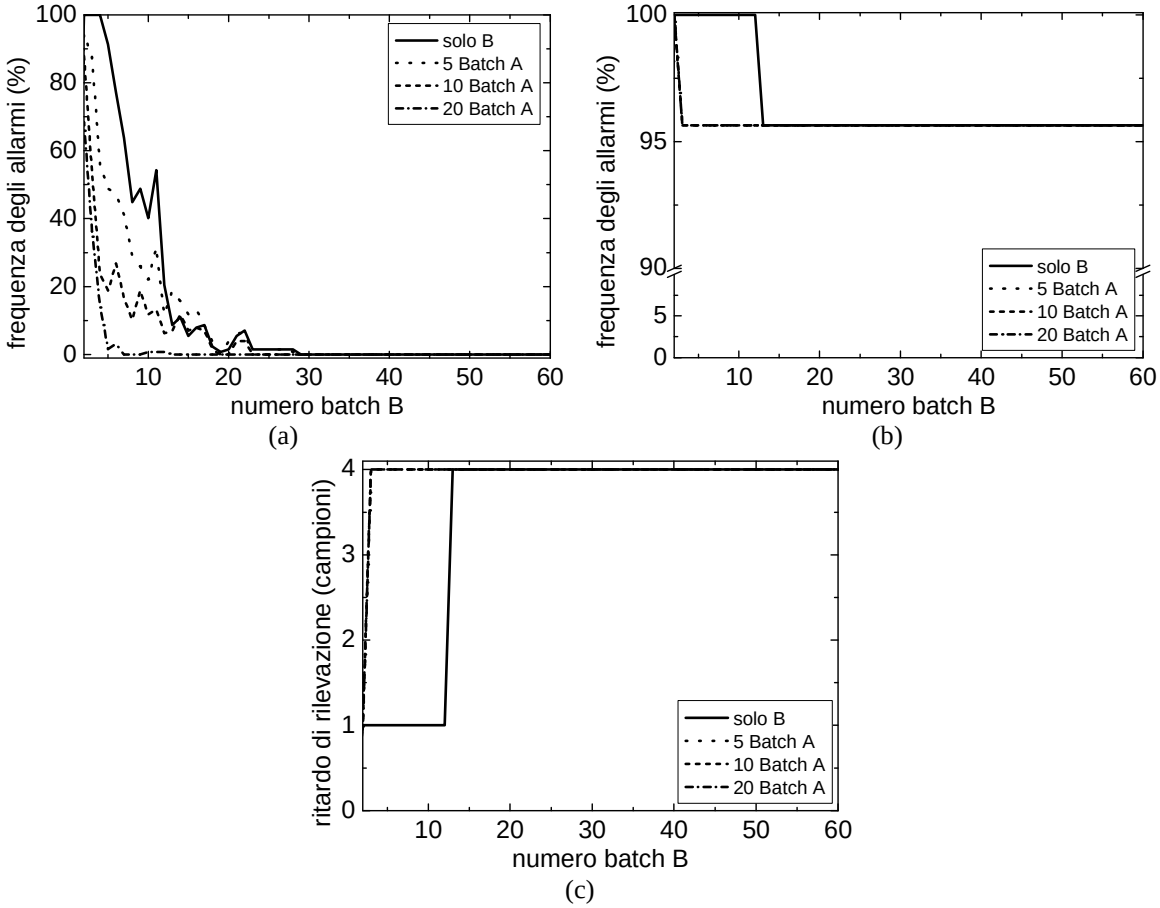
4.2 Trasferimento di modelli per il monitoraggio

Il problema del trasferimento è già stato affrontato nel Capitolo 3, per il caso di un processo continuo. In questo Capitolo viene applicato al caso batch, per il quale si assume che ogni nuovo batch disponibile dall'impianto B possa essere utilizzato nel *set* di calibrazione assieme ad un numero pre-definito di batch A già disponibili. Ad ogni adattamento del modello ai nuovi batch di B, si proiettano i dati dell'anomalia e si determinano le diagnostiche per il nuovo modello. In questo modo è possibile stabilire quanti batch di A e B sono sufficienti per il monitoraggio.
I metodi utilizzati sono PCA e JY-PLS, rispettivamente nel caso siano utilizzate per il monitoraggio solo variabili comuni o anche variabili non comuni. In entrambi i casi, i batch degli impianti A e B sono sincronizzati rispetto la concentrazione di substrato, che rappresenta la fase batch, e la concentrazione di penicillina, caratterizzante la fase fed-batch. In modo analogo si sincronizzano i dati dell'anomalia. La procedura di sincronizzazione è definita al Paragrafo 1.1.3.

4.2.1 Trasferimento con metodo PCA

Per dimostrare i vantaggi del trasferimento, introdotto al Paragrafo 4.1, si costruisce un modello PCA a 2 componenti principali utilizzando per il *set* di calibrazione i dati delle variabili comuni $\mathbf{v}'^{B} = \mathbf{v}'^{A} = \{1, 2, 3, 4, 8, 9, 10, 14, 15\}$. La Figura 4.3 riporta gli andamenti

delle diagnostiche con il numero di batch dell'impianto B in funzione del numero di batch A disponibili (5, 10 o 20).



PCA_variBatchA_caso1.opj

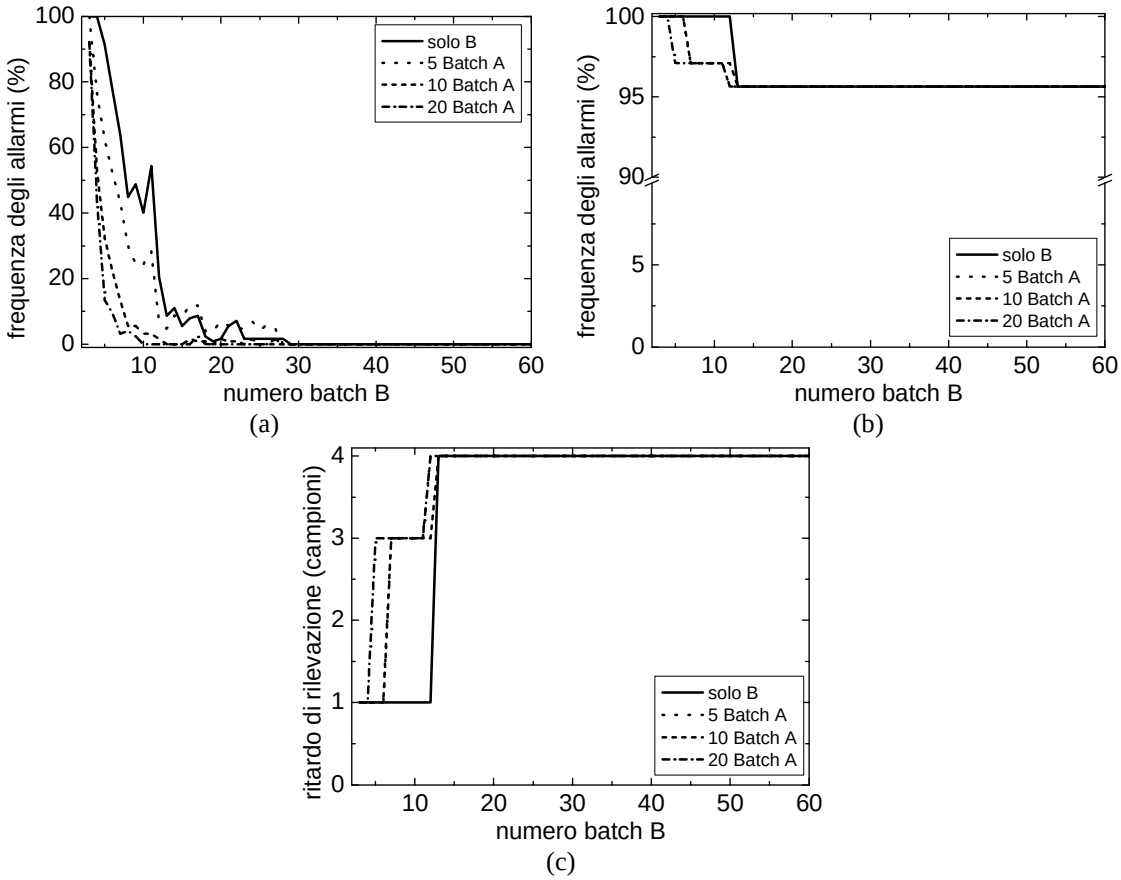
Figura 4.3. *Risultati del trasferimento di un modello PCA a 2 componenti principali costruito con le variabili* **v**'*: effetto del numero di batch A sulla frequenza degli allarmi in (a) fase 1, (b) fase 2 e sul (c) ritardo di rilevazione dell'anomalia.*

Dalla Figura 4.3 si nota come il trasferimento con il metodo PCA permetta di costruire modelli per il monitoraggio più efficienti della PCA sui soli dati di B. Infatti, la frequenza degli allarmi nella fase 1, mostrata in Figura 4.3a, raggiunge molto rapidamente il valore atteso di 0% con un numero di batch A pari a 20. Già con 10 batch di A la frequenza degli allarmi scende al di sotto del 20% con pochi batch B completati. In riferimento alla frequenza degli allarmi nella fase 2, invece, analizzando la Figura 4.3b si nota che con 3 batch dell'impianto B, indipendentemente dai batch disponibili dall'impianto A, la diagnostica assume il valore prossimo al 100%, in accordo con il criterio di segnalazione dell'anomalia. Il ritardo di rilevazione, nelle stesse condizioni, assume anch'esso il valore atteso di 4 campioni, come visibile in Figura 4.3c. La variabilità introdotta dai dati dei batch A ha un ruolo importante nel monitoraggio. Infatti, il modello costruito con essi può essere utilizzato per il monitoraggio di B con buone prestazioni. Aumentando il numero di batch A disponibili oltre

a 20, le curve ottenute sono molto simili a quelle relative al caso 20 batch A, per cui quest'ultimo rappresenta un valore limite oltre il quale i miglioramenti sono marginali per il caso di studio affrontato in questa Tesi.
In definitiva, applicando il metodo PCA per il trasferimento, per disporre di un modello efficiente, si riduce la richiesta di dati di batch B dal valore di 25 a 5 se sono disponibili 20 batch dell'impianto A.

4.2.2 Trasferimento con metodo JY-PLS

Per considerare anche le informazioni aggiuntive apportate dalle variabili non comuni, viene applicato il trasferimento con metodo JY-PLS. Per questo metodo l'insieme delle variabili comuni è dato da $\mathbf{v}'^{A}$ e $\mathbf{v}'^{B}$; per l'insieme delle variabili non comuni si ha $\mathbf{v}''^{A}$ nell'impianto A e $\mathbf{v}''^{B}$ nell'impianto B. In questo metodo, le carte di monitoraggio per SPE sono costruite sullo spazio delle variabili non comuni (SPE^{x}) e su quello delle variabili comuni (SPE^{y}). In particolare, i limiti di fiducia di queste ultime sono calcolati considerando i residui sui dati sia dell' impianto A sia dell'impianto B.
I risultati del trasferimento sono riportati in Figura 4.4.



JY_variBatchA_caso1_autosc.opj

Figura 4.4. *Risultati del trasferimento di un modello JY-PLS a 2 componenti principali: effetto del numero di batch A sulla frequenza degli allarmi in (a) fase 1, (b) fase 2 e sul (c) ritardo di rilevazione dell'anomalia.*

In Figura 4.4 si vede che anche il metodo JY-PLS permette di costruire modelli per il monitoraggio più efficienti rispetto ad un modello adattativo su soli dati di B. Infatti, la frequenza degli allarmi in fase 1 (Figura 4.4a) decresce molto velocemente e, già quando 10 batch di B sono disponibili, è prossima a 0%. In questo caso, le prestazioni sono migliori se sono disponibili più di 5 batch di A. Per quanto riguarda la fase 2, la frequenza degli allarmi presenta dei miglioramenti, seppur non rilevanti, indipendentemente dal numero di batch di A. Analogamente, il ritardo di rilevazione (Figura 4.4c) si avvicina al valore atteso di 4 campioni.

Per tutta l'analisi del processo batch, le matrici di dati, per il metodo JY-PLS, sono state trattate mediante *autoscaling*. Seguendo il metodo proposto da García-Muñoz *et al.* (2005) per il pretrattamento di dati nella costruzione di modelli JY-PLS, quando le matrici dei dati vengono autoscalate devono anche essere divise per la traccia della matrice di varianza-covarianza della matrice stessa. Questo equivale a dividere per il numero di elementi della matrice, per cui aumentando il numero di batch A, la variabilità dei dati diminuisce, al punto che risulta troppo piccola per portare nuove informazioni per il modello costruito sui nuovi dati.

In sintesi, per il trasferimento vale che:

- il metodo PCA è più veloce ad adattarsi. Infatti, perché il modello di monitoraggio sia efficiente sono richiesti 5 batch B avendo disponibili più di 10 batch A;
- il metodo JY-PLS generalmente è più lento rispetto al trasferimento con PCA. Esso richiede almeno 10 batch di B con più di 5 batch di A disponibili, ma in ogni caso il modello per il monitoraggio è efficiente e il trasferimento è conveniente rispetto al modello adattativo sui soli dati di B;
- per il metodo JY-PLS si esegue un trattamento delle matrici secondo *autoscaling* per poter considerare la variabilità introdotta da molti dati.

4.3 Anomalie su variabili non comuni

Quando si presenta un'anomalia su una variabile non comune il modello PCA non è in grado di rilevarla, a meno che essa non si ripercuota sulle variabili comuni. In questo caso i dati per l'anomalia ottenuti dalla simulazione sono tali da soddisfare questa condizione, per cui l'anomalia interessa una variabile non comune e non altera la correlazione tra le variabili comuni. Il modello PCA adattativo sui soli dati dei batch B è in grado di rilevare l'anomalia. Infatti, esso è costruito sia con le variabili comuni sia con quelle non comuni dei batch B. Differentemente, applicando il trasferimento, se si utilizza il metodo PCA, l'anomalia non viene rilevata. Infatti, il modello è costruito considerando solo le variabili comuni; non include quindi la variabile dell'anomalia ed essa non si ripercuote sulle variabili incluse nel

modello. Tali considerazioni si notano dalla Figura 4.5, in cui sono rappresentate le frequenze degli allarmi in fase 1 e 2.

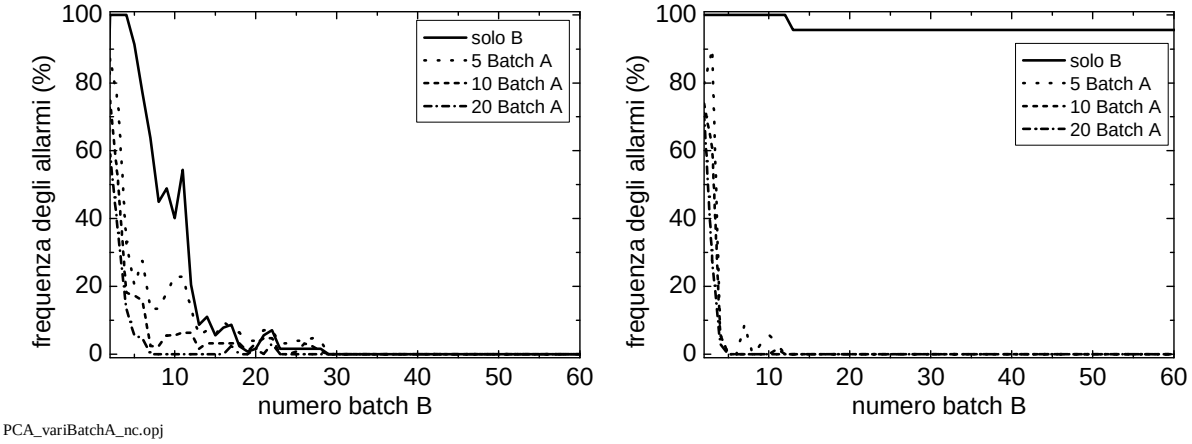
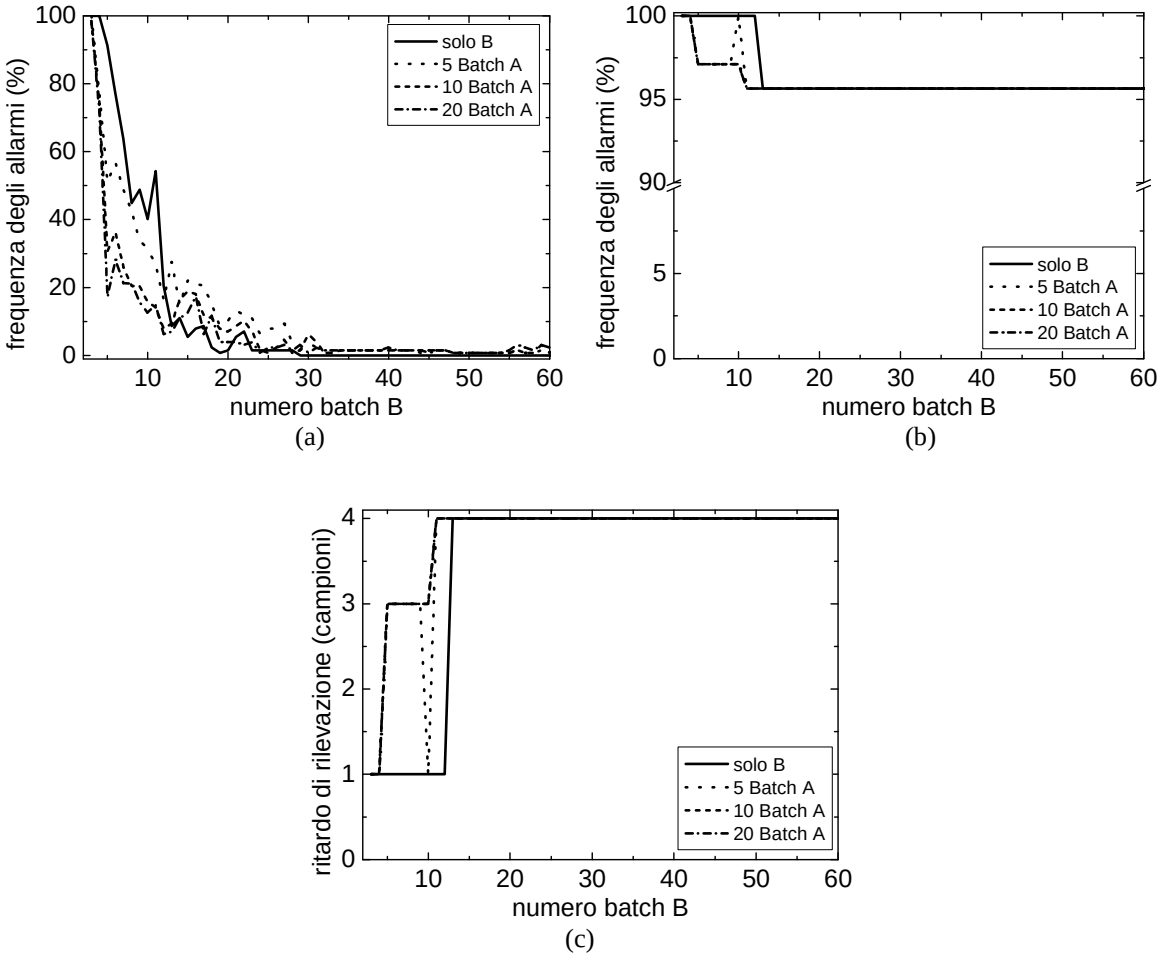


Figura 4.5. *Risultati del trasferimento di un modello PCA a 2 componenti principali in cui l'anomalia è in una variabile non comune: effetto del numero di batch A sulla frequenza degli allarmi in (a) fase 1, (b) fase 2.*

La situazione per la fase 1 (Figura 4.5a) è molto simile al caso in cui la variabile su cui si sviluppa l'anomalia è nell'insieme comune. Però, analizzando la fase 2 (Figura 4.5b), si nota che non vengono generati allarmi e l'anomalia non è rilevata, indipendentemente dal numero di batch A o di B. Il metodo PCA non è in grado di rilevare anomalie che si sviluppano esclusivamente su variabili non comuni.

Invece, se come metodo di trasferimento basato su JY-PLS, la stessa anomalia viene rilevata. Esso sviluppa un modello sulle variabili comuni e non comuni e il monitoraggio avviene anche sugli SPE definiti nell'insieme delle variabili non comuni. La Figura 4.6 riporta i risultati del trasferimento.



JY-PLS_variBatchA_nc.opj

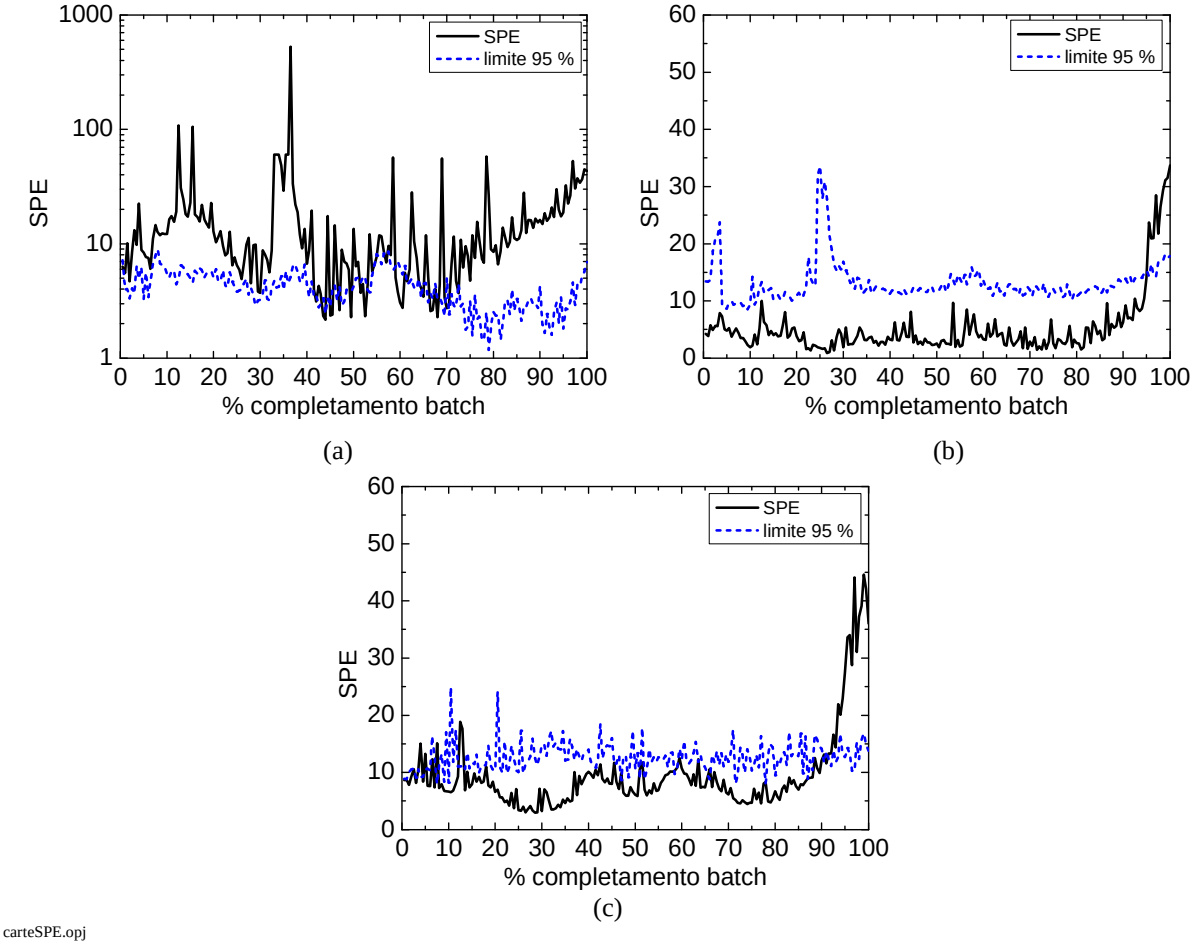
Figura 4.6. *Risultati del trasferimento di un modello JY-PLS a 2 componenti principali in cui l'anomalia avviene su una variabile non comune: effetto del numero di batch A sulla frequenza degli allarmi in (a) fase 1, (b) fase 2 e sul (c) ritardo di rilevazione dell'anomalia.*

La Figura 4.6 mostra che il trasferimento con il metodo JY-PLS permette di costruire modelli per il monitoraggio molto efficienti. In particolare, dalla Figura 4.6a si può affermare che la frequenza degli allarmi in fase 1 è più bassa se sono disponibili almeno 10 batch A. Il miglioramento apportato dal trasferimento, inoltre, è visibile per i primi 15 batch B disponibili. Successivamente un modello costruito con soli dati di B risulta più efficiente. Analizzando la fase 2 dell'anomalia nella Figura 4.6b si vede che, diversamente dal metodo PCA, l'anomalia viene rilevata e la frequenza degli allarmi è elevata. Ciò è confermato dalla Figura 4.6c, in cui il ritardo di rilevazione assume il valore atteso di 4 campioni, adattandosi più velocemente rispetto al modello definito da soli dati di B.

Generalmente si può dire che il trasferimento con metodo JY-PLS permette di definire modelli che rilevino le anomalie che si sviluppano su variabili non comuni, poiché le informazioni di queste variabili sono considerate per la costruzione del modello stesso.

4.4 Considerazioni generali su diverse anomalie

L'analisi sviluppata nei Paragrafi precedenti viene generalizzata per diversi tipi di anomalie. In particolare, si analizza l'anomalia 2, relativa ad una diminuzione a rampa della velocità di aerazione. In questo caso un modello PCA, costruito con due componenti principali sui soli dati di batch B, è efficiente se sono disponibili 50 batch B. Però la natura dell'anomalia genera elevati ritardi di rilevazione e, di conseguenza, la frequenza degli allarmi in fase 2 è molto bassa. Se si applica il trasferimento a questo caso, sia con metodo PCA che con JY-PLS, il modello per il monitoraggio è efficiente anche con soli 10 batch B se sono disponibili 20 batch A. Questi effetti si possono vedere analizzando le carte di monitoraggio per SPE, riportate in Figura 4.7. Si consideri che l'anomalia si presenta quando il batch è completato per il 67% e a questo punto inizia la fase 2 dell'anomalia.



carteSPE.opj

Figura 4.7. *Carte di monitoraggio per SPE nel caso di monitoraggio dell'impianto B con un modello costruito su (a) 5 batch B, (b) 50 batch B e (c) 10 batch B e 20 batch A (caso trasferimento).*

La Figura 4.7a mostra come un modello costruito con pochi batch di B non sia efficiente in quanto si realizzano numerosi falsi allarmi per la fase 1 e la frequenza degli allarmi è elevata. Aumentando il numero di batch di B, la Figura 4.7b rappresenta la carta di monitoraggio di

SPE nel caso in cui siano disponibili 50 batch B con cui costruire un modello PCA. In questo caso, la fase 1 è ben rappresentata poiché non si individuano punti al di sopra del limite che genererebbero falsi allarmi. L'anomalia viene segnalata con un ritardo di 63 campioni, infatti gli SPE eccedono il limite solo al 95% di completamento del batch anomalo. Il ritardo è dovuto alla natura dell'anomalia, che è una rampa di pendenza ridotta (-0.5). Infine, se si considera il trasferimento a partire da 10 batch B e 20 batch A, le prestazioni del modello migliorano. Infatti, la Figura 4.7c risulta molto simile alla situazione di Figura 4.7a per cui la fase 1 è ben rappresentata senza la generazione di falsi allarmi e l'anomalia si manifesta a percentuali di completamento del batch prossime a 95%.
Sviluppando l'analisi per le anomalie simulate, le prestazioni dei modelli ottenuti dal trasferimento sono riportate in Tabella 4.2. In particolare, nella Tabella 4.1 sono stati analizzati modelli costruiti con 4 batch di B. Per poter confrontare i risultati del trasferimento, però, si devono considerare le prestazioni di un modello costruito su dati di soli 10 batch B, per cu si fa riferimento alla Tabella 4.3.

Tabella 4.2. *Frequenza degli allarmi e ritardo di rilevazione dopo il trasferimento per le diverse anomalie.*

	Trasferimento PCA			**Trasferimento JY-PLS**		
Anomalia	**AR fase 1 (%)**	**AR fase 2 (%)**	**TD (campioni)**	**AR fase 1 (%)**	**AR fase 2 (%)**	**TD (campioni)**
1	0	97.1	3	0	95.7	4
2	0	23.1	54	3.1	15.9	59
3	14.1	76.8	1	14.1	24.6	53
4	0.7	95.7	4	0	97.1	3
5	5.5	0	200	3.9	1.4	64

Tabella 4.3. *Frequenza degli allarmi e ritardo di rilevazione per un modello costruito con soli dati di 10 batch di B per le diverse anomalie.*

Anomalia	**AR fase 1 (%)**	**AR fase 2 (%)**	**TD (campioni)**
1	1.6	95.7	4
2	18.1	37.7	44
3	48.8	42.0	1
4	40.1	100	1
5	26.8	23.2	30

Dal confronto dei risultati di Tabella 4.2 e di Tabella 4.3, si possono trarre delle considerazioni generali:

- per tutti i casi il trasferimento migliora le prestazioni del modello di monitoraggio rispetto al modello sui soli dati di B, sia che si utilizzi un metodo PCA che un metodo JY-PLS. Ciò è visibile confrontando i risultati del trasferimento rispetto a quelli ottenuti dall'analisi

con soli 10 batch di B (Tabella 4.3), in cui la frequenza degli allarmi per la fase 1 è elevata;

- generalmente, il metodo di trasferimento PCA è più veloce ad adattarsi ai nuovi dati, ma in questo caso le prestazioni, sia per il modello PCA che per quello JY-PLS, sono abbastanza confrontabili;
- le anomalie 1 e 4 (step) sono rilevate in modo efficace da entrambi i modelli, PCA e JY-PLS, mentre le altre anomalie (rampe) vengono rilevate con ritardo anche se vi è il trasferimento. Ciò conferma i risultati dell'analisi precedentemente condotta sull'anomalia 2;
- considerando la natura dell'anomalia, se essa è caratterizzata da una forma a rampa, anomalie con rampe di pendenza maggiore presentano ritardi di rilevazione minori. Per tutti i casi, però, il trasferimento porta miglioramenti e quindi il comportamento della fase 2 dipende solo dalla natura dell'anomalia e non dal metodo con cui si definisce il modello per il monitoraggio.

4.5 Effetto delle variabili impiegate sul modello

Al fine di avere una valutazione più obiettiva sui risultati del trasferimento, si è pensato di cambiare la ripartizione tra variabili comuni e non comuni per mostrare che i risultati non sono dipendenti dal *set* di dati utilizzato. Modificando le variabili comuni e non comuni tra i due impianti, le prestazioni del modello per il monitoraggio sono differenti, ma i vantaggi apportati dal trasferimento sono mantenuti, come mostrato nella seguente analisi.

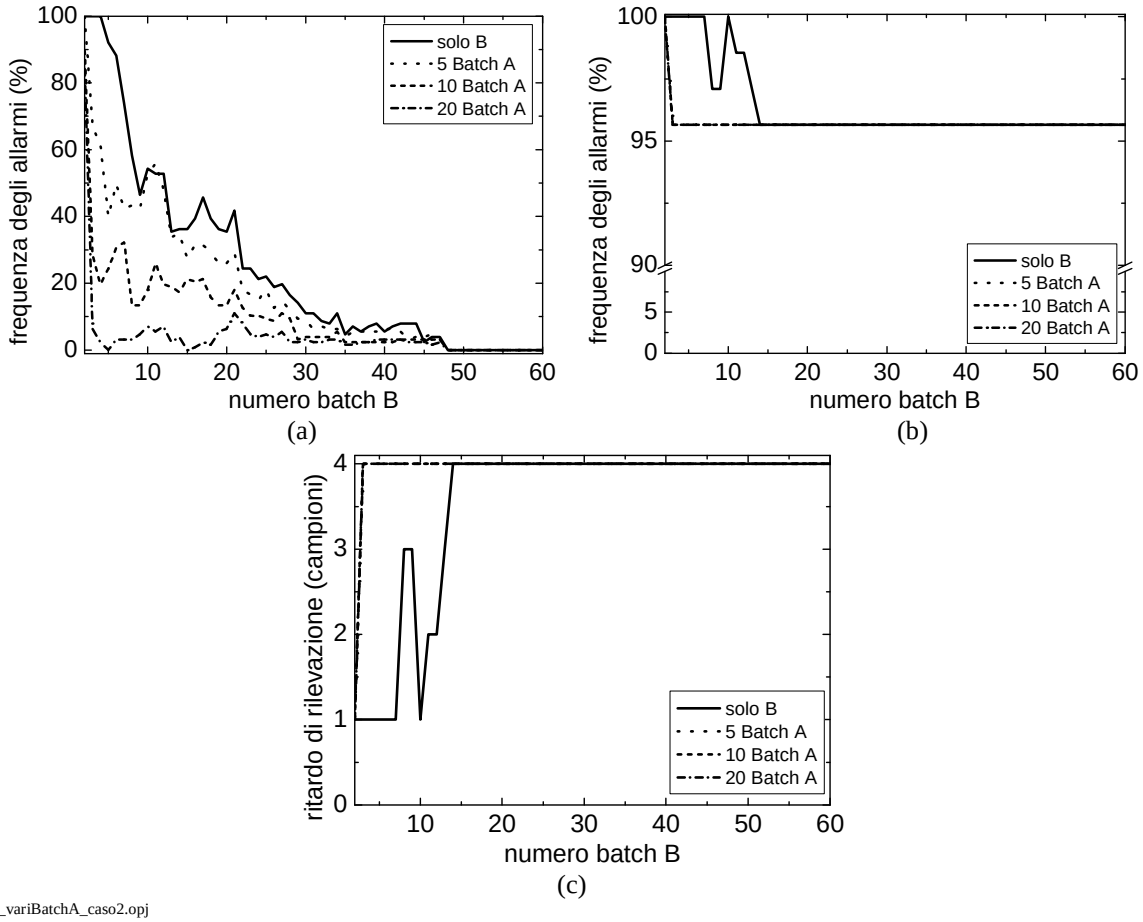
4.5.1 Variabili comuni in cui compare la temperatura del reattore

La temperatura del reattore è una variabile soggetta al sistema di controllo e il suo valore deve rimanere costante per tutta la durata del batch. In realtà, poiché il regolatore è di tipo PID, essa presenta delle piccole oscillazioni attorno al valore medio. Costruendo un modello PCA sui soli dati dei batch B, in cui le variabili sono $\mathbf{v}^{B} = \{1, 2, 3, 5, 6, 7, 8, 9, 12, 13, 14, 15\}$, le condizioni per definire un modello di monitoraggio efficiente sono determinabili dall'analisi delle diagnostiche al variare del numero di batch B. Dall'analisi deriva che il modello con prestazioni migliori è costruito con due componenti principali e in questo caso sono necessari 40/50 batch B perché sia autosufficiente. La richiesta di dati è elevata, per cui è necessario il trasferimento.

4.5.1.1 Trasferimento PCA

Applicando il metodo PCA per il trasferimento del modello, si considera l'insieme delle variabili comuni $\mathbf{v}' = \{1, 2, 3, 8, 9, 12, 14, 15\}$. Le prestazioni nel monitoraggio migliorano,

come si può vedere dalla Figura 4.8, in cui sono rappresentate le diagnostiche del modello per il monitoraggio.



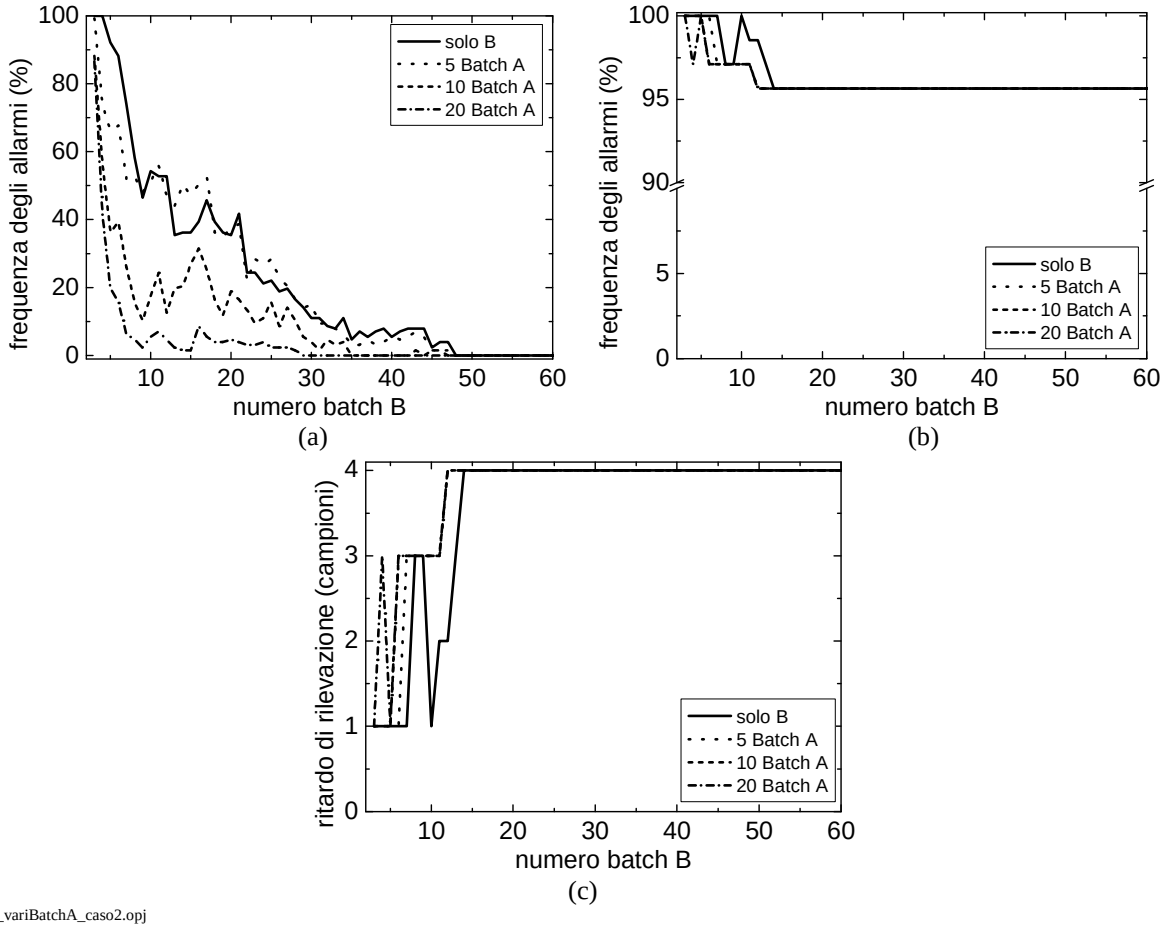
PCA_variBatchA_caso2.opj

Figura 4.8. *Risultati del trasferimento di un modello PCA a 2 componenti principali costruito con le nuove variabili* **v***': effetto del numero di batch A sulla frequenza degli allarmi in (a) fase 1, (b) fase 2 e sul (c) ritardo di rilevazione dell'anomalia.*

La Figura 4.8a mostra come all'aumentare del numero di batch A disponibili, e in particolare con più di 10/20 batch A, la frequenza degli allarmi in fase 1 decresce rapidamente tanto che sono sufficienti 5 batch B e 20 di A perché il modello sia efficiente. Tale considerazione può essere estesa alla fase 2, per la quale con 5 batch B la frequenza degli allarmi (Figura 4.8b) e il ritardo di rilevazione (Figura 4.8c) assumono i valori attesi in accordo con il criterio di generazione degli allarmi.

4.5.1.2 Trasferimento JY-PLS

Quando si considerano anche variabili non comuni si usa il metodo per il trasferimento JY-PLS. Si definiscono gli insiemi delle variabili comuni $\mathbf{v}'$, analogo al metodo PCA, e di quelle non comuni $\mathbf{v}''^{\mathrm{A}} = \{4,10,11,16\}$ per l'impianto A e $\mathbf{v}''^{\mathrm{B}} = \{5,6,7,13\}$ per l'impianto B. Per questo caso, i risultati del trasferimento sono riportati in Figura 4.9.



JY_variBatchA_caso2.opj

Figura 4.9. *Risultati del trasferimento di un modello JY-PLS a 2 componenti principali costruito con le nuove variabili: effetto del numero di batch A sulla frequenza degli allarmi in (a) fase 1, (b) fase 2 e sul (c) ritardo di rilevazione dell'anomalia.*

Anche per questo metodo, il trasferimento migliora le prestazioni rispetto al modello per il monitoraggio su soli B. Infatti, la frequenza degli allarmi in fase 1 (Figura 4.9a) assume valori prossimi allo 0%, se sono disponibili più di 10 batch sia dall'impianto A che dall'impianto B. Analoghe considerazioni si possono fare per la frequenza degli allarmi in fase 2 (Figura 4.9b) e il ritardo di rilevazione (Figura 4.9c).

In definitiva:

- modificando le variabili dell'insieme comune, le prestazioni del modello sono differenti, ma i miglioramenti apportati dal trasferimento sono analoghi;
- per il caso particolare, il metodo PCA è più veloce e richiede 5 batch B con 20 batch A disponibili, mentre JY-PLS è più lento ad adattarsi necessitando di 10 batch B a parità di batch A disponibili;
- inserendo la temperatura del reattore, è richiesto un maggior numero di batch B perché il modello costruito solo con essi sia efficiente. Infatti, la temperatura è una variabile controllata che porta una variabilità che il metodo non è in grado di rappresentare. Questo

si riflette su SPE che ha un andamento oscillante e fa aumentare la frequenza degli allarmi in fase 1.

4.5.2 Effetto del pH

Il pH del reattore, come pure la temperatura, è una variabile controllata. Nei due impianti i sistemi di regolazione del pH sono differenti: di tipo on-off per l'impianto A, di tipo PID per l'impianto B. Di conseguenza, i profili delle variabili controllate (pH) e manipolate (portate di acido e base) sono diversi tra i due impianti, come si mostra in Figura 4.10, nella quale sono riportati dei generici profili del pH per il primo batch simulato in entrambi gli impianti. Tali andamenti sono analoghi per tutti i batch.

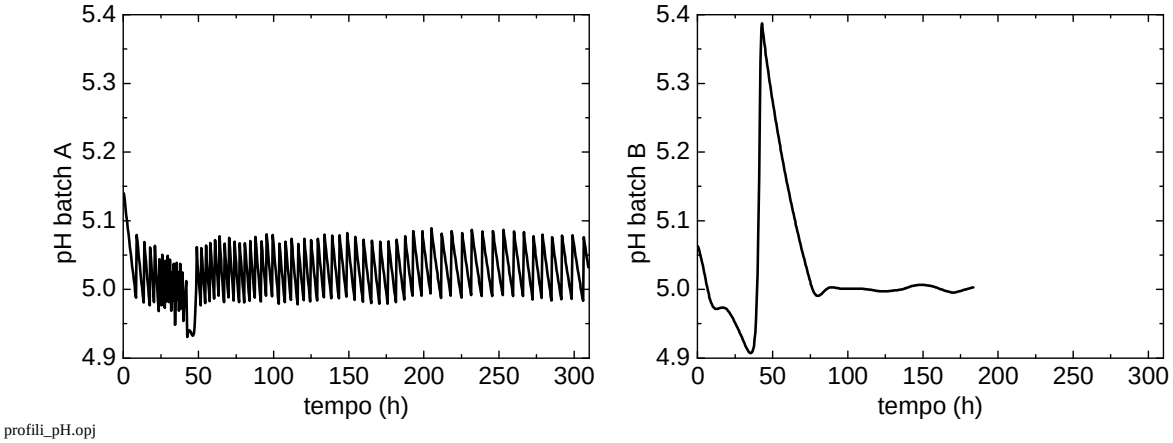


Figura 4.10. *Profilo del pH nel primo batch di (a) impianto A e (b) impianto B.*

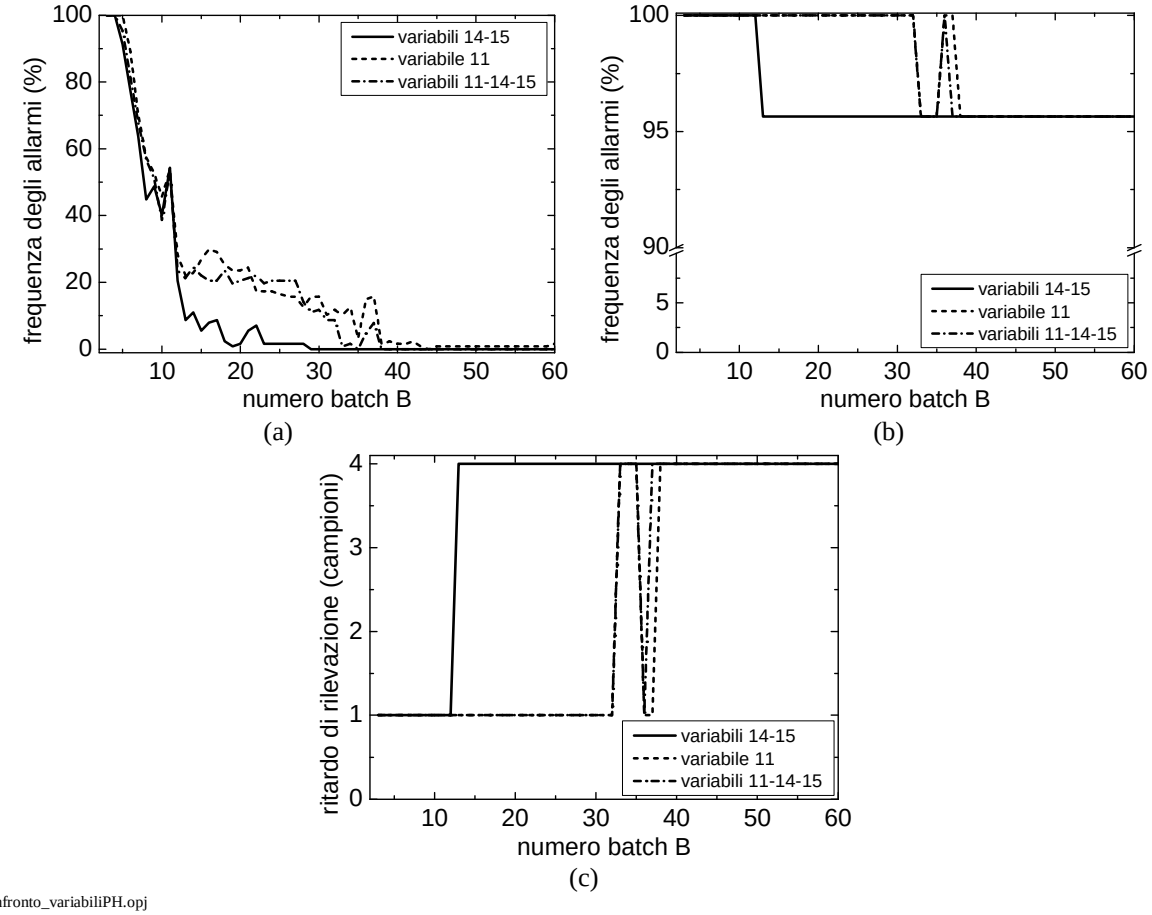
I profili del pH rappresentati in Figura 4.10a, relativo al batch A, e in Figura 4.10b, relativo al batch B, sono molto diversi. Infatti, nonostante per entrambi si realizzi la discontinuità in prossimità del momento di passaggio dalla fase batch alla fase fed-batch, il profilo del pH è diverso. Questo è dovuto alle caratteristiche del regolatore. Di conseguenza, la scelta delle variabili del sistema di controllo del pH da includere per la costruzione del modello per il monitoraggio ne influenza le prestazioni.

L'analisi condotta permette di stabilire l'effetto della scelta delle variabili legate al controllo del pH sul modello per il monitoraggio. Inizialmente si sono valutate le prestazioni del modello nel caso in cui si dispongano di soli dati di batch B e successivamente si è introdotto il trasferimento. In ogni caso, si considerano tre possibilità di variabili dell'insieme comune che rappresentano il pH, da cui derivano i tre casi:

1. includere solo le variabili manipolate dal regolatore 14 (portata di acido) e 15 (portata di base);
2. includere solo la variabile del pH (11);
3. includere tutte le variabili controllate (11) e manipolate (14, 15).

Per ogni caso, il modello costruito sia con PCA sia con JY-PLS è un modello a due componenti principali, in quanto, se queste aumentano, le prestazioni del modello peggiorano. Le variabili degli insiemi comuni sono $\mathbf{v}'^{A} = \mathbf{v}'^{B} = \{1, 2, 3, 4, 8, 9, 10,$ variabili del pH$\}$ e le non comuni $\mathbf{v}''^{A} = \{11, 12, 16\}$ e $\mathbf{v}''^{B} = \{5, 6, 7, 13\}$, in cui le variabili che rappresentano il pH sono diverse a seconda del caso considerato.

Una prima analisi riguarda il confronto delle prestazioni di modelli PCA costruiti con i soli dati dei batch B per i tre casi. Le variabili considerate sono quelle dell'insieme $\mathbf{v}^{B}$ che include tutte le variabili misurate nell'impianto B. I risultati del monitoraggio all'aumentare del numero di batch B disponibili sono visibili in Figura 4.11 per i tre casi.



confronto_variabiliPH.opj

Figura 4.11. *Risultati del monitoraggio con modello PCA a 2 componenti principali: effetto delle variabili pH (11), portata di acido (14) e portata di base (15) sulla frequenza degli allarmi in (a) fase 1, (b) fase 2 e sul (c) ritardo di rilevazione dell'anomalia.*

Dalla Figura 4.11, si nota come la scelta di includere o no le variabili 11, 14, 15 modifichi le prestazioni del modello di monitoraggio. In particolare, la Figura 4.11a mostra come la frequenza degli allarmi in fase 1 è minore se nel modello non si considera la variabile pH del reattore. Considerazioni analoghe si possono fare per la frequenza degli allarmi in fase 2 (Figura 4.11b) e per il ritardo di rilevazione (Figura 4.11c). Generalmente, la variabile pH del reattore peggiora le prestazioni del modello e lo rende più lento ad adattarsi ai dati,

aumentando la richiesta di batch B per un modello autosufficiente. Anche se si considerano le tre variabili (11, 14, 15) per lo stesso insieme, non si raggiungono le prestazioni che si hanno per un modello che abbia le sole variabili delle portate di acido e base.

Il comportamento è dovuto alla diversa importanza delle variabili sul modello. Infatti, analizzando i *loading* per la prima e seconda componente principale dei due metodi PCA, tutte le variabili (11, 14, 15) sono rappresentate dalla prima componente principale. Le variabili 14 e 15 hanno peso maggiore; infatti, hanno una maggiore variabilità rispetto la variabile 11. Ciò spiega il peggioramento delle prestazioni del modello costruito considerando la sola variabile 11, ma non è sufficiente per il caso in cui il modello consideri tutte le tre variabili in questione (11, 14, 15).

In questo caso, le prestazioni non migliorano a causa dell'utilizzo di tutte le variabili del sistema controllo, per cui il metodo deve modellare anche il sistema di controllo, oltre alla variabilità dei dati di processo. Ciò rende il modello più lento ad adattarsi ai nuovi dati.

In definitiva:

- considerare solo le variabili manipolate nell'insieme delle variabili comuni è più conveniente; infatti, qualora venga modellata la variabilità delle variabili manipolate e controllate, includere queste ultime non sempre porta miglioramenti al trasferimento del modello di monitoraggio;
- il trasferimento con metodo PCA, applicato per ogni caso, riduce il numero di batch B richiesti all'aumentare dei batch A disponibili;
- il trasferimento con metodo JY-PLS, sebbene più lento, migliora comunque le prestazioni del modello, all'aumentare dei batch A;

Per giustificare le considerazioni sul trasferimento includendo la variabile pH, in Figura 4.12 si riportano gli andamenti delle diagnostiche con il numero di batch di B per il caso PCA.

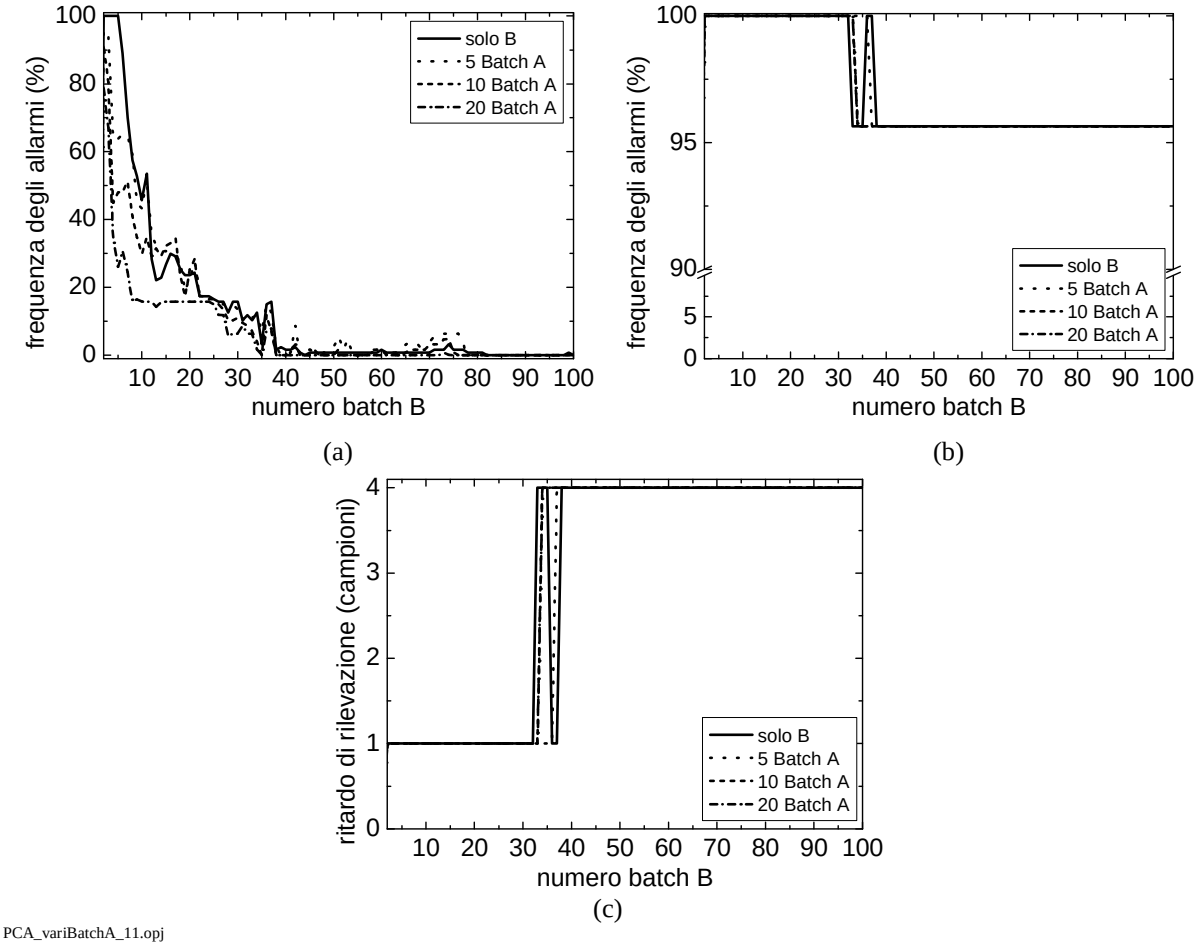


Figura 4.12. *Risultati del trasferimento di un modello PCA a 2 componenti principali costruito con la variabile 11 del pH: effetto del numero di batch A sulla frequenza degli allarmi in (a) fase 1, (b) fase 2 e sul (c) ritardo di rilevazione dell'anomalia.*

La Figura 4.12 mostra come il trasferimento con metodo PCA migliori le prestazioni del modello per il monitoraggio, nonostante il miglioramento non sia elevato. Infatti, dalla Figura 4.12a si vede che la frequenza degli allarmi della fase 1 diminuisce velocemente con 20 batch A nei primi batch di B, ma le prestazioni ottimali si hanno con molti batch di B. Analogamente, la frequenza degli allarmi nella fase 2 (Figura 4.12b) e il ritardo di rilevazione (Figura 4.12c) sono migliori se sono disponibili 20 batch A, ma il vantaggio non è eccessivo. Questo conferma che la variabile pH peggiora le prestazioni del modello di monitoraggio e anche gli effetti positivi del trasferimento sono smorzati.

4.6 Conclusioni sul trasferimento per il processo di produzione di penicillina

L'analisi condotta in questo Capitolo mostra che il trasferimento riduce la richiesta di dati per ottenere un modello di monitoraggio efficiente per un impianto in cui il numero di batch conclusi è basso e si hanno pochi dati di processo, qualora siano disponibili dati da un

impianto simile al primo che abbiano la struttura di correlazione simile. Dunque, anche per il caso di processo batch, conviene applicare il trasferimento negli istanti iniziali di lavoro di un nuovo impianto. L'analisi ha portato ad una serie di considerazioni che sono del tutto analoghe a quanto ottenuto nel trasferimento per processi continui (Capitolo 3):

- il trasferimento con metodo PCA permette di definire modelli per il monitoraggio efficienti e l'adattamento a nuovi batch è più veloce. Modelli costruiti con questo metodo, però, non sono in grado di rilevare anomalie che si sviluppano esclusivamente su variabili non comuni, senza ripercuotersi su quelle comuni;
- il trasferimento con metodo JY-PLS costruisce modelli efficienti, come PCA, ma l'adattamento ai nuovi dati è più lento. Diversamente da PCA, considerando per il modello le variabili non comuni, si è in grado di rilevare anomalie che si ripercuotono su queste.

L'inserimento di problemi di sincronizzazione e gestione della dinamica non modifica i vantaggi che porta il trasferimento. Anzi, nelle fasi iniziali di lavoro di un nuovo impianto, questo approccio è conveniente per ridurre la richiesta di dati per un modello di monitoraggio efficiente, qualunque sia la natura del processo.

Infine, nel caso del processo batch, inserire nel modello le variabili controllate dal sistema di regolazione ne peggiora le prestazioni. Infatti, il metodo deve modellare la variabilità dei dati delle variabili di processo e anche quella dovuta al sistema di controllo.

Conclusioni

Nella Tesi è stato affrontato il problema del trasferimento di modelli per il monitoraggio di un processo che si trova nelle fasi iniziali di funzionamento. È noto che, se sono disponibili pochi dati, i metodi statistici non permettono di costruire modelli di monitoraggio efficienti. Nella Tesi il problema è stato risolto proponendo delle metodologie per il trasferimento di modelli per il monitoraggio da un impianto di cui un'ampia serie di dati è già disponibile ad un impianto che ha iniziato a marciare da breve tempo.

Metodi con cui è possibile sviluppare il trasferimento erano stati proposti da Facco *et al.* (2012) e Tomba *et al.* (2012). In questa Tesi tali metodi sono stati confrontati e modificati per i casi di studio di un processo industriale continuo di *spray-drying* dell'industria farmaceutica e di un processo batch simulato per la produzione di penicillina.

Innanzitutto, le tecniche presentate sono state confrontate in riferimento al caso particolare dello *spray-drying*. Il modello per il monitoraggio può essere costruito considerando le variabili comuni, cioè quelle variabili di simile significato fisico che vengono misurate in entrambi gli impianti in questione (tecnica PCA), o considerando anche quelle non comuni, cioè quelle che vengono misurate esclusivamente in uno dei due impianti (tecnica JY-PLS). In ogni caso, l'ipotesi alla base dell'utilizzo di questi metodi, che rende possibile il trasferimento, è la simile struttura di correlazione tra le variabili comuni tra gli impianti. Dal confronto operato, si è visto come i metodi PCA che utilizzano solo le variabili comuni tra gli impianti per il modello richiedano meno dati per costruire un modello efficiente. Tuttavia, poiché modellano la correlazione delle sole variabili comuni tra gli impianti, non sono in grado di rilevare anomalie che si manifestano su variabili non comuni e che non hanno influenza su quelle comuni. Il metodo JY-PLS che modella entrambi i *set* di variabili, comuni e non comuni, è più lento, ma più efficace nel rilevare anomalie sulle variabili non comuni. In questo caso, l'adattamento ai nuovi dati acquisiti dal nuovo impianto può dare benefici se, oltre alle variabili di processo, si considerano anche informazioni fisiche, caratteristiche del processo stesso.

Il contributo principale della Tesi è relativo ad una seconda applicazione, che riguarda lo sviluppo di metodi di trasferimento per un processo batch di produzione di penicillina. I metodi sviluppati utilizzano le variabili di processo comuni (PCA) o anche le variabili non comuni (JY-PLS). Per questo caso sono stati simulati dati di processo che rappresentassero le traiettorie temporali delle variabili. Nella Tesi i dati sono stati trattati, risolvendo problemi di sincronizzazione e di gestione della dinamica, per poter applicare i metodi statistici. Come per il processo continuo, sono state verificate le capacità di monitoraggio di modelli costruiti su un numero crescente di batch disponibili per definire le condizioni migliori per il

monitoraggio. Inoltre, è stato rilevato che la richiesta di dati per un modello efficiente è elevata e comporta settimane di produzione per l'acquisizione di dati. Tale richiesta diminuisce in modo marcato con il trasferimento. Nel particolare caso di studio, con il trasferimento i dati necessari per un modello efficiente passano da dati di 20 a 5 batch del nuovo impianto, se sono disponibili dati di almeno 20 batch di un altro impianto simile. Dall'analisi condotta nella Tesi è risultato che valgono le considerazioni del caso continuo relativamente ai metodi che utilizzano le variabili comuni e a quelli che considerano anche le variabili non comuni. Inoltre, è stato assodato che i vantaggi apportati dal trasferimento sono validi per riuscire a rilevare ogni tipo di anomalia.

Infine, è stata condotta un'ulteriore analisi sulle variabili che vengono incluse nell'insieme delle variabili comuni. Si è visto come i modelli più efficienti siano quelli per cui le variabili comuni non includono le variabili controllate dal sistema di regolazione. Ciò è stato giustificato considerando che il metodo deve modellare sia la variabilità dei dati di processo che quella del sistema di controllo ed è quindi più lento ad adattarsi ai nuovi dati.

In conclusione, dallo studio condotto nella Tesi deriva che, qualunque sia la natura del processo, il trasferimento di modelli tra impianti rende possibile il monitoraggio degli impianti nelle fasi iniziali di funzionamento, laddove i modelli statistici costruiti sui dati del singolo impianto non sono efficienti.

Appendice

Figure e codici contenuti nella Tesi

Nell'Appendice vengono riportate le Tabelle che forniscono una lista delle Figure presenti nei Capitoli della Tesi, reperibili nella cartella `\Tesi_MLargoni\Grafici`. Inoltre sono riportati i codici di calcolo contenuti nella Tesi. Essi sono *file* `.m` presenti nella cartella `\Tesi_MLargoni\Programmi.`

A.1 Figure del Capitolo 1

In Tabella A.1 sono riportati i riferimenti delle Figure del Capitolo 1.

Tabella A.1. *Figure del Capitolo 1.*

Figura	***File***
Figura 1.1	`PrinComp.vsd`
Figura 1.2	`Figura1.2.vsd`
Figura 1.3	`Figura1.3.vsd`
Figura 1.4a	`sincronizzazione_a.vsd`
Figura 1.4b	`sincronizzazione_b.vsd`
Figura 1.5	`trasferimento.vsd`

A.2 Figure del Capitolo 2

In Tabella A.2 sono riportati i riferimenti delle Figure del Capitolo 2.

Tabella A.2. *Figure del Capitolo 2.*

Figura	***File***
Figura 2.1	`Figura2.1`
Figura 2.2	`variabile9_fault.opj`
Figura 2.3	`processo.vsd`
Figura 2.4	`penicillina.opj`
Figura 2.5	`variabilibatch.opj`
Figura 2.6	`variabilicontrollo.opj`
Figura 2.7	`sincronizzazione.opj`

A.3 Figure del Capitolo 3

In Tabella A.3 sono riportati i riferimenti delle Figure del Capitolo 3.

Tabella A.3. *Figure del Capitolo 3.*

Figura	***File***
Figura 3.1	`Figura3.1.vsd`
Figura 3.2	`Figura3.2.vsd`
Figura 3.3	`Figura3.3.vsd`
Figura 3.4	`monitoraggioB.opj`
Figura 3.5	`AR_scenario1.opj`
Figura 3.6	`TD_scenario1.opj`
Figura 3.7	`AR_scenario3.opj`
Figura 3.8	`TD_scenario3.opj`
Figura 3.9	`AR_scenario2.opj`
Figura 3.10	`TD_scenario2.opj`
Figura 3.11	`AR_scenario4.opj`
Figura 3.12	`TD_scenario4.opj`
Figura 3.13	`AR_noncomuni_scenario1.opj`
Figura 3.14	`AR_noncomuni_scenario2.opj`
Figura 3.15	`TD_noncomuni_scenario2.opj`
Figura 3.16	`AR_noncomuni_scenario3.opj`
Figura 3.17	`AR_noncomuni_scenario4.opj`
Figura 3.18	`TD_noncomuni_scenario4.opj`

A.4 Figure del Capitolo 4

In Tabella A.4 sono riportati i riferimenti delle Figure del Capitolo 4.

Tabella A.4. *Figure del Capitolo 4.*

Figura	***File***
Figura 4.1	`variabili.vsd`
Figura 4.2	`PCAbatch_variePC_caso1.opj`
Figura 4.3	`PCA_variBatchA_caso1.opj`
Figura 4.4	`JY_variBatchA_caso1_autosc.opj`
Figura 4.5	`PCA_variBatchA_nc.opj`
Figura 4.6	`JY-PLS_variBatchA_nc.opj`
Figura 4.7	`carteSPE.opj`
Figura 4.8	`PCA_variBatchA_caso2.opj`
Figura 4.9	`JY_variBatchA_caso2.opj`
Figura 4.10	`profili_pH.opj`
Figura 4.11	`confronto_variabiliPH.opj`
Figura 4.12	`PCA_variBatchA_11.opj`

A.5 Codici di calcolo

In Tabella A.5 e Tabella A.6 sono riportati i codici di calcolo e i *file* da cui sono presi i relativi dati di *input*.

Tabella A.5. *Codici di calcolo per il Capitolo 3.*

Codici di calcolo	**Dati di *input***	**Descrizione**
`Monitoring_B`	`NOCimpiantoind.mat` `faultimpiantoind.mat`	Codice per il monitoraggio dell'impianto B con un modello costruito sui suoi soli dati

`Sc1_PCA.m`	`NOCpilota.mat`
	`NOCimpiantoind.mat`
	`realizations_medianTRISv2.mat`
		Codici per i metodi per il trasferimento di modelli per il monitoraggio del processo di *spray-drying*

`Sc2_JYPLS.m`		
`Sc3_PCAfisico.m`		
`Sc4_JYPLSfisico.m`		

`Sc1_perJY.m`	`NOCpilota.mat`
	`NOCimpiantoind.mat`
	`faultind_rampa50_realizz.mat`
		Codici per i metodi per il trasferimento di modelli per il monitoraggio del processo di *spray-drying* nel caso di anomalia sulle variabili non comuni

`Sc2_perJY.m`		
`Sc3_perJY.m`		
`Sc4_perJY.m`		

Tabella A.6. *Codici di calcolo per il Capitolo 4.*

Codici di calcolo	**Dati di *input***	**Descrizione**
`CostruzBatch1.m`	`-`	Codici per la simulazione dei dati dei processi batch
`CostruzBatch2.m`	`-`	
`Faultbatch2.m`	`batch2.mat`	Codice per la simulazione dei dati delle anomalie del processo batch B

`TrasfPCA.m`	`BATCH1_Penicillina.mat`
	`BATCH2_Penicillina.mat`
	`FaultsBatch2.mat`
		Codici per il trasferimento di modelli per il monitoraggio del processo batch (con possibilità di scelta delle variabili dei *set* comuni e non comuni)

`TrasfJYPLS.m`		

Riferimenti bibliografici

Birol, G., C. Undey e A. Çinar (2002). A modular simulation package for fed-batch fermentation: penicillin production. *Comp. Chem. Eng.*, **26**, 1553-1565.

Birol, G., C. Undey, S. J. Parulekar e A. Çinar (2002). A morphologically structured model for penicillin production. *Biotechnology and bioengineering*, **77**, 538-552.

Box, G. E. P., W. G. Hunter e J. S. Hunter (1978). *Statistics for experiments.* Wiley, New York.

Çinar, A., J. S. Parulekar, C. Undey e G. Birol (2003). *Batch fermentation. Modeling, monitoring, and control.* Marcel Dekker, Inc., New York (U.S.A.).

Dobry, D. E., D. M. Settell, J. M. Baumann, R. J. Ray, L. J. Graham e R. A. Beyerinck (2009). A model-based methodology for spray-drying process development. *J. Pharm. Innov.*, **4**, 133-142.

Jackson, J. E. (1991). *A user's guide to principal components.* John Wiley & Sons Inc., New York (U.S.A.).

Facco, P., E. Tomba, F. Bezzo, S. García-Muñoz, e M. Barolo (2012). Transfer of process monitoring models between different plants using latent variable techniques. *Ind. Eng. Chem. Res.*, **51**, 7327-7339.

Feudale, R. N., N. A. Woody, H. Tan, A. J. Myles, S. D. Brown e J. Ferré (2002). Transfer of multivariate calibration models: a review. *Chemom. Intell. Lab. Syst.*, **64**, 181-192.

García-Muñoz, S., T. Kourti e J. F. MacGregor (2003). Troubleshooting of an industrial batch process using multivariate methods. *Ind. Eng. Chem. Res.*, **42**, 3592-3601.

García-Muñoz, S., J. F. MacGregor e T. Kourti (2005). Product transfer between sites using Joint-Y PLS. *Chemom. Intell. Lab. Syst.*, **79**, 101-114.

García-Muñoz, S. e D. Settell (2009). Application of multivariate latent variable modeling to pilot-scale spray drying monitoring and fault detection: monitoring with fundamental knowledge. *Comp. Chem. Eng.*, **33**, 2106-2110.

Kassidas, A., J. F. MacGregor e P. Taylor (1998). Synchronization of batch trajectories using dynamic time warping. *AIChE J.*, **44**, 864-875.

Li, W., H. H. Yue, S. Valle-Cervantes and S. J. Qin (2000). Recursive PCA for adaptive process monitoring. *J. Process Control*, **10**, 471-486.

Lu, J. e F. Gao (2008). Process modeling based on process similarity. *Ind. Eng. Chem. Res.*, **47**, 1967-1974.

Lu, J., K. Yao e F. Gao (2009). Process similarity and developing new process models through migration. *AIChE J.*, **55**, 2318-2328.

Martens, H. e T. Naes (1989). *Multivariate calibration.* John Wiley & Sons Inc., New York (U.S.A.).

Nomikos, P. e J. F. MacGregor (1994). Monitoring batch processes using multiway principal component analysis. *AIChE J.*, **40**, 1361-1375.

Nomikos, P. e J. F. MacGregor (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics*, **37**, 41-58.

Qin, S. J. (1998). Recursive PLS algorithms for adaptive data modeling. *Computers Chem. Eng.*, **22**, 503-514.

Rännar, S., J. F. MacGregor and S. Wold (1998). Adaptive batch monitoring using hierarchical PCA. *Chemom. Intell. Lab. Syst.*, **41**, 73-81.

Tomba, E., P. Facco, F. Bezzo, S. García-Muñoz e M. Barolo (2012). Combining fundamental knowledge and latent variable techniques to transfer process monitoring models between plants. *Chemom. Intell. Lab. Syst.*, **116**, 66-67.

Wise, B. M. e N. B. Gallagher (1996). The process chemometrics approach to process monitoring and fault detection. *J. Process Control*, **6**, 329-348.

Siti web

`http://simulator.iit.edu/web/pensim/bground.html` (ultimo accesso: 16/10/2012)