UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE

CORSO DI LAUREA IN INGEGNERIA CHIMICA E DEI PROCESSI INDUSTRIALI

Tesi di Laurea Magistrale in

Ingegneria Chimica e dei Processi Industriali

SVILUPPO DI UN SENSORE VIRTUALE MULTIVARIATO E MULTI-RISOLUZIONE IN FERMENTATORI FED-BATCTH PER LA PRODUZIONE DI PENICILLINA

Relatore: Dr. Pierantonio Facco

Laureando: EMANUELE SPONTON

ANNO ACCADEMICO 2019 - 2020

Alla mia famiglia

Riassunto

In questo Tesi viene proposta una nuova metodologia per lo sviluppo di sensori virtuali che si ispira a una metodologia comunemente utilizzata nell'analisi multivariata d'immagine (Facco et al., 2010). La metodologia proposta è in grado di stimare le variabili di qualità di processi batch in tempo reale mediante un modello di regressione multivariata, la proiezione su strutture latenti. Questa utilizza come regressori alcuni comuni indici statistici (media, deviazione standard, momenti) calcolati dalla funzione di distribuzione di probabilità dei segnali della decomposizione multi-risoluzione. I segnali vengono decomposti per mezzo della trasformata wavelet discreta dei profili temporali delle variabili di processo misurate in linea, senza richiedere peraltro la sincronizzazione artificiale dei batch, come comunemente avviene nelle metodologie più accurate di Letteratura.

In particolare, la metodologia viene valutata nel caso di un processo fed-batch simulato per la produzione di penicillina, *IndPenSim* (Goldrick et al., 2015). Il risultato è che essa garantisce un'ottima accuratezza di stima (errore relativo di stima di circa 6%), dello stesso ordine di grandezza dei migliori metodi di letteratura, ma con un grado di semplificazione maggiore (in quanto non richiede sincronizzazione artificiale dei segnali), e con una percentuale di batch per cui la stima è ritenuta statisticamente attendibile molto maggiore (20% al posto di 30-50%). Infine, anche per i batch la cui stima risulta non attendibile la stima è accurata (errore relativo ~8%) a fronte del completo fallimento delle stime dei metodi di letteratura (errore relativo >100%).

Indice

Introduzione

L'utilizzo di processi batch è molto diffuso nei settori dell'industria di processo, specialmente nei settori in cui si realizzano prodotti ad alto valore aggiunto, come lo sono il farmaceutico, il settore della chimica fine e l'alimentare. Monitorare questi processi si rivela di fondamentale importanza per assicurare la realizzazione di prodotti di elevata qualità (Nomikos & MacGregor, 1994). In molti contesti, tuttavia, le variabili che determinano la qualità del prodotto possono essere misurate solo attraverso strumentazioni fuori linea, con costose analisi di laboratorio che sono dispendiose anche dal punto di vista del tempo richiesto e dei ritardi sulla valutazione della qualità e degli interventi correttivi. In queste condizioni, risulta molto difficile realizzare una efficace azione di controllo sulla qualità del prodotto.

Per fornire una soluzione a questo problema sono addottati dei sensori virtuali, modelli matematici in grado di stimare in tempo reale la qualità del prodotto dalle variabili di processo comunemente misurate in linea (Wise & Gallagher, 1996). Tra le metodologie basate su dati descritte in letteratura, la proiezione su strutture latenti (PLS) ricopre un ruolo di grande importanza, in quanto, permette lo sviluppo di modelli multivariati in grado correlare i valori delle misure fuori linea alle variabili misurate in linea (Geladi & Kowalski, 1986)(Nomikos & MacGregor, 1995)(Cinar et al., 2003). Per lo sviluppo dei modelli PLS i dati raccolti durante i batch vengono riordinati in matrici tridimensionali che possono essere trattati solo se i dati sono allineati e sincronizzati, anche artificialmente (Camacho et al., 2008) (Rothwell et al., 1998). In alcune applicazioni tuttavia, la sincronizzazione non è possibile (Facco et al., 2009). Per questo, risulta efficace una strategia di modellazione basata su indici statistici (Rendall et al., 2017)(Facco et al., 2010).

In questa Tesi si propone una nuova metodologia per lo sviluppo di sensori virtuali. Il metodo proposto utilizza come predittori di un PLS gli indici statistici (Facco et al., 2010) ricavati della decomposizione multi-risoluzione mediante trasformata wavelet (Addison, 2017) dei profili temporali delle variabili di processo.

La metodologia proposta viene quindi confrontata con le più accurate metodologie di Letteratura, nel caso studio di un processo fed-batch su scala industriale per la di produzione di penicillina, simulato utilizzando il software *IndPenSim* (Goldrick et al., 2015).

Capitolo 1

Metodi Matematici

In questo capitolo vengono presentati i metodi matematici utilizzati in questa Tesi per lo sviluppo dei sensori virtuali. Nella prima parte, si riportano e discutono criticamente alcuni tra i metodi di letteratura più utilizzati per lo sviluppo di modelli predittivi per processi batch (cioè, su dati che includono la dinamica del sistema). Inoltre, viene proposta una nuova strategia per lo sviluppo di un modello predittivo basato sulla trasformata wavelet.

1.1 Modelli MLR e PCR

Per sviluppare un modello che descriva una relazione tra due gruppi di variabili, ad esempio tra misure di processo **X** e variabili che identificano la qualità del prodotto **Y,** possono essere utilizzate diverse tecniche di regressione. La tecnica più comune per lo sviluppo di modelli predittivi è la regressione lineare multivariata (MLR). L'equazione di modello è

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E} \qquad \text{dove, } \boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{1.1}$$

dove $\mathbf{X}$ [$I \times J$] è una matrice di I osservazioni e J variabili indipendenti, $\mathbf{y}$ [$I \times 1$] è il vettore di I osservazioni della variabile dipendente, $\boldsymbol{\beta}$ [$J \times 1$] è la matrice dei coefficienti di regressione del modello e $\mathbf{E}$ [$I \times J$] è la matrice dei residui, l'apice T denota la trasposta della matrice a cui è riferito.

La MLR presenta un forte limite quando si modella sistemi multivariati le cui variabili sono fortemente collineari Infatti, in questi casi, si verifica un problema numerico nel calcolo dell'inversa $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ rendendo impossibile il calcolo dei coefficienti $\boldsymbol{\beta}$ attraverso il metodo dei minimi quadrati. I problemi che presentano queste caratteristiche vengono detti mal condizionati.

La regressione per componenti principali (*principal component regression*, PCR) è una delle tecniche per risolvere i problemi di questo genere. La modellazione PCR si basa su una decomposizione in autovettori della matrice di covarianza (o correlazione) delle variabili di processo $\mathbf{X}$.

La matrice di covarianza di una matrice di dati $\mathbf{X}$ è definita come:

$$\mathrm{cov}(\mathbf{X}) = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{m-1} \tag{1.2}$$

assumendo che le colonne di **X** siano centrate sulla media, ossia che la media di ogni colonna sia stata sottratta a ciascun elemento della colonna stessa,
Se la matrice **X** è stata invece autoscalata, cioè se le colonne sono state centrate sulla media e ciascun elemento di ogni colonna è stato diviso per la deviazione standard della colonna stessa, l'Equazione (1.2) fornisce la matrice di correlazione di **X**.
Attraverso l'ortogonalizzazione, la matrice originale dei dati viene decomposta come la somma del prodotto esterno dei vettori $\mathbf{t}_i$ $[M \times 1]$ e $\mathbf{p}_i$ $[N \times 1]$ più la matrice dei residui **E**:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \cdots + \mathbf{t}_{\mathrm{A}}\mathbf{p}_{\mathrm{A}}^{\mathrm{T}} + \mathbf{E} \tag{1.3}$$

dove $\mathbf{p}_i$ sono gli autovettori della matrice $\mathrm{cov}(\mathbf{X})$ definiti come *loadings* del modello e contengono informazioni su come le variabili sono correlate tra di loro. I vettori $\mathbf{t}_i$ sono noti come *scores* del modello e rappresentano la proiezione delle osservazioni (righe di **X**) sullo spazio definito dai loadings. Il numero delle componenti principali del modello A, viene scelto in modo da non lasciare importanti informazioni nella matrice dei residui **E**, matrice che quindi rappresenta solo errori casuali.
La relazione tra **T** $[I \times A]$, matrice cui le A colonne sono i vettori $\mathbf{t}_i$, **P** $[J \times A]$, matrice le cui A colonne sono gli autovettori $\mathbf{p}_i$, e **X** può essere riassunta nelle espressioni:

$$\mathbf{T} = \mathbf{XP} \quad , \tag{1.4}$$

$$\mathbf{X} = \mathbf{TP}^{\mathbf{T}} + \mathbf{E} \quad . \tag{1.5}$$

La regressione su componenti principali risolve il problema del mal condizionamento regredendo la variabile indipendente **Y** sugli scores delle componenti principali, ortogonali tra di loro, invece che sulle colonne della matrice originale **X**; inoltre, la capacità di eliminare alcune delle componenti principali, le meno rilevanti, permette di effettuare una riduzione del rumore nelle misure.
L'equazione di regressione del modello PCR diventa quindi:

$$\mathbf{y} = \mathbf{TB} + \mathbf{E} \quad , \tag{1.6}$$

dove la matrice dei coefficienti di regressione **B** $[j \times 1]$ è ottenuta come

$$\widehat{\mathbf{B}} = (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{y} \quad . \tag{1.7}$$

1.2 Proiezione su strutture latenti (PLS)

Il metodo di regressione lineare basto sulla proiezioni su strutture latenti (*projection on latent structures o partial least squares*, PLS) (Wold *et al.*, 1983), è più stabile dei metodi MLR e PCR, come osservato nei lavori di Wold *et al.*, (1984) e di Otto e Wegscheider (1985). Il metodo PLS rappresenta quindi una buona alternativa per lo sviluppo di modelli predittivi.

La proiezione su strutture latenti può essere considerata come un metodo intermedio tra le regressioni MLR e PCR. MLR ricerca i coefficienti che meglio correlano le variabili indipendenti **X** con le variabili dipendenti **Y**, mentre il metodo PCR individua le direzioni di maggiore variabilità tra i predittori.
PLS ricerca i fattori che eseguono entrambe le operazioni contemporaneamente, massimizzando la covarianza di **X** e **Y**, (Wise e Gallagher, 1996). Più semplicemente, il metodo PLS invece di focalizzarsi solo sulla variabilità della **X** (come nella PCR), considera la variabilità della **X** che è più è utile per predire i valori delle variabili dipendenti **Y**, (Nomikos e MacGregor 1995).
Il metodo PLS si basa su una decomposizione delle matrici autoscalate di *I* campioni, *J* variabili indipendenti e *M* variabili dipendenti, **X** $[I \times J]$ e **Y**$[I \times M]$:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{1.8}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \tag{1.9}$$

dove **T**$[I \times A]$ e **U**$[I \times A]$ sono le matrici composte da A score, rispettivamente di **X** e **Y**, essendo A il numero di variabili latenti utilizzate nel modello, mentre **P**$[J \times A]$ e **Q**$[M \times A]$ sono le matrici costituite da A vettori di loading delle due matrici, ed **E** $[I \times J]$ e **F**$[I \times M]$ sono le matrici dei residui.
Le Equazioni (1.8) e (1.9) sono chiamate le relazioni esterne del modello PLS; un'ulteriore relazione è richiesta per descrivere la relazione tra **X** e **Y**, chiamata relazione interna, che nella sua forma più semplice è data da:

$$\mathbf{u}_a = b_a \mathbf{t}_a \tag{1.10}$$

dove gli scalari $b_a = \mathbf{u}_a^{\mathrm{T}}\mathbf{t}_a / \mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a$ sono i coefficienti di regressione tra l' a-esimo vettore degli score di **X**, $\mathbf{t}_a$ e l' a-esimo vettore degli score di **Y**, $\mathbf{u}_a$.
I parametri del modello PLS riportati nelle equazioni precedenti possono essere calcolati in modo iterativo mediante l'algoritmo NIPALS (*nonlinear iterative partial least squares*). NIPALS calcola gli *scores* e i *loadings* (in analogia a quanto visto nel modello PCR) e una serie aggiuntiva di vettori chiamati *weights*, necessari per mantenere gli scores tra loro ortogonali. Diversamente dai modelli PCR e MLR, i modelli PLS posso essere utilizzati per predire più di una variabile indipendente **Y**. Scores **U** e loading **Q** vengono quindi calcolati anche per la matrice **Y**. La decomposizione PLS inizia selezionando una delle colonne di **Y**, $\mathbf{y}_m$ (con $m = 1, \dots, M$) come stima iniziale degli scores $\mathbf{u}_1$ (solitamente viene utilizzata la colonna di **Y** che presenta la variabilità più grande). Nel caso di una **y** univariata $\mathbf{u_1} = \mathbf{y}$.
L'algoritmo NIPALS si sviluppa nei seguenti passaggi:

$$\mathbf{w}_1 = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{u}_1}{\|\mathbf{X}^{\mathrm{T}}\mathbf{u}_1\|} \tag{1.11}$$

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1 \tag{1.12}$$

per la variabile dipendente $\mathbf{y}$:

$$\mathbf{q}_1 = \frac{\mathbf{u}_1^{\mathrm{T}}\mathbf{t}_1}{\|\mathbf{u}_1^{\mathrm{T}}\mathbf{t}_1\|} \tag{1.13}$$

$$\mathbf{u}_1 = \mathbf{Y}\mathbf{q}_1 \tag{1.14}$$

si verifica la convergenza del modello comparando $\mathbf{t}_1$ dell'Equazione (1.12) con il vettore calcolato nella precedente iterazione. Se differiscono meno di un certo errore (tipicamente 10^{-8}-10^{-10}), si procede con l'Equazione (1.15); se non ha convergenza si ritorna all'Equazione (1.8) e si usa il vettore $\mathbf{u_1}$ dell'equazione (1.14) come nuova $\mathbf{u}$. Se la $\mathbf{Y}$ è univariata, le Equazioni (1.13) e (1.14) possono essere omesse e $\boldsymbol{q_1} = 1$; non vengono in questo caso richieste iterazioni. Vengono poi calcolati i loadings della matrice $\mathbf{X}$ riscalati gli scores e i weights nel modo seguente:

$$\mathbf{p}_1 = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{t}_1}{\|\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1\|} \tag{1.15}$$

$$\mathbf{p}_{1\,\mathrm{new}} = \frac{\mathbf{p}_{1\,\mathrm{old}}}{\|\mathbf{p}_{1\,\mathrm{old}}\|} \tag{1.16}$$

$$\mathbf{t}_{1\,\mathrm{new}} = \mathbf{t}_{1\,\mathrm{old}}\|\mathbf{p}_{1\,\mathrm{old}}\| \tag{1.17}$$

$$\mathbf{w}_{1\,\mathrm{new}} = \mathbf{w}_{1\,\mathrm{old}}\|\mathbf{p}_{1\,\mathrm{old}}\| \quad , \tag{1.18}$$

Infine, vengono trovati i coefficienti di regressione della relazione interna,

$$\mathrm{b}_1 = \frac{\mathbf{u}_1^{\mathrm{T}}\mathbf{t}_1}{\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1} \quad , \tag{1.19}$$

vengono calcolati i residui delle matrici $\mathbf{X}$ e $\mathbf{Y}$:

$$\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} \tag{1.20}$$

$$\mathbf{F}_1 = \mathbf{Y} - \mathrm{b}_1\mathbf{u}_1\mathbf{q}_1^{\mathrm{T}} \tag{1.21}$$

L'intera procedura viene ripetuta per la successiva variabile latente ripartendo dall'Equazione (1.11). **X** e **Y** vengono rispettivamente rimpiazzati dai residui $\mathbf{E}_1$ e $\mathbf{F}_1$ e tutti gli indici vengono incrementati di 1.
Si può dimostrare che il modello PLS forma la matrice inversa di **X**:

$$\mathbf{X}^{+} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}} \tag{1.22}$$

dove **W**, **P** e **T** sono rispettivamente le matrici costituite dall'insieme dei vettori dei *weights*, dei *loadings* e degli *scores* del modello.
Si noti che gli scores e i loadings calcolati nel modello PLS non sono uguali a quelli calcolati nel modello PCR, essi sono stati ruotati per essere più predittivi su di **y** (dove **y** rappresenta un vettore colonna della matrice **Y**).
Riassumendo, PLS individua i fattori (o LVs) che meglio correlano e descrivono congiuntamente la variabilità di **X** e **Y**. Questo è in contrasto con il modello PCR dove i fattori (o PCs) vengono unicamente scelti per spiegare la massima variabilità di **X**, senza considerare quindi la capacità predittiva che tali fattori hanno sulla variabile dipendente.

1.2.1 Diagnostica sulle osservazioni

I modelli PLS sono ampiamente utilizzati come sensori virtuali nei processi batch in quanto permettono sia di eseguire predizioni sui parametri di qualità sia di eseguire azioni di monitoraggio.
L'azione di monitoraggio può essere condotta con diversi gradi di complessità, in applicazioni in cui rappresenta l'obbiettivo principale per lo sviluppo del sensore virtuale, vengono definite delle carte di controllo in grado di determinare l'istante in cui si verificano eventuali anomali e le variabili a causa dell'allontanamento dalle normali condizioni operative (Nomikos & MacGregor, 1995).
In questa Tesi, l'obbiettivo principale dei sensori virtuali utilizzati è la predizione in tempo reale dei parametri di qualità dei batch. L'azione di monitoraggio viene eseguita unicamente per determinare l'attendibilità di una nuova predizione.
Idealmente, il miglior modo per determinare l'attendibilità delle predizioni è eseguire un confronto con i valori reali determinati attraverso sensori fisici, tuttavia le difficoltà nel reperire queste informazioni attraverso metodi convenzionali sono alla base dell'utilizzo dei sensori virtuali.
Il metodo utilizzato si basa quindi sulla valutazione in tempo reale di alcuni parametri statistici direttamente determinabili da parametri ricavati nello sviluppo del modello, quindi senza conoscere alcuna informazione sul valore della variabile predetta. Vengono presentati di seguito i parametri utilizzati.

1.2.1.1 SPE (Q) e Hotelling

È possibile utilizzare degli indici per calcolare una statistica della rappresentatività del modello per ogni osservazione *i*. A questo scopo si definisce l'indice *squared prediction error* SPE:

$$\mathrm{SPE}_i = \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}} \quad , \tag{1.23}$$

SPE_i è una misura della variabilità in ogni campione che non viene catturata dalle A componenti principali utilizzate nel modello (Wise & Gallagher, 1996).
Una misura della variazione all'interno del modello PLS è invece fornita dalla statistica T^2 di Hotelling. T^2 è la somma degli score al quadrato normalizzati sulla varianza di ciascun vettore colonna di score ed è definita come:

$$T_i^2 = \mathbf{t}_i \mathbf{\Lambda}^{-1} \mathbf{t}_i^{\mathrm{T}} \quad , \tag{1.24}$$

anche qui i indica un'osservazione, quindi $\mathbf{t}_i[A \times 1]$ sono i vettori degli score relativi all'i-esima osservazione $\mathbf{\Lambda}$ è una matrice diagonale che contiene gli autovalori $\Lambda_1, \dots, \Lambda_A$ associati agli A loading (autovettori) considerati nel modello.

1.2.1.2 Limiti di confidenza

Vengono inoltre calcolati dei limiti di controllo su T^2 di Hotelling e su SPE. Dati gli autovalori $(\Lambda_1, \dots, \Lambda_A)$ della matrice di covarianza di $\mathbf{X}$, il limite di confidenza per SPE viene calcolato utilizzando l'equazione (Jackson & Mudholkar, 1979):

$$\mathrm{SPE}_a = \Theta_1 \left[\frac{c_\alpha \sqrt{2\Theta_2 h_0^2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0 (h_0 - 1)}{\Theta_1^2} \right] \tag{1.25}$$

$$\Theta_l = \sum_{g=A+1}^{Z} \Lambda_g^l \qquad \text{for } l = 1, 2, 3 \tag{1.26}$$

$$h_0 = 1 - \frac{2\Theta_1 \Theta_3}{3\Theta_2^2} \quad , \tag{1.27}$$

nell'Equazione (1.25) c_α è la deviazione standar normale corrispondente al percentile $(1 - \alpha)$. nell'equazione (1.5), e Z è il numero totale delle variabili latenti (equivalente al più piccolo tra i numeri delle variabili J o dei campioni I in $\mathbf{X}$).
Il limite di confidenza per Hotelling invece viene calcolato utilizzando l'equazione

$$T_{A,I,\alpha}^2 = \frac{k(i-1)}{i-k} F_{A,I-A,\alpha} \tag{1.28}$$

dove F è la distribuzione di Fisher.
Segue una rappresentazione grafica dei parametri appena introdotti. SPE rappresenta la variabilità dei dati non catturata dal modello. Immaginando per un momento un processo la cui

matrice **X** contiene solo tre variabili di processo, quindi tre colonne, e I misurazioni; ogni misurazione i può essere rappresentata nello spazio tridimensionale come accade in Figura 1.1

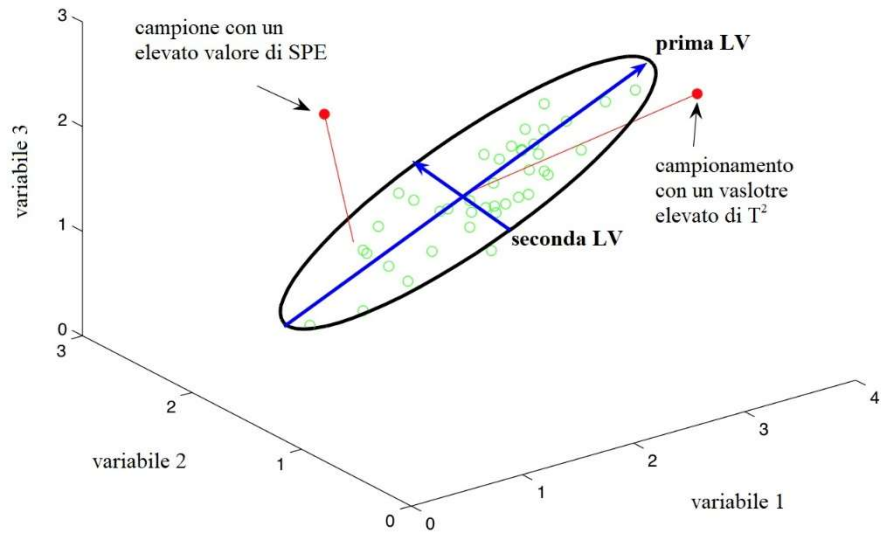


Figura 1.1. *Interpretazione grafica di SPE e Hotelling per un modello PLS*

Si osserva che le misurazioni giacciono approssimativamente su di un piano nello spazio. I campioni di **X** possono essere dunque ben descritti utilizzando un modello PLS con due variabili latenti. In questo schema, SPE rappresenta la distanza euclidea tra il piano definito dalle due variabili latenti del modello e la proiezione del campione i (misurazione, riga della matrice **X**) nello spazio degli scores. Il limite di confidenza di SPE calcolato attraverso l'Equazione (1.25) definisce la distanza dal piano che viene considerata usuale per le normali condizioni operative. T^2 rappresenta invece la distanza dalla media delle proiezioni sul piano degli scores, ossia dall'intersezione dei due assi che rappresentano le variabili latenti del modello.
Il limite di confidenza di Hotelling definisce un'elisse sul piano del quale giacciono le proiezioni che rappresentano le normali condizioni operative. Un campione di misure la cui proiezione ricade a una distanza dal piano superiore a quella definita da SPE limite oppure a una distanza dall'intersezione degli assi superiore a quella definita dal T^2 limite, determina una condizione anomala; la misura del parametro di qualità effettuata in corrispondenza ti tale campionamento viene ritenuta inaffidabile.

1.2.2 Predizioni nei modelli PLS

Determinati i parametri $\mathbf{p}_a$, $\mathbf{q}_a$, $\mathbf{w}_a$ del modello per ogni variabile latente ($a = 1, .., A$), i modelli PLS possono essere utilizzati per eseguire predizioni sui valori delle variabili dipendenti. Definita una nuova osservazione $\mathbf{x}_{\mathrm{NEW}}[1 \times J]$, la stima della matrice delle variabili

indipendenti predette $\hat{\mathbf{y}}_{\mathrm{NEW}}$ $[1 \times M]$ avviene mediante una proiezione della matrice $\mathbf{x}_{\mathrm{NEW}}$ e una costruzione della matrice $\hat{\mathbf{y}}_{\mathrm{NEW}}$ secondo:

$$\hat{\mathbf{t}}_a = \mathbf{x}_{\mathrm{NEW}}\mathbf{w}_a \tag{1.29}$$

$$\hat{\mathbf{y}}_{\mathrm{NEW}} = \sum_{a=1}^{A} b_a \hat{\mathbf{t}}_a \mathbf{q}_a^{\mathrm{T}} \tag{1.30}$$

dove $\hat{\mathbf{t}}_a[1 \times A]$è il vettore degli score della proiezione di $\mathbf{x}_{\mathrm{NEW}}$ e b_a è il coefficiente di regressione del modello per la variabile latente a.

1.2.3 Determinazione del numero di variabili latenti

Nei modelli a proiezione su strutture latenti PLS si possono calcolare tante componenti del modello PLS quanto è il rango della matrice delle variabili indipendenti **X**, tuttavia, non tutte le variabili latenti vengono usualmente usate; le ragioni principali sono in generale le necessità di comprime la dimensionalità del problema, il superamento del problema relativo al mal condizionamento delle matrici e l'eliminazione della componente di rumore nelle misure considerate.

La scelta del numero di variabili latenti ottimale per un modello PLS tuttavia è spesso un problema non banale.

Quando i modelli PLS vengono utilizzati per la predizione delle variabili indipendenti, la scelta del numero di variabili latenti viene effettuata utilizzando una classe di metodi chiamata *cross-validazione*; in questi metodi i dati a disposizione vengono suddivisi in due sottoinsiemi, un insieme utilizzato per la determinazione dei parametri del modello e un insieme per la determinazione delle capacità predittive del modello. L'errore commesso nelle predizioni può essere calcolato attraverso molti di parametri, di seguito vengono riportati alcuni esempi.

Sommatoria dei quadrati dei residui (*prediction residual sum of squares*, PRESS):

$$\mathrm{PRESS} = \sum_{i=1}^{I} (\hat{y}_i - y_i) \quad ; \tag{1.31}$$

Coefficiente di determinazione (R^2)

$$\mathrm{R}^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \tag{1.32}$$

$$\mathrm{RSS} = \sum_{i=1}^{I} (\hat{y}_i - y_i)^2 \tag{1.33}$$

$$\mathrm{TSS} = \sum_{i=1}^{I} (\bar{y} - y_i)^2 \quad ; \tag{1.34}$$

Media della sommatoria della radice dei quadrati degli errori di predizione (*root mean squares error of prediction*, RMSEP)

$$\mathrm{RMSEP} = \sqrt{\frac{\mathrm{PRESS}}{I}} \quad ; \tag{1.35}$$

nelle equazioni $\hat{y}_i$ rappresenta l'i-simo valore delle I misurazioni dei parametri di qualità contenuti in una delle M colonne della matrice $\mathbf{Y}$ $[I \times M]$, e y_i la predizione corrispondente determinata attraverso il modello di regressione.

L'errore di predizione, rappresentato dal parametro di valutazione scelto, viene calcolato per ogni variabile latente aggiunta al modello. In linea teorica, all'aggiunta di ogni nuova variabile che apporta un contributo sostanziale nella rappresentazione dei dati, l'errore di predizione diminuisce, mentre all'aggiunta di ogni nuova variabile latente che descrive unicamente il rumore presente nelle misure, l'errore di predizione aumenta. Il numero di variabili latenti deve essere scelto inoltre anche in relazione alla variabilità spiegata delle matrici **X** e **Y** dal modello. La scelta del numero di variabili latenti viene quindi supportata da due regole generali: 1) si scelgono solo i fattori in grado di migliorare il parametro di valutazione scelto del 2%, 2) si sceglie il minor numero di variabili latenti possibile.

1.2.4 Determinazione delle prestazioni predittive dei modelli

Per valutare e confrontare le prestazioni predittive dei modelli MPLS sviluppati possono essere utilizzate diverse metriche. Vengono riportati di seguito i parametri utilizzati in questa Tesi.

Errore relativo medio di predizione:

$$ErR_{medio,m} = \frac{1}{I} \sum_{i=1}^{I} \left(\frac{y_{i,m} - \hat{y}_{i,m}}{y_{i,m}} \right) \quad , \tag{1.36}$$

dove $y_{i,m}$ rappresenta l'i-sima osservazione, e $\hat{y}_{i,m}$ la corrispondente predizione determinata attraverso il modello di regressione.

Errore assoluto medio di predizione su deviazione standard:

$$ErA_{medio,m} = \frac{1}{\sigma_m} \left[\frac{1}{I} \sum_{i=1}^{I} \left(y_{i,m} - \hat{y}_{i,m} \right) \right] \tag{1.37}$$

dove σ rappresenta la deviazione standard delle misure contenute nell' m-simo vettore colonna y_m.

1.3 Proiezioni multidirezionali su strutture latenti (MPLS)

Il metodo di regressione lineare basto sulla proiezione su strutture latenti (PLS) non tiene in considerazione della dimensione tempo nelle matrici **X** e **Y**. Esistono, tuttavia, metodi che considerano esplicitamente il modo in cui i dati sono ordinati nel tempo, questi metodi vengono detti multidirezionali (*multi-way*). I modelli multidirezionali si rivelano particolarmente utili nell'analisi dei dati di processi non stazionari. Si consideri la matrice tridimensionale $\underline{\mathbf{X}}$ mostrata in Figura 1.2.

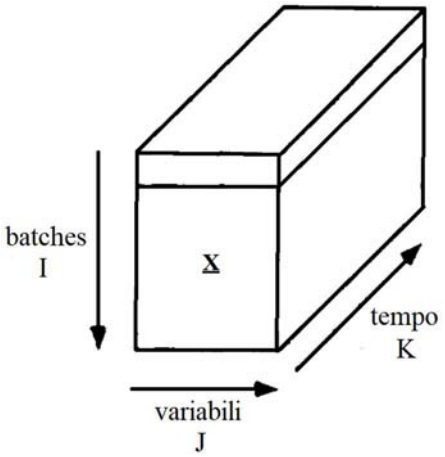


Figura 1.2. *Matrice di dati tridimensionale* (Wise & Gallagher, 1996).

Una matrice di dati di questo tipo è costituita da una serie di misure temporali delle variabili raccolte durante diversi batch. Come si vede in Figura 1.2 i batch ($i = 1,2,3,\dots,I$) vengono disposti sulla dimensione verticale, le variabili di processo ($j = 1,2,3,\dots,J$) vengono disposte sulla dimensione orizzontale e il tempo ($k = 1,2,3,\dots,K$) viene disposto lungo la terza dimensione, in profondità rispetto al piano del foglio. Le facce orizzontali $[J \times K]$ della matrice rappresentano la storia temporale di ogni variabile per un determinato batch. Le facce verticali, parallele alla faccia frontale del cubo, sono matrici $[I \times J]$che rappresentano il valore di ogni variabile presa allo stesso istante temporale per ogni batch; le facce verticali, parallele al lato del cubo (quindi sull'asse del tempo) sono matrici $[I \times K]$ che descrivono la storia delle misure di una variabile nei diversi batch.
Multiway PLS (MPLS) è una estensione del metodo PLS necessaria per gestire dati disposti in matrici tridimensionali. In particolare, si utilizza un metodo di sfogliamento della matrice (*unfolding*) per costruire un modello PLS standard su una versione bidimensionale della matrice tridimensionale.

1.3.1 Batch-Wise unfolding

La matrice tridimensionale $\underline{\mathbf{X}}$ $[I \times J \times K]$ può essere decomposta in una matrice bidimensionale di dimensioni $[I \times JK]$ mediante *batch-wise unfolding* (Nomikos & MacGregor, 1994) in modo che ogni faccia verticale $[I \times J]$ sia posta fianco a fianco verso destra, iniziando con la faccia

corrispondente al primo valore temporale $k = 1$. Questo particolare modo per decomporre la matrice permette di analizzare la variabilità delle misure rispetto ai loro profili temporali, Figura 1.3.

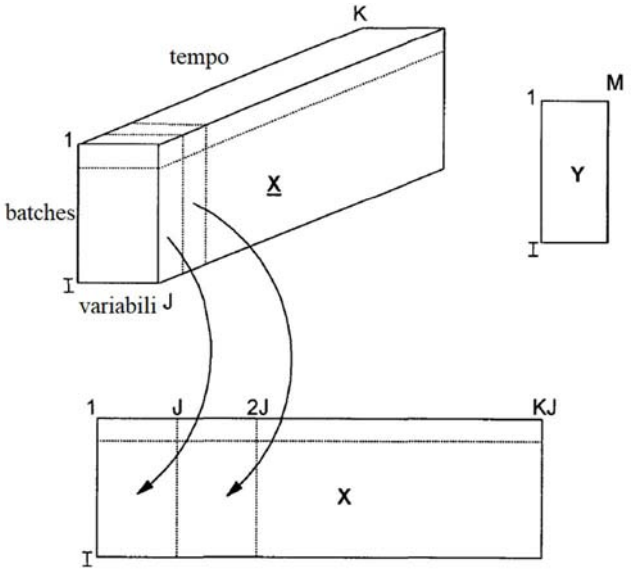


Figura 1.3. *Batch-wise unbfolding della matrice tridimensionale nel modello MPLS* (Nomikos & MacGregor, 1994).

Successivamente all'operazione di sfogliamento (*unfolding*) le matrici vengono autoscalate (ad ogni elemento della matrice viene sottratto il valore medio dei valori della colonna di appartenenza e viene diviso per la deviazione standard della stessa colonna) e utilizzate come matrici bidimensionali nella costruzione di un modello PLS bilineare.
Mediante batch-wise unfolding, i modelli MPLS ottenuti sono in grado di considerare eventuali variazioni di correlazione tra le variabili misurate durante il processo (Camacho et al., 2008). Per poter essere applicata l'operazione di sfogliamento batch-wise unfolding, richiede che i batch considerati siano tutti della stessa durata. Inoltre, richiede che nella strategia *batch-wise unfolding* i batch siano sincronizzati e abbiano la stessa durata. Deve esserci, in altre parole, una completa corrispondenza temporale tra i dati contenuti in ogni colonna delle matrici bidimensionali.

1.3.1.1 Sincronizzazione mediante variabile indicatrice

Spesso i processi batch sono utilizzati per la loro caratteristica versatilità che rende possibile arrestare i processi al raggiungimento di una predeterminata specifica. In relazione alla sensibilità dei processi considerati alle diverse condizioni iniziali e a disturbi esterni, la durata dei batch può quindi subire delle variazioni di diversa entità.
La maggior parte dei processi batch, inoltre, attraversa durante la propria evoluzione diverse fasi, ognuna di esse caratterizzata da differenti fenomeni chimico-fisici o fisiologici nel caso di bioreattori. Quando processi simili hanno durate differenti, anche le fasi contenute in essi

subiscono un disallineamento temporale e questi slittamenti possono influenzare negativamente il monitoraggio di processo e generare falsi allarmi (Cinar et al., 2003). Condizioni di questo tipo si verificano, ad esempio, nei processi di fermentazione: interrotti al raggiungimento di determinati parametri di qualità, la durata dei processi è molto diversa tra un batch e l'altro. La causa è la forte sensibilità della biomassa alle condizioni iniziali e ai disturbi.
In tutti questi scenari risulta impossibile confrontare in modo immediato i dati raccolti, allocare le misurazioni in modo omogeneo in una matrice tridimensionale, e sviluppare un modello MPLS senza prima uniformare i dati attraverso una sincronizzazione. Per questo motivo la sincronizzazione delle misure di processo ricopre una importanza cruciale nel rendere possibile l'utilizzo di tecniche di analisi multivariate, come lo sviluppo di un modello MPLS.
Il modo più semplice per sincronizzare batch di durata differente, in cui le misure sono effettuate con cadenza regolare, è troncare i dati disponibili in corrispondenza del processo di durata più breve. In tal modo si rende possibile la disposizione dei dati in una matrice tridimensionale. Ciò nonostante, troncare i dati non permette di allineare misurazioni effettuate con frequenze diverse o l'allineamento delle fasi di un batch; questo comporta la perdita di una considerevole quantità di informazioni, per questi motivi non è una metodologia consigliabile. Si rende quindi necessario l'utilizzo di tecniche di sincronizzazione più avanzati (Rothwell et al., 1998). Uno di questi metodi è la tecnica a variabile indicatrice.
In questa Tasi viene utilizzata la tecnica di sincronizzazione a variabili indicatrice, la quale si basa sulla scelta di una variabile in grado di sostituire il tempo come indicatore della percentuale di completamento del batch. Le misure di processo vengono quindi sincronizzate in relazione al progresso di questa variabile.
Una variabile di processo per essere scelta come variabili indicatrice, deve possedere i seguenti requisiti: deve essere continua e monotona per tutta la durata del batch e deve mantenere queste caratteristiche in tutti i batch considerati; inoltre, la variabile indicatrice deve essere scelta in modo da rappresentare la maturità del processo e la sua percentuale di completamento, quindi deve avere lo stesso valore iniziale e il medesimo valore finale. L'intervallo tra questi due valori viene successivamente suddiviso in punti percentuali che corrispondo alla progressione del batch. La sincronizzazione si realizza facendo corrispondere misure che, sebbene siano effettuate in tempi diversi, hanno lo stesso valore della variabile indicatrice.
La scelta della variabile indicatrice dipende dalla natura del processo. In un processo di fermentazione fed-batch in cui il substrato viene continuamente alimentano per mantenerne costante la concentrazione, si considera come variabile indicatrice la quantità cumulativa di substrato alimentato al processo.
Dopo la sincronizzazione i dati possono essere disposti in una matrice tridimensionale come in Figura 1.1 in cui ogni colonna della matrice raggruppa le misure allo stesso grado di avanzamento del batch.

1.3.2 Strategie multi-modello per la stima della qualità di prodotto

I processi batch industriali generano una enorme quantità di dati che possono essere raccolti in matrici tridimensionali utilizzabili per la stima della qualità in tempo reale. Questa può seguire due strategie: lo sviluppo di un modello unico per tutta la durata del processo o lo sviluppo di più sotto-modelli, ad esempio un sotto modello per ogni istante temporale k in cui si ha a disposizione una nuova misurazione. L'insieme dei K sotto-modelli sviluppati seguendo la seconda metodologia viene definito come tecnica multi-modello (K-models, (Camacho et al., 2008)). Entrambe le strategie possono essere sviluppate sia su batch-wise unfolding che su variable-wise unfolding.

La metodologia multi-modello, anche se più dispendiosa dal punto di vista computazionale, permette di costruire modelli più localizzati; permette di trascurare le informazioni meno rilevanti e di eseguire una stima più precisa con un numero minore di predittori, risulta essere quindi più vantaggiosa rispetto alla strategia a modello unico.

In letteratura vengono proposte molte alternative per lo sviluppo di strategie multi-modello, le quali differiscono principalmente per la quantità e tipologia dei dati contenuti nei sotto-modelli. Vengono riportate di seguito le due metodologie usate in questa Tesi.

1.3.2.1 Modelli evolutivi

La strategia multi-modello in cui ognuno dei K sotto-modelli viene sviluppato utilizzando tutte le misurazioni a disposizione al tempo di campionamento k fin dall'inizio del batch, vengo definiti modelli evolutivi. In Figura 1.4 viene riportata una rappresentazione schematica delle variabili utilizzate dal k-simo sotto-modello evolutivo. All'istante $k = 1$ si considerano tutte le variabili per quell'istante, all'istante 2 si affiancano le nuove osservazioni allo stesso modo che nel batch-wise unfolding le J variabili dell'istante $k = 2$ a quelle dell'istante $k = 1$. All'istante k si considerano in questo modo le variabili kJ valori delle variabili.

1.3.2.2 Modelli a finestra mobile

La strategia multi-modello a finestra mobile, invece, è costituiti da sotto-modelli che utilizzano la misura corrente e le misurazioni delle J variabili negli N istanti di tempo precedenti all'istante di campionamento considerato k. A differenza di quanto accade nei modelli evolutivi, il numero di misurazioni utilizzate da ogni sotto-modello rimane costante nel tempo. In Figura 1.5 viene riportata una rappresentazione schematica delle variabili utilizzate dal k-simo sotto-modello a finestra mobile per $N = 2$, all'istante 3, in cui si considerano le osservazioni delle variabili J relative agli istanti $k = 2$ e $k = 3$, successivamente per l'istante k, si considerano le variabili J relative agli istanti k-1 e k,.

Le misure utilizzate nei sotto-modelli rappresentano un insieme di misure effettuate in una finestra temporale che trasla con il tempo.

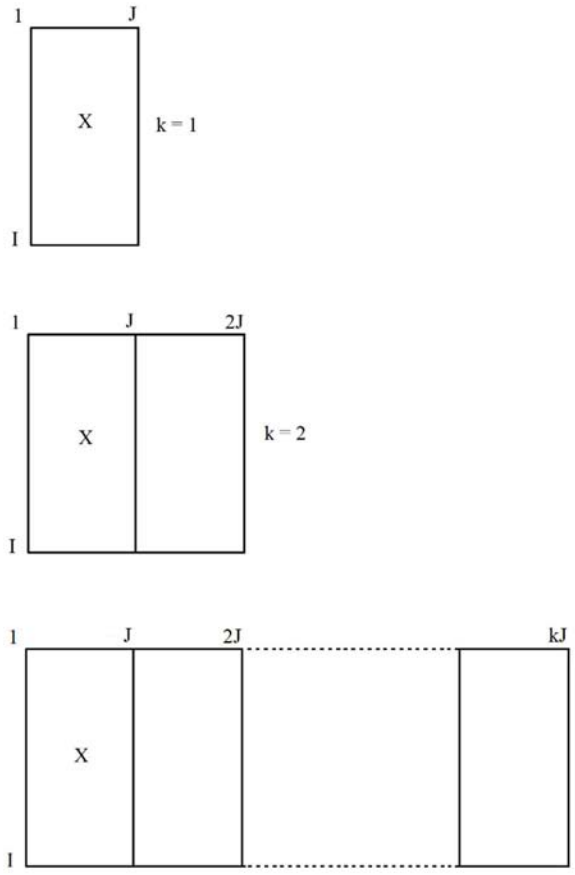


Figura 1.4. *Dati della matrice* $\boldsymbol{X}[I \times KJ]$*, ottenuta attraverso batch-wise unfolding della matrice tridimensionale* $\underline{\boldsymbol{X}}[I \times K \times J]$*, utilizzati nello sviluppo del sotto-modello a evolutivo corrispondente al campionamento k.*

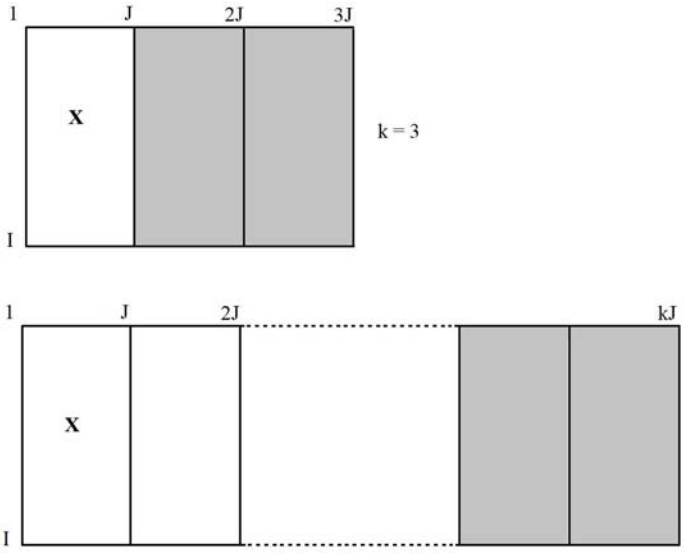


Figura 1.5. *Dati della matrice* $\boldsymbol{X}[I \times KJ]$*, ottenuta attraverso batch-wise unfolding della matrice tridimensionale* $\underline{\boldsymbol{X}}[I \times K \times J]$*, utilizzati nello sviluppo del sotto-modello a finestra mobile corrispondente al campionamento k, per N=2.*

Entrambe le strategie multi-modello presentate richiedono comunque che i batch considerati siano tutti della stessa durata, se questo non avviene si rende necessario utilizzare una sincronizzazione dei dati.

1.3.3 Variable-Wise unfolding

In alternativa la decomposizione della matrice tridimensionale $\underline{\mathbf{X}}$ ad una bidimensionale può essere ottenuta mediante *variable-wise unfolding* preservando cioè la direzione delle variabili (Svante Wold et al., 1998). In questo caso le facce $[I \times J]$ vengono disposte una sopra l'altra, Figura 1.6.
Le matrici bidimensionali ottenute vengono successivamente autoscalate; infine, si utilizza l'algoritmo NIPALS per ottenere i parametri del modello PLS bilineare.

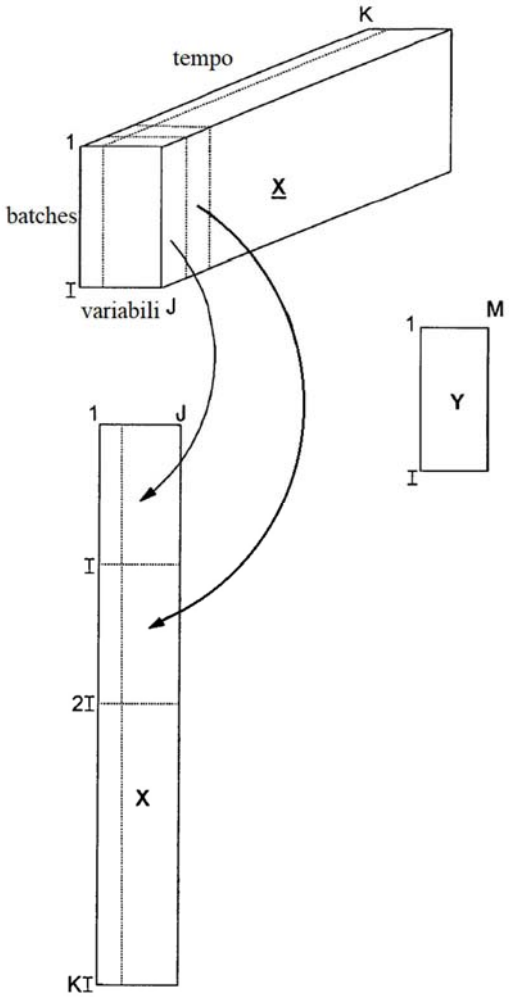


Figura 1.6. *Scomposizione della matrice tridimensionale nel modello MPLS.*

I modelli MPLS sviluppati attraverso la metodologia *variable-wise unfolding* (diversamente dai modelli MPLS sviluppati con *batch-wise unfolding*) determinano una correlazione tra le

variabili mediata su tutta la durata del processo e non permettono di individuare variazioni di correlazione nel tempo (Camacho et al., 2008).

1.4 Modelli PLS a indici statistici multi-risoluzione

I modelli PLS descritti nei paragrafi precedenti determinano una correlazione tra le variabili di processo e i parametri di qualità utilizzando i profili delle misurazioni come regressori. I modelli a indici statistici invece, sono sviluppati mediante una strategia differente. I regressori utilizzati sono parametri statistici (*features*) calcolati sui profili temporali delle misurazioni (Rendall et al., 2017). In Figura 1.7 viene riportata una rappresentazione schematica dello sviluppo di un modello a indici statistici.

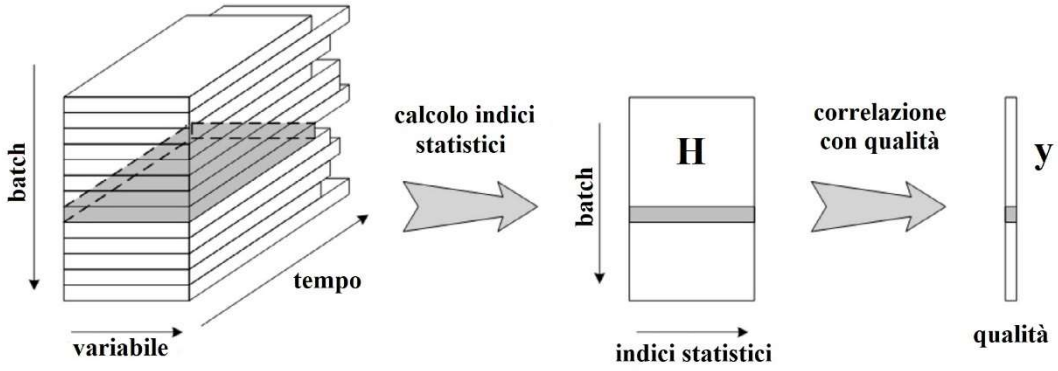


Figura 1.7. *Descrizione schematica dello sviluppo di un modello di regressione a indici statistici per processi non sincronizzati.*(Suthar et al., 2019).

Inizialmente vengono estratti degli indici statistici dalle traiettorie delle variabili misurate attraverso l'operazione:

$$\rho: \mathbf{X} \rightarrow \mathbf{H} \tag{1.38}$$

Dove ρ rappresenta l'operatore che trasforma le informazioni contenute nei dati di processo della matrice $\mathbf{X}[I \times J]$, di I campioni e J variabili nella matrice di indici statistici $\mathbf{H}[I \times S]$, di I campioni e S indici (Suthar et al., 2019). Senza perdere generalità si può estendere il medesimo discorso a dati dinamici di processi batch.

Gli indici possono essere parametri che caratterizzino le singole variabili (come la media e la varianza), l'interazione tra diverse variabili (come la correlazione incrociata) o il processo stesso (come l'integrale nel tempo del calore scambiato).

Come si può osservare in Figura 1.7 l'utilizzo dei modelli PLS a indici statistici permette di gestire il problema della sincronizzazione dei dati appiattendo a un singolo valore (di uno o più indici per variabile) le misure contenute nei profili temporali delle variabili anche per batch di

durata differente. In altre parole, viene eliminata la necessità della sincronizzazione dei dati, problema ancora aperto nel caso della stima in linea della qualità.
In questa tesi si propone una nuova metodologia per lo sviluppo di modelli predittivi MPLS utilizzabili per stimare in tempo reale le variabili di qualità in processi batch. I nuovi modelli ad indici statistici multi-risoluzione (ISMR) intendono combinare i vantaggi introdotti dall'utilizzo di indici statistici e dalla decomposizione multi-risoluzione dei segnali ottenuta con l'utilizzo della trasformata wavelet discreta.
In letteratura lo studio e l'applicazione della decomposizione wavelet è già stata considerevolmente esplorata, riscontrando un particolare successo nel monitoraggio di processi industriali continui (Bakshi, 1998) (Yoon & MacGregor, 2004), profili spaziali (in entrambi i casi, 1D e 2D)(Reis & Saraiva, 2006)(Facco et al., 2010) e nel settore manifatturiero (Lada et al., 2002).

1.4.1 Trasformata wavelet

Molti metodi sofisticati per il monitoraggio e la caratterizzazione di processo e della qualità di prodotto utilizzano la trasformata wavelet in quanto questa tecnica matematici in grado di analizzare i segnali simultaneamente nel dominio del tempo.
La trasformata wavelet si basa sull'utilizzo di funzioni simili a ondine chiamate *wavelet* per una più utile rappresentazione delle informazioni in esso contenute, Figura 1.8. Le wavelet possono essere manipolate in due modi: possono essere posizionate in diversi punti del segnale e possono essere allargate o compresse, Figura 1.9.

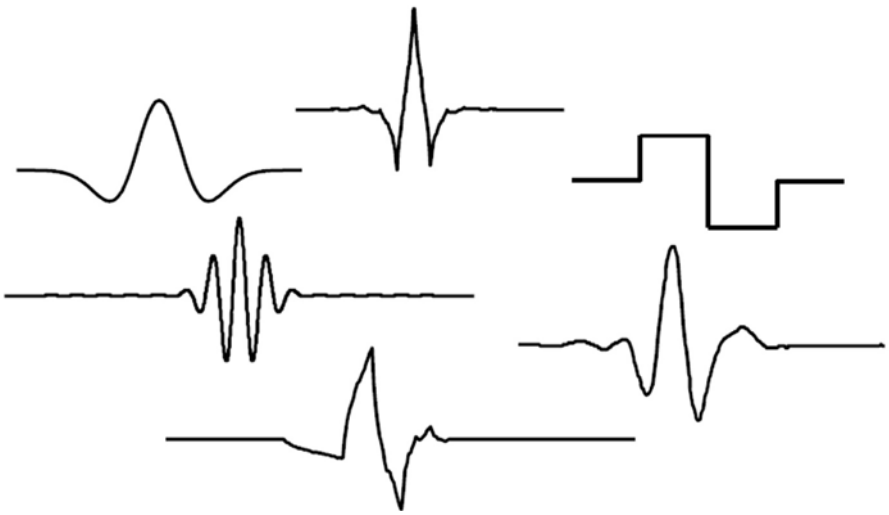


Figura 1.8. *Esempi di alcune funzioni wavelet madre utilizzate nella trasformata wavelet.*

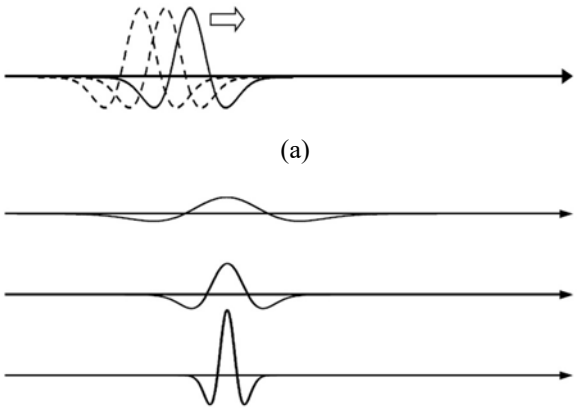


Figura 1.9. *Esempio della traslazione lungo la coordinata temporale di una wavelet* (a) *e di una wavelet in diverse scale* (b).

Si consideri un segnale, ad esempio una variabile $x(t)$ misurata nel tempo. Se per una determinata scala e posizione la wavelet è correlata (cioè, si sovrappone con buona corrispondenza) alla forma del segnale analizzato, si ottiene un elevato valore della trasformata; se invece, la wavelet non combacia con la forma del segnale, si ottiene un basso valore della trasformata. I valori ottenuti della TW per ogni posizione temporale e scala della wavelet madre vanno a costituire una superficie bidimensionale che giace sopra il piano descritto dal dominio della trasformata (Figura 1.10).

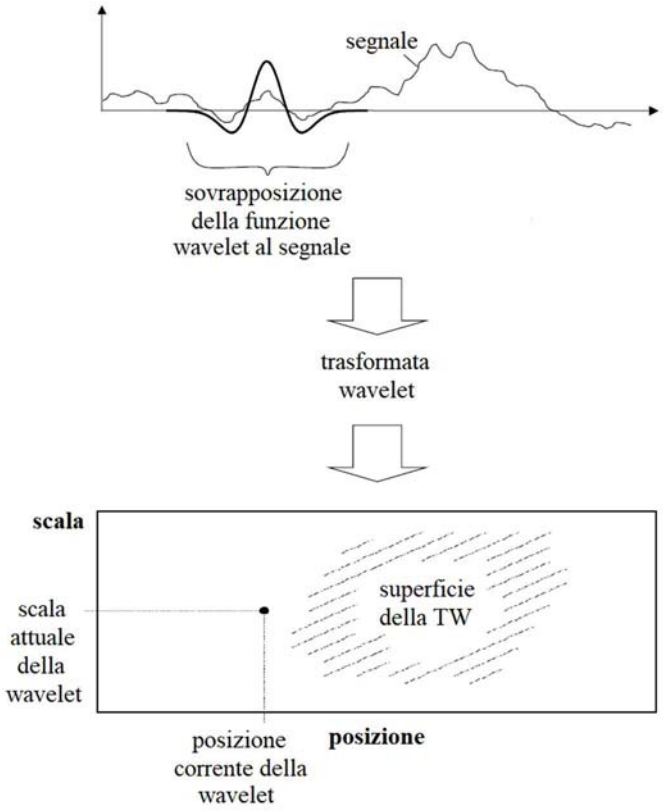


Figura 1.10. *Rappresentazione schematica della applicazione della trasformata wavelet.*

La trasformata wavelet è rappresentata in termini matematici come:

$$W(v,w) = \frac{1}{\sqrt{|a|}} \int x(t) \Psi \left(\frac{t-w}{v} \right) dt \tag{1.39}$$

dove Ψ rappresenta la *wavelet madre*, $x(t)$ è il segnale originale, v e w sono i parametri che determinano rispettivamente la scala (dilatazione) della wavelet e la sua traslazione, il fattore $1/\sqrt{|a|}$ viene utilizzato per assicurarsi che l'energia del segnale scalato e traslato sia la stessa di quella della wavelet madre (Addison, 2017). L'Equazione (1.39) mostra la trasformata wavelet continua (*continue wavelet transform*, CWT) e può essere interpretata come il prodotto interno del segnale $x(t)$ con la versione scalata e traslata della wavelet madre Ψ , (Daubechies, 1992):

$$W(v,\mathrm{w}) = \int x(t) \Psi_{(v,w)}(t) dt \tag{1.40}$$

$$\Psi_{(v,w)}(t) = \frac{1}{\sqrt{|a|}} \Psi \left(\frac{t-w}{v} \right) \quad . \tag{1.41}$$

La versione scalata e traslata della wavelet madre viene rappresentata nell'Equazione (1.41).

1.4.1.1 Trasformata wavelet discreta

Calcolare la TW in modo continuo per ogni punto del dominio del tempo e della frequenza comporta un elevato costo computazionale.

In applicazioni pratiche si preferisce la trasformata wavelet discreta (*discrete wavelet transform*, DWT) che permette di ridurre la quantità di operazioni matematiche necessarie senza perdere rilevanti informazioni nel calcolo della trasformata del segnale. Le wavelet discrete sono ottenute considerando dei valori discreti dei parametri v e w, al posto dei rispettivi valori continui.

Per ottenere la discretizzazione della DTW, comunemente si utilizza una discretizzazione diadica dei parametri di scala e traslazione, $v = 2^m$ e $w = 2^m n$ in modo da costituire una base ortonormale. La funzione wavelet madre discretizzata diventa:

$$\Psi_{(m,n)}(t) = 2^{-\frac{m}{2}} \Psi \left(2^{-m} t - n\right) \tag{1.42}$$

m e n sono rispettivamente i parametri di scala e traslazione. Utilizzando le wavelet definite in questo modo, la trasformata wavelet discreta che si ottiene è:

$$T_{(m,n)} = \int x(t) \, \Psi_{(m,n)}(t) dt \quad . \tag{1.43}$$

Scegliendo una base wavelet ortonormale è possibile ricostruire il segnale originale semplicemente invertendo il processo:

$$x(t) = \sum_{m=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} T_{(m,n)}\, \Psi_{(m,n)}\,(t) \quad . \tag{1.45}$$

In accordo con la teoria di (Mallat, 1989), ogni funzione a quadrato integrabile (una funzione a valori reali o complessi, si dice a quadrato sommabile, o anche a quadrato integrabile, in un determinato intervallo se l'integrale del quadrato del suo modulo è finito) può essere decomposta in diversi gradi di risoluzione attraverso una funzione di scala (funzione wavelet padre) e una funzione wavelet madre. Le funzioni di scala hanno la stessa forma delle funzioni wavelet e sono associate allo smorzamento del segnale. Inoltre, sono ortogonali alle proprie traslazioni ma non alle dilatazioni. Le funzioni di scala, definite con $\Phi_{(m,n)}$, vengono calcolate in modo analogo alle funzioni wavelet $\Psi_{(m,n)}$ (definite quindi con (1.42) con le dovute sostituzioni). La funzione di scala può essere convoluta con il segnale per produrre il coefficiente di approssimazione definito:

$$S_{(m,n)}(t) = \int x(t)\, \Phi_{(m,n)}\,(t)dt \quad . \tag{1.46}$$

I coefficienti di approssimazione sono le medie pesate del segnale continuo fattorizzate con $2^{m/2}$, inoltre ad una specifica scala m rappresentano le approssimazioni discrete del segnale nella medesima scala.

È possibile ottenere un'approssimazione continua del segnale alla scala m attraverso la seguente sommatoria:

$$x_m(t) = \sum_{m=-\infty}^{\infty} S_{(m,n)}\ \Phi_{(m,n)}(t) \quad . \tag{1.47}$$

dove $x_m(t)$ è la versione "smussata" del segnale originale $x(t)$ alla scala m. L'approssimazione si avvicina alla forma del segnale per un valore di $m \to -\infty$.

È possibile ricostruire il segnale originale $x(t)$ combinando le due serie, usando quindi il coefficiente di approssimazione e la wavelet, definita anche come coefficiente di dettaglio:

$$x_m(t) = \sum_{m=-\infty}^{\infty} S_{(m,n)}\ \Phi_{(m,n)}(t) + \sum_{m=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} T_{(m,n)}\, \Psi_{(m,n)}\,(t) \tag{1.48}$$

Un segnale quadrato integrabile può essere decomposto in più risoluzioni seguendo le equazioni descritte in modo iterativo.

Il segnale in entrata x viene filtrato in parallelo attraverso un filtro passa basso L(n), costituito dalla funzione di scala $\Phi_{(m,n)}$, e un filtro passa alto H(n), costituito dalla funzione wavelet $\Psi_{(m,n)}$ Successivamente, a diverse scale di risoluzione i filtri passa basso e passa alto vengono applicati al segnale filtrato in uscita da L(n) definito come $A1$(l'approssimazione al primo livello di risoluzione), in quanto l'uscita dal filtro passa basso possiede la maggior parte delle

informazioni contenute nel segnale. Applicando questo algoritmo iterativamente vengono individuati i coefficienti di approssimazione Ai e di dettaglio Di per diverse scale (risoluzioni). La Figura 1.11 rappresenta il processo di filtrazione passa alto e passa basso appena descritto per una scala (grado di risoluzione).

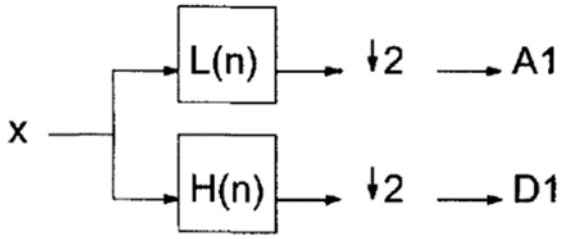


Figura 1.11. *Rappresentazione schematica dei filtri passa basso e passa alto che descrivono la trasformata wavelet discreta di un segnale x.*

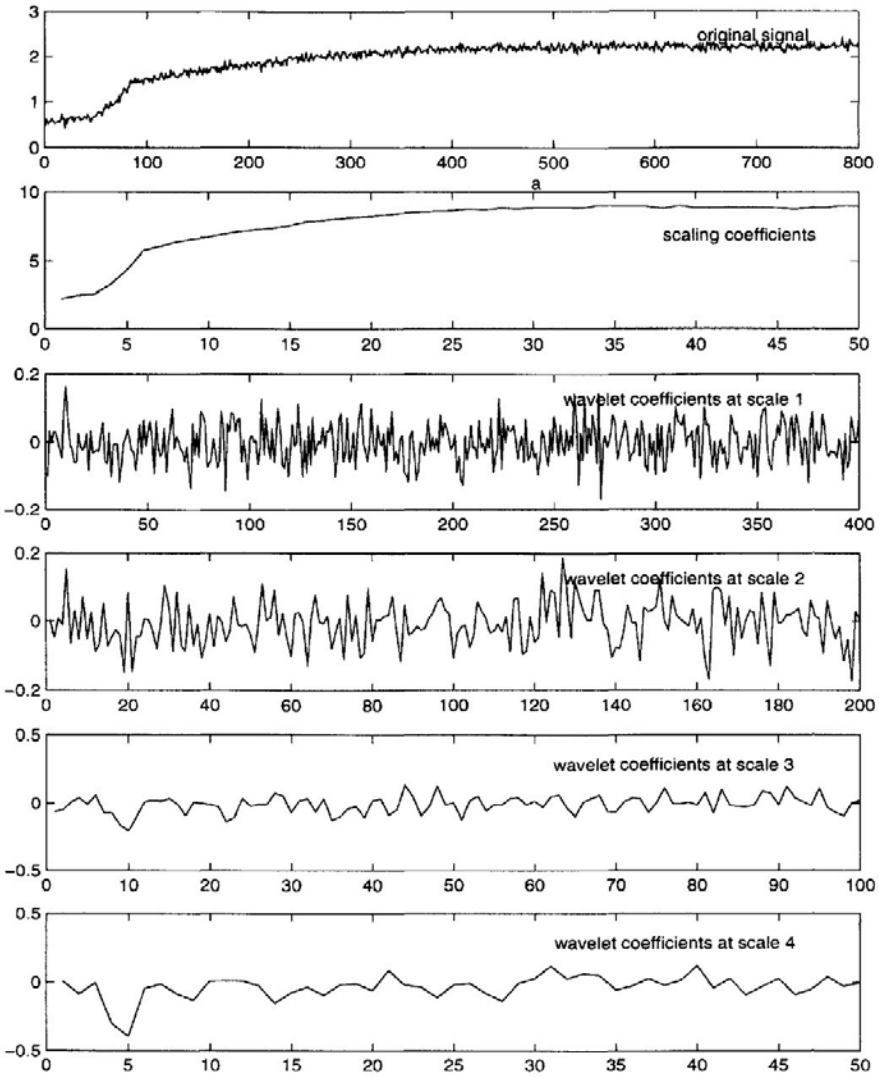


Figura 1.12. *Rappresentazione della decomposizione multi-risoluzione di un segnale mediate l'applicazione della trasformata wavelet. Il segnale originale viene riportato nel primo riquadro in alto. La decomposizione avviene mediate la determinazione dei coefficienti di approssimazione e dei coefficienti di dettaglio per quattro gradi di risoluzione.*

1.4.2 Metodologia proposta

Lo sviluppo di un modello ISMR si articola in due passaggi principali: 1) una decomposizione dei segnali, in cui la trasformata wavelet discreta viene applicata al profilo di ogni variabile, 2) lo sviluppo di un modello MPLS a indici statistici sui coefficienti di approssimazione e dettaglio ottenuti dalla decomposizione wavelet.

Lo sviluppo di un modello ISMR avviene più dettagliatamente secondo i seguenti i passaggi che si ripetono per ogni istante *k*:

- selezionare i valori delle k misurazioni disponibili per le J variabili di processo;
- applicare la trasformata wavelet discreta ai profili delle variabili fino all'istante k per ottenere i coefficienti di approssimazione e dettaglio per ogni grado di risoluzione R.
- determinare S indici statistici dai coefficienti di dettaglio e approssimazione di 6 gradi di risoluzione successivi. La metodologia proposta comprende $S = 7$ indici statistici: media, deviazione standard, skewness, kurtosis, momento quinto, momento sesto, entropia.

I valori degli indici statistici calcolati per ogni grado di risoluzione per tutti gli istanti K del batch sono allocati in una matrice $\mathbf{H}[I \times (S \cdot 2R \cdot J) \times K]$ eseguendo un *unfolding variable-wise* e si costruisce un modello MPLS. Si ottiene in questo modo un modello globale in grado di descrivere l'intera durata del processo e che non necessita di sincronizzazione dei batch. Inoltre, i modelli ISMR globali possono calcolare i valori degli indici statistici utilizzando tutte le misure precedenti per ogni instante di campionamento k senza richiedere una struttura multi-modello, in quanto il numero dei descrittori statistici resta sempre il medesimo per ogni valore di k.

1.4.2.1 Modello ISMR locale

Per migliorare ulteriormente le capacità predittive dei i modelli ISMR, il modello MPLS globale viene utilizzato per lo sviluppo in tempo reale di un modello locale.

Il modello ISMR locale, differentemente da tutti i modelli descritti nei paragrafi precedenti, viene sviluppato in tempo reale durante il monitoraggio di processo. Lo sviluppo del modello ISMR locale avviene si articola in tre passaggi: 1) la proiezione del nuovo campione nello spazio degli scores definito dal modello ISMR globale; 2) l'individuazione e la selezione dei campioni più simili a quello dell'istante corrente (cioè più vicini nello spazio degli score); 3) lo sviluppo di un modello PLS utilizzando come regressori solo gli indici statistici contenuti nei campioni selezionati.

Nello spazio degli scores l'identificazione dei campioni più vicini alla misurazione corrente avviene seguendo una metodologia *nearest neighbor*: quando una nuova misura viene proiettata nell'iperspazio degli scores, viene calcolata la distanza euclidea da tutti i punti utilizzati in calibrazione e vengono selezionati solo gli N_N valori più vicini.

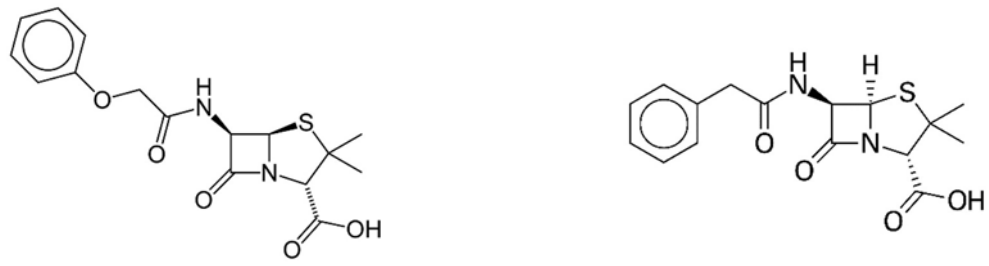
Capitolo 2

Processo di produzione della Penicillina

In questo capitolo viene presentata una descrizione riassuntiva del processo di produzione della penicillina su scala industriale, processo per il quale è stato sviluppato il sensore virtuale oggetto di questo lavoro. Viene inoltre riportata una descrizione del simulatore di processo utilizzato per raccogliere i dati per sviluppo del sensore virtuale.

2.1 Penicillina

La penicillina è stato il primo antibiotico isolato e prodotto a livello industriale. Più in generale con il termine penicillina si intende la classe di antibiotici caratterizzata dalla presenza di un anello beta lattamico nella loro struttura (Figura 2.1). La famiglia delle penicilline è composta quindi da diverse varianti della molecola di base, tra cui la penicillina V (per uso orale), penicillina G (per uso intravenoso), la penicillina oltre a molti altri derivati sintetici e semisintetici come l'ampicillina e l'amoxicillina. Attualmente, globale di penicillina è di circa 25 mila tonnellate annue, di cui il 70% viene venduto direttamente sia come penicillina V per somministrazione orale sia come penicillina G per uso animale o come sale sterile, il rimanente viene convertito in amoxicillina e ampicillina via 6-APA (acido (+)-6-aminopenicillanico).



(a) (b)

Figura 2.1. *Strutture delle penicilline più comuni,* (a) *struttura della penicillina V (per uso orale) di formula chimica* $C_{16}H_{18}N_2O_5S$*, e* (b) *della penicillina G (per uso intravenoso) di formula chimica* $C_{16}H_{18}N_2O_4S$*.*

Industrialmente la penicillina viene prodotta mediante la fermentazione di funghi del genere *Penicillium*, tra cui vengono utilizzati principalmente il *penicilllium chrysogenum* e alcune sue

mutazioni Questo microorganismo biosintetizza la molecola e la secerne come metabolita secondario.
Il processo avviene principalmente in reattori batch o fed-batch in cui la crescita attiva della popolazione microbiologica viene inibita; in condizioni di crescita normali infatti, i funghi del tipo penicillium non producono penicillina.
Il mezzo di fermentazione detto anche brodo di coltura, contiene tipicamente una fonte organica di azoto, carboidrati di fermentazione (saccarosio, fruttosio, glucosio) e carbonato di calcio come tampone. Deve inoltre essere fornito all'ambiente di reazione ossigeno, evitando di eccedere oltre il 40% della concentrazione di saturazione. Ciò corrisponde a una richiesta volumetrica di ossigeno di circa 0.4 - 0.8 mmol/(L·min). Il pH e la temperatura di coltivazione, specifiche per ogni microorganismo, sono abitualmente compresi tra i valori di 5-7 e 23°-28° C, rispettivamente. Il brodo di coltura viene inoltre agitato. In queste circostanze la massima resa in penicillina rispetto al glucosio fornito è di 0.12 g penicillina/g glucosio.
Nella produzione industriale del principio attivo il processo di fermentazione viene seguito da processi di estrazione, filtrazione e purificazione.

2.2 Simulatore

Il processo di produzione della penicillina è stato simulato utilizzando il software *IndPenSim* basato su un modello a principi primi (Goldrick et al., 2015). I dati raccolti durante le simulazioni costituiscono il punto di riferimento per lo sviluppo di modelli empirici in grado di stimare le variabili di qualità del processo.

2.2.1 Modello

Una fermentazione in sospensione in un liquido coinvolge fenomeni che avvengono in tre fasi: gas, liquido e solido. Per descrivere il processo di fermentazione si richiede quindi l'applicazione del principio di conservazione della massa per ognuno dei componenti coinvolti in ognuna delle tre fasi. Nel processo di produzione della penicillina: i nutrienti sono presenti nella fase liquida; la fase solida è principalmente composta dalla biomassa; la fase gas invece è richiesta per l'apporto di ossigeno necessario per la crescita aerobica dei microorganismi e per la rimozione della CO_2 prodotta durante la fermentazione. Ognuna delle tre fasi viene considerata come perfettamente mescolata.
I modelli a principi primi definiti per descrivere la fermentazione all'interno di bioreattori variano da modelli strutturati molto complessi che considerano la struttura interna del fungo *Penicillium Chrysogenum*, a modelli più semplici non strutturati, basati sulle espressioni cinetiche di crescita della concentrazione di penicillina e della biomassa.
Il modello su cui è stato sviluppato il simulatore di scala industriale *IndPenSim* è un'estensione del modello sviluppato da (Paul & Thomas, 1996), modello strutturato in grado di considerare

l'evoluzione della struttura interna del *Penicillium chrysogenum fungus* (fungo il cui metabolismo produce la penicillina) durante il processo. Si riportano in Figura 2.2 lo schema della biomassa e, di seguito, i bilanci di materia relativi a ogni regione del *P. chrysogenum* in fase liquida.

Nel modello la struttura interna del fungo viene suddivisa in quattro regioni separate,(Paul & Thomas, 1996):

- regione di crescita attiva A_0;
- regione di non-crescita A_1;
- regione degenerata A_3;
- regione formata attraverso vacuolizzazione e autolisi A_4.

Per ognuna di esse viene applicato un bilancio di conservazione della materia.

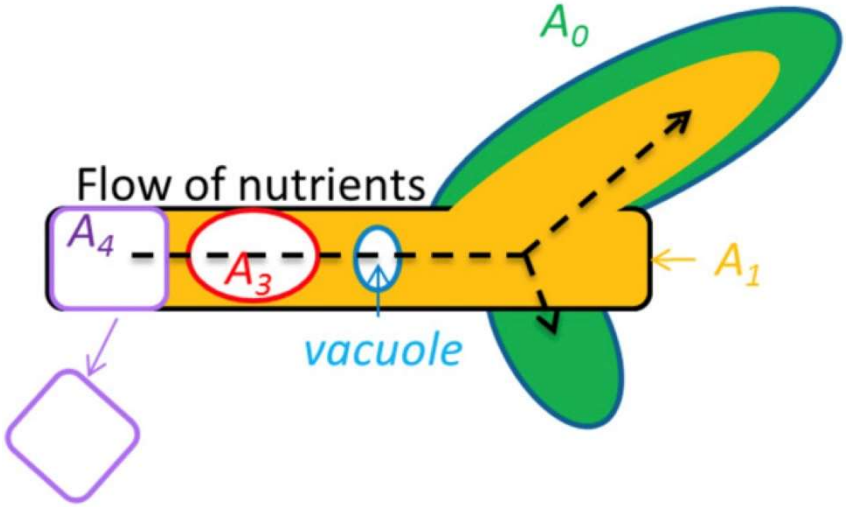


Figura 2.2. *Schema della struttura della biomassa rappresentata nel modello strutturato utilizzato all'interno del simulatore Indpensim,*(Goldrick et al., 2015)*. Si notano la regione di crescita attiva A_0, regione di non-crescita A_1, regione degenerata A_3, regione formata attraverso vacuolizzazione e autolisi A_4 e il flusso di nutrienti attraverso la cellula.*

Regione di crescita attiva (A_0),

$$\frac{dA_0}{dt} = r_b - r_{diff} - \frac{F_{in}A_0}{V} \tag{2.1}$$

Regione di non-crescita (A_1),

$$\frac{dA_1}{dt} = r_e - r_b + r_{diff} - r_{deg} - \frac{F_{in}A_1}{V} \tag{2.2}$$

Regione degenerata (A_3),

$$\frac{dA_3}{dt} = r_{deg} - r_a - r_{diff} - \frac{F_{in}A_3}{V} \tag{2.3}$$

Regione formata attraverso l'autolisi (A_4),

$$\frac{dA_4}{dt} = r_a - \frac{F_{in}A_4}{V} \tag{2.4}$$

Biomassa totale,

$$B_x = A_1 + A_2 + A_3 + A_4 \tag{2.5}$$

Formazione del prodotto, penicillina (P),

$$\frac{dP}{dt} = r_p - r_h - \frac{F_{in}P}{V} \tag{2.6}$$

Consumo di substrato (s),

$$\frac{ds}{dt} = -Re_{s/B_x}\, r_e - Re_{s/B_x}\, r_b - m_s r_m - Y_{s/P}\, r_P + \frac{F_s c_s}{V} + \frac{F_{oil} c_{oil}}{V} - \frac{F_{in} s}{V} \tag{2.7}$$

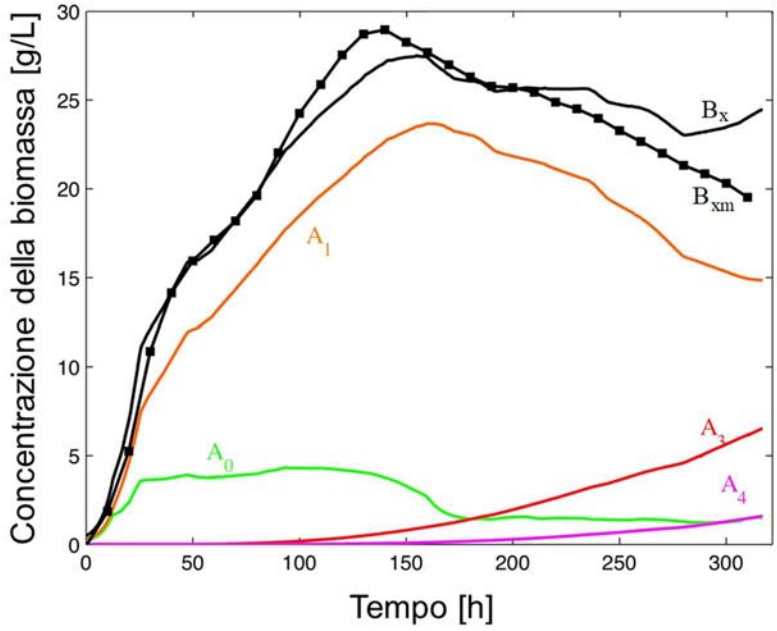
dove, r_b è la cinetica di ramificazione, r_{diff} è la cinetica di differenziazione, r_e è la cinetica di elongazione, r_{deg} è la cinetica di degenerazione, r_aè la cinetica di autolisi, r_Pè la cinetica di formazione di prodotto, r_h è la cinetica di idrolisi del prodotto e r_m è la cinetica di mantenimento della biomassa. Il tempo del batch è rappresentato da t. Re_{s/B_x} e $Re_{s/P}$ rappresentano le rese di substrato in biomassa e di substrato in prodotto, m_s è il termine di mantenimento della biomassa. F_s , F_{oil} , c_s , e c_{oil} rappresentano le portate di alimentazione di zucchero e olio di soia e le relative concentrazioni. Per semplicità l'addizione di olio di soia è stata combinata con quella di zucchero per formare un unico substrato s. F_{in} rappresenta l'insieme di tutte le alimentazioni del processo.

Il volume del fermentatore varia secondo il bilancio,

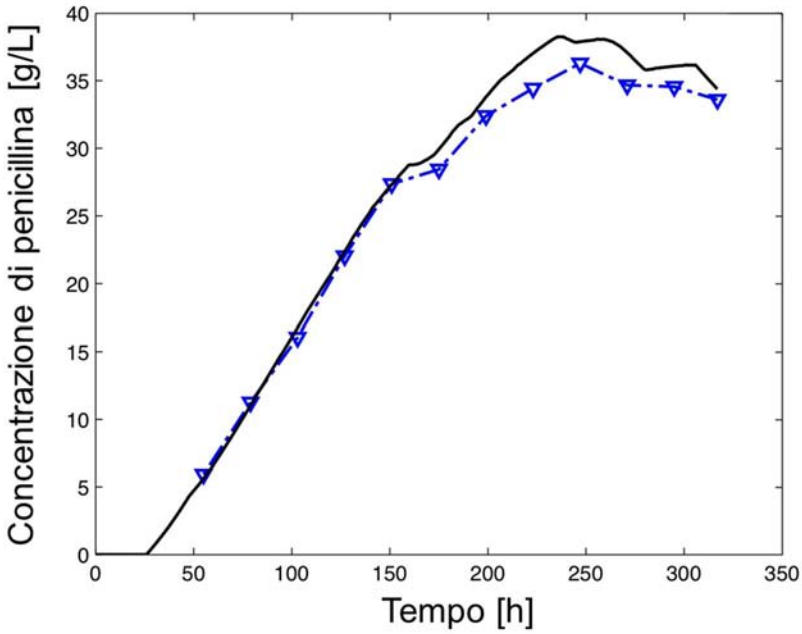
$$\frac{dV}{dt} = F_s + F_{oil} + F_{PAA} + F_a + F_b + F_w - F_{evp} - F_{dis} \tag{2.8}$$

dove F_{PAA} è la portata di acido fenilacetico e F_a e F_b sono le portate di acido e base utilizzate per controllare il pH. F_wè la portata di acqua di processo la quale viene tipicamente usata per ridurre la viscosità del brodo di coltura. F_{evp} è la portata in uscita per evaporazione dal volume di liquido del fermentatore e F_{dis} è la portata di scarico del reattore che viene saltuariamente impostata diversa da zero per mantenere il volume del brodo di coltura all'interno del reattore nei limiti previsti.

In Figura 2.3 (a) viene riportata una rappresentazione dei tipici profili di concentrazione nel tempo delle quattro regioni che compongono la biomassa: regione di crescita attiva A_0 in verde, regione di non-crescita A_1in arancione, regione degenerata A_3 in rosso e regione formata attraverso vacuolizzazione e autolisi A_4 in viola. In linea continua nera sono riportati i valori reali della concentrazione della biomassa B_x a confronto con i valori determinati attraverso il simulatore di processo *IndPenSim*, B_{xm}. In Figura 2.3 (b) viene riportato il confronto tra i valori delle concentrazioni di penicillina reali per un batch rappresentativo, in linea nera continua, e la concentrazione stimata attraverso il simulatore *IndPenSim*, in linea blu.



(a)



(b)

Figura 2.3. *Profilo delle concentrazioni nel tempo delle quattro regioni della biomassa* (a)*, regione di crescita attiva (A_0), regione di non-crescita (A_1), regione degenerata (A_3), regione formata attraverso vacuolizzazione e autolisi (A_4), della concentrazione totale della biomassa stimata attraverso il modello (B_{xm}) e reale (B_x);* (b) *profilo nel tempo della concentrazione della penicillina reale (linea continua),*(Goldrick et al., 2015).

2.2.1.1 Sistema di controllo di temperatura

Le fluttuazioni di temperatura hanno grande influenza sull'attività microbiologica per cui il controllo della temperatura è di fondamentale importanza in bioreattori di scala industriale.
Il profilo dinamico della temperatura viene calcolata nel modello mediante un bilancio di energia su tutti gli ingressi e le uscite nel bioreattore:

$$\frac{dT_b}{dt} = \frac{1}{VC_{P_b}\rho_b}\left(F_s\rho_s C_{P_s}(T_s - T_b) + F_w\rho_w C_{P_w}(T_w - T_b) - \Delta H_{evp}\rho_w F_{evp} + P_w - U_{jacket}A_w(T_b - T_{air}) - Q_c + Q_{xrn}\right) \quad (2.9)$$

dove C_{Pb} è il calore specifico del brodo di coltura; C_{Ps} è il calore specifico del substrato e C_{Pw} è il calore specifico dell'acqua di processo; T_s è la temperature delle portata di alimentazione di substrato; T_w è la temperature delle portata di alimentazione acqua di processo, T_b è la temperature del brodo di coltura; F_{evp} è la portata evaporata all'interfaccia tra la fase liquida e gas nel reattore. P_w è il calore generato come risultato della agitazione meccanica e di quella dovuta all'aerazione. Il calore scambiato attraverso la superficie del reattore con l'ambiente viene calcolato assumendo una differenza di temperature tra quella dell'aria T_{air} e quella del bioreattore T_b utilizzando un coefficient di scambio termico U_{jacket} e superficie del reattore A_w costanti(Goldrick et al., 2015). Q_c é il calore rimosso attraverso la serpentina per il raffreddamento interno e Q_{rxn} è il calore di reazione prodotto dalla crescita della popolazione microbiologica e dal loro metabolismo.

Per una descrizione più approfondita del modello attraverso le equazioni di bilancio relative a tutte le specie presenti nel fermentatore, le equazioni per il calcolo di viscosità e pH, la derivazione completa del calore prodotto dalla reazione e scambiato con il circuito di raffreddamento interno, si rimanda alla lettura del lavoro di Goldrick *et al.* (2015).

2.2.2 Organizzazione del simulatore

Il simulatore IndPenSim è costituito da un insieme di 14 MATLAB R2018b, di cui in Figura 2.4 viene riportato lo schema: il file principale per avviare la simulazione è `Generate_Production_Batch_data_V4.m`. Ad ogni simulazione devono essere obbligatoriamente specificate quattro opzioni:
1) presenza o meno di guasti e loro tipologia nei processi che si vogliono simulare;
2) tipologia di controllo dei processi (su ricetta o su controllo dell'operatore);
3) tempo di produzione dei singoli batch (tempo fisso o variabile);
4) utilizzo della spettroscopia Raman per il controllo di processo.
Nel file `Indpensim_Run.m` inoltre possono venire specificate alcune opzioni aggiuntive per il processo simulato, come le condizioni di inibizione della crescita microbiologica, i disturbi nelle portate e concentrazioni di alimentazioni, il verificarsi di irregolarità dovute alla presenza

di guasti come un'improvvisa interruzione dell'aerazione, un aumento repentino di pressione all'interno del reattore, una interruzione delle portate di substrato, base (per il controllo del pH), acqua di raffreddamento oppure guasti nei sensori di pH e temperatura. È possibile ancora definire l'intervallo di tempo nel raccogliere le misure off-line è il loro tempo morto di risposta dovuto alle tempistiche di analisi in laboratorio.

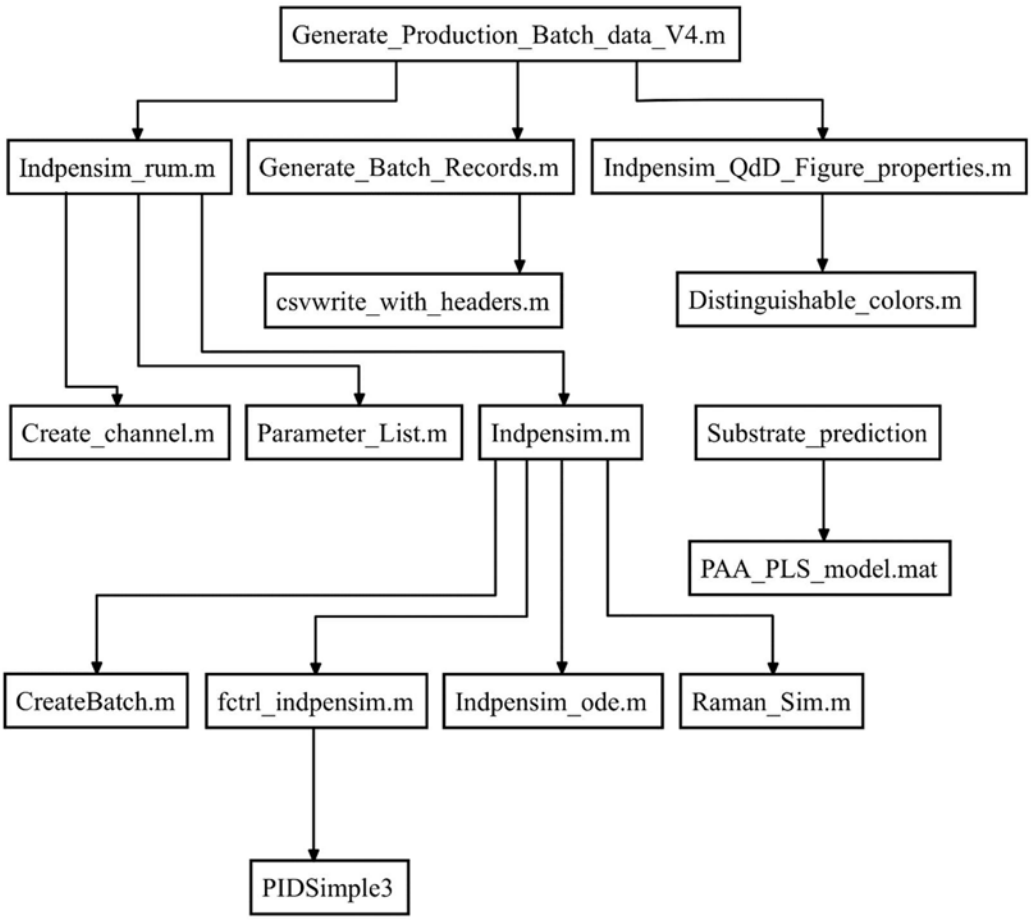


Figura 2.4. *Schema di funzionamento degli script MATLAB del simulatore Indpensim. Il software è composto da un totale di 14 script il cui titolo viene riportato nei riquadri. Si osserva come il file principale* `Generate_Production_Batch_data_V4.m` *sia in grado di avviare la simulazione.*

I file `Indpensim_QdD_Figure_properties.m` e `Distinguishable_colors.m` permettono la gestione dei grafici generati dal simulatore.

I file `Generate_Batch_Records.m`, `csvwrite_withheaders.m`, `Create_channel.m` e `CreateBtach.m` creano le strutture di in cui vengono salvati i dati. Nel file `Parameter_List.m` sono gestiti i parametri del processo. Nel file `fctrl_indepensim.m` possono essere gestiti nel dettaglio i guasti simulabili durante il processo e i disturbi sulle alimentazioni che sono normalmente presenti nei processi industriali. `PIDSimple.m` gestisce l'azione del controllore PID per il controllo di temperatura e pH. `Raman.sim.m` genera la

simulazione dello spettro Raman del brodo di coltura. `Indpensim.m` e `Indpensim_ode.m` risolvono numericamente le equazioni differenziali come i bilanci di massa ed energia del modello. `Substrate_prediction` e `PAA_PLS_model.mat` gestiscono il modello PLS che correla lo spettro Raman a una misura quantitativa dell'acido fenilacetico nel reattore.

2.3 Processo

Il processo di fermentazione è condotto in condizioni fed-batch, e segue una predeterminata ricetta per l'alimentazione di nutrienti.
Questo modo di operare permette di ottenere un controllo più efficace sulle diverse fasi di vita della biomassa come ad esempio la crescita attiva, la richiesta di nutrienti e la produzione del metabolita considerato come prodotto principale del processo. L'alimentazione di nutrienti segue quindi il profilo ottimale per aumentare la produzione di penicillina; limitando l'apporto di substrato. Infatti, viene favorita la produzione di metaboliti secondari e limitata la crescita della biomassa. Durante il processo viene inoltre previsto di eseguire saltuariamente lo scarico del contenuto del reattore seguendo una cadenza prestabilita, ciò permette di mantenere il livello all'interno del reattore entro il limite stabilito e di condurre processi di maggiore durata.

2.3.1 Reattore e condizioni operative

Il processo avviene in un reattore di scala industriale del volume di V_r=1·10^5 L, e raggio r = 2.1 m. Il mescolamento viene fatto utilizzando tre agitatori Rushton coassiali di raggio interno r_{imp}= 0.85 m. Le operazioni sono condotte con una velocità di agitazione fissa di 100 rpm.

Il reattore è equipaggiato con sensori per la misura in linea di: pH; temperatura T; ossigeno disciolto DO_2; pressione; presenza di schiuma; concentrazione di ossigeno $O_{2,og}$ e anidride carbonica $CO_{2,og}$ della fase gassosa, monitorate analizzando la corrente in uscita (off-gas) posta nell'estremità superiore dell'apparecchiatura; alimentazione di zucchero (substrato) F_s, fonte principale di carbonio per la biomassa; alimentazioni di olio di soia F_{oil} utilizzato come antischiuma e fonte secondaria di carbonio; portata di areazione F_g; portata di acido fenilacetico F_{PAA}; portata di acqua di raffreddamento F_c; portata di acido F_a; portata di base F_b; peso del reattore W; porta di scarico F_{dis}. Le misure in linea sono disponibili ogni 12 minuti.
La portata di acqua di raffreddamento F_c, la portata di acido F_a e la portata di base F_b sono usate per controllare pH e la temperatura e mantenerli rispettivamente a 6.5 e 298 K mediante l'uso di un controllore PID.
La porta di scarico F_{dis}, in accordo con il modello fed-batch, assume valore zero per la maggior parte della durata del processo.
Sono eseguite, inoltre, alcune misure fuori linea come la misura della concentrazione di penicillina c_P, azoto c_N , misura della viscosità del mezzo, e della concentrazione di acido

fenilacetico c_{PAA} usato come precursore per la penicillina. L'azoto è presente nel mezzo inizialmente introdotto nel batch (inoculo) e viene monitorato durante il processo attraverso misure fuori linea. Le misure fuori linea vengono effettuate ogni 4 ore e la risposta dello strumento di analisi è disponibile con un ritardo di altre 4 ore.
Uno schema del reattore è riportato in Figura 2.3.

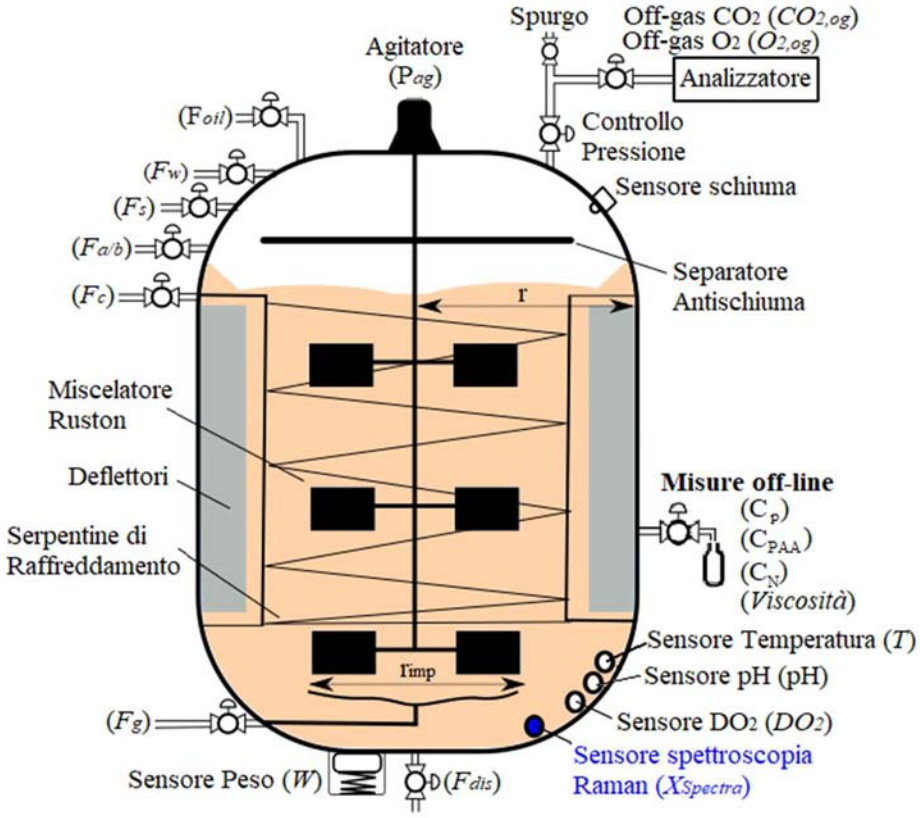


Figura 2.5. *Schema del fermentatore utilizzato nel processo fed-batch di scala industriale per la produzione di penicillina utilizzato nel simulatore Indpensim, Goldrick* et al. *(2015). Vengono riportate in modo schematico anche le portate e i sensori principali collegati al reattore.*

Sono disponibili anche delle variabili calcolate. Le misure di ossigeno e anidride carbonica nella portata di scarico dei gas dal reattore permettono di calcolare due parametri molto importanti per il monitoraggio della fermentazione in tempo reale: il consumo di ossigeno (*OUR*, *Oxygen Upatke Rate*) e il consumo di carbonio *(CER, Carbon Evolution Rate)* da parte della biomassa; il calcolo viene fatto mediante le Equazioni seguenti:

$$OUR = \frac{32}{22.4} F_{g_{in}} \left(O_{2_{in}} - O_{2_{out}} \frac{N_{2_{in}}}{1 - O_{2_{out}} - CO_{2_{in}}} \right) \tag{2.9}$$

$$CER = \frac{44}{22.4} F_{g_{in}} \left(CO_{2_{out}} \frac{N_{2_{in}}}{1 - O_{2_{out}} - CO_{2_{out}}} - CO_{2_{in}} \right) \tag{2.10}$$

Tabella 2.1. *Variabili di processo, produzione di penicillina in un fermentatore di scala industriale.*

numero	nome	simbolo	unità di misura	misura	tipo variabile
1	portata di aerazione	F_g	L/h	on-line	manipolata da ricetta
2	portata olio di soia	F_{oil}	L/h	on-line	manipolata da ricetta
3	portata substrato	F_s	L/h	on-line	manipolata da ricetta
4	portata acido fenilacetico	F_{PAA}	L/h	on-line	manipolata da ricetta
5	portata acqua raffreddamento	F_c	L/h	on-line	manipolata attraverso PID
6	portata acqua riscaldamento	F_h	L/h	on-line	manipolata attraverso PID
7	portata di acido	F_a	L/h	on-line	manipolata attraverso PID
8	porta di base	F_b	L/h	on-line	manipolata attraverso PID
9	portata di scarico	F_{dis}	L/h	on-line	manipolata da ricetta
10	portata totale alimentata	F_{in}	L/h	on-line	manipolata
11	portata acqua di processo	F_w	L/h	on-line	manipolata da ricetta
12	raggio reattore	r	m	fissa	-
13	raggio agitatore	r_{imp}	m	fissa	-
14	volume reattore	V_r	L	fissa	-
15	volume processato	V	L	on-line	manipolata da ricetta
16	pH	pH	pH	on-line	monitorata
17	temperatura	T	K	on-line	monitorata
18	concentrazione ossigeno disciolto	$DO2$	mg/L	on-line	monitorata
19	pressione	$pressione$	bar	on-line	monitorata
20	concentrazione di penicillina	c_P	g/L	off-line	monitorata
21	concentrazione di azoto	c_N	g/L	off-line	monitorata
22	viscosità	$viscosità$	cP	off-line	monitorata
23	concentrazione di acido fenilacetico	c_{PAA}	g/L	off-line	monitorata
24	concentrazione di O_2 nei gas	$O_{2,og}$	% vol.	on-line	monitorata
25	concentrazione di CO_2 nei gas	$CO_{2,og}$	% vol.	on-line	monitorata
26	concentrazione della biomassa	B_x	g/L	off-line	monitorata
27	consumo di carbonio	CER	g/h	on-line	monitorata (derivata)
28	consumo di ossigeno	OUR	g/min	on-line	monitorata (derivata)

Capitolo 3

Sensore virtuale per la stima in tempo reale della concentrazione di penicillina in fermentazioni fed-batch

In questo Capitolo si presenta la metodologia proposta per lo sviluppo di sensori virtuali, basata su modelli di regressione a variabili latenti costruiti su indici statistici multi-risoluzione. Inoltre, si riporta un confronto tra la metodologia proposta e alcune tra le metodologie di letteratura maggiormente efficaci e affidabili. Le prestazioni predittive dei modelli sono testate nel caso di studio della stima in tempo reale della concentrazione di penicillina in processi fermentativi.

3.1 Misure di processo

In questa Tesi vengono valutati due scenari per il soft-sensing:

- Scenario #1: batch della stessa durata e sincronizzati;
- Scenario #2: batch di durata differente.

3.1.1 Scenario #1: batch della stessa durata

Nel caso in cui si considerino batch della stessa durata e sincronizzati il dataset utilizzato (Dataset-1) è composto da 120 batch di fermentazione per la produzione di penicillina della durata di 230 ore simulati mediante il software IndPenSim (Goldrick et al., 2015). I batch del dataset rappresentano processi in condizioni operative normali (*normal opertive condition*, NOC), in cui la quantità di penicillina raccolta supera la specifica di produzione, fissata a 2000 Kg.

Le variabili di processo considerate sono undici (Tabella 3.1), e sono variabili monitorate in linea durante la fermentazione. La frequenza di registrazione è di una misura ogni 12 minuti; durante un batch si ottengono quindi 1150 misurazioni. Le misure delle undici variabili di processo monitorate in linea durante la fermentazione vengono disposte in una matrice tridimensionale $\underline{\mathbf{X}}$ $[144 \times 11 \times 1150]$.

La concentrazione di penicillina viene monitorata fuori linea con una frequenza di monitoraggio di una misura ogni 4 ore; perciò nelle 230 ore di processo vengono effettuate 59 misurazioni. La matrice di dati della variabile di qualità diventa così $\underline{\mathbf{Y}}$ $[120 \times 1 \times 59]$.
Tutti i modelli riportati nei paragrafi che seguono sono stati sviluppati utilizzando i 120 batch considerati in condizioni normali. Per la convalida del sensore virtuale si è utilizzata una procedura Monte Carlo su 25 iterazioni in cui il dataset viene suddiviso in due parti mediante una selezione casuale dei batch in: 100 batch utilizzati per calibrare il modello PLS e 20 batch utilizzati per la convalida.

Tabella 3.1. *Elenco delle variabili del processo di fermentazione misurate in linea e utilizzate come predittori nei sensori virtuali per la stima in tempo reale della penicillina.*

numero	nome	simbolo	unità di misura
1	portata acido	F_a	L/h
2	portata base	F_b	L/h
3	portata acqua raffreddamento	F_c	L/h
4	portata acqua riscaldamento	F_h	L/h
5	concentrazione ossigeno disciolto	$DO2$	mg/L
6	volume processato	V	L
7	peso	W	kg
8	pH	pH	[-]
9	temperatura	T	K
10	concentrazione di CO_2 nei gas	$CO_{2,og}$	% vol.
11	concentrazione di O_2 nei gas	$O_{2,og}$	% vol.

3.1.2 Scenario #2: batch di durate diverse

Anche nel caso di batch di durata differente viene considerato un dataset (Dataset-2) di 120 batch per la produzione di penicillina, simulati sempre mediante il software IndPenSim; la durata minima è di 123 ore e 12 minuti, mentre la durata massima è di 187 ore e 12 minuti.
Le variabili di processo considerate sono undici e coincidono con quelle di Tabella 3.1. La frequenza di registrazione è di una misura ogni 12 minuti; le misure delle undici variabili di processo monitorate in linea durante la fermentazione vengono disposte in una matrice tridimensionale $\underline{\mathbf{X}}$ $[120 \times 11 \times \Omega]$ in cui il numero di osservazioni varia da Ω=634 a Ω=975 in relazione alla durata del batch considerato.
La concentrazione di penicillina viene monitorata fuori linea con una frequenza di monitoraggio di una misura ogni quattro ore; la matrice di dati della variabile di qualità diventa così $\underline{\mathbf{Y}}$ $[120 \times 1 \times \eta]$, in cui η varia da 33 a 50 e dipende dalla durata del batch considerato.
In Figura 3.1 vengono riportati tre batch rappresentativi del Dataset-2; si riportano il batch di minore durata (linea nera continua), un batch di durata intermedia (linea nera tratteggiata) e il batch di maggiore durata (linea grigia continua).

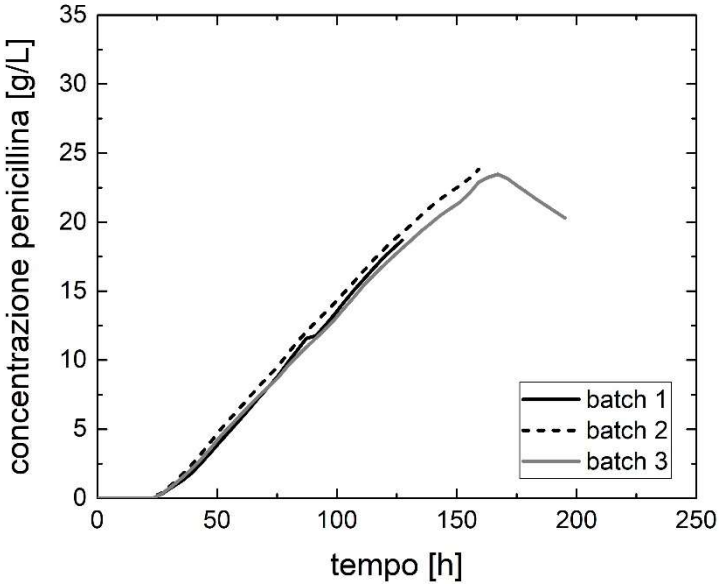


Figura 3.1. *Batch di durate diverse: profili della concentrazione di penicillina per tre batch: in linea nera continua il batch di minor durata (126 ore e 48 minuti); in linea nera tratteggiata un batch di durata intermedia (159 ore e 12 minuti); in linea grigia continua il batch di durata maggiore (195 e 12 minuti).*

3.2 Modello ad indici statistici multi-risoluzione

Il modello proposto in questa Tesi è un modello ad indici statistici multi-risoluzione ISMR (i cui dettagli matematici si trovano nel Paragrafo 1.4). In questa strategia modellistica i profili delle misure ad ogni istante di tempo $k = 1, 2, \ldots, K$ vengono decomposti in sei risoluzioni utilizzando la wavelet Daubachies-2. Quindi, i regressori del modello PLS globale sono 924 indici statistici:

$$11 \text{ variabili} \times [6 \text{ risoluzioni} \cdot (1 \text{ approssimazione} + 1 \text{ dettaglio})] \times 7 \text{ indici stattistici} = 924 \text{ regressori}$$

Secondo la metodologia introdotta al Paragrafo 1.4.2. Il modello ISMR globale (ISMR-g1) si basa su un *variable-wise unfolding* dei dati e gli indici statistici permettono di riassumere le informazioni contenute nei profili temporali delle variabili, comprimendone la dimensione tempo ad un'unica dimensione per tutti i gradi di risoluzione.

Indipendentemente dalla durata di un batch e dal numero di osservazioni disponibili, la metodologia ISMR permette di mantenere invariato il numero di regressori utilizzati nel modello. In questo modo non vi è la necessità di una sincronizzazione quando si considerano batch di durate diverse.

3.2.1 Prestazioni del sensore virtuale nello Scenario #1

Nello Scenario #1 si utilizza il Dataset-1 in cui i batch sono tutti della stessa durata e sono caratterizzati da 59 osservazioni della variabile di qualità. Si ottiene una matrice di calibrazione $\mathbf{X}$ $[5900 \times 924]$.

3.2.1.1 Modello globale

Il modello PLS globale ISMR-g1 utilizza 6 variabili latenti in grado di spiegare: in calibrazione una variabilità cumulata della **X** del 49.32% e della **Y** del 98.74%; in convalida con RMSEP = 1.45 e $R^2 = 0.98$, quindi prestazioni di stima assolutamente soddisfacenti nella procedura Monte Carlo descritta sopra. Le variabili latenti considerate nel modello sono state scelte attraverso una metodologia di convalida incrociata.

Per una valutazione più approfondita delle prestazioni predittive del modello ISMR-g1 vengono inoltre utilizzate le seguenti quattro metriche: *i*) l'errore relativo medio (ErR_{totale}) calcolato su tutti i batch e tutte le iterazioni; *ii*) l'errore relativo ($ErR_{parziale}$) medio calcolato su tutte le iterazioni, ma solo sui batch contemporaneamente dentro i limini di confidenza del 95% di T^2 e SPE; *iii*) il rapporto tra l'errore assoluto medio sulla deviazione standard calcolato per i batch entro i limiti di confidenza del 95 % di T^2 e SPE (ErA/σ); *iv*) la percentuale media di batch considerati oltre i limiti di confidenza del 95 % di T^2 o SPE, definiti come batch di cui la stima non è affidabile perché troppo discostati dalla media o dalla struttura di correlazione del dataset di calibrazione. Di conseguenza, la differenza tra ErR_{totale} e $ErR_{parziale}$ permette di definire la perdita di prestazioni predittive del modello per i batch la cui stima è ritenuta inaffidabile. Quindi, ErR_{totale}, $ErR_{parziale}$, ed ErA/σ forniscono una valutazione dell'accuratezza del modello, mentre la percentuale di batch oltre i limiti di confidenza del 95 % di T^2 o SPE e la differenza tra ErR_{totale} e $ErR_{parziale}$ forniscono una valutazione dell'affidabilità del modello.

Il modello ISMR-g1 presenta un ErR_{totale}[1]=13.9 %, un $ErR_{parziale}$[1] = 13.41 %, un ErA/σ = 1.30, e una percentuale di batch anomali del 20.32 %. Quindi, un modello globale, che consideri l'intero dataset **X** per la costruzione del modello PLS, può essere costruito ottenendo una accuratezza accettabile e una buona affidabilità nelle predizioni. Si ritiene comunque il modello ISMR-globale sia troppo ampio per descrivere con accuratezza le diverse fasi dei batch considerati. Si ricerca quindi la costruzione di un modello più specifico che possa migliorare le prestazioni del modello ISMR-g1.

3.2.1.2 Modello locale

Si sviluppa per queste ragioni un modello locale (ISMR-loc1) mediante l'utilizzo una strategia *nearest neighbor* come quella descritta nel Paragrafo 1.4.2.1. Si richiama che il modello locale

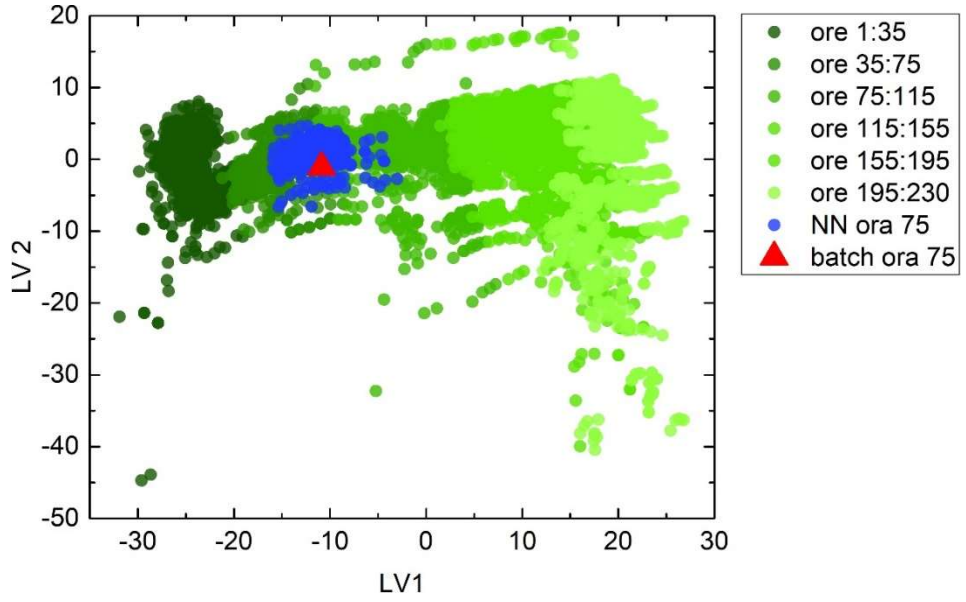
[1] Valore calcolato sulle osservazioni successive l'ora 31 e 12 minuti dall'inizio del batch, le prime trenta ore vengono escluse dal calcolo perché l'errore relativo assume valori molto elevati, in quanto il valore della concertazione di penicillina nella prima fase del batch è trascurabile (dell'ordine di 0.001 g/L).

funziona nel modo seguente: 1) gli indici statistici multi-risoluzione vengono estratti per il batch di convalida; 2) questa osservazione viene proiettata sul modello globale e si selezionano le 300 osservazioni più simili; 3) si calibra un modello PLS locale sui nearest neighbor selezionati al punto 2 e si esegue la stima mediante l'Equazione (1.30).

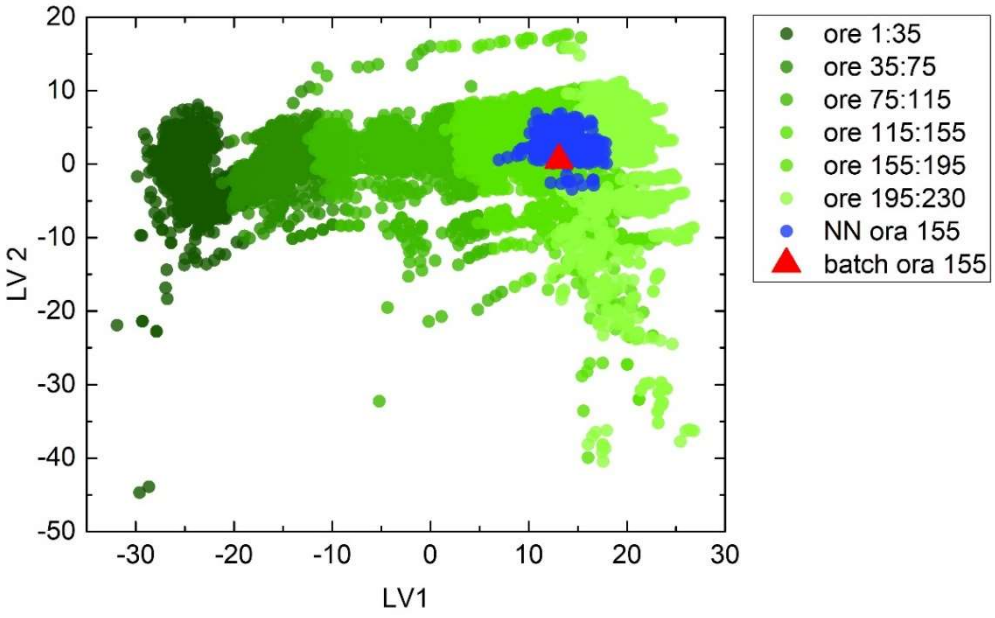
Il modello ISMR-loc1 su A = 6 variabili latenti in grado di spiegare: in calibrazione una variabilità cumulata della **X** del 48.29% e della **Y** del 98.82%; in convalida con RMSEP = 1.64 e R^2 = 0.97, quindi prestazioni di stima assolutamente soddisfacenti nella procedura Monte Carlo descritta sopra. Nonostante i valori di RMSEP e R^2 risultano dello stesso ordine di grandezza che nel modello precedente e il valore dei batch fuori dai limiti di confidenza salga lievemente (26.86 %), nelle predizioni ritenute affidabili l'accuratezza del modello migliora significativamente, ErR_{totale} = 8.09%, e $ErR_{parziale}$ = 6.20%, un ErA/σ = 0.71. In particolare, $ErR_{parziale}$ viene dimezzato rispetto a quello del modello globale.

In Figura 3.2a si riportano in verde gli score delle osservazioni utilizzate per la costruzione del modello globale sulle prime due variabili latenti: con il verde scuro i punti relativi all'inizio del batch, con il verde chiaro le fasi seguenti del batch. Valori crescenti di LV1 e decrescenti di LV2 danno una indicazione nello score plot della strategia *variable-wise* del decorso del tempo. Un batch di convalida all'istante k = 376 (cioè, al tempo 75 ore e 12s), per cui a posteriori si viene a sapere che la concertazione di penicillina reale risulta y =10.35 g/L, viene proiettato nello spazio degli score (triangolo rosso) e vengono selezionate le 300 osservazioni più simili (le più vicine nello spazio degli score), i punti blu. Queste 300 osservazioni vengono quindi utilizzate per la costruzione del modello locale. La stima in tempo reale risulta essere $\hat{y} = 10.34\ \mathrm{g/L}$, assolutamente soddisfacente perché molto vicina al valore reale di y =10.35 g/L, determinando così un errore relativo di solo 0.14 %.

Analogamente, in Figura 3.2b in si evidenziano in blu le 300 osservazioni più simili all'osservazione dello stesso batch all'istante k = 761 (cioè al tempo 155 ore e 12 minuti); in questo caso in cui la concentrazione di penicillina reale è aumentata al valore di 25.12 g/L; la stima risulta essere assolutamente accurata $\hat{y} = 23.77\ \mathrm{g/L}$, determinando così un errore relativo di solo 5.37 %



(a)



(b)

Figura 3.2. *Modello ad indici statistici multi-risoluzione locale: score plot delle prime due LV. In diverse tonalità di verde gli score del modello globale: in verde scuro, nella parte sinistra del diagramma, gli score sono relativi alla prima fase del batch (dall'ora 1 all'ora 35), in verde chiaro nella zona a destra della figura gli score relativi alla fase finale del batch (dall'ora 1195 all'ora 230). Sono evidenziati in blu le 300 osservazioni individuate dalla strategia nearest neighbor per l'osservazione di un batch di convalida all'istante:* (a) *k = 376 corrispondente al tempo 75 ore e 12 dall'inizio del batch, che viene riportata in figura di colore rosso in forma triangolare;* (b) *k = 761 corrispondente al tempo 155 ore e 12 minuti dall'inizio del batch, che viene riportata in figura di colore rosso in forma triangolare.*

In entrambe le figure, si può osservare l'efficienza della strategia *nearest neighbor* nell'individuare gli score più vicini alla nuova osservazione (si sottolinea che gli score selezionati non sono in assoluto i più vicini, in quanto il modello utilizza A = 6 LV e la distanza viene quindi calcolata su un iperpiano di sei dimensioni, mentre in Figura 3.2 viene riportata solo la proiezione su piano descritto dalle prime due variabili latenti.

In Figura 3.3a viene riportata la distribuzione dei valori reali della concentrazione di penicillina corrispondente alle osservazioni selezionate secondo la strategia *nearest neighbors* per l'istante k = 376 al tempo 75 ore 12 minuti, in cui la concertazione di penicillina reale risulta essere y = 10.35 g/L, il valore predetto è $\hat{y}$=10.34 g/L, il valore medio della concentrazione di penicillina nelle osservazioni selezionate risulta essere 9.60 g/L, la deviazione standard è σ=1.88 g/L. Di qui si comprende anche come questa strategia, quindi, fornisca una procedura di stima che è in un certo modo probabilistica.

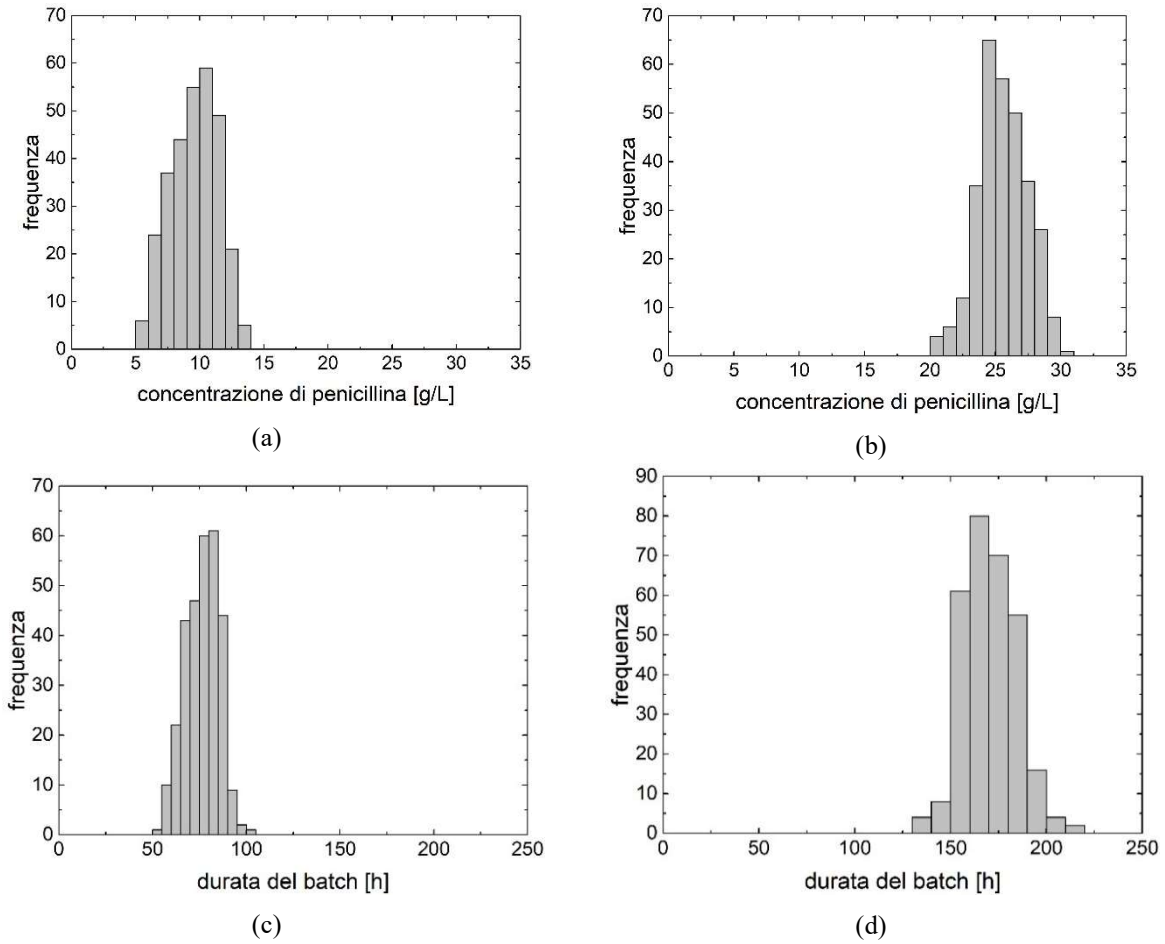
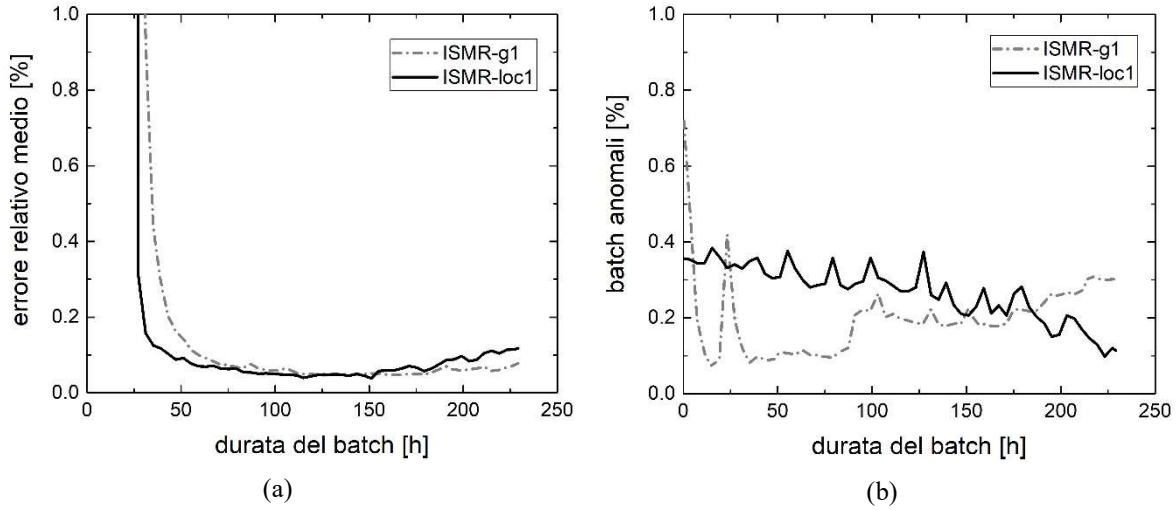


Figura 3.3. *Modello ad indici statistici multi-risoluzione locale: distribuzione della concentrazione di penicillina per i 300 campioni selezionati dalla strategia nears-neighbors per gli istanti:* (a) *k=376*; (b) *k=761; distribuzione dei tempi di osservazione per i 300 campioni selezionati dalla strategia nears-neighbor per l'istante k=376* (c) *e per l'istante k=761* (d).

In Figura 3.3b, in modo analogo, viene riportata la distribuzione dei valori reali della concentrazione penicillina delle osservazioni selezionate per l'istante k = 761 al tempo 152 ore e 15 minuti, in cui la concertazione di penicillina reale risulta essere y = 25.12 g/L, il valore predetto è $\hat{y}$=23.77 g/L il valore medio della concentrazione di penicillina nelle osservazioni selezionate risulta essere 25.56 g/L, la deviazione standard è σ = 1.90 g/L. In entrambi i grafici si osserva quindi che le osservazioni selezionate hanno valori della concentrazione di penicillina assolutamente adatti, perché molto simili, alla stima della concentrazione del batch corrente all'istante k-esimo.

Il modello locale permette infatti di utilizzare non solo le osservazioni più simili registrate in istanti (k = 1, 2, …, k_{cor}) precedenti a quello del batch corrente k_{cor}, ma anche le osservazioni successive (k = k_{cor}+1, k_{cor}+2, …, K). Questo aspetto viene evidenziato in Figura 3.3c, in cui viene riportata la distribuzione temporale delle osservazioni selezionate secondo la strategia *nearest neighbor* per l'istante k = 376 al tempo 75 ore e 12 minuti. Il valore medio dei tempi in cui vengono effettuate le registrazioni delle osservazioni selezionate risulta essere di 76 ore 34 minuti con una a deviazione standard è di σ = 8.88 ore. In Figura 3.3d, in modo analogo, viene riportata la distribuzione dei tempi in cui vengono effettuate le osservazioni selezionate secondo la strategia *nearest neighbor* per l'istante k = 761 al tempo 152 ore e 15 minuti. Il valore medio dei tempi in cui vengono effettuate le registrazioni risulta essere 170 ore e 26 minuti con una deviazione standard è di17 ore e 45 minuti.

In Figura 3.4a viene riportato il profilo di $ErR_{parziale}$ per i modelli ISMR-g1 e ISMR-loc1; si osserva che il modello locale migliora in accuratezza rispetto al modello globale praticamente su tutta la durata del batch. In Figura 3.4b viene riportato il profilo della percentuale di batch considerati anomali per i due modelli ISMR-g1 e ISMR-loc1. Si osserva che anche se il modello locale migliora in accuratezza viene leggermente peggiorata l'affidabilità del modello.



(a) (b)

Figura 3.4. *Modello ad indici statistici multi-risoluzione locale: Confronto tra dei profili dell'errore relativo medio* (a) *e della percentuale di batch considerati anomali* (B) *per i modelli ISMR-globale, ISMR-locale.*

3.2.2 Prestazioni del sensore virtuale nello Scenario #2

Nello Scenario #2 si utilizza il Dataset-2 in cui i batch non sono tutti della stessa durata e sono caratterizzati da η osservazioni della variabile di qualità.

3.2.2.1 Modello globale

Le prestazioni dei modelli ISMR vengono inoltre studiate nelle condizioni i cui batch considerati non siano di uguale durata, scenario #2. Utilizzando il dataset-2, la matrice di calibrazione che si ottiene risulta essere $[\delta \times 924]$ in cui il numero delle righe δ dipende dalla durata dei batch selezionati ad ogni iterazione della procedura di Monte Carlo.
Il modello globale (ISMR-g2) su $A = 6$ variabili latenti in grado di spiegare: in calibrazione una variabilità cumulata della **X** del 48.72% e della **Y** del 99.29%; in convalida con RMSEP = 1.39 e $R^2 = 0.98$, quindi prestazioni di stima assolutamente soddisfacenti nella procedura Monte Carlo descritta sopra
Il modello presenta le seguenti prestazioni di stima: $ErR_{totale} = 9.31\%$, $ErR_{parziale} = 8.03\%$, $ErA/\sigma = 0.75$, e una percentuale molto ridotta di batch al di fuori dei limiti del 12.77 %. Quindi, l'accuratezza del modello ISMR-g2 è buona (poco inferiore al termine di riferimento che è il modello globale ISMR-g1, che però è costruito su batch della stessa durata), e il modello presenta inoltre una ottima affidabilità.

3.2.2.2 Modello locale

Un modello locale è stato sviluppato mediante l'utilizzo una strategia *nearest neighbor* su $A = 6$ variabili latenti, utilizzando le 300 osservazioni più simili a quella del batch corrente all'istante *k*-esimo. Il modello ISMR locale (ISMR-loc2) che ne deriva in grado di spiegare una variabilità cumulata della **X** del 49.32% e della **Y** del 99.26% con RMSEP = 1.42 e $R^2 = 0.97$.
Il modello ISMR-loc2 presenta un $ErR_{totale} = 9.17\%$, un $ErR_{parziale} = 4.79\%$, un $ErA/\sigma = 0.63$, denotando una accuratezza ottimale e una affidabilità accettabile, data percentuale di batch fuori dai limiti di controllo del 23.73%.
In sintesi, si può affermare che, confrontando le prestazioni di predizione nei due scenari, si osserva che la metodologia ISMR non risente della disomogeneità nella durata dei batch, permettendo di mantenere elevata accuratezza e buona affidabilità senza richiedere la sincronizzazione artificiale dei batch considerati.

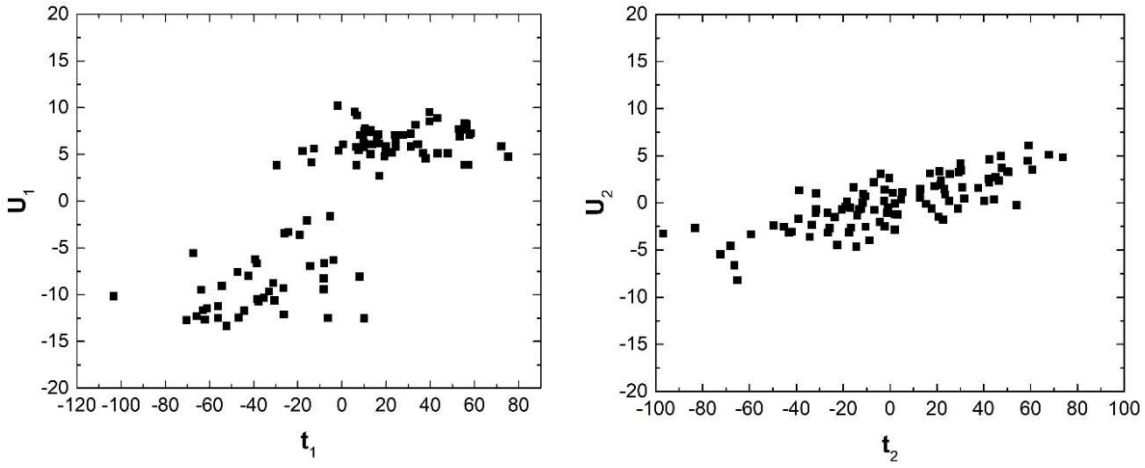
3.3 Modelli evolutivi

I modelli evolutivi vengono sviluppati utilizzando una strategia multi-modello evolutiva su un *batch-wise unfolding* della matrice tridimensionale **X**. La metodologia di letteratura utilizzata nello sviluppo dei modelli non permette la loro applicazione a batch di durate diverse, senza che prima venga effettuata una sincronizzazione dei dati. I modelli evolutivi sono costruiti

utilizzando il dataset-1. Vengono dunque sviluppati 59 sotto-modelli in corrispondenza dei campionamenti delle misure fuori linea.

3.3.1 Modello Evolutivo E-1150

Il modello E-1150 viene sviluppato utilizzando tutte le misure disponibili dalla prima alla corrente, la *k*-esima; sono considerate due variabili latenti in grado di spiegare una variabilità della **X** di 22.51 % e della **Y** di 55.04 %, e RMSEP=0.97 e R^2 = 0.99. I valori degli errori di predizione $ErR_{totale} > 100\%$, $ErR_{parziale} = 6.47\%$, $ErA/\sigma = 0.65$, e la percentuale di batch anomali del 49.92 %, indicano una buona accuratezza del modello per i batch per cui è ritenuta affidabile, ma una bassissima affidabilità delle predizioni effettuate in quanto per metà dei batch di convalida la stima è ritenuta inaffidabile e gli errori risultano altissimi. Si nota quindi un notevole mancanza di affidabilità del modello E-1150 rispetto al modello ISMR-loc1. A causa della strategia evolutiva utilizzata nella costruzione del modello E-1150, il numero di regressori considerato aumenta con il numero *k* delle osservazioni disponibili ottenendo nell'ultimo istante $k = K = 1150$ una matrice di calibrazione $\mathbf{X}$ $[100 \times (11 \cdot 1150)]$. Matrici di questo tipo, in cui la dimensione delle colonne è ordini di grandezza maggiore della dimensione delle righe, vengono dette matrici larghe e, quando utilizzate per sviluppare modelli PLS hanno la caratteristica di dare un numero eccessivo di batch oltre i limiti di confidenza di T^2 e *SPE* (Wise & Gallagher, 1996). E questo è ciò che accade nei modelli E-1150, in cui all'aumentare del numero di misure disponibili, dalla metà del batch in poi, il numero di batch considerato come anomalo aumenta fino a raggiungere valori oltre il 60%, nonostante la struttura del modello sia adeguata. Si riportano infatti gli score di **X** e **Y** del modello E-1150 per la prima e sulla seconda variabile latente (Figura 3.5). Si osserva che la relazione tra gli score è lineare su entrambe le variabili latenti e può quindi essere adeguatamente descritta da un modello lineare, come E-1150.



(a) (b)

Figura 3.5. *Modello Evolutivo: diagramma degli scores di* **X** *e di* **Y** *per i il modello* E-1150 sulla: (a) *prima variabile latente;* (b) *seconda variabile latente*.

I modelli evolutivi sviluppati su un *batch-wise unfolding* della matrice tridimensionale dei dati **X**, presentano il vantaggio di poter individuare agevolmente le variabili più rilevanti per predire la concentrazione di penicillina. L'individuazione avviene confrontando il profilo dei *weight* di ciascuna variabile. Ad esempio, in Figura 3.6a vengono riportati i profili dei *weight* del modello E-1150 sulla prima variabile latente per l'istante k = 376 corrispondente al tempo 75 e 12 minuti. In Figura 3.6b vengono riportati i *weight* dello stesso modello per l'istante k=1146 corrispondente al tempo 229 e 12 minuti. Si osserva che, sia nella prima fase del batch che nella fase finale, le variabili di processo più rilevanti per la predizione della concentrazione del prodotto sono la concentrazione di ossigeno (O_2og) e la concentrazione di anidride carbonica (CO_2og) nei gas di uscita dal reattore (*off-gas*). In particolare, queste due variabili sono anti-correlate alla concentrazione della penicillina. Confrontando i profili riportati in Figura 3.6 con il profilo della concentrazione della biomassa e della penicillina (si vedano le Figure 2.3a e Figura 2.3b del Paragrafo 2.2.1), è ragionevole pensare che quando la concertazione della biomassa e della penicillina aumentano, aumenti anche il consumo di ossigeno e la produzione di anidride carbonica da parte dei microorganismi contenuti nel fermentatore; di conseguenza la concentrazione di ossigeno e di anidride carbonica in uscita dal reattore risultano non solo anti-correlate, ma aumentano di rilevanza nella fase finale del batch, ad indicare l'aumento della concentrazione della biomassa nel reattore (e di conseguenza l'aumento della concentrazione di penicillina). Inoltre, nelle prime 75 ore risulta essere rilevante anche il volume processato all'interno del reattore, mentre nella fase finale della fermentazione diminuisce di peso. Si osserva infine che la portata di base, manipolata da un controllore PID per mantenere il pH a set-point, assume una condizione di particolare rilevanza alla fine del batch ed è correlata alla concentrazione di anidride carbonica nei gas di uscita. Questo indica che maggiore è la quantità della biomassa e della penicillina nel reattore, maggiore è il peso della portata della base nelle predizioni.

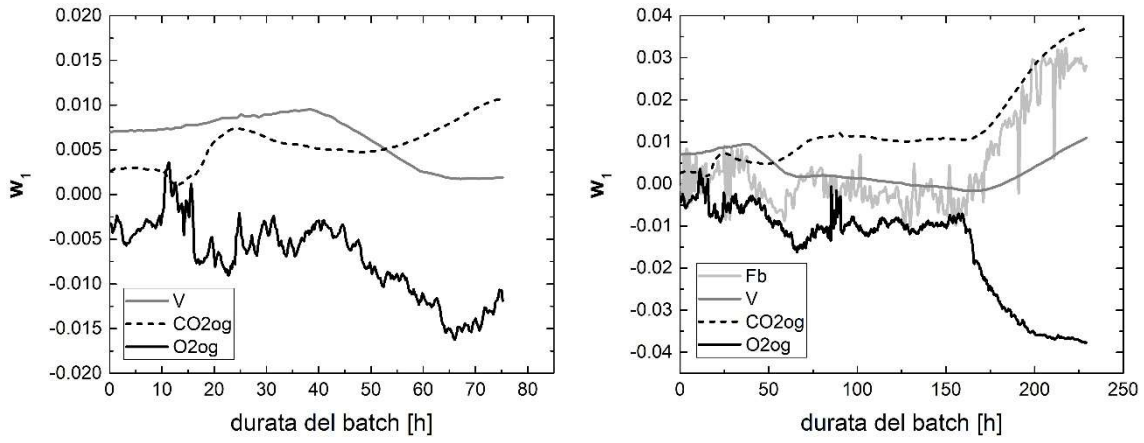


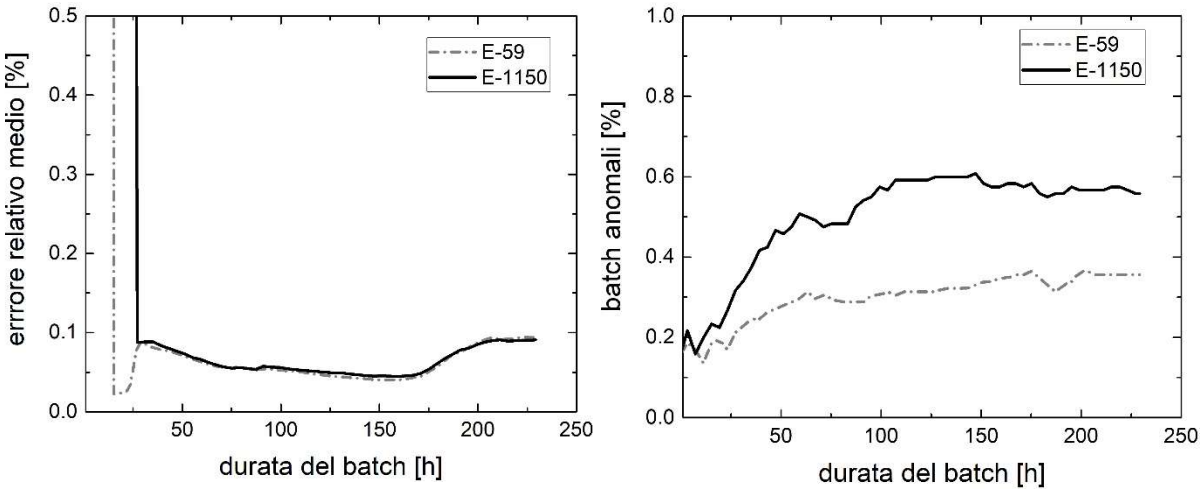
Figura 3.6. *Modello Evolutivo: confronto tra i valori dei weight più rilevanti sulla prima variabile latente per l'istante k=376 corrispondente all'ora 75 e 12 (a) e per l'istante k=1146 corrispondete all'ora 229 e 12 minuti (b).*

3.3.2 Modello Evolutivo E-59

Al fine di valutare se il numero eccessivo di batch oltre i limiti di confidenza di T^2 e *SPE*, riscontrato nel modello E-1150, sia effettivamente causato da un numero elevato di regressori, viene costruito il modello E-59, in cui si considerano solo le misure delle variabili in linea corrispondenti a campionamenti della concentrazione di penicillina, pena la perdita di parte delle informazioni sulla storia del batch.

Le dimensioni della matrice di calibrazione in questo modello diventano $\mathbf{X}\,[100 \times (11 \cdot 59)]$. Nel modello E-59 sono considerate $A = 2$ variabili latenti in grado di spiegare: in calibrazione una variabilità della $\underline{\mathbf{X}}$ di 22.51% e della $\underline{\mathbf{Y}}$ di 55.01%; in convalida con RMSEP = 3.59 e R^2 = 0.86, quindi prestazioni di stima accettabili nella procedura Monte Carlo descritta sopra. L'accuratezza è data da: ErR_{totale} = 20.97%, $ErR_{parziale}$ = 6.29 %, ed ErA/σ = 0.61, e risulta paragonabile a quella del modello E-1150. La percentuale media di batch fuori dai limiti invece diminuisce sensibilmente, raggiungendo il 29.91 %, si osserva quindi una effettiva dipendenza della affidabilità del modello dal numero di regressori considerato.

In Figura 3.7a si osserva il confronto tra i valori di $ErR_{parziale}$ al variare del tempo per i modelli E-1150 ed E-59; i valori di $ErR_{parziale}$ antecedenti l'ora 31 assumo valori molto elevati a causa della bassa valore reale della concentrazione di penicillina nella fase iniziale del processo. In Figura 3.7b viene riportato il confronto tra la percentuale di batch al di fuori del limite al variare del tempo per i modelli E-1150 e d E-59. Il profilo relativo a E-59 si mantiene sempre sotto al profilo di E-1150; inoltre si osserva che effettivamente la percentuale di batch anomali dipende dal numero di osservazioni considerate, e quindi dal numero dei regressori del modello.



(a) (b)

Figura 3.7. *Modelli Evolutivi: Confronto tra dei profili dell'errore relativo medio* (a) *e della percentuale di batch considerati anomali* (B) *per i modelli E-1150, E-59.*

3.4 Modelli Synchronized Moving Window (SMW)

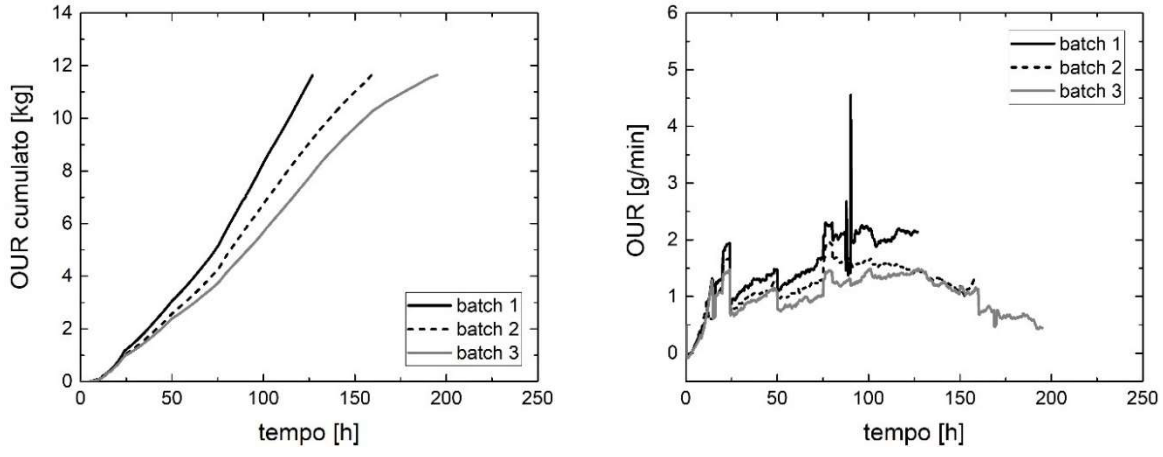
I modelli *Synchronized Moving Window* SMW sono sviluppati utilizzando prevalentemente una strategia a finestra mobile applicata a batch di durate diverse nello scenario #2, sincronizzati artificialmente utilizzando la tecnica della variabile indicatrice. I modelli SMW sono costruiti utilizzando il Dataset-2. La strategia a finestra mobile viene scelta per mantenere un numero limitato di regressori nel tempo; tuttavia, questa strategia non permette di effettuare predizioni quando il numero delle osservazioni $\omega = 1, 2, \ldots, \Omega$ a disposizione è inferiore a quelle necessarie per completare la finestra Ω_f. Nei modelli SMW, per poter predire il valore della concentrazione di penicillina anche nella prima fase del batch, quando $\omega < \Omega_f$, si utilizza una strategia a finestra mobile.

La variabile indicatrice per la sincronizzazione dei batch di diversa durata è l'integrale dell'*Oxygen Uptake Rate* (Goldrick et al., 2015), il quale corrisponde alla portata cumulata di ossigeno utilizzato dalla biomassa per la fermentazione, quindi l'ossigeno totale utilizzato dai microorganismi:

$$O_2 = OUR_{cumulato} = \int_{t_1}^{t_2} \frac{32}{22.4} F_{g_{in}} \left(O_{2_{in}} - O_{2_{out}} \frac{N_{2_{in}}}{1 - O_{2_{out}} - CO_{2_{in}}} \right) dt \quad . \qquad (3.1)$$

L' *Oxygen Uptake Rate* non è misurabile direttamente ma è calcolabile attraverso i valori della concentrazioni di ossigeno $(O_{2_{in}})$ e $(N_{2_{in}})$ azoto nella portata gassosa di alimentazione $(F_{g_{in}})$, dalla concentrazione di ossigeno $(O_{2_{out}})$ e anidride carbonica $(CO_{2_{in}})$nella corrente di uscita dei gas dal fermentatore, le quali però sono tutte variabili misurate in linea.

Si analizzano in Figura 3.8 i profili delle grandezze principali utilizzati per la sincronizzazione: l'*Oxygen Uptake Rate (OUR)* e l'integrale dell'*OUR*.



(a) (b)

Figura 3.8. *Modello Siunchronized Moving Window: profili della portata cumulata di ossigeno* (a) *e dell'oxygen uptake rate, OUR (b) per tre batch di diversa durata.*

Per ragioni di chiarezza espositiva non vengono riportati tutti i 120 profili, ma solo quelli relativi a 3 batch caratteristici: il batch di minore durata (123 ore e 12 minuti) in linea nera continua un batch di durata intermedia (155 ore e 12 minuti) in linea nera tratteggiata e il batch di maggiore durata (187 ore e 12 minuti) in linea grigia continua.
Tutti i batch considerati si fermano al raggiungimento di 11.65 kg di portata cumulata di *Oxygen Uptake Rate*; il raggiungimento di tale valore di *OUR* viene considerato come la conclusione del processo. La durata espressa attraverso la variabile indicatrice viene suddivisa in 600 intervalli. In queste condizioni l'incremento di un punto percentuale nell'evoluzione del batch corrisponde a un incremento di 1165 g della portata di ossigeno cumulata.
La dimensione temporale della finestra mobile è in questo caso di 320 intervalli, e corrisponde approssimativamente a un intervallo di tempo di 85 ± 18 ore. Le dimensioni della matrice di calibrazione sono $[100 \times (11 \cdot 320)]$. Il modello SMW utilizza $A = 6$ variabili latenti in grado di spiegare in calibrazione: una variabilità della **X** di 44.67% e della **Y** di 90.88%; in convalida con RMSEP = 69.48 e $R^2 = -8.51$. I valori di RMSEP e coefficiente di determinazione risultano disastrosi in convalida per le pessime predizioni dei batch fuori dai limiti di confidenza; se si considerano le sole predizioni affidabili RMSEP = 0.25 e $R^2 = 0.99$. Si nota quindi che le predizioni non attendibili sono caratterizzate da un errore molto elevato. Infatti, l'accuratezza, descritta dai valori degli errori di predizione risulta: $ErR_{totale} > 100\%$, $ErR_{parziale} = 5.37\%$, $ErA/\sigma = 056$. Quindi l'accuratezza risulta ottima (e paragonabile a quella del modello ISMR-loc2 descritto nel Paragrafo 3.2.2.2) per le predizioni considerate attendibili. Però anche in questo caso la percentuale media di batch fuori dai limiti aumenta sensibilmente, raggiungendo il 47.63%. L'elevato numero di batch anomali e l'elevato errore nelle predizioni di questi batch identificano una bassa affidabilità del modello SMW sviluppato secondo metodologie di letteratura.

3.5 Confronto tra modelli e conclusione

In Tabella 3.2 e in Tabella 3.3 si riporta un confronto riassuntivo tra i modelli presentati. In entrambe le tabelle viene riportato il numero e la sigla identificativa di ogni modello. Nelle colonne Tabella 3.2 si specificano: la strategia multi-modello utilizzata, il tipo di unfolding dei dati per gestire la matrice tridimensionale dei data **X**, la tipologia di regressori considerati. Nella ultima colonna della tabella vengono specificate le dimensioni massime che può assume la matrice di calibrazione di ciascun modello. In Tabella 3.2 si riportano le principali metriche utilizzate per confrontare le prestazioni dei modelli.
Si osserva infine che la metodologia proposta permette non solo di superare il problema della sincronizzazione ma è in grado di fornire prestazioni predittive migliori rispetto alle metodologie di letteratura, permettendo di descrivere con accuratezze migliori o nei peggiori dei casi confrontabili un numero maggiore di batch analizzati.

Tabella 3.2. *Riassunto strutture di diversi modelli.*

numero	modello	sotto modello	strategia	unfolding	regressori	dimens. Xcalib.
1	ISMR	ISMR-g1	evolutiva	variable-wise	features	5900×924
2	ISMR	ISMR-loc1	nearest neighbors	variable-wise	features	300×924
3	ISMR	ISMR-g2	evolutiva	variable-wise	features	δ×924
4	ISMR	ISMR-loc2	nearest neighbors	variable-wise	features	300×924
5	E	E-1150	evolutiva	batch-wise	misure	100×(11·1150)
6	E	E-59	evolutiva	batch-wise	misure	100×(11·59)
7	SMW	SMW	finestra mobile	batch-wise	misure	100×(11·320)

Tabella 3.3. *Riassunto accuratezze e affidabilità tra diversi modelli.*

numero	modello	sotto modello	ErR tot. [%]	ErR par. [%]	ErA/σ [adim]	batch anomali [%]
1	ISMR	ISMR-g1	13.93	13.41	1.30	20.32
2	ISMR	ISMR-loc1	8.09	6.20	0.71	26.86
3	ISMR	ISMR-g2	9.31	8.03	0.75	12.77
4	ISMR	ISMR-loc2	9.17	4.79	0.63	23.73
5	E	E-1150	>100	6.47	0.65	49.92
6	E	E-59	20.97	6.28	0.61	29.91
7	SMW	SMW	>100	5.73	0.56	47.63

Conclusioni

In questa Tesi è stata proposta una nuova metodologia per lo sviluppo di sensori virtuali. Il metodo proposto utilizza la proiezione su strutture latenti (PLS) per la stima delle variabili di qualità di un prodotto sfruttando le misure in linea delle variabili di processo. Tuttavia, mentre le metodologie di Letteratura più accurate sono comunemente applicabili solo a batch sincronizzati, la metodologia proposta permette di analizzare batch di durate diverse o numero di fasi operative differenti senza ricorrere all'utilizzo di complicate tecniche di sincronizzazione. I modelli PLS sviluppati infatti utilizzano una strategia a indici statistici multi-risoluzione che vengono ricavati dalla decomposizione multi-risoluzione dei i profili temporali delle variabili di processo mediante trasformata wavelet discreta.

Per verificare le prestazioni predittive della metodologia proposta è stato utilizzato un processo industriale simulato per la produzione fed-batch di penicillina. Il software usato per la simulazione è *IndPenSim* (Goldrick et al., 2015). In particolare, la metodologia proposta è stata utilizzata per stimare la concentrazione della penicillina in tempo reale, principale variabile di qualità del processo.

La metodologia proposta ha mostrato ottime prestazioni predittive: l'errore relativo medio di stima sono del 6.20%, con una percentuale di batch per cui la stima non viene considerata affidabile del 26.86%, laddove i metodi di Letteratura assicurano al meglio un errore relativo medio di stima del 5.73% e una percentuale di batch per cui la stima non viene ritenuta affidabile del 47.63%. Inoltre, la perdita di prestazioni porta ad errori del 8.09% nel caso di batch per cui la stima è considerata non affidabile a fronte di errori maggiori al 100% dei metodi di letteratura.

Le stesse considerazioni qualitative sono confermate sia nel caso si utilizzino strategie evolutive che a finestra mobile per realizzare i modelli, sia per batch della stessa durata che di durata diversa.

In conclusione, la metodologia proposta risulta essere migliorativa rispetto alle metodologie di letteratura non solo per le capacità predittive, ma anche perché non richiede la sincronizzazione artificiale dei batch, che sovente non è possibile, specialmente in tempo reale.

Riferimenti bibliografici

Addison, P. S. (2017). The illustrated wavelet transform handbook. In *Biomedical Instrumentation and Technology*. https://lccn.loc.gov/2016033578

Bakshi, B. R. (1998). Multiscale PCA with Application to Multivariate Statistical Process Monitoring. *AICHE JOURNAL*, *44*, 1596--1610. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.658

Camacho, J., Picó, J., & Ferrer, A. (2008). Bilinear modelling of batch processes. Part I: Theoretical discussion. *Journal of Chemometrics*, *22*(5), 299–308. https://doi.org/10.1002/cem.1113

Cinar, A., Parulekar, S. J., Undey, C., & Birol, G. (2003). Batch Fermentation: Modeling: Monitoring, and Control. In *Journal of Chemical Information and Modeling* (1st Editio, Vol. 1, Issue 1). CRC Press; 1 edition (April 1, 2003). https://doi.org/10.1017/CBO9781107415324.004

Facco, P., Doplicher, F., Bezzo, F., & Barolo, M. (2009). Moving average PLS soft sensor for online product quality estimation in an industrial batch polymerization process. *Journal of Process Control*, *19*(3), 520–529. https://doi.org/10.1016/j.jprocont.2008.05.002

Facco, P., Tomba, E., Roso, M., Modesti, M., Bezzo, F., & Barolo, M. (2010). Automatic characterization of nanofiber assemblies by image texture analysis. *Chemometrics and Intelligent Laboratory Systems*, *103*(1), 66–75. https://doi.org/10.1016/j.chemolab.2010.05.018

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, *185*(C), 1–17. https://doi.org/10.1016/0003-2670(86)80028-9

Goldrick, S., Ştefan, A., Lovett, D., Montague, G., & Lennox, B. (2015). The development of an industrial-scale fed-batch fermentation simulation. *Journal of Biotechnology*, *193*, 70–82. https://doi.org/10.1016/j.jbiotec.2014.10.029

Ingrid Daubechies. (1992). *Ten Lectures on Wavelets* (society for industrial and applied mathematics (ed.)). https://books.google.it/books?hl=en&lr=&id=cwdjT3CWY1kC&oi=fnd&pg=PP2&dq=i.+daubechies+ten+lectures+on+wavelets&ots=RTkRcE82Sx&sig=WJ7FOa64vlNtnslwVA0MhSIh_9I#v=onepage&q=i. daubechies ten lectures on wavelets&f=false

Jackson, J. E., & Mudholkar, G. S. (1979). Control procedures for residuals associated with principal component analysis. *Technometrics*, *21*(3), 341–349. https://doi.org/10.1080/00401706.1979.10489779

Lada, E. K., Lu, J. C., & Wilson, J. R. (2002). A wavelet-based procedure for process fault detection. *IEEE Transactions on Semiconductor Manufacturing*, *15*(1), 79–90. https://doi.org/10.1109/66.983447

Mallat, S. G. (1989). A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674–693. https://doi.org/10.1109/34.192463

Nomikos, P., & MacGregor, J. F. (1994). Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, *40*(8), 1361–1375. https://doi.org/10.1002/aic.690400809

Nomikos, P., & MacGregor, J. F. (1995). Multi-way partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems*, *30*(1), 97–108. https://doi.org/10.1016/0169-7439(95)00043-7

Otto, M., & Wegscheider, W. (1985). Spectrophotometric Multicomponent Analysis Applied to Trace Metal Determinations. *Analytical Chemistry*, *57*(1), 63–69. https://doi.org/10.1021/ac00279a020

Paul, G. C., & Thomas, C. R. (1996). A structured model for hyphal differentiation and penicillin production using Penicillium chrysogenum. *Biotechnology and Bioengineering*, *51*(5), 558–572. https://doi.org/10.1002/(SICI)1097-0290(19960905)51:5<558::AID-BIT8>3.0.CO;2-B

Reis, M. S., & Saraiva, P. M. (2006). Multiscale SPC in the presence of multiresolution data. *Computer Aided Chemical Engineering*, *21*(C), 1359–1364. https://doi.org/10.1016/S1570-7946(06)80236-1

Rendall, R., Lu, B., Castillo, I., Chin, S. T., Chiang, L. H., & Reis, M. S. (2017). A Unifying and Integrated Framework for Feature Oriented Analysis of Batch Processes. *Industrial and Engineering Chemistry Research*, *56*(30), 8590–8605. https://doi.org/10.1021/acs.iecr.6b04553

Rothwell, S. G., Martin, E. B., & Morris, A. J. (1998). Comparison of Methods for Handling Unequal Length Batches. *IFAC Proceedings Volumes*, *31*(11), 67–72. https://doi.org/10.1016/s1474-6670(17)44908-1

Suthar, K., Shah, D., Wang, J., & He, Q. P. (2019). Next-generation virtual metrology for semiconductor manufacturing: A feature-based framework. *Computers and Chemical Engineering*. https://doi.org/10.1016/j.compchemeng.2019.05.016

Wise, B. M., & Gallagher, N. B. (1996). The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, *6*(6), 329–348. https://doi.org/10.1016/0959-1524(96)00009-1

Wold, S., Martens, H., & Wold, H. (1983). *The multivariate calibration problem in chemistry solved by the PLS method* (pp. 286–293). Springer, Berlin, Heidelberg. https://doi.org/10.1007/bfb0062108

Wold, S., Ruhe, A., Wold, H., & Dunn, III, W. J. (1984). The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM Journal on Scientific and Statistical Computing*, *5*(3), 735–743. https://doi.org/10.1137/0905052

Wold, Svante, Kettaneh, N., Fridén, H., & Holmberg, A. (1998). Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemometrics and Intelligent Laboratory Systems*, *44*(1–2), 331–340. https://doi.org/10.1016/S0169-7439(98)00162-2

Yoon, S., & MacGregor, J. F. (2004). Principal-component analysis of multiscale data for process monitoring and fault diagnosis. *AIChE Journal*, *50*(11), 2891–2903. https://doi.org/10.1002/aic.10260