UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA CHIMICA E DEI PROCESSI INDUSTRIALI

Tesi di Laurea Magistrale
in Ingegneria Chimica e dei Processi Industriali

Sensori virtuali per la predizione della qualità in tempo reale del prodotto in processi continui e batch - Confronto tra modelli di regressione multivariata

Relatore: Dr. Pierantonio Facco
Correlatore: Dott. Francesco Sartori

Laureando: LOUIS TONION

ANNO ACCADEMICO 2019-2020

Riassunto

Il monitoraggio e il controllo di sistemi biologici sono problemi complessi a causa della difficoltà che si incontrano nello sviluppo di modelli a principi primi affidabili e nella stima dei loro parametri. La digitalizzazione dell'industria rende disponibile una grande quantità di dati riguardo le condizioni operative dei processi, rendendo possibile lo studio di questi sistemi attraverso modelli basati su dati. I modelli statistici multivariati di regressione sono, quindi, uno strumento utile a supportare il monitoraggio ed il controllo di processo.

In questa Tesi sono state implementate e confrontate cinque tecniche di regressione multivariata per il soft-sensing, cioè la predizione in tempo reale della qualità di un prodotto dai dati registrati in linea dal processo: *partial least squares*, Ridge, *least absolute shrinkage and selection operator* (LASSO), Kriging e Regression Trees. Le prestazioni dei sensori virtuali sono state valutate in due casi studio, il primo simulato e il secondo reale da Letteratura: una fermentazione fed-batch per la produzione di penicillina (sia per la predizione della qualità finale, che per la stima in tempo reale della qualità) e un trattamento di acque reflue.

Le prestazioni in termini di accuratezza di stima sono risultate soddisfacenti, robuste e con un basso impegno computazionale. Il modello *least absolute shrinkage and selection operator* si è dimostrato il miglior modello in termini di accuratezza, nonostante sia quello a maggior peso computazionale, mentre il metodo di proiezione su strutture latenti è risultato, mediamente, un buon compromesso tra accuratezza e peso computazionale.

Introduzione

I processi industriali offrono grandi opportunità riguardo monitoraggio e immagazzinamento dei dati lungo il profilo temporale o lungo differenti sezioni di un impianto, tali che l'implementazione di tecniche basate su dati permettono sia di monitorare un processo dalle variabili sistema in tempo reale, sia di inferire il valore di variabili qualità prodotto, qualora non ne sia possibile, o risulti troppo costosa, la misurazione in tempo reale (Birol *et al.,* 2003).

Questa Tesi riguarda la comparazione di differenti tecniche multivariate di regressione applicate al caso del soft-sensing, sensori virtuali che sono in grado di stimare/predire la qualità del prodotto dalle misure comunemente disponibili in linea dal processo (Kadlec *et al.,* 2008). In particolare, sono state confrontate e valutate le prestazioni di stima di cinque tecniche di regressione multivariata note in Letteratura: *partial least squares* (Nomikos, 1994) Ridge (Snee, 1975), *least absolute shrinkage and selection operator* (Rasmussen, 2012), Kriging (Krige, 1951) e Regression Trees (Wei-win, 2011). Il confronto ha valutato in modo obiettivo le prestazioni in termini di accuratezza e precisione di stima (tramite valutazione dell'errore), robustezza parametrica, e interpretabilità dei modelli (tramite analisi dei coefficienti di regressione).

Queste tecniche sono state applicate a due casi di studio: un processo simulato per la produzione in semicontinuo della penicillina e il processo reale continuo di trattamento di acque reflue.

La tesi si articola in tre capitoli: nel primo vengono descritti i metodi matematici, utilizzati nel secondo capitolo, si descrivono i due casi studio affrontati: un caso studio simulato di produzione della penicillina (Birol *et al.,* 2002) ed un caso studio reale di trattamento delle acque reflue (Anter *et al.,* 2019), nel terzo capitolo si espongono i risultati ottenuti sui due casi studio.

Capitolo 1

Modelli matematici

In questo Capitolo vengono presentati i metodi matematici utilizzati in questa Tesi per lo sviluppo di sensori virtuali. Nel primo paragrafo sono riportati e discussi i passaggi preliminari da adottare, mentre dal secondo in poi i metodi di regressione più comunemente utilizzati in letteratura per lo sviluppo di modelli predittivi sia per processi continui che batch.

1.1 Procedure matematiche preliminari

1.1.1 Determinazione delle prestazioni predittive dei modelli

Si consideri una matrice $\mathbf{Y}[I \times K]$ contenente le risposte con le quali si vuole costruire un modello predittivo utilizzando **X** come matrice dei regressori. Con il termine convalida incrociata si definisce la procedura che consente di determinare il numero delle variabili latenti utilizzate dal modello. La metrica più utilizzata per quantificare le prestazioni del modello in convalida incrociata è *RMSECV* (*Root-Mean-Square Error of Cross-Validation*, radice dell'errore quadratico medio in convalida incrociata, mostrata in Equazione 1.3), il quale a sua volta dipende dal valore di *PRESS*, errore di predizione sulla somma dei quadrati (*Prediction Error of Sum of Squares*, *equazione 1.*):

$$RMSECV_j = \sqrt{PRESS_j / I}, \tag{1.3}$$

$$PRESS_j = \sum_{j=1}^{I} (y_j - \hat{y}_j)^2 , \tag{1.4}$$

Con *I*, dimensione legata al numero di campioni, y indicante il valore reale della risposta del batch in posizione *j*, con $\hat{y}$ valore della risposta predetta dal modello. In questa Tesi è stata considerata una convalida incrociata su 10 iterazioni, la cui rappresentazione grafica è esposta in Figura 1.2.

Come mostrato in Figura 1.2, la convalida incrociata consiste nel suddividere in 10 gruppi contenenti lo stesso numero di campioni il dataset. Vengono quindi utilizzati nove gruppi per la calibrazione del modello di regressione, e uno per la convalida. Questa procedura viene ripetuta cambiando il gruppo di convalida fino a che ogni gruppo è stato utilizzato una volta per la convalida del modello. Questa tipologia di convalida incrociata è stata utilizzata in questa tesi per ogni tecnica di regressione utilizzata, in maniera tale da ottenere risultati comparabili da diversi modelli.

Figura 1.3: *rappresentazione grafica dei passaggi per ottenere una convalida incrociata a 10-folds. (Buhagiar, 2017).*

1.1.2 Valutazione delle prestazioni del modello

I parametri statistici attraverso cui valutare le prestazioni predittive dei modelli sono la radice dell'errore quadratico medio (*root mean square error,* RMSE), l'errore relativo medio (*mean relative error,* MRE) e l'errore assoluto medio standardizzato (standardized mean absolute error, SMAE). RMSE è definito come:

$$RMSE = \sqrt{\frac{\sum_1^t (\hat{y}_t - y_{t,m})}{n}} \quad , \tag{1.5}$$

mentre, MRE è definito come:

$$MRE = \frac{1}{n}\sum_1^t \frac{|\hat{y}_t - y_{t,m}|}{y_{t,m}} \cdot 100 \quad , \tag{1.6}$$

e SMAE:

$$SMAE = \frac{1}{n}\sum_1^t \frac{|\hat{y}_t - y_{t,m}|}{\sigma(y_{t,m})} \quad , \tag{1.7}$$

Con $\sigma(y_{t,m})$, deviazione standard all'instante t calcolato sulla variabile in uscita misurata.

I parametri statistici definiti serviranno come giudizio oggettivo della prestazione delle varie tecniche di regressione, e per analizzare quali metodi abbiano maggiori deviazioni rispetto ai dati misurati, sia in termine percentuale (errore relativo medio), che indica in maniera immediata con il grado di differenza tra predizione e misurazione, che in termine di distanza geometrica (RMSE), tramite la media di quanto differisce ogni punto predetto rispetto al valore reale. L'errore assoluto medio standardizzato misura la generale capacità di predizione della tecnica nei confronti delle variabili in uscita in questione, più questo parametro tenderà a zero, migliore sarà il metodo.

1.2 Proiezione su strutture latenti

La proiezione su strutture latenti (Wold, 1985), anche definita come PLS, è un metodo di regressione che ha lo scopo di risolvere problemi di regressione con dati multicollineari. Questo metodo prevede

la costruzione di nuove variabili, definite come variabili latenti, che meglio correlano le variabili di ingresso, dette predittori raccolte in un dataset $\mathbf{X}[I \times JK]$**,** con le variabili di uscita, $\mathbf{Y}[I \times K]$, cercando allo stesso tempo di spiegare il massimo della variabilità congiunta dei datasets. I dati che vengono utilizzati per calibrare un modello PLS vengono prima *autoscalati*, vengono cioè centrate tutte le colonne sulla loro media e poi ciascun elemento della colonna viene diviso per la deviazione standard della colonna stessa.

Le equazioni che costituiscono il modello PLS sono:

$$\mathbf{X} = \mathbf{R}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \quad , \tag{1.8}$$

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

dove, la matrice **X** dei predittori viene decomposta in *score* e *loading*, $\mathbf{R}[I \times L]$ e $\mathbf{S}[JK \times L]$, rispettivamente, con L, numero di variabili latenti selezionate. Gli *score* contengono le coordinate delle proiezioni dei campioni nello spazio delle variabili latenti, mentre i *loading* sono la proiezione di questo stesso spazio sulle colonne di **X**. Analogamente, **U** e **Q** sono le matrici degli scores e loadings della matrice **Y**, rispettivamente, mentre **E** ed **F** sono le matrici dei residui. Lo scopo del modello PLS è quello di massimizzare la varianza tra *score* di **X** e di **Y** tramite algoritmo di NIPALS.

1.3 Regressione Ridge

Il metodo di regressione tramite regolarizzazione Ridge (Snee, 1975), è un metodo che deriva dal metodo di regressione ai minimi quadrati ordinario (OLS, *ordinary least squares*) ma permette di risolvere problemi di regressione in cui è presente multicollinearità tra variabili. Questa caratteristica viene conferita alla regressione Ridge dall'applicazione di una penalizzazione con un termine chiamato norma L2. Questa permette la stima dei coefficienti di regressione, anche in presenza di dati con collinearità.

I coefficienti di regressione $\widehat{\boldsymbol{\beta}}_{RIDGE}$ della regressione Ridge tra i predittori $\mathbf{X}[I \times JK]$ e risposta **y** sono calcolati utilizzando (1.9):

$$\widehat{\boldsymbol{\beta}}_{RIDGE} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathrm{k}\mathbf{I})^{-\mathbf{1}}\mathbf{X}^{\mathrm{T}}\mathbf{y}\,, \tag{1.10}$$

dove k è identificabile come parametro di penalizzazione di Ridge e $\mathbf{I}[JK \times JK]$ è la matrice identità. Il parametro di Ridge viene calcolato tramite metodo di Hoerl-Kennard-Baldwin, che cerca il compromesso tra la varianza spiegata e il *bias*, ovvero la differenza tra il valore atteso di uno stimatore rispetto al valore reale del parametro da stimare. In generale sono sufficienti valori bassi (anche nell'ordine dei decimi) del parametro di Ridge per migliorare notevolmente il condizionamento di stima dei parametri (Hansen, 1998).

1.4 Regressione LASSO

La regressione LASSO (*least absolute shrinkage and selection operator*, Rasmussen, 2012) è una tecnica di regressione che seleziona i predittori più significativi e forza i coefficienti di regressione degli altri predittori al valore zero. Differentemente dalla regressione Ridge, la soluzione della regressione LASSO viene ottenuta tramite un algoritmo iterativo, che rende questo metodo generalmente più dispendioso dal punto di vista computazionale. La tecnica di regressione LASSO si basa sulla minimizzazione della funzione obiettivo in Equazione 1.11:

$$min\left(\frac{\mathbf{1}}{\boldsymbol{2N}}\sum_{n=1}^{N}(\mathbf{y}_n - \beta_0 - \mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta})^2 + \lambda\sum_{n=1}^{\mathrm{p}}|\boldsymbol{\beta}_{\mathrm{n}}|\right) \quad , \tag{1.11}$$

dove:

- N è il numero di osservazioni
- $\mathrm{y_k}$, è il valore della risposta all'instante k-esimo.
- $\mathbf{X}_k$ $[J \times K]$, con J numero di variabili, all'instante k-esimo.
- λ, parametro di regolarizzazione LASSO, non-negativo.
- β_0 *e* $\boldsymbol{\beta}[1 \times JK]$, che rappresentano, rispettivamente, l'intercetta e i coefficienti di regressione LASSO scalari.

La regressione LASSO restituisce un vettore di coefficienti di regressione, in cui molti tra questi saranno pari a zero: più il valore del parametro di regolarizzazione, λ, aumenta, più coefficienti di regressione tenderanno a zero. Questo indica che solamente una ristretta parte di predittori viene considerata significativa per la predizione della risposta.

1.5 Kriging

La tecnica di regressione Kriging è nata principalmente per l'utilizzo nel campo della geostatistica (Krige, 1951). È stata implementata successivamente come tecnica di regressione (Agterberg, 1974) non parametrica, dove, al posto di calcolare la distribuzione di probabilità dei parametri di una funzione specifica, viene valutata, invece, la distribuzione di probabilità su tutte le funzioni possibili che si adattano ad un dataset. Per sviluppare un modello tramite Kriging è richiesto di specificare una stima delle seguenti caratteristiche.

- funzione della covarianza: parametrizzata in funzione del parametro del Kernel, viene basata sul principio che punti nel sistema con predittori x_i simili (utile nel caso di calibrazione e convalida) abbiano una risposta su una variabile target, y_i, simile a loro volta. La funzione della covarianza in un processo di regressione Gaussiano cercherà esattamente di agire su questa similarità, specificando la tipologia di covarianza tra variabili latenti, e come la risposta del punto x_i sarà influenzato dai punti vicini, x_j. Esistono molti *kernel* della funzione di covarianza (Wackernagel, 2003). In questa Tesi viene utilizzato il *kernel esponenziale*, in Equazione 1.12:

$$\mathrm{k}\left(\mathrm{x_i}, \mathrm{x_j}, \theta\right) = \sigma^2 \exp\left(-\frac{\mathrm{r}}{\sigma_\mathrm{f}}\right) , \tag{1.12}$$

dove, σ_f, è la lunghezza caratteristica propria di ogni predittore pari a $\exp(\theta)$ con θ, vettore parametrico non-vincolato, mentre *r* è calcolato come la distanza euclidea tra due punti adiacenti da un punto di vista temporale, come mostrato in Equazione 1.13:

$$r = \sqrt{(X_i - X_{i+1})^T (X_i - X_{i+1})} \quad , \tag{1.13}$$

- varianza del rumore, σ^2;
- vettore contenente i coefficienti di regressione, $\beta[KJ \times 1]$ determinato come:

$$\beta_{Kriging} = (X_i{}^T C^{-1} X_i)^{-1} X_i{}^T C^{-1} y_m \quad , \tag{1.14}$$

Con C, matrice dei residui della covarianza, e y_m, vettore dei valori misurati della variabile in uscita. In questo modo si ottiene un modello di regressione con rumore Gaussiano del tipo:

$$y = f(x) + N(k, \sigma^2) \quad , \tag{1.15}$$

Lo scopo del Kriging è quindi modellare la risposta stimata mediante l'aggiunta di rumore alla risposta di calibrazione su cui si potevano riscontrare predittori simili, più la quantità $(x_i - x_j)$ tenderà a zero, maggiormente la funzione delle variabili in uscita calcolata sarà ottimale, per via dell'autocorrelazione tra predittori, secondo il principio per cui le osservazioni più vicine sono rispetto alle osservazioni lontane in uno spazio euclideo. Secondo Kleijnen (2017), l'accuratezza di questa tecnica dipende principalmente dalla scelta della funzione di covarianza.

1.6 Regression trees

La tecnica dei *regression trees* (Wei-Win Loh, 2011) permette di costruire modelli predittivi, frammentando lo spazio dei predittori in un numero inferiore di regioni sulla quale vengono compiute decisioni binarie. L'algoritmo iterativo parte da dataset completo raggruppato in un unico spazio, eseguendo un'analisi completa su tutte le possibili suddivisioni binarie di questo ad ogni iterazione, in maniera tale da trovare la combinazione che permetta di massimizzare la prestazione di stima.

Le dimensioni ottimali vengono identificate attraverso la minimizzazione di:

Supponendo di avere un dataset di calibrazione con *N* osservazioni, e J, numero di predittori in ingresso, l'algoritmo prevede:

- divisione della matrice $\underline{\mathbf{X}}$ in *s* sottoinsiemi distinti $(\mathbf{A_1}, \mathbf{A_2}, \dots, \mathbf{A_s})$ tali da costituire "l'albero", costruito sulla base di caratteristiche di classificazione comuni. Il numero di sottoinsiemi viene calcolato minimizzando la funzione obiettivo:

$$\min\left(\sum_{i:\, x_i \in R_1} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2} (y_i - \hat{y}_{R_2})^2\right) \quad , \tag{1.16}$$

  Previo *recursive binary splitting* associato al minor RSS (*residual sum of squares*).

$$R_1 = \{X | X_j \leq s\};\ R_2 = \{X | X_j > s\}, \tag{1.17}$$

- nodo per nodo, la decisione di andare su "ramo" destro o sinistro viene basata sulla similarità dei predittori in convalida con quelli in calibrazione su cui si è costruito il modello.

- una volta terminati i nodi, l'estrema sezione del modello predittivo, ovvero "foglia", restituisce il valore predetto come la media delle variabili qualitative associate al dataset di calibrazione, incontrate ad ogni nodo.

1.7 Gestione della dinamica in processi batch

1.7.1 Unfolding

L'unfolding è un pretrattamento operato su matrici tridimensionali, sia in calibrazione che in convalida che consente di ristrutturare la matrice in due dimensioni per trattarla mediante regressione. Una tecnica di unfolding è rappresentata graficamente in Figura 1.1, e serve per trattare dati che tipicamente caratterizzano processi batch.

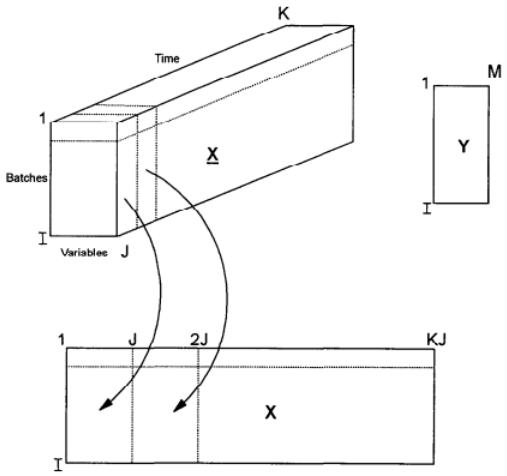


Figura 1.1: *Unfolding tridimensionale di dati da processo batch (Nomikos, 1994).*

Come mostrato in Figura 1.1 la matrice $\underline{\mathbf{X}}[I \times K \times J]$**,** dove I è il numero di batch presi in considerazione, K è il numero degli istanti di tempo al quale vengono effettuate le misurazioni e J il numero totale di variabili, viene sottoposta ad *unfolding*: le sottomatrici bidimensionali contenenti i dati per ogni istante di tempo vengono affiancate in senso orizzontale, ottenendo $\mathbf{X}[I \times KJ]$, in cui ogni riga rappresenta i dati misurati per ogni variabile in un unico batch a tutti gli istanti temporali. In questa Tesi, il pretrattamento di *unfolding* è risultato necessario anche sulla matrice delle variabili in uscita, $\underline{\mathbf{Y}}[I \times K]$**,** in modalità analoghe a quanto esposto in precedenza per la matrice tridimensionale delle variabili di processo in ingresso.

L'utilizzo di questa tecnica permette di analizzare dati contenenti le traiettorie temporali delle variabili con tecniche standard di regressione multivariata.

1.7.2 Metodo evolutivo

Un metodo evolutivo è un metodo che permette di predire delle grandezze utilizzando dati che provengono da diversi istanti di tempo, per esempio di un processo batch. Esso consiste nel calibrare

un diverso modello per ogni istante di tempo al quale sono disponibili nuovi dati. La metodologia verrà implementata prevederà la seguente procedura:

- costruzione del primo modello, al primo istante temporale, e valutare la predizione sulle variabili qualità, calcolata da una matrice $\mathbf{X_{val}}$ di dimensioni pari a $\mathbf{X_{cal}}$ in ingresso, con matrice $\mathbf{X_{cal}}$, $[I \times JK]$, e $\mathbf{Y_{cal}}[I \times 1]$, e I, numero di batch.
- proseguire instante temporale per instante temporale, per costruire su ognuno di questi un modello differente. Valutare le nuove $\mathbf{Y_{val}}$ di conseguenza.
- conclusione al giungere del valore massimo temporale previsto dal campionamento in calibrazione, tenendo in considerazione che all'instante finale si avrà una matrice in ingresso completata di tutto il profilo temporale, K su cui predire la risposta

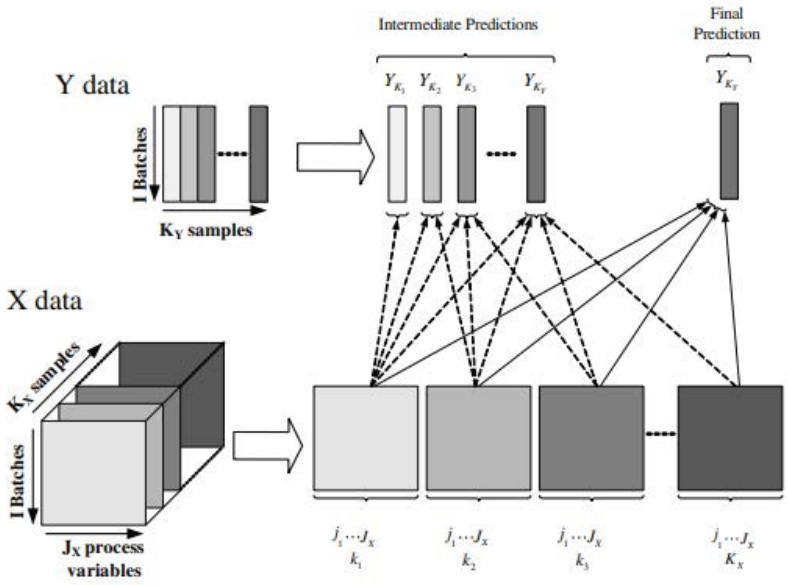


Figura 1.2: *Rappresentazione grafica dell'approccio time-evolving (Gunther, 2009).*

Nel modello che comprende tutti gli istanti di tempo, $\mathbf{X_{cal}}$ $[I \times JK]$ e $\mathbf{Y_{cal}}$ $[I \times 1]$, valore della variabile qualità nel dataset di calibrazione all'instante finale. In Figura 1.3, la procedura iterativa legata al completamento della matrice in ingresso $\mathbf{X_{cal}}$ con il progressivo ottenimento e misurazione dei dati istante per istante. Ogni istante sarà associato alla predizione della variabile in uscita solamente fino a quello stesso istante temporale disponibile.

1.7.3 Modelli dinamici applicati a tecniche di regressione

Un modello dinamico prevede la dipendenza delle variabili in uscita non solo dai valori dei predittori allo stesso istante temporale, ma anche dai predittori agli istanti temporali precedenti alla misurazione (Georgakis *et al.,* 1995). È necessario determinare una finestra temporale H, che definisce quante misurazioni precedenti si suppone influiscano sullo stato corrente del sistema, tale da costruire una matrice dinamica dei predittori $\mathbf{X_{din}}$ concatenando orizzontalmente i vettori delle misure di ogni istante a quelle degli istanti precedenti per una finestra temporale di lunghezza H nel modo seguente:

$$\mathbf{X}_{din} = \begin{bmatrix} \mathbf{x}_{1,k}^{\mathrm{T}} & \mathbf{x}_{1,0}^{\mathrm{T}} & \cdots & \mathbf{x}_{1,k-H}^{\mathrm{T}} \\ \mathbf{x}_{2,k}^{\mathrm{T}} & \mathbf{x}_{2,0}^{\mathrm{T}} & \cdots & \mathbf{x}_{2,k-H}^{\mathrm{T}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{j,k}^{\mathrm{T}} & \mathbf{x}_{j,0}^{\mathrm{T}} & \cdots & \mathbf{x}_{j,k-H}^{\mathrm{T}} \end{bmatrix} \quad , \tag{1.18}$$

dove $\mathbf{x}_{i,k}^{\mathrm{T}}$ è il vettore delle misure delle *J* variabili per l'*i*-esimo batch all'istante *k*. Per ogni predittore all'instante k, esisteranno H vettori simili ma traslati di un instante temporale ciascuno.

Risultando in un algoritmo del tipo:

$$y_i = \mathbf{X}_{\boldsymbol{din}}\boldsymbol{\beta}_{\boldsymbol{din}} + \mathbf{e} \quad , \tag{1.19}$$

con $\beta_{din}[1 \times hJ]$, ad indicare il vettore dei coefficienti di regressione con dipendenza alla dinamica del sistema.

Capitolo 2

Casi studio

In questo Capitolo vengono presentati i casi studio analizzati in questa Tesi. Il primo caso studio riguarda un processo batch simulato di produzione della Penicillina (Cinar, 1998). Il secondo caso studio riguarda un processo continuo per il trattamento di acque reflue urbane (Poch, 1993).

2.1 Caso studio 1: processo per la produzione di penicillina

Nel Caso studio 1 viene considerato un processo simulato di produzione della penicillina. La simulazione di questo processo avviene mediante l'utilizzo di un modello a principi primi implementato nella versione 2.0 del simulatore *PenSim* (Cinar, 1998). In Tabella 2.1 sono mostrate le variabili considerate dal modello e il loro intervallo consigliato per la simulazione

Tabella 2.1. *Caso studio 1: Elenco variabili di ingresso utilizzate dal simulatore PenSim (Cinar, 1998) per la produzione batch di penicillina.*

Variabile	Unità di misura	Intervallo consigliato
Concentrazione di substrato	g/L	[5;50]
Concentrazione di ossigeno	mmol/L	[1.10 1.20]
Concentrazione di biomassa	g/L	[0;0.2]
Concentrazione di penicillina	g/L	0
Volume di coltura	L	[100;200]
Concentrazione di CO_2	mmol/L	[0.5;1.0]
pH	-	[4:6]
Temperatura	K	[298;300]
Calore generato	Cal	0
Portata di aerazione	L/h	[3;10]
Potenza di agitazione	W	[20;50]
Portata di alimentazione di substrato	L/h	[0.035; 0.045]
Temperatura di alimentazione del substrato	K	[296;298]
pH set point	-	[5;6]
Temperatura set point	K	[298;300]

In Figura 2.1, è mostrato lo schema del processo studiato. Sono mostrati inoltre i cicli di controllo del pH e della temperatura. In particolare, la temperatura all'interno del bioreattore viene controllata mediante manipolazione delle portate d'acqua di raffreddamento e riscaldamento, alimentata nella

camicia del bioreattore, mentre il pH viene controllato manipolando la portata di un acido ed una base alimentati all'interno del bioreattore stesso.

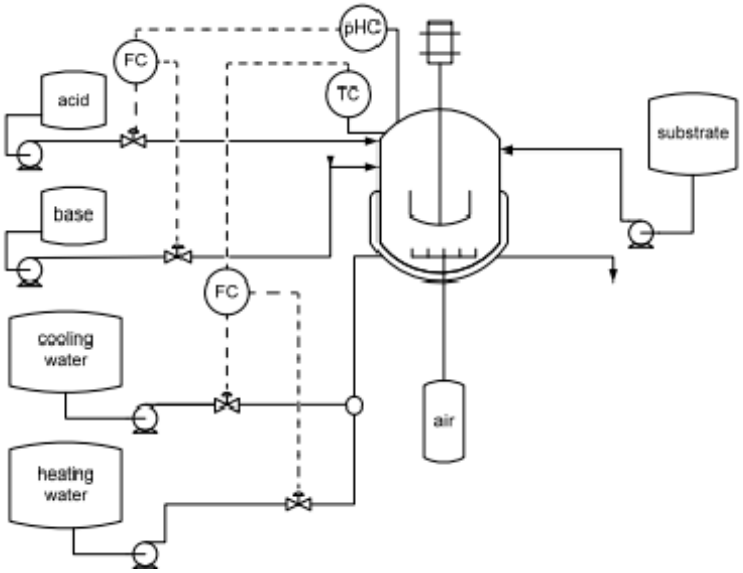


Figura 2.1. *Caso studio 1: Schema di processo per la produzione industriale di penicillina (Facco et al., 2014).*

Il controllo di queste due variabili è fondamentale in un processo di fermentazione, infatti i batteri coinvolti nella produzione di penicillina sono estremamente sensibili a sbalzi di temperatura e pH. La temperatura del sistema deve essere mantenuta costante e vicina a temperatura ambiente per non incorrere in fenomeni di degradazione, mentre il pH deve essere mantenuto tra 6 e 6.8 per evitare di inibire la crescita microbiotica. Il processo può essere suddiviso temporalmente (Birol *et al.*, 2003) in tre differenti fasi:

- fase di iniziazione: nessuna crescita microbiotica (da 0 a 30-50 ore).
- fase di crescita esponenziale: aumento di biomassa in modalità esponenziale (da 30-50 a 200 ore). Il processo inizia ad essere operato in modalità semicontinua. In particolare, vengono aggiunti aminoacidi e acido fenilacetico per incrementarne ulteriormente la produzione. La concentrazione di biomassa deve tuttavia rimanere entro valori moderati (tra i 4 e 10 g/L), per avere consumi di ossigeno su unità di volume contenuti. Nel caso la concentrazione di biomassa fosse eccessivamente alta il processo potrebbe andare fuori controllo, non garantendo il mantenimento di condizioni aerobiche costanti all'interno del reattore.
- fase stazionaria: crescita stabile. Non vengono più riscontrati cambiamenti significativi nella misurazione delle variabili. Gli andamenti temporali tendono, quindi, ad un comportamento asintotico (da 200 ore in poi, fino al termine della simulazione, posta a 300 ore in questo caso studio). Si deve garantire una costante alimentazione di ossigeno, tenendo conto che una elevata concentrazione di quest'ultimo richiede anche un efficace sistema di miscelazione con eventuale impiego di agenti antischiuma (che tenderanno ad inibire il trasporto dell'ossigeno).

2.1.1 Modello matematico e simulatore

Le variabili considerate nel modello sono mostrate in Tabella 2.2.

Tabella 2.2. *Caso studio 1: variabili di processo misurate in linea e variabili qualità di prodotto nel processo di produzione della Penicillina simulato con PenSim.*

Variabile di processo misurate in linea
Portata di aerazione, 1
Potenza di agitazione, 2
Portata di alimentazione di substrato, 3
Temperatura di alimentazione di substrato, 4
Concentrazione di substrato, 5
Concentrazione di ossigeno, 6
Volume di coltura, 7
Concentrazione di $\boldsymbol{CO_2}$, 8
pH, 9
Temperatura, 10
Calore generato, 11
Portata di controllo pH acido, 12
Portata di controllo pH alcalino, 13
Portata di acqua di raffreddamento, 14
Variabili di qualità del prodotto
Concentrazione di biomassa
Concentrazione di penicillina

Le Equazioni da 2.1 a 2.6 mostrano le relazioni funzionali tra le variabili nel modello a principi primi sul quale è basato il simulatore (Birol *et al.*, 2002):

$$\mathrm{B} = \mathrm{f}(\mathrm{B}, \mathrm{S}_c, \mathrm{C}_L, \mathrm{H}, \mathrm{T}), \tag{2.1}$$

$$\mathrm{S}_c = \mathrm{f}(\mathrm{B}, \mathrm{S}_c, \mathrm{C}_L, \mathrm{H}, \mathrm{T}), \tag{2.2}$$

$$\mathrm{C}_L = \mathrm{f}(\mathrm{B}, \mathrm{S}_c, \mathrm{C}_L, \mathrm{H}, \mathrm{T}), \tag{2.3}$$

$$\mathrm{P} = \mathrm{f}(\mathrm{B}, \mathrm{S}_c, \mathrm{C}_L, \mathrm{H}, \mathrm{T}, \mathrm{P}), \tag{2.4}$$

$$\mathrm{CO}_2 = \mathrm{f}(\mathrm{B}, \mathrm{H}, \mathrm{T}), \tag{2.5}$$

$$\mathrm{H} = \mathrm{f}(\mathrm{B}, \mathrm{H}, \mathrm{T}), \tag{2.6}$$

dove:

- B è la concentrazione di biomassa;
- S_c è la concentrazione di substrato;
- C_O è la concentrazione di ossigeno nel sistema;
- P è la concentrazione di penicillina;
- CO_2 è la concentrazione di anidride carbonica;
- H è la concentrazione di ioni H;
- T è la temperatura.

Il modello a principi primi implementato nel simulatore consiste in un sistema di equazioni differenziali nelle quali sono presenti le variabili elencate in Tabelle 2.1 e 2.2 (i valori dei parametri sono riportati in Appendice A). La crescita della biomassa nel reattore viene calcolata come:

$$\frac{dB}{dt} = \mu B - \frac{B}{V}\frac{dV}{dt} \quad , \tag{2.7}$$

dove μ, è il tasso specifico di crescita,

$$\mu = \mu_x \frac{\mathbf{S_c}}{(K_X B + \mathbf{S_c})} \frac{C_L}{(K_{OX} B + C_L)} \quad . \tag{2.8}$$

Il pH agisce sul sistema:

$$\frac{d[H^+]}{dt} = \gamma\left(\mu B - \frac{FB}{V}\right) + \left[\frac{-Bt + \sqrt{(Bt^2 + 4\times10^{-14})}}{2} - [H^+]\right]\frac{1}{\Delta t} \quad , \tag{2.9}$$

dove Bt è definita come:

$$Bt = \frac{\left[{}^{10^{-14}}/_{[H^+]} - [H^+]\right]V - C_{a,b}(F_a - F_b)\Delta t}{V + (F_a - F_b)\Delta t} \quad , \tag{2.10}$$

con F_a, portata di alimentazione acida, mentre F_b è la portata di alimentazione basica. $C_{a,b}$, è la concentrazione in entrambe le soluzioni e viene considerata pari a 3M.

La concentrazione di penicillina da:

$$\frac{dP}{dt} = \mu_{pp} B - KP - \frac{P}{V}\frac{dV}{dt} \quad , \tag{2.11}$$

dove μ_{pp} è il tasso specifico di produzione penicillina:

$$\mu_{pp} = \mu_p \frac{\mathbf{S_c}}{(K_p + \mathbf{S_c} + \mathbf{S_c}^2/K_1)} \frac{c_L^p}{(K_{op} B + c_L^p)} \quad . \tag{2.12}$$

L'utilizzo del substrato a base di glucosio è la causa della crescita della biomassa, *X*, e della formazione di penicillina, *P*::

$$\frac{d\mathbf{S_c}}{dt} = -\frac{\mu}{Y_{x;s}} B - \frac{\mu_{pp}}{Y_{p;s}} B - m_x B + \frac{F s_f}{V} - \frac{\mathbf{S_c}}{V}\frac{dV}{dt} \quad , \tag{2.13}$$

$$\frac{dC_L}{dt} = -\frac{\mu}{Y_{x,o}} B - \frac{\mu_{pp}}{Y_{p,o}} B - m_o B + K_{la}(C_L^* - C_L) - \frac{C_L}{V}\frac{dV}{dt}, \tag{2.14}$$

Dove K_{la} è il coefficiente di trasferimento di massa generale e viene indicato come funzione sia della potenza di agitazione di input P_w, che del flusso alimentazione di ossigeno f_g.

Si noti che, considerato che K_{la} dipende dal volume del fluido all'interno del reattore:

$$K_{la} = \alpha\sqrt{f_g}\left(\frac{P_W}{V}\right)^{\beta} \quad , \tag{2.15}$$

e che il processo opera in modalità semicontinua, devono essere calcolati con Equazione 2.16 i cambiamenti di volume:

$$\frac{dV}{dt} = F + F_{a;b} - F_{loss} \quad , \tag{2.16}$$

dove F_{loss} è un termine che tiene conto dei processi di perdita legati all'evaporazione che avvengono durante la fermentazione:

$$F_{loss} = V\lambda\left(e^{5((T-T_0)/T_v-T_0)} - 1\right) \quad , \tag{2.17}$$

con T_0 e T_v, rispettivamente le temperature di congelamento e di ebollizione della coltura del sistema. Le rimanenti relazioni sono per il calore generato, la temperatura e la concentrazione di CO_2:

$$\frac{dQ_{react}}{dt} = r_{q1}\frac{dB}{dt}V + r_{q2}BV \quad , \tag{2.18}$$

$$\frac{dT}{dt} = \frac{F}{s_f}(T_f - T) + \frac{1}{V\rho c_p}\left[Q_{react} - \frac{aF_c^{b+1}}{aF_c^b/2\rho_c c_{pc}}\right] \quad , \tag{2.19}$$

$$\frac{dCO_2}{dt} = \alpha_1\frac{dB}{dt} + \alpha_2 B + \alpha_3. \tag{2.20}$$

Ogni parametro è stato riportato con nominativo e valore in Appendice A.

Le soluzioni delle equazioni differenziali cinetiche forniscono le variabili di processo, la qualità finale di un prodotto (concentrazione di penicillina e biomassa) di Tabella 2.2. In particolare, concentrazione di biomassa e penicillina sono variabili risposta, le cui misure sono laboriose e costose, che però identificano la qualità del prodotto. La loro stima in tempo reale è quindi importante perché determina una diminuzione di costi e durata del processo, nonché una maggior efficienza. I dati simulati riguardano 166 batch, a condizioni di ingresso scelte in modo casuale negli intervalli consigliati, illustrati in Tabella 2.1**.** I dati disponibili sono raccolti in una matrice di regressori $\underline{\mathbf{X}}$ $[166 \times 14 \times 600]$.all'interno della quale sono presenti 14 variabili campionate per 600 istanti temporali. La matrice delle risposte è $\underline{\mathbf{Y}}$ $[166 \times 2 \times 600]$.

2.2 Caso studio 2: processo di trattamento acque reflue

Nel Caso studio 2, vengono considerati dati provenienti da un impianto reale di trattamento di acque reflue. Per trattamento acque reflue si intende il processo di rimozione di contaminanti, che costituiscono un pericolo sotto un punto di vista o chimico o biologico, da scarichi o industriali e domestici (Anter, 2019). Il caso studio presenta un processo di trattamento di acque reflue (mostrato in Figura 2.2) operato in modalità continua.

Il processo è suddiviso in 3 sezioni.

- prima sezione: decantazione per la rimozione di materia organica tramite separazione fisica; i solidi in sospensione verranno quindi raccolti e rimossi nel fondo del primo decantatore (trattamento primario in Figura 2.2).
- seconda sezione: rimozione della materia organica suscettibile a degradazione biologica (facilitata tramite insufflazione di ossigeno).

- terza sezione: separazione per gravità dei liquami e fanghi di ricircolo dal refluo chiarificato.

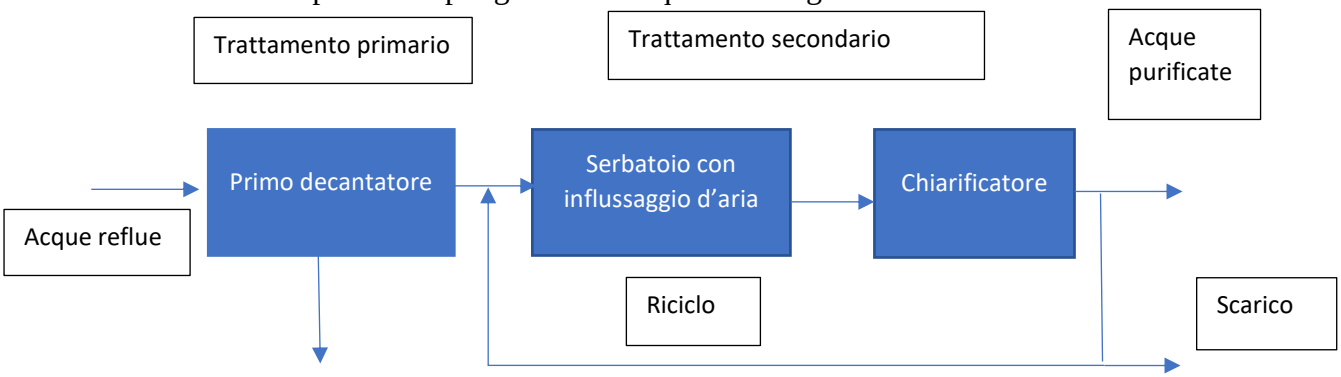


Figura 2.2. *Caso di studio 2: schema di processo per il trattamento di acque reflue, (Fuente, 2004).*

Nell'impianto vengono misurate 29 variabili di processo (Tabella 2.3) su un arco temporale di 527 giorni, e sono disponibili le misurazioni di 9 variabili di qualità (Tabella 2.4), di cui:

- 3 riguardano la linea in uscita dal primo decantatore.
- 2 riguardano la linea in uscita dal secondo decantatore. e
- i rimanenti sono variabili di qualità globali.

Tabella 2.3. Caso studio 2: *riassunto delle variabili di processo misurate.*

Variabile di ingresso	Sigla	Media	Deviazione standard
Input flow to plant, 1	Q-E	37372	6851.7
Input zinc to plant, 2	ZN-E	2.277	2.329
Input pH to plant, 3	PH-E	7.825	0.2366
Input biological demand O_2 to plant, 4	DBO-E	189.63	61.374
Input chemical demand O_2 to plant, 5	DQO-E	403.97	117.86
Input suspended solids to plant, 6	SS-E	226.80	118.18
Input volatile suspended solids to plant, 7	SSV-E	60.858	12.661
Input sediments to plant, 8	SED-E	4.689	2.908
Input conductivity to plant, 9	COND-E	1468.4	393.61
Input pH to primary settler, 10	PH-P	7.851	0.2240
Input biological demand O_2 to primary settler, 11	DBO-P	209.67	71.917
Input suspended solids to primary settler, 12	SS-P	257.39	147.33
Input volatile suspended solids to primary settler, 13	SSV-P	59.680	12.733
Input sediments to primary settler, 14	SED-P	5.117	3.578
Input conductivity to primary settler, 15	COND-P	1484.6	398.60
Input pH to secondary settler, 16	PH-D	7.835	0.1952
Input biological demand O_2 to secondary settler, 17	DBO-D	122.59	36.529
Input chemical demand O_2 to secondary settler, 18	DQO-D	272.98	70.953
Input suspended solids to secondary settler, 19	SS-D	93.911	23.439
Input volatile suspended solids to secondary settler, 20	SSV-D	72.723	10.394
Input sediments to secondary settler, 21	SED-D	0.4142	0.3782
Input conductivity to secondary settler, 22	COND-D	1477.7	401.09
Output pH, 23	PH-S	7.721	0.1534
Output biological demand O_2, 24	DBO-S	18.734	9.607
Output chemical demand O_2, 25	DQO-S	84.132	33.663
Output suspended solids, 26	SS-S	20.913	11.547
Output volatile suspended solids, 27	SSV-S	79.684	9.238
Output sediments, 28	SED-S	0.0361	0.2003
Output conductivity, 29	COND-S	1481.5	384.28

Tabella 2.4: *Variabili in uscita per processo di trattamento acque reflue*

Variabile di uscita	Sigla
Performance input biological demand of oxygen primary settler	RD-DBO-P
Performance input suspended solids primary settler	RD-SS-P
Performance input sediments primary settler	RD-SED-P
Performance input biological demand of oxygen secondary settler	RD-DBO-S
Performance input chemical demand of oxygen secondary settler	RD-DQO-S
Global performance input biological demand of oxygen	RD-DBO-G
Global performance input chemical demand of oxygen	RD-DQO-G
Global performance input suspended solids	RD-SS-G
Global performance input sediments	RD-SED-G
Performance input biological demand of oxygen primary settler	RD-DBO-P
Performance input suspended solids primary settler	RD-SS-P
Performance input sediments primary settler	RD-SED-P
Performance input biological demand of oxygen secondary settler	RD-DBO-S
Performance input chemical demand of oxygen secondary settler	RD-DQO-S
Global performance input biological demand of oxygen	RD-DBO-G
Global performance input chemical demand of oxygen	RD-DQO-G
Global performance input suspended solids	RD-SS-G

Capitolo 3

Confronto delle prestazioni di stima dei sensori virtuali

In questo Capitolo vengono presentati i risultati dell'applicazione dei metodi di regressione PLS, Ridge, LASSO, Regression Trees e Kriging ai due casi studio descritti nel Capitolo 2. In particolare, sono state confrontate le prestazioni dei modelli sviluppati su un processo semibatch e su un processo continuo. Per quanto riguarda il primo sono stati sviluppati modelli per risolvere due tipologie di problemi: la predizione del valore finale e la stima della traiettoria di due variabili di qualità del prodotto. Nel secondo caso studio vengono sviluppati modelli per la stima dei parametri di qualità in tempo reale.

3.1 Risultati e discussione per il caso studio 1

I modelli sviluppati per il Caso studio 1 hanno come obiettivi:

i. la previsione del valore finale di due variabili di qualità (*end-point prediction*), la concentrazione di biomassa e la concentrazione di prodotto (Penicillina).
ii. la stima in linea delle due variabili qualità.

3.1.1 Stima del valore all'instante finale

La stima del valore all'instante finale esegue a batch completo, avendo quindi a disposizione una matrice in ingresso **X** completa sia di tutte le variabili in ingresso (14, nel caso studio) che di tutti gli istanti temporali (300 ore, campionate ogni mezzora). Tramite cinque differenti tecniche di regressione, sono stati sviluppati differenti sensori virtuali per cercare di predire la qualità finale del prodotto (concentrazione di biomassa e concentrazione di penicillina), ovvero il valore raggiunto a 300 ore.

3.1.1.1 Caratteristiche dei modelli di regressione analizzati

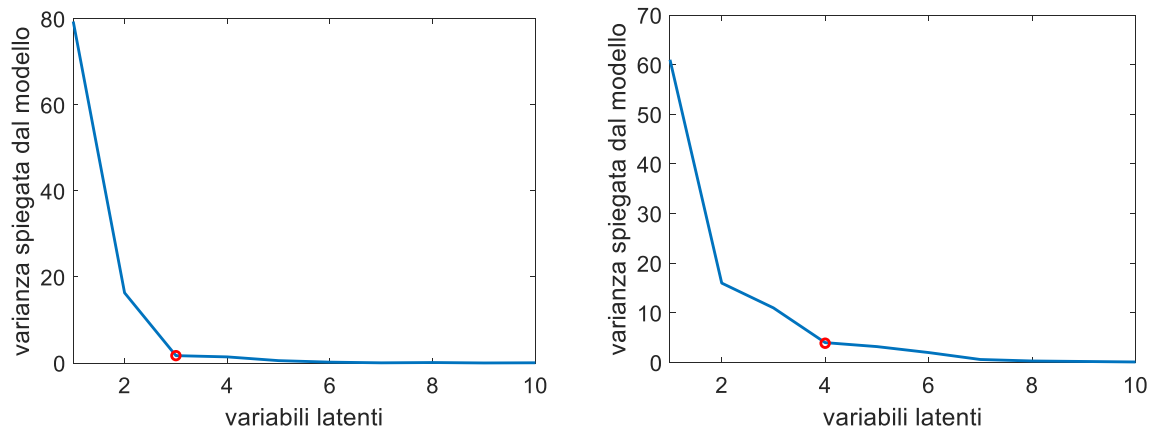
I risultati sono stati convalidati per tutti i sensori virtuali mediante una procedura di convalida incrociata in 10 iterazioni con una ripartizione dei batch disponibili del 67% in calibrazione e 33% in

convalida. Due modelli di regressione PLS separati sono stati selezionati per la stima della qualità del prodotto (rispettivamente con 3 e 4 variabili latenti), la cui varianza spiegata dalla variabile di risposta al variare del numero di variabili latenti considerate viene mostrata in Figura 3.1. Nei modelli di regressione Ridge e LASSO, i parametri ottimali sono riportati in Tabella 3.2. Il modello di regressione Kriging ottimale è stato selezionato su base lineare con forma (*kernel*) esponenziale della funzione di covarianza.

Il modello Regression Trees ha un totale di 39 nodi per la predizione della concentrazione finale di biomassa, e 43 per la predizione della concentrazione finale di penicillina, entrambi valori ottenuti tramite ottimizzazione interna a MatLab (insieme a numero minimo di elementi per "foglia", pari a dieci circa in entrambi i modelli) per garantire sia omogeneità che profondità "all'albero" del modello.

Tabella 3.1. *Caso studio 1, previsione del valore finale di qualità: parametri dei modelli Ridge e LASSO.*

	Ridge (k)	**LASSO (λ)**
Biomassa	0.013	0.0248
Penicillina	0.024	0.0025



(a) *(b)*

Figura 3.1. *Caso studio 1***:** *varianza spiegata dal modello PLS per la stima del valore finale di: (a) concentrazione di biomassa; (b) concentrazione di penicillina, in funzione del numero di variabili latenti.*

Nel modello PLS i coefficienti di determinazione sono $R^2_{\mathbf{X},\mathbf{1}} = 52.54\%, R^2_{\mathbf{Y},\mathbf{1}} = 97.68\%$ e $R^2_{\mathbf{X},\mathbf{2}} = 56.86\%, R^2_{\mathbf{Y},\mathbf{2}} = 95.14\%$ per la predizione delle concentrazioni di biomassa e penicillina rispettivamente. Il modello mostra di dare ottimi risultati in calibrazione con poche variabili latenti che sfruttano circa metà della variabilità dei predittori per la stima. Un confronto con gli altri modelli è mostrato in Tabella 3.1, da cui si possono trarre i primi risultati: i modelli PLS, Ridge e LASSO approssimano molto bene i dati, Kriging stima bene la biomassa, ma non la penicillina mentre Regression Trees mostrano le prestazioni peggiori.

Tabella 3.2. *Caso studio 1, previsione del valore finale di qualità***:** *confronto dei coefficienti di determinazione per i sensori virtuali costruiti con diverse tecniche di regressione per la previsione di entrambe le variabili di qualità.*

	PLS	Ridge	LASSO	Kriging	R. Trees
$R^2_{biomassa}$	0.9768	0.9921	0.9946	0.9738	0.8228
$R^2_{penicillina}$	0.9514	0.9857	0.9886	0.7418	0.6893

3.1.1.2 Confronto delle prestazioni predittive dei sensori virtuali

Le prestazioni dei sensori virtuali per la stima del valore finale di concentrazione di biomassa e penicillina sono confrontate in termini di indici RMSE, MRE e SMAE (equazioni 1.3, 1.4 e 1.5) in Tabella 3.3.

Tabella 3.3. *Caso studio 1, previsione del valore finale di qualità***:** *confronto delle prestazioni predittive in convalida dei diversi sensori virtuali per la previsione della variabile concentrazione di biomassa (a) e della concentrazione di penicillina (b)*

(a)

	PLS	Ridge	LASSO	Regression trees	Kriging
RMSE (g/L)	0.2816	0.1881	**0.1512**	0.7523	0.3051
MRE (%)	1.9839	1.2627	**1.0240**	5.1616	1.5989
SMAE	0.0667	0.0427	**0.0354**	0.1750	0.0560

(b)

	PLS	Ridge	LASSO	Regression trees	Kriging
RMSE (g/L)	**0.0334**	0.0646	0.0384	0.1787	0.1828
MRE (%)	2.1878	3.5360	**1.3874**	8.5768	11.711
SMAE	**0.0417**	0.1011	0.0609	0.2485	0.3184

La concentrazione finale di biomassa è mediamente più facile da predire, infatti, tutti gli indici d'errore ottenuti per differenti tecniche di regressione sono minori per la concentrazione di biomassa che per la concentrazione di penicillina. Sia il sensore virtuale basato sulla regressione Ridge che quello costruito sulla regressione LASSO sono molto accurati, ancorché dispendiosi da un punto di vista computazionale, per via dell'ottimizzazione dei parametri di Ridge e LASSO, a differenza dei metodi Kriging e Regression Trees, che hanno in questo caso, però, prestazioni non soddisfacenti per accuratezza. Il metodo PLS è un ottimo compromesso tra accuratezza predittiva e velocità d'esecuzione, in quanto il peso computazionale è molto inferiore a quello richiesto da Ridge e LASSO.

Per la convalida incrociata, il tempo medio d'esecuzione, a parità di hardware (AMD quad core 3.66 GHz, 8 GB RAM), è stato riportato in Tabella 3.5.

Tabella 3.5. *Caso studio 1, previsione del valore finale di qualità: confronto della velocità d'esecuzione per singola iterazione delle tecniche di soft-sensing.*

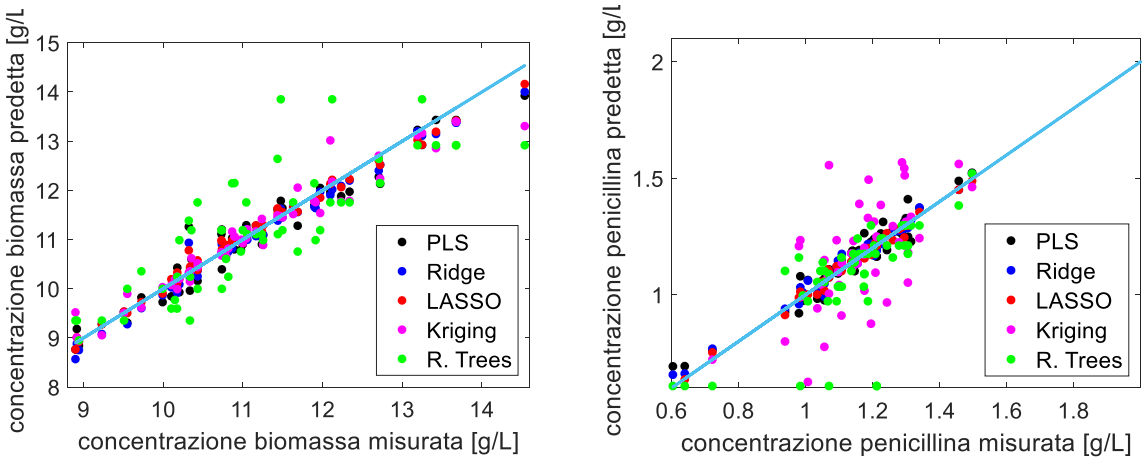
	PLS	Ridge	LASSO	Regression trees	Kriging
Tempo (sec)	19.97	114.11	207.05	**9.52**	**3.51**

Per capire quale tra i metodi di regressione presi in considerazione sia quello più robusto si studia anche la variabilità dei risultati, assumendo che un modello robusto presenti una minor variabilità a seconda della ripartizione utilizzata nella convalida. A questo scopo, sono state analizzate diverse ripartizioni tra set di calibrazione e convalida, ovvero 10%, 33%, 50%, 66% e 80% in calibrazione e la parte rimanente in convalida. In Tabella 3.6 sono stati proposti per tutti i sensori virtuali media e deviazione standard calcolati su differenti ripartizioni. Ci si aspetta che la tecnica di regressione maggiormente robusta abbia una minor deviazione standard, laddove la media è riportata per valutare anche la bontà delle prestazioni predittive.

Tabella 3.6. *Caso studio 1, previsione del valore finale di qualità: confronto di accuratezza (in termini di errore medio) e robustezza (in termini di deviazione standard degli errori) dei sensori virtuali per differenti ripartizioni dei batch tra calibrazione e convalida.*

	Errore relativo			
	Concentrazione di biomassa		**Concentrazione di penicillina**	
	Deviazione standard	**Media**	**Deviazione Standard**	**Media**
Kriging	2.1321	3.1504	**3.3519**	**11.6380**
Ridge	1.7063	3.8031	**3.0078**	**11.3921**
LASSO	1.3189	1.7546	6.8891	10.3326
PLS	**0.8786**	**1.9740**	6.2145	8.6305
R. trees	1.1099	5.2615	6.5825	11.4760
	MSE			
	Concentrazione di biomassa		**Concentrazione di penicillina**	
	Deviazione standard	**Media**	**Deviazione standard**	**Media**
Kriging	0.2422	0.2292	**0.0145**	**0.0250**
Ridge	0.2457	0.4365	**0.0133**	**0.0268**
LASSO	0.1500	0.0970	0.0269	0.0279
PLS	**0.1154**	**0.1126**	0.0256	0.0182
R. trees	0.2570	0.5804	0.0236	0.0454
	Errore assoluto			
	Concentrazione di biomassa		**Concentrazione di penicillina**	
	Deviazione standard	**Media**	**Deviazione Standard**	**Media**
Kriging	0.1891	0.2836	**0.0234**	**0.1053**
Ridge	0.1776	0.4170	**0.0318**	**0.1062**
LASSO	0.1490	0.1986	0.0451	0.0854
PLS	**0.1090**	**0.2252**	0.0367	0.0745
R. trees	0.1263	0.5810	0.0388	0.1303

Sulla predizione della concentrazione finale di biomassa il metodo PLS appare come il più robusto (marginalmente migliore rispetto a LASSO) su tutti gli indici di accuratezza, in quanto risultano minimi, sia in termini di media che di deviazione standard, calcolate sulle diverse ripartizioni. I metodi Kriging e Ridge appaiono come i più robusti nella predizione della concentrazione finale di penicillina, anche se non si riesce ad ottenere allo stesso tempo ottima robustezza e accuratezza, avendo Kriging e Ridge, i quali sono comunque i metodi più accurati seppur meno robusti.



(a) *(b)*

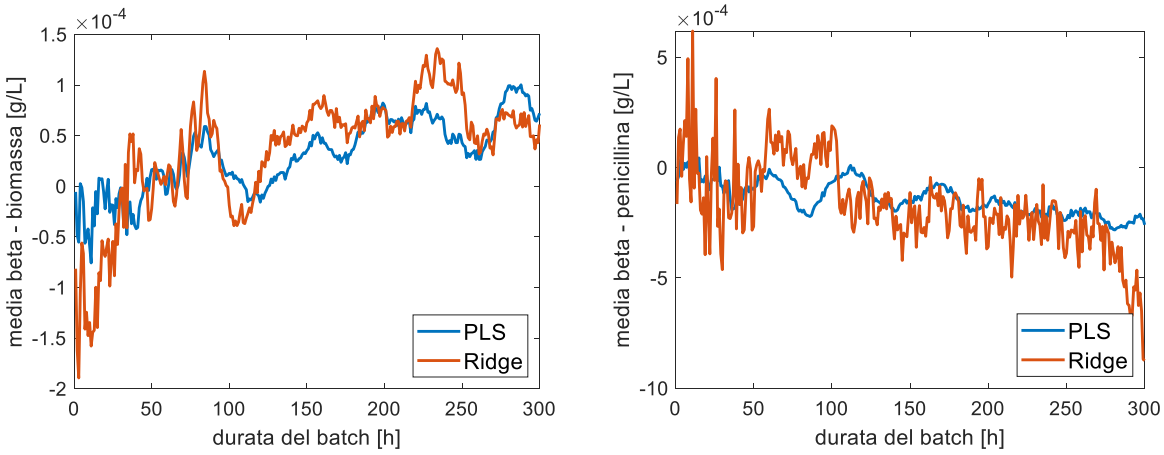
Figura 3.2. *Caso studio 1, previsione del valore finale di qualità***:** *parity plot dei modelli di regressione per la predizione dei valori finale di: (a) concentrazione di biomassa; (b) concentrazione di penicillina.*

Un confronto visuale delle prestazioni predittive dei modelli viene fatto mediante i parity plot di Figura 3.2. L'incapacità dei modelli Kriging e Regression Trees di predire in maniera appropriata la qualità finale del prodotto è evidente, notando come questi due modelli risultino inaccurati per la maggior parte delle osservazioni ed evidenziando forti discostamenti tra predizione e misura. I modelli PLS, LASSO e Ridge non evidenziano particolari deviazioni rispetto ai rispettivi valori reali, e quindi si confermano accurati.

In conclusione, si osserva che in termini di prestazioni predittive i modelli migliori sono: LASSO per la stima della concentrazione finale di biomassa; PLS per la stima della concentrazione finale di penicillina. PLS fornisce anche il miglior compromesso tra prestazioni predittive e velocità d'esecuzione.

3.1.1.3 Interpretazione e discussione dei modelli di regressione

L'interpretazione dei modelli richiede l'analisi dei coefficienti di regressione in termini di profilo temporale medio (mediando tutti i contributi dei diversi predittori) e di contributo medio di ciascun predittore (mediando tutti i contributi dei diversi istanti di tempo). Il profilo temporale dei coefficienti di regressione mostrato in Figura 3.3 per i modelli PLS e Ridge permette di visualizzare gli istanti di campionamento durante il batch più significativi per la stima della qualità finale.

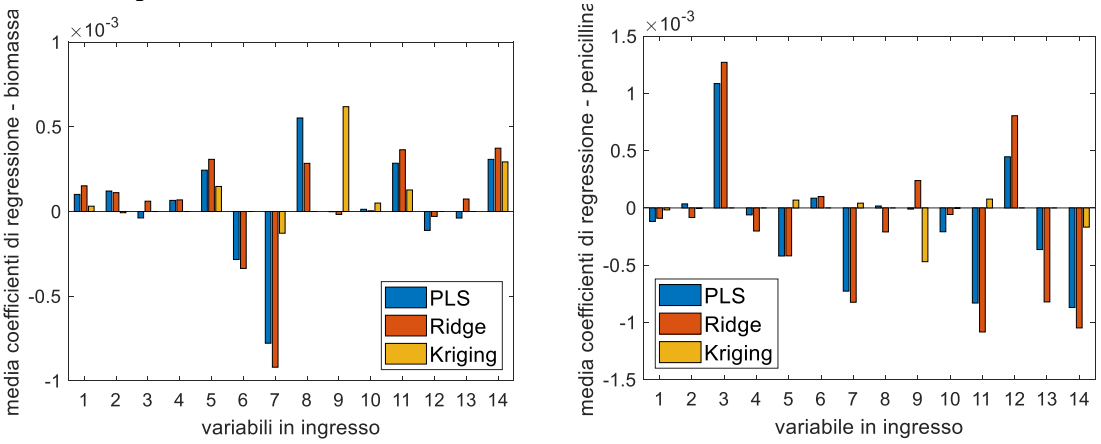


(a) *(b)*

Figura 3.3. *Caso studio 1, previsione del valore finale di qualità: traiettoria temporale della media dei coefficienti di regressione di diverse variabili di processo, sia del modello PLS che Ridge, di (a) concentrazione di biomassa e (b) concentrazione di penicillina.*

I due modelli danno indicazioni simili (con PLS più stabile e meno rumoroso nel tempo) e sottolineano che:

- i predittori nelle prime 30 ore del batch sono inversamente correlati alla concentrazione finale di biomassa, che invece è correlata positivamente ai predittori nella seconda parte del batch;
- i predittori sembrano inversamente correlati alla concentrazione finale di penicillina nella seconda parte del batch.



(a) *(b)*

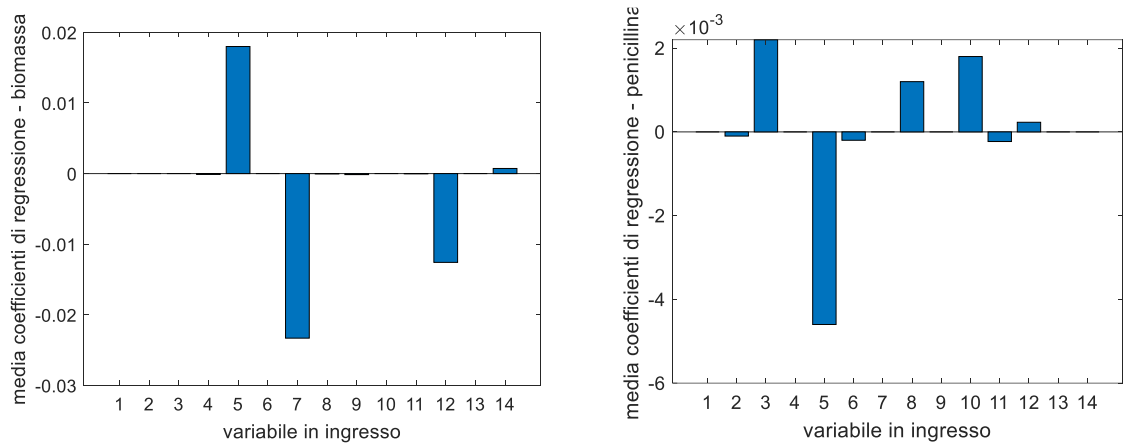
Figura 3.4. *Caso studio 1, predizione del valore finale di qualità: valore medio, in funzione della variabile in ingresso, dei coefficienti di regressione del modello PLS, Ridge, LASSO e Kriging su (a) concentrazione di biomassa e (b) concentrazione di penicillina.*

Uno studio (Figure 3.4) che mostra il contributo medio su tutta la durata del batch di ciascuna variabile di processo (ciascun predittore mediato nel tempo per tutta la durata del batch) del modello mostra che:

- le variabili che più influiscono sulla concentrazione finale della biomassa (Figura 3.4a) sono: concentrazione di substrato (5), concentrazione di ossigeno (6), volume di coltura (7), concentrazione di CO_2 (8), calore generato (11) e portata di acqua di raffreddamento (14); il volume di coltura e concentrazione di ossigeno sono correlate negativamente, mentre concentrazione di substrato, concentrazione di CO_2, calore generato e portata di acqua di raffreddamento sono correlate positivamente.
- le variabili che più influiscono sulla concentrazione finale della penicillina (Figura 3.4b) sono: portata di alimentazione di substrato (3), concentrazione di substrato (5), volume di coltura (7) e tutte le variabili di controllo (12, 13 e 14) su temperatura e pH. Portata di alimentazione di substrato e portata di controllo pH acido sono correlate positivamente, mentre volume di coltura, calore generato, portata di controllo pH base e portata di acqua di raffreddamento sono correlate negativamente.

Questi risultati hanno senso da un punto di vista fisico, in quanto ci si aspetta che la concentrazione di biomassa finale sia influenzata maggiormente dai volumi e substrato disponibile ad inizio processo, come ci si aspetta pure che la buona riuscita del processo sia legato al controllo del sistema, in quanto ogni antibatterico è fortemente sensibile sia a variazioni della temperatura che del pH. Nelle successive Figure (3.6 e 3.7) verranno ripetuti, questa volta in dettaglio, i coefficienti di regressione del modello LASSO, ad evidenziarne il comportamento in funzione delle variabili in ingresso. Siccome lo scopo della regressione LASSO è quello di forzare il maggior numero di predittori ad essere nullo, ci si aspetta che molte delle variabili non saranno significative in maniera compatibile l'una all'altra, ma esisteranno dei predittori maggiormente significativi nei confronti del valore finale delle variabili qualità, peculiarità che si può notare immediatamente sia in Figura 3.6 che Figura 3.7.

In questo caso studio, si possono subito distinguere sia le variabili in ingresso (Figura 3.6a-b) che i predittori temporali (Figura 3.7) maggiormente significativi, infatti si riscontra una maggior importanza individuata dal modello verso le variabili 5 (concentrazione di substrato), variabile 7 (volume di coltura) e variabile 12 (portata di controllo pH acido) per quanto riguarda la predizione della concentrazione finale di biomassa, mentre variabile 3 (portata di alimentazione di substrato), variabile 5 (concentrazione di substrato), variabile 8 (concentrazione di anidride carbonica) e variabile 10 (temperatura del reattore) sulla predizione della concentrazione finale di penicillina.



(a) *(b)*

Figura 3.6. *Caso studio 1, predizione del valore finale di qualità: valore medio, in funzione della variabile in ingresso, dei coefficienti di regressione del modello LASSO sulla concentrazione di biomassa (a) e concentrazione di penicillina (b)*

Il modello LASSO identifica come uniche variabili di processo più correlate alla qualità finale: volume di coltura (7) e portata di controllo pH acido (12) per la concentrazione di biomassa, mentre portata di alimentazione del substrato (3), concentrazione di substrato (5), concentrazione di CO_2 (8) e temperatura del fermentatore (10) per la concentrazione di penicillina.

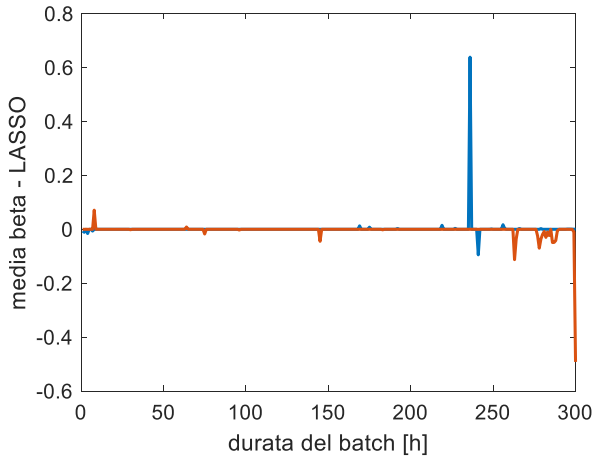


Figura 3.7. *Caso studio 1, previsione del valore finale di qualità: traiettoria della media dei coefficienti di regressione del modello LASSO su concentrazione di biomassa e concentrazione di penicillina*

In figura 3.7 si osserva che l'andamento temporale dei predittori mostra maggiore significatività nella fase di stazionarietà nella fase stazionaria finale del batch, sia per la concentrazione finale di biomassa che per la concentrazione finale di penicillina, ma la selezione delle variabili operata da LASSO mostra un risultato aleatorio.

3.1.1.4 Indici di prestazione per l'identificazione del metodo più accurato

Per cercare di comparare le varie tecniche di regressione in maniera il più oggettiva possibile, si è deciso di introdurre un KPI *(Key Performance Index),* basato sui risultati e mostrati sopra. In particolare, al metodo di regressione migliore ciascun indice di accuratezza, errore relativo medio, RMSE, e SMAE, vengono attribuiti 2 punti, mentre 1 punto viene dato al secondo migliore. 0.5 punti ai rimanenti, così da ottenere un contributo *KPI* totale pari ad un massimo di 6.

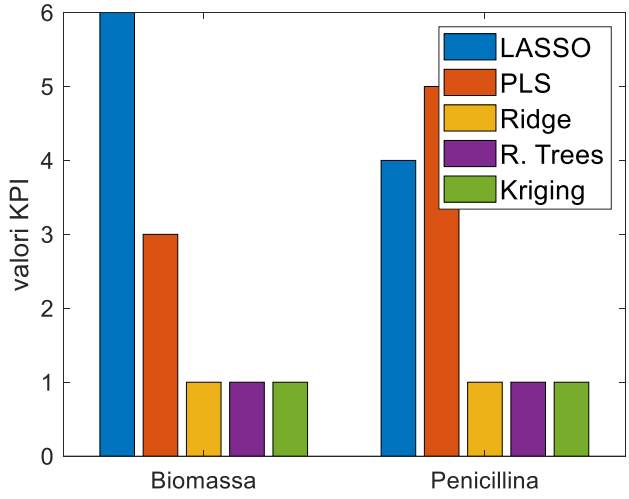


Figura 3.8. *Caso studio 1, previsione del valore finale di qualità: confronto tra indici di prestazione su ogni sensore virtuale.*

Figura 3.8 mostra la superiorità del metodo LASSO in termini di accuratezza di predizione rispetto alle rimanenti tecniche di regressione, nel caso della predizione della concentrazione di biomassa, mentre per la concentrazione di penicillina, il metodo PLS risulta migliore.

3.1.2 Stima della traiettoria temporale delle concentrazioni di biomassa e penicillina

In questo paragrafo vengono confrontate le prestazioni predittive di diversi sensori virtuali nel problema della stima della traiettoria temporale per tutta la durata del batch mediante approccio evolutivo.

3.1.2.1 Confronto delle prestazioni dei modelli di regressione sulla stima della traiettoria temporale di concentrazione di biomassa e penicillina

In Tabella 3.7 e Tabella 3.8 sono riportati gli indici delle prestazioni predittive per i diversi sensori virtuali al fine di permetterne un confronto oggettivo su entrambe le variabili di qualità del prodotto. Il confronto è stato fatto distinguendo due fasi operative: la fase di crescita iniziale (tra le 30 e 110 ore) e la fase stazionaria e finale (dopo le 150 ore), della traiettoria temporale sia per la concentrazione di biomassa che per la concentrazione di penicillina. La fase antecedente le 30 ore non è stata ritenuta

significativa, in quanto non avviene produzione di penicillina, mentre la fase tra 110 e 150 ore è una fase di transizione tra fase di crescita esponenziale microbiotica e fase di stazionarietà.

Tabella 3.7. *Caso studio 1, stima del profilo temporale della qualità in tempo reale***:** *confronto delle prestazioni predittive in convalida dei diversi sensori virtuali per la predizione di (a) concentrazione di biomassa e (b) concentrazione di penicillina.*

(a)

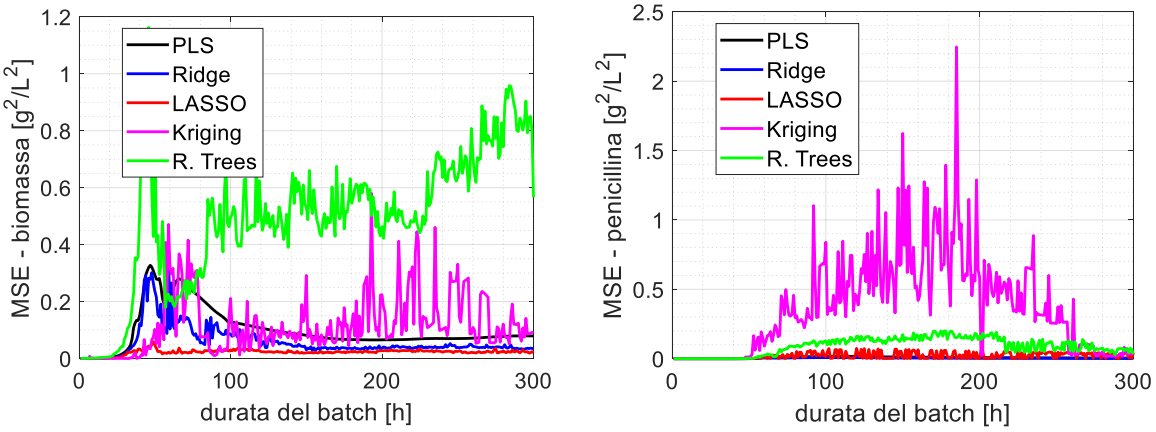
Concentrazione di biomassa						
	30<ore<110			**Ore >150**		
	MRE	**RMSE**	**SMAE**	**MRE**	**RMSE**	**SMAE**
PLS	**1.3809**	0.2267	0.0582	1.1636	0.1825	**0.0667**
Ridge	6.8138	0.8665	0.2642	3.0766	0.4981	0.0924
LASSO	1.5062	**0.1658**	**0.0419**	**1.0548**	**0.1546**	0.0751
Kriging	3.8263	0.3644	0.1299	1.3376	0.2335	0.0994
R. Trees	5.0665	0.5229	0.1778	4.5958	0.6417	0.1945

(b)

Concentrazione di penicillina						
	30<ore<110			**Ore >150**		
	MRE	**RMSE**	**SMAE**	**MRE**	**RMSE**	**SMAE**
PLS	27.748	0.1285	0.2843	7.5555	0.0608	0.1223
Ridge	123.902	0.1353	0.2839	9.4280	0.1185	0.3535
LASSO	33.546	0.0938	0.2203	5.7415	0.1217	0.1302
Kriging	54.763	0.1523	0.2206	9.0230	0.1183	0.1720
R. Trees	29.434	0.1367	0.2267	8.2038	0.1895	0.4023

3.1.2.2 Comparazione dell'andamento dell'errore quadratico medio

La prestazione di ogni modello è stata valutata nella predizione di ogni istante temporale in tempo reale. A questo scopo vengono mostrati i profili temporali di MSE come indicatore dello scostamento tra predizione e misurazione.



(a) *(b)*

Figura 3.9. *Caso studio 1, stima del profilo temporale della qualità in tempo reale***:** *confronto tra profili temporali del MSE calcolati da ogni tecnica di regressione, su (a) concentrazione di biomassa e (b) concentrazione di penicillina.*

Per quanto riguarda la concentrazione di biomassa si può notare che:

- gli errori più alti sono nella fase iniziale del processo, intorno all'ora 50 dall'inizio del batch

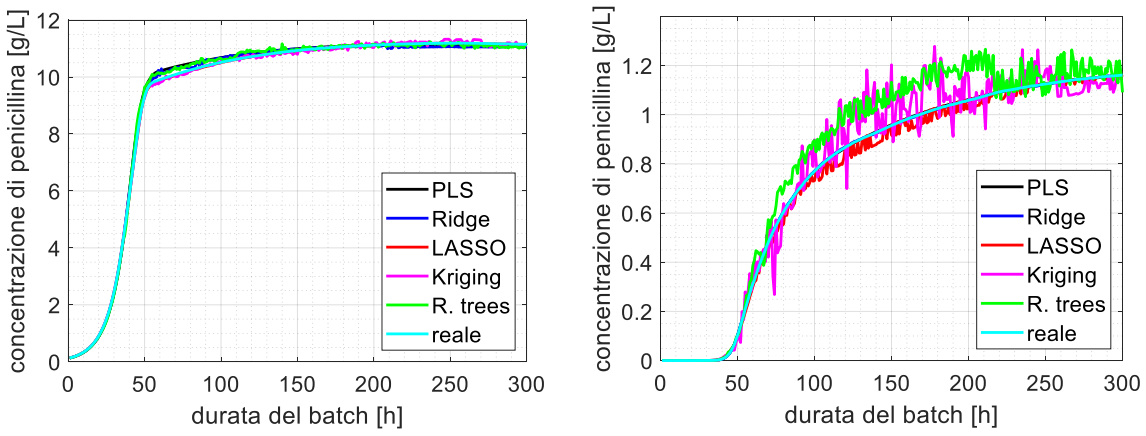
.

Questa criticità, comune a tutti i modelli di regressione, è generata dal passaggio di fase nel reattore da continuo a batch per effetto dell'alimentazione del substrato. Sulla concentrazione di Penicillina non esistono criticità analoghe, infatti i contributi degli scostamenti tra predizione e misurazione sono dispersi in maniera omogenea su tutto il profilo temporale di processo, tranne che nella parte iniziale in cui la penicillina non viene ancora prodotta.

- Regression Trees e Kriging sono meno accurati e anche meno precisi, in quanto danno stime più rumorose.
- LASSO, PLS e Ridge sono molto accurati e precisi.

A dimostrazione di quanto detto sopra, si mostrano come esempio in Figura 3.10 un caso di predizione della traiettoria temporale stimata per la concentrazione di biomassa (Figura 3.10a) e per la concentrazione di penicillina (Figura 3.10b) da ciascun sensore virtuale. Si osserva che:

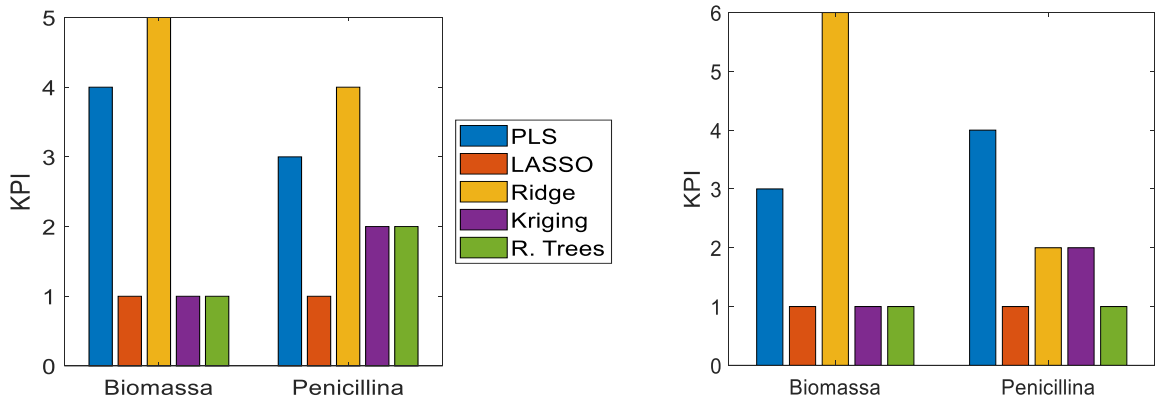
- la stima della concentrazione di biomassa è più accurata e precisa di quella della penicillina, che evidentemente è più critica da stimare.
- la stima mediante PLS, LASSO e Ridge è molto più accurata e precisa.



(a) *(b)*

Figura 3.10. *Caso studio 1, stima del profilo temporale della qualità in tempo reale: confronto tra profili temporali medi di (a) concentrazione di biomassa e di (b) concentrazione di penicillina, secondo ogni sensore virtuale.*

3.1.2.3 Indici di prestazione per l'identificazione del metodo più accurato



(a) *(b)*

Figura 3.11. *Caso studio 1, stima della traiettoria temporale della qualità in tempo reale***:** *confronto tra indici di prestazione di ogni sensore virtuale per traiettoria temporale di concentrazione di biomassa e concentrazione di penicillina, in (a) fase di crescita e in (b) fase di stazionarietà.*

I risultati dei KPI in Figura 3.11 sono compatibili con quelli legati alla stima del valore finale, infatti il metodo di regressione LASSO rimane superiore sulla concentrazione di biomassa lungo tutto l'arco temporale (sia in fase di crescita, che in fase di stazionarietà), come anche il metodo di regressione PLS per quanto riguarda la concentrazione di penicillina in fase di stazionarietà. Nella fase di crescita della concentrazione di penicillina non si osservano metodi che prevalgono sugli altri.

3.2 Risultati e discussione per il Caso studio 2

Per quanto riguarda il secondo caso studio, che consiste in un processo di trattamento acque reflue. Le stesse tecniche di soft-sensing del paragrafo 3.1 vengono paragonate e le relative prestazioni predittive discusse in modo critico.

3.2.1 Stima dei valori delle variabili qualità in uscita dal processo e ad ogni sezione

Per la valutazione delle capacità predittive dei sensori virtuali si è adottata una procedura di convalida su 50 iterazioni in cui la partizione casuale dei dati disponibili $\mathbf{X}[380 \times 29]$, con 380 campionamenti giornalieri e 29 variabili di ingresso, Tabella 2.3, è di 67% in calibrazione e 33% in convalida (cioè, 255 osservazioni in calibrazione e 125 in convalida). Le variabili di qualità del refluo trattato (variabili di risposta) sono 9 e sono riportate in Tabella 2.4 nel secondo capitolo. Come nel caso precedente gli indici considerati per valutare la prestazione di stima delle tecniche di regressione utilizzate sono: *MRE (mean relative error), SMAE (standardized mean square error)* e *RMSE (root mean square error).* Siccome i fenomeni di decantazione possono protrarsi per giorni, si è deciso di implementare un approccio dinamico alle tecniche di regressione, così da introdurre una dipendenza

delle variabili qualità del refluo trattato fino a due giorni prima, tenendo conto dei tempi di permanenza nelle singole operazioni unitarie.

In Tabella 3.10, il sensore virtuale si è dimostrato tendenzialmente meno accurato per le operazioni unitarie a monte del processo (si ricordi infatti che la sigla *P* indica il primo sedimentatore, *S*, il secondo, mentre, *G*, l'uscita dall'impianto). Le prestazioni predittive dei modelli di regressione miglioreranno (secondo tutte le metriche di analisi) lungo il processo: infatti, nelle operazioni unitarie iniziali del processo si osserva un errore relativo del 15-20%, mentre l'errore scende fino a 1-3% nelle sezioni finali dell'impianto. Siccome nelle Tabelle 3.11 e 3.12, l'indice SMAE e l'indice RMSE non mostrano difformità significative di prestazioni questo significa che l'errore relativo è principalmente alto per via della mancanza di un numero adeguato di predittori a spiegare il sistema.

Tabella 3.10. *Caso studio 2, stima della qualità del refluo trattato: confronto dell'errore relativo medio di convalida per i sensori virtuali considerati.*

	MRE				
	PLS	**Ridge**	**LASSO**	**Kriging**	**R. Trees**
DBO-P	27.99 ± 13.19	16.48 ± 5.98	**13.52 ± 4.56**	30.34 ± 1.73	17.63 ± 7.32
SS-P	11.48 ± 1.82	**10.31 ± 1.88**	13.53 ± 2.89	16.90 ± 2.07	22.06 ± 2.03
SED-P	8.98 ± 2.15	*8.79 ± 2.54*	9.21 ± 2.46	**8.68 ± 2.29**	16.03 ± 2.51
DBO-S	3.73 ± 1.11	*3.12 ± 1.43*	**1.83 ± 0.27**	3.46 ± 0.17	8.11 ± 2.23
DQO-S	6.40 ± 0.86	*5.47 ± 0.90*	**2.72 ± 0.62**	9.81 ± 0.37	15.97 ± 0.83
DBO-G	2.10 ± 0.86	2.80 ± 0.79	**0.98 ± 0.46**	2.21 ± 0.12	5.04 ± 1.30
DQO-G	4.38 ± 0.73	3.51 ± 0.53	**2.12 ± 0.32**	6.32 ± 0.22	9.72 ± 0.60
SS-G	2.37 ± 0.40	2.12 ± 0.32	**1.72 ± 0.20**	3.14 ± 0.20	5.97 ± 0.50
SED-G	1.92 ± 0.53	1.77 ± 0.71	**0.43 ± 0.20**	1.06 ± 0.17	3.57 ± 0.53

Tabella 3.11. *Caso studio 2, stima della qualità del refluo trattato: confronto dell'errore assoluto medio standardizzato di convalida per i sensori virtuali considerati.*

	SMAE				
	PLS	**Ridge**	**LASSO**	**Kriging**	**R. Trees**
DBO-P	0.42 ± 0.03	0.37 ± 0.03	**0.26 ± 0.01**	0.31 ± 0.11	0.49 ± 0.05
SS-P	0.35 ± 0.04	**0.22 ± 0.04**	0.34 ± 0.08	0.35 ± 0.26	0.31 ± 0.09
SED-P	0.35 ± 0.07	**0.20 ± 0.06**	0.32 ± 0.06	0.34 ± 0.25	0.29 ± 0.08
DBO-S	0.30 ± 0.03	0.35 ± 0.07	**0.19 ± 0.07**	0.30 ± 0.16	0.32 ± 0.07
DQO-S	0.29 ± 0.04	0.22 ± 0.05	**0.22 ± 0.01**	0.30 ± 0.11	0.35 ± 0.05
DBO-G	0.23 ± 0.10	0.27 ± 0.04	**0.12 ± 0.06**	0.16 ± 0.06	0.17 ± 0.07
DQO-G	0.19 ± 0.05	0.16 ± 0.05	**0.12 ± 0.02**	0.28 ± 0.10	0.23 ± 0.06
SS-G	0.23 ± 0.05	0.21 ± 0.04	**0.17 ± 0.05**	0.21 ± 0.09	0.22 ± 0.08
SED-G	0.18 ± 0.02	0.19 ± 0.04	**0.10 ± 0.01**	0.16 ± 0.06	0.21 ± 0.06

Tabella 3.12. *Caso studio 2, stima della qualità del refluo trattato: confronto della radice dell'errore medio quadratico di convalida per i sensori virtuali considerati.*

	RMSE				
	PLS	**Ridge**	**LASSO**	**Kriging**	**R. Trees**
DBO-P	10.80 ± 0.50	6.03 ± 0.55	**4.87 ± 0.36**	12.29 ± 3.74	7.56 ± 0.94
SS-P	8.86 ± 0.86	**7.53 ± 0.74**	9.00 ± 1.55	12.24 ± 7.87	14.88 ± 1.26
SED-P	6.62 ± 0.87	**6.29 ± 1.06**	6.44 ± 1.04	7.77 ± 3.95	11.47 ± 1.46
DBO-S	4.05 ± 0.49	3.37 ± 0.87	**1.98 ± 0.52**	3.59 ± 1.93	8.84 ± 1.71

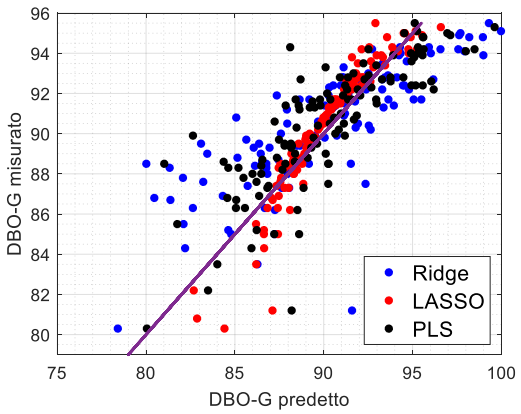
DQO-S	5.80 ± 0.54	4.72 ± 0.62	**2.79 ± 0.49**	8.46 ± 0.57	13.73 ± 0.71
DBO-G	2.79 ± 0.77	3.37 ± 0.81	**1.21 ± 0.50**	2.45 ± 0.72	5.82 ± 1.81
DQO-G	4.44 ± 0.59	3.38 ± 0.47	**2.13 ± 0.45**	6.30 ± 0.82	9.68 ± 0.68
SS-G	2.86 ± 0.48	2.55 ± 0.44	**2.14 ± 0.29**	3.62 ± 0.89	7.06 ± 0.77
SED-G	2.84 ± 0.85	3.05 ± 0.49	**0.83 ± 0.39**	1.45 ± 0.69	4.56 ± 1.42

In Tabella 3.13, i coefficienti di determinazione in convalida di ogni modello.

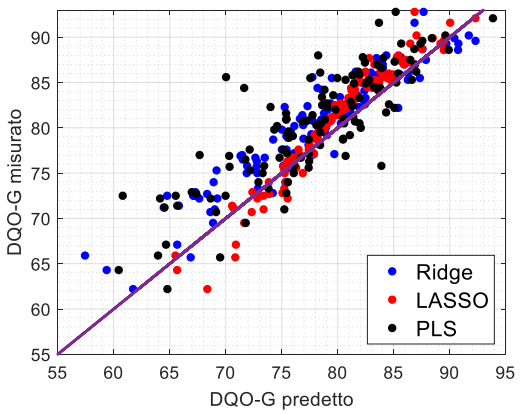
Tabella 3.13. *Caso studio 2, stima della qualità del refluo trattato: confronto dei coefficienti di determinazione di convalida per i sensori virtuali considerati.*

	R^2				
	PLS	**Ridge**	**LASSO**	**Kriging**	**R. Trees**
DBO-P	0.61	**0.94**	0.89	0.69	0.47
SS-P	0.73	**0.81**	0.72	0.71	0.40
SED-P	0.79	**0.87**	0.85	0.89	0.42
DBO-S	0.85	0.90	**0.97**	0.90	0.62
DQO-S	0.91	0.94	**0.95**	0.70	0.81
DBO-G	0.82	0.86	**0.97**	0.92	0.68
DQO-G	0.84	0.95	**0.97**	0.74	0.59
SS-G	0.87	0.88	**0.89**	0.79	0.50
SED-G	0.89	0.94	**0.96**	0.83	0.52

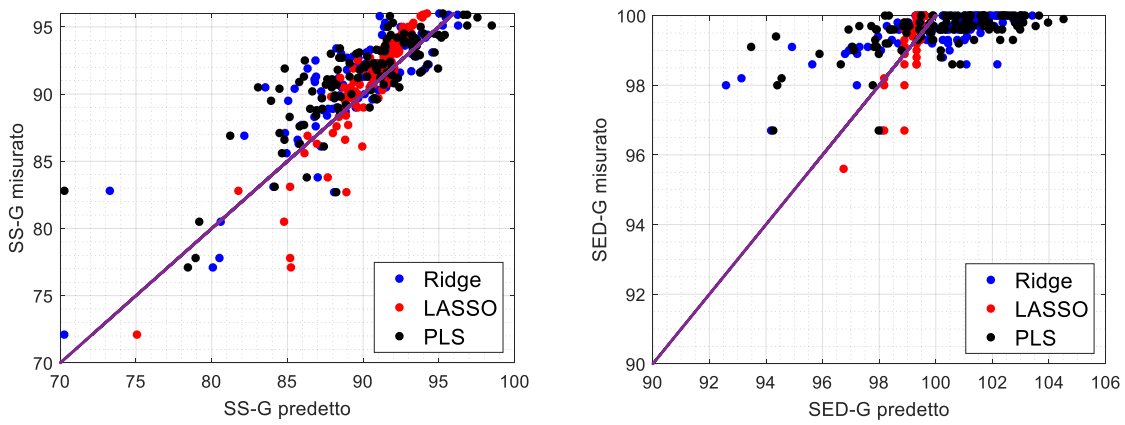
Visti i risultati migliori nelle variabili legate agli indici globali di processo, ovvero DBO-G, DQO-G, SS-G e SED-G, si è deciso, per brevità, di analizzare attraverso *parity plot* solamente le tecniche LASSO, Ridge e PLS.



(a)



(b)



(c) *(d)*

Figura 3.12. *Caso studio 2, stima della qualità del refluo trattato***:** *parity plot dei modelli di regressione per la predizione dei valori (a) DBO-G, (b) DQO-G, (c) SS-G e (d) SED-G.*

In Figura 3.12a-b-c-d, i *parity plot* sulle quattro variabili di prestazione a fine processo, rispettivamente sul consumo di ossigeno da parte di fenomeni biologici, sul consumo di ossigeno da parte di fenomeni chimici, presenza di solidi in sospensione e presenza di sedimenti. I modelli di regressione Kriging si evidenzia ancora una volta come non ottimale nella predizione delle variabili qualità prodotto, rispetto ai modelli LASSO, Ridge o PLS. Sono stati ottenuti risultati scadenti con il metodo Regression Trees.

3.2.2 Analisi robustezza e discussione delle caratteristiche dei modelli di regressione

Tenendo in considerazione solamente le variabili legate alla sezione finale di processo (successiva al secondo sedimentatore), nelle Figure 3.13, 3.14, 3.15 e 3.16 vengono riportate i coefficienti di regressione mediati, in maniera tale da comprendere quali tra le variabili in ingresso influiscano maggiormente sulla predizione di una variabile di risposta. Sono considerati i risultati ottenuti mediante LASSO, in quanto indicativi dei risultati ottenuti con le altre metodiche. Inoltre, sono state selezionate unicamente le variabili più importanti (circa 15 su 29 variabili in ingresso), tralasciando le variabili per cui i coefficienti di regressione sono trascurabili.

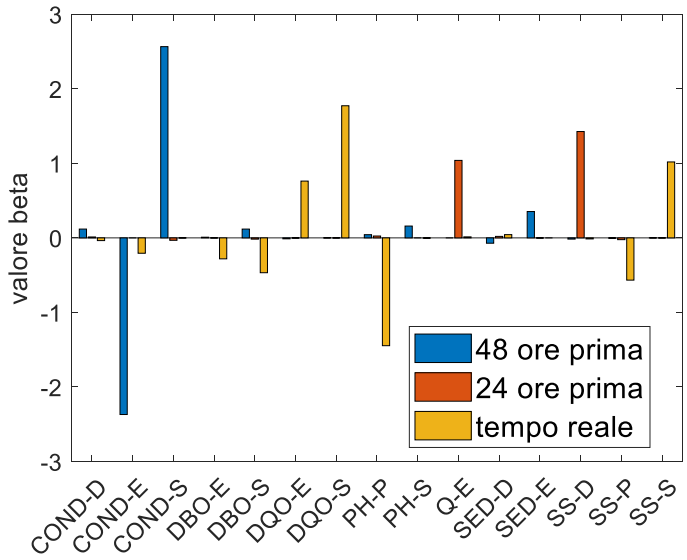


Figura 3.13. *Caso studio 2, stima della qualità del refluo trattato***:** *coefficienti di regressione in funzione delle variabili in ingresso su DBO-G.*

Per quanto riguarda l'indice di consumo di ossigeno legato a trattamenti biologici (Figura 3.13), le variabili maggiormente correlate sono quelle i cui valori dei coefficienti di regressione appaiono mediamente più elevati; esse sono le variabili DQO-S e PH-P (rispettivamente, correlata e anticorrelata) ovvero rispettivamente la portata di ossigeno in uscita dal secondo sedimentatore e il pH in ingresso al primo sedimentatore. Per quanto riguarda le misurazioni effettuate nei due giorni precedenti, si evidenzia la conduttività in ingresso e uscita all'impianto (COND-E e COND-S, rispettivamente anticorrelato e correlato), mentre nelle misurazioni del giorno precedente si evidenzia una correlazione di Q-E (portata in ingresso all'impianto) e SS-D (presenza di solidi in sospensione al secondo sedimentatore).

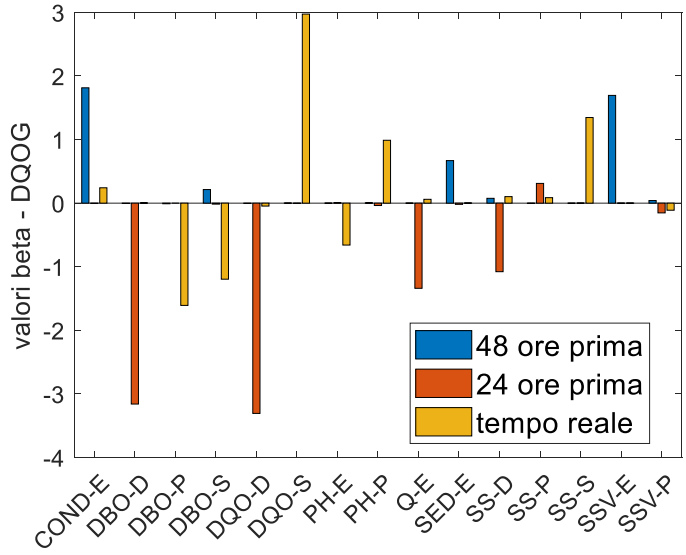


Figura 3.14. *Caso studio 2, stima della qualità del refluo trattato***:** *coefficienti di regressione in funzione delle variabili in ingresso su DQO-G.*

Le variabili in ingresso che maggiormente influiscono sull'indice globale di consumo di ossigeno su trattamenti chimici sono le portate di ossigeno in uscita dall'impianto legato a trattamenti chimici (DQO-S), in ingresso all'impianto legato a trattamenti biologici (DBO-P) e solidi in sospensione in uscita all'impianto (SS-S), per quanto riguarda i valori misurati in tempo reale. Le misurazioni del giorno precedente sono DBO-D, DQO-D, Q-E e SS-D, tutte anticorrelate, mentre le misurazioni effettuate due giorni prima sono COND-E e SSV-E, correlate.

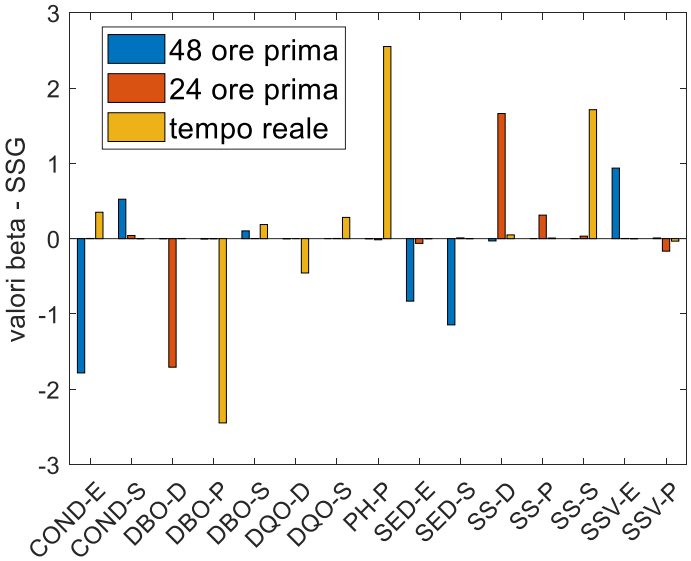


Figura 3.15. *Caso studio 2, stima della qualità del refluo trattato***:** *coefficienti di regressione in funzione delle variabili in ingresso su SS-G.*

Nel caso della predizione sull'indice di abbattimento dei solidi in sospensione lungo il processo, le variabili più significative in tempo reale sono la portata di ossigeno in entrata per trattamenti biologici (DPO-P) e pH (PH-P) al primo decantatore, rispettivamente anticorrelata e correlato. Nei predittori

di riferimento 24 ore prima, si evidenziano unicamente la portata di ossigeno dedicato a trattamenti biologici (DBO-D, anticorrelato) e la quantità di solidi in sospensione (SS-D, correlato) al secondo sedimentatore. Analogamente, si individuano COND-E (conduttività all'ingresso dell'impianto, anticorrelato), SED-E (quantità di sedimenti in ingresso all'impianto, anticorrelato), SED-S (anticorrelato) e SSV-E (solidi volatili in sospensione in ingresso all'impianto, correlato) per quanto riguarda i predittori misurati nei due giorni precedenti.

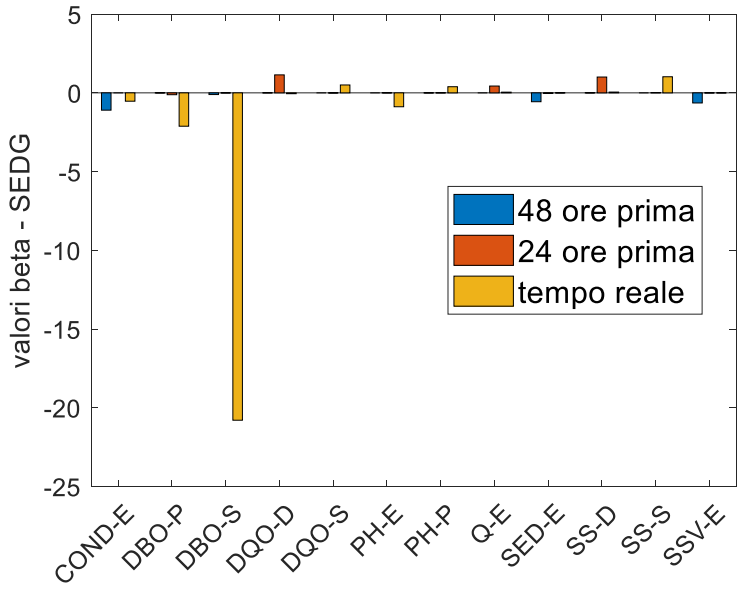


Figura 3.16. *Caso studio 2, stima della qualità del refluo trattato***:** *coefficienti di regressione in funzione delle variabili in ingresso su SED-G.*

La variabile più correlata all'indice globale di performance dei sedimenti nel processo (SED-G) è la portata di ossigeno dedicato a trattamenti biologici, in uscita dall'impianto (DBO-S).

3.2.3 Analisi KPI

Il confronto dei sensori virtuali è fatto anche in questo caso in termini del *key performance index* definito al Paragrafo 3.1.1.4.

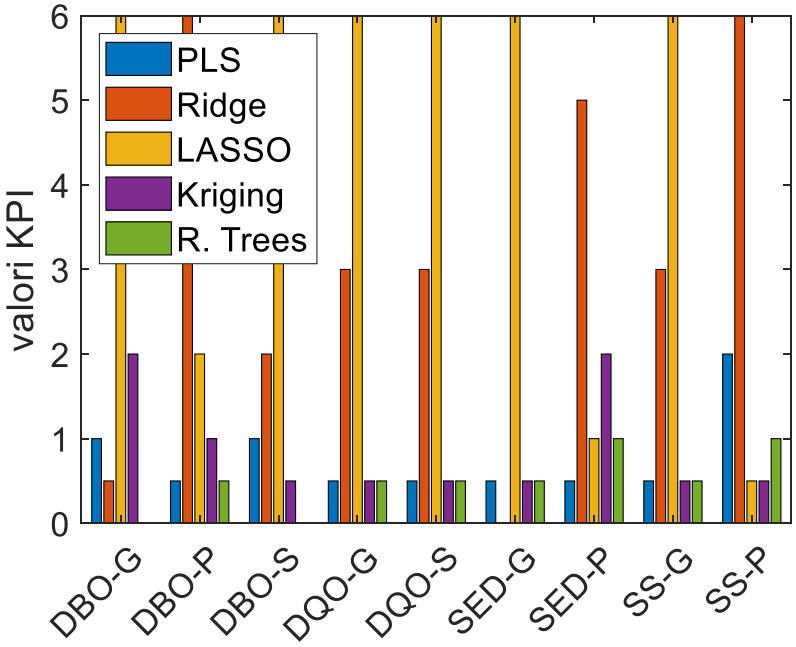


Figura 3.17. *Caso studio 2, stima della qualità del refluo trattato***:** *confronto tra indici di prestazione su ogni sensore virtuale.*

Figura 3.17 conferma la superiorità del metodo LASSO, che mostra prestazioni migliori su tutti gli indici presi in considerazione, e per tutte le variabili in uscita legate alla prestazione globale di processo e al secondo sedimentatore, mentre il metodo Ridge prevale nella stima delle variabili in uscita legate al primo sedimentatore, solamente a seguito dell'implementazione di un approccio dinamico con dipendenza temporale delle misurazioni ai due giorni precedenti.

Conclusioni

L'obiettivo di questa tesi è stato il confronto tra cinque metodologie di modellazione multivariata basata su dati per il soft-sensing in sistemi chimici e biologici, evidenziandone criticità, vantaggi e similarità. Le tecniche implementate e confrontate sono: proiezione su strutture latenti, regressione Ridge, regressione LASSO, Kriging e Regression Trees. Queste sono state confrontate su due casi di studio: la produzione semi-continua simulata di penicillina mediante fermentazione e un processo continuo reale per il trattamento di acque reflue.

Per quanto riguarda il primo caso studio, riguardante la produzione semibatch di penicillina, è stata implementata una metodologia evolutiva per la stima della traiettoria temporale, mentre nel secondo caso studio, riguardante un processo continuo di trattamento di acque reflue, sono state applicate per la stima di parametri di qualità dell'effluente in uscita dal processo delle metodologie di modellazione dinamica. Per determinare la tecnica di regressione migliore in termini di accuratezza è stato implementato un metodo per l'identificazione di un *key performance index* (KPI).

Il metodo LASSO si è rivelato come più accurato nella maggior parte delle stime, sebbene a scapito di una bassa velocità d'esecuzione. Il metodo PLS è risultato essere un buon compromesso tra costo computazionale ed accuratezza, poiché, a fronte di prestazioni di stima poco inferiori rispetto a LASSO e Ridge, il tempo richiesto da questo metodo è di circa un decimo rispetto alla tecnica LASSO. Inoltre, si è rivelato come il più robusto, adattandosi particolarmente bene ad ogni ripartizione analizzata. I metodi Kriging e Regression Trees non hanno avuto risultati comparabili ai precedenti, in termini di accuratezza.

Su entrambi i casi studio è stato possibile identificare quali siano le variabili di processo che più sono correlate con la qualità del prodotto finale. Nel primo caso studio, infatti, si sono identificate come molto significative per la stima concentrazione di biomassa e penicillina finali il volume di coltura, la concentrazione iniziale di substrato e la portata di alimentazione, oltre alle variabili controllate del sistema (temperatura e pH). Riguardo il secondo caso studio, le correlazioni individuate sono in funzione non solo dei predittori, ma anche della dinamicità, avendo implementato un approccio dinamico per la modellazione dei sensori virtuali.

Nel caso degli indici globali di processo, si possono individuare le seguenti correlazioni più significative: su DBO-G, le misurazioni in tempo reale più importanti sono DQO-S e PH-P, nei due giorni precedenti COND-S e COND-E, mentre per il giorno precedente Q-E e SS-D.

Su DQO-G, le misurazioni in tempo reale più importanti sono DQO-S, DBO-P e SS-S, nei due giorni precedenti SSV-E e COND-E, mentre per il giorno precedente DBO-D e DQO-D. L'indice di rimozione sedimenti è principalmente legato alla quantità di sedimenti in entrata e uscita da ogni sezione, analogamente all'indice di rimozione composti volatili.

Si suggerisce come lavoro futuro, dati gli alti errori di predizione nelle sezioni del primo e secondo decantatore, l'implementazione di metodi non-lineari, come ad esempio *ensemble learning* o reti neurali, al fine di migliorare dove sono stati riscontrati problemi di accuratezza con metodi ai minimi quadrati.

Appendice A

Tabella A.1.Caso studio 1: *valori e unità di misura per i parametri del modello di simulazione della produzione di Penicillina mediante PenSim.*

Condizioni iniziali del sistema	**Identificazione variabile e unità di misura**	**Valore**
Concentrazione di substrato	S_C (g/L)	15
Concentrazione di ossigeno disciolta nel sistema	C_L (C_L^* in saturazione) (g/L)	1.16
Concentrazione di biomassa	B (g/L)	0.1
Concentrazione di penicillina	P (g/L)	0
Volume di coltura	V (L)	100
Concentrazione di anidride carbonica	CO_2(mmole/L)	0.5
Concentrazione di ioni idrogeno	$[H^+]$ (mol/L)	$10^{-5.1}$
Temperatura	T (K)	297
Calore generato	Q_{react} (cal)	0
Variabili e parametri cinetici		
Concentrazione di substrato in alimentazione	s_f (g/L)	600
Portata di alimentazione di substrato	F (L/h)	0.040
Temperatura di alimentazione di substrato	T_f (K)	298
Costante di resa 1	$Y_{x;s}$(g biomassa/g glucosio)	0.45
Costante di resa 2	$Y_{x,o}$ (g biomassa/g ossigeno)	0.04
Costante di resa 3	$Y_{p;s}$ (g penicillina/g glucosio)	0.90
Costante di resa 4	$Y_{p,o}$ (g penicillina/g ossigeno)	0.20
Costante cinetica 1	K_1 (mol/L)	10^{-10}
Costante cinetica 2	K_2 (mol/L)	$7\cdot 10^{-5}$
Coefficiente di mantenimento sul substrato	m_x (per h)	0.014
Coefficiente di mantenimento su ossigeno	m_o (per h)	0.467
Costante di relazione tra CO_2 e crescita	α_1 (mmol CO_2/g biomassa)	0.143
Costante di relazione tra CO_2 e mantenimento energia	α_2 (mmol CO_2/g biomassa orari)	$4\cdot 10^{-7}$
Costante di relazione tra CO_2 e produzione di penicillina	α_3 (mmol CO_2/L orari)	10^{-4}
Tasso specifico di crescita massimo	μ_x(per h)	0.092
Costante di saturazione di Contois	K_x (g/L)	0.15
Costante di limitazione ossigeno (senza limitazioni)	K_{ox}; K_p	0
Costante di limitazione ossigeno (con limitazioni)	K_{ox}	$7\cdot 10^{-5}$
	K_p	$5\cdot 10^{-4}$
Tasso specifico di produzione penicillina	μ_p (per h)	0.005
Costante di inibizione	K_p (g/L)	0.0002
Costante di inibizione per la produzione del prodotto	K_l (g/L)	0.10
Costante	p	3
Costante rateo di idrolisi per la penicillina	K (per h)	0.04
Costante di Arrhenius di crescita	k_g	$7\cdot 10^{3}$
Energia attivazione di crescita	E_g (cal/mol)	5100
Costante di Arrhenius di morte cellulare	k_d	10^{33}
Energia attivazione di morte cellulare	E_d (cal/mol)	50000
Densità su capacità termica media	ρ c_p (per $1^{\circ}C$)	1/1500
Densità su capacità termica di liquido di raffreddamento	ρ_c c_{pc} (per $1^{\circ}C$)	1/2000
Resa per la generazione di calore	r_{q1} (cal/g biomassa)	60
Costante di generazione del calore	r_{q2} (cal/g biomassa orari)	$1.6783\cdot 10^{-4}$
Coefficiente di scambio termico con il liquido di raffreddamento	a (cal/h $^{\circ}C$)	1000
Portata di acqua di raffreddamento	F_c (L/h)	

Costante		b	0.60
Costanti nella K_{la}		α	70
		β	0.4
Costante in F_{loss}	λ (per h)		$2.5 \cdot 10^{-4}$
Costante di proporzionalità	γ (mol[H^+]/g biomassa)		10^{-5}
Parametri di controllo			
pH basico:			
		K_c	$8 \cdot 10^{-4}$4.2
		τ_1 (h)	0.2625
		τ_d (h)	
pH acido:			
		K_c	10^{-4}8.4
		τ_1 (h)	0.125
		τ_d (h)	
Temperatura in raffreddamento			
		K_c	70
		τ_1 (h)	0.5
		τ_d (h)	1.6
Temperatura in riscaldamento			
		K_c	5
		τ_1 (h)	0.8
		τ_d (h)	0.05

Riferimenti bibliografici

Anter, A., D. Gupta e O. Castillo (2019). A novel parameter estimation in dynamic model via fuzzy swarm intelligence and chaos theory for faults in wastewater treatment plant. Springer Nature.

Agterberg, F. P (1974), Geomathematics, Mathematical background and Geo-science applications, Elsevier Scientific publishing company, Amsterdam.

Birol G., C. Undey, A. Cinar (1998). PenSim v2.0 User's guide. Process Modeling, Monitoring and Control Research Group. Department of Chemical and Environmental Engineering Illinois Institute of Technology (USA)

Birol G., C. Undey, A. Cinar (2002). A modular simulation package for fed-batch fermentation: penicillin production. Department of Chemical and Environmental Engineering, Illinois Institute of Technology, Chicago (USA)

Birol G., A. Cinar, S.J. Parulekar, C. Undey (2003). Batch fermentation: modeling, monitoring and control, CRC press publisher.

Buhagiar, J (2017), Automatic segmentation of indoor and outdoor scenes from visual lifelogging, bachelor's degree thesis in *Artificial Intelligence*, University of Amsterdam.

Facco P., M. Largoni, E. Tomba, F. Bezzo, M. Barolo (2014), Transfer of process monitoring models between plants: Batch systems. chemical engineering research and design 92, 273-284, Dipartimento di ingegneria industriale, Università degli studi di Padova.

Fuente. C (2004) Supervisory systems in waste-water treatment plants: systematize their implementation. Ph.D., Università di Girona.

Georgakis, C., R. H. Storer, W. Ku (1995), Disturbance rejection and isolation by dynamic principal component analysis, Chemical Process Modeling and Control Research Center, Lehigh University, Bethlehem, USA.

Gunther J, D. E. Seborg, (2009). Process monitoring and quality variable prediction utilizing PLS in industrial fed-batch cell culture.

Hansen, P.C (1998), Rank-deficient and Discrete ill-posed problems, SIAM.

Kleijnen, J.P.C (2017), Kriging: methods and applications, SRNN Electronic Journal, Tilburg University

Krige, D.G (1951), A statistical approach to some mine valuations and allied problems at the Witwatersrand, Master's thesis of the University of Witwatersrand.

Nomikos P, J. F. MacGregor (1994). Multi-way partial least squares in monitoring batch processes. Department of Chemical Engineering, McMaster University, Hamilton, Ontario, Canada

Poch M. (1993), "A classification Methodology for ill-structured domains", Università di Barcellona.

Rasmussen, M.A. (2012), A tutorial on the Lasso approach to sparse modelling, Elsevier, Copenhagen University.

Snee, R. D. e Marquardt D. W. (1975), Ridge regression in practice, The American statistician, Febbraio 1975.

Wackernagel, Hans (2003), Multivariate geostatistics, Springer.

Wei-Win Loh, (2011), Classification and regression trees, John Wiley volume 1.

Wold, H. (1985), Partial least squares, Encyclopaedia of statistical sciences, pg. 581-591.

Siti web:

https://archive.ics.uci.edu/ml/datasets/Water+Treatment+Plant