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

TECNICHE DIAGNOSTICHE PER MODELLI A PRINCIPI PRIMI BASATE SU ANALISI DI DATI

Relatore: Prof. Massimiliano Barolo
Correlatrice: Dott. Natascia Meneghetti

Laureando: DANIELE CANAGLIA

ANNO ACCADEMICO 2015-2016

Riassunto

La determinazione delle cause delle discrepanze tra i dati storici di un processo e le stime calcolate tramite il modello a principi primi sviluppato per rappresentarlo (*process/model mismatch*, PMM) che spesso si riscontrano durante l'utilizzo o lo sviluppo del modello stesso, richiede spesso molto tempo e una discreta quantità di nuove campagne sperimentali. In questa Tesi, è stata analizzata una possibile soluzione a questo problema, considerando come caso studio la diagnosi di un modello a principi primi sviluppato per descrivere un processo di fermentazione di penicillina.

Come già rilevato dal lavoro di Ibrahim (2016), anche in questa Tesi è stato dimostrato che l'utilizzo della metodologia sviluppata da Meneghetti *et al*. (2014) per identificare la causa principale di un PMM tramite l'utilizzo dell'analisi alle componenti principali (PCA), presenta alcune limitazioni nell'analisi di variabili fortemente correlate tra loro. Per questo motivo è stata analizzata una seconda soluzione basata sulla metodologia sviluppata da Rato e Reis (2015) nell'ambito di monitoraggio di processo, già preliminarmente testata da Ibrahim (2016) e basata sull'analisi di coefficienti di correlazione parziale per confrontare la struttura di correlazione di un set di dati storici di processo e di un set di dati ottenuti tramite il modello a principi primi analizzato.

In particolare in questa Tesi sono stati identificati diversi limiti legati sia alla correlazione delle variabili indagate sia alla difficoltà dell'implementazione di tale tecnica in condizioni diverse da quelle per cui è stata sviluppata. Infine sono state proposte alcune soluzioni a questi problemi che riguardano da un lato l'analisi preliminare e l'eventuale pretrattamento dei dati in esame e dall'altro l'applicazione di una tecnica di decorrelazione delle variabili analizzate.

Indice

Introduzione

La modellazione di processo riveste un ruolo fondamentale nella nel supportare diverse attività industriali (come ad esempio nel controllo di processo, ottimizzazione e progettazione di prodotto e di processo) sia in ambito chimico e biologico che farmaceutico. La modellazione di processo può essere basata o sull'utilizzo di dati storici (modelli empirici, *data-driven*, DD) o sulla conoscenza fisica del processo (modelli a principi primi, *first-principles*, FP). I modelli empirici, grazie alla possibilità di poter essere costruiti rapidamente e al fatto di richiedere in generale una minore quantità di dati rispetto a quelli a principi primi, risultano molto più economici; d'altra parte tali modelli possono essere utilizzati solo nelle condizioni operative in cui sono stati raccolti i dati. I modelli a principi primi, invece, richiedono tempi realizzazione più lunghi e solitamente un numero molto elevato di prove sperimentali ma offrono una conoscenza più approfondita dei fenomeni alla base del sistema (Seborg 2010) e per questo sono spesso preferiti rispetto ai primi. Infatti questi modelli sono realizzati a partire dai bilanci di materia, di energia e di quantità di moto, traducendo i fenomeni chimici e fisici coinvolti in un processo industriale in equazioni matematiche.

Durante la fase di validazione del modello a principi primi o quando il modello viene utilizzato in condizioni diverse da quelle in cui è stato sviluppato (per esempio nel caso di scale-up di apparecchiature), è possibile che il modello dimostri scarsa aderenza alla realtà sperimentale causando un disallineamento tra i dati di processo raccolti, che rappresentano l'evoluzione effettiva del processo, e le predizioni del modello costruito (*process-model mismatch*, PMM). In generale, il PMM può essere:

- strutturale: se le equazioni del modello sono inadeguate, ad esempio perché la conoscenza del processo è limitata oppure perché la complessità dei fenomeni fisici coinvolti nel processo è stata semplificata in modo eccesivo;
- parametrico: se ad alcuni dei parametri del modello sono stati assegnati valori non appropriati a causa di eccessive approssimazioni di fenomeni fisici o a causa di adattamenti di parametri relativi a processi diversi da quelli in esame.

Attualmente, le tecniche più utilizzate per la correzione di un modello si basano sulla progettazione di esperimenti basata su modello (*model-based design of experiment*, MBDoE) (Franceschini *et al.,* 2008, Marquardt *et al.,* 2005). Queste tecniche possono essere molto dispendiose se non si conosce in anticipo quali equazioni o parametri è necessario analizzare. Una diagnosi preliminare in grado di individuare quali equazioni o parametri sono maggiormente responsabili del PMM permetterebbe quindi di ridurre o addirittura evitare l'utilizzo di tali tecniche con notevole risparmio di tempi e dei costi di intervento. (Meneghetti *et al.,* 2014). Una possibile soluzione è stata proposta da Meneghetti *et al.* (2014), sfruttando

l'analisi delle componenti principali (*principal component analysis*, PCA) per confrontare le strutture di correlazione dei dati di modello e dei dati di processo disponibili. In particolare, il *mismatch* viene analizzato tramite l'introduzione di nuove variabili, dette variabili ausiliarie, che rappresentano diverse combinazioni di variabili in uscita dal modello e dal processo e di parametri del modello a principi primi, selezionate in base alla struttura del modello stesso. I casi di studio indagati hanno tuttavia evidenziato alcuni limiti nell'applicazione di tale metodologia, soprattutto nel caso dell'analisi di modelli a principi primi in cui le variabili ausiliarie siano strettamente correlate tra loro. Per risolvere tale problema, è stata recentemente testata una soluzione alternativa (Ibrahim 2016) basata sul confronto delle strutture di correlazione dei dati di modello e dei dati di processo tramite l'utilizzo di coefficienti di correlazione parziale. Tale tecnica è stata sviluppata da Rato e Reis (2015) allo scopo di ottimizzare il monitoraggio di processo di sistemi stazionari. I primi risultati ottenuti, hanno confermato le potenzialità di tale tecnica, ma anche la presenza di numerosi aspetti da indagare. L'obiettivo di questa Tesi è quindi di valutare le eventuali limitazioni di questa seconda soluzione e delineare alcune linea guida generali che ne permettano l'appropriata implementazione allo scopo di una corretta diagnosi di modello.

A tal scopo è stato analizzato un modello a principi primi, che descrive un processo di fermentazione per la produzione di penicillina (Birol *et al.*, 2002). Tale modello è stato utilizzato sia per generare i dati 'storici' che i dati di modello. Tre diversi tipi di *mismatch* parametrici sono stati forzati introducendo diversi errori nel modello allo scopo di testare la nuova metodologia. Diversamente da ciò che è stato fatto da Ibrahim (2016) nell'adattamento della metodologia proposta da Rato e Reis (2015), in questa Tesi l'analisi è stata focalizzata nell'analizzare la sensibilità della tecnica diagnostica rispetto alle caratteristiche dei dati disponibili e nell'introduzione di una tecnica decorrelativa proposta dagli stessi autori (Rato e Reis, 2015) che permetta la corretta identificazione delle possibili cause di un PMM anche per variabili altamente correlate tra loro.

La Tesi è organizzata in cinque capitoli. Nel Capitolo 1 sono descritte le tecniche statistiche multivariate utilizzate per l'applicazione della prima metodologia, e i fondamenti matematici alla base della seconda metodologia utilizzata. Nel Capitolo 2 è riportato il modello a principi analizzato in questo studio e vengono discussi i tre casi studio analizzati, in ognuno dei quali viene modificato un diverso parametro. Nel Capitolo 3 sono presentati i risultati dell'applicazione della prima metodologia analizzata, considerando tutti e tre i casi studio in esame. Il Capitolo 4 e il Capitolo 5 riguardano l'applicazione della seconda metodologia, le cui prestazioni sono state testate considerando diversi dataset, diversi tipi di variabili (variabili misurate o variabili ausiliarie) e l'effetto del rumore nei dati disponibili.

CAPITOLO 1

Richiami di matematica e statistica multivariata

In questo Capitolo vengono presentate le tecniche statistiche alla base delle due metodologie analizzate in questa Tesi per l'identificazione delle possibili cause di diversi casi di visibili disallineamenti tra modello e processo (*process/model mismatch*, PMM). In particolare, la prima metodologia si basa sull'utilizzo di tecniche di statistica multivariata, mentre la seconda sull'analisi dei coefficienti di correlazione.

1.1 Modelli a variabili latenti

Attività industriali come il monitoraggio e il controllo di processo comportano la raccolta di dataset molto estesi specialmente per i processi batch. Spesso, la quantità di informazioni raccolta è ridondante e l'analisi dei dati forniti diviene molto complessa. Al fine di semplificare l'analisi di questi dati in modo da estrarre le informazioni necessarie, sono state proposte metodologie di analisi statistica multivariata.

Data una matrice di dati **X** di dimensione [$N \times K$], in cui N è il numero di campioni disponibili per K variabili misurate, i modelli a variabili latenti permettono di riassumere l'informazione contenuta in **X**, con un numero minore di variabili dette variabili latenti, (Eriksson *et al.*, 2001) rispetto al numero di variabili originali. Le variabili latenti vengono definite con l'obbiettivo di catturare la maggior parte della variabilità dei dati in analisi. Maggiore è la correlazione tra le K variabili originali, tanto minore sarà il numero di variabili latenti A rispetto a K. I modelli a variabili latenti vengono utilizzati sia per descrivere le relazioni tra le variabili di uno stesso set di dati **X** che le relazioni tra un set di regressori **X** [$N \times K$] e variabili risposta **Y** (e.g., specifiche di prodotto, variabili in uscita di un processo, etc.) [$N \times M$].

Grazie alla loro flessibilità e generalità, questi metodi sono già largamente utilizzati in diverse applicazioni industriali: *process understanding* (Soh *et al.* 2008), progettazione di prodotto e processo (Gabrielsson *et al.* 2002), monitoraggio e controllo di processo (Chew e Sharratt, 2010).

1.1.1 Analisi delle componenti principali (PCA)

L'analisi delle componenti principali (PCA, *principal component analysis*; Jackson, 1990), è un metodo statistico multivariato in grado di rappresentare in modo sintetico i dati di una matrice **X** [$N \times K$] dove N è il numero di campioni disponibili, e K il numero di variabili da analizzare (misure di processo o di qualità di un prodotto per esempio, normalmente correlate) in un nuovo spazio vettoriale le cui direzioni (ortogonali tra loro) sono chiamate componenti principali (*principal component*, PC). Tali variabili sono determinate in modo da massimizzare la variabilità dei dati catturata da ognuna di esse. Prima di applicare la PCA, ad ogni elemento della matrice **X** viene sottratta la media della rispettiva colonna (*mean centering*) e il risultato viene diviso per la deviazione standard della stessa (*scaling*). Questa trasformazione permette di traslare i dati all'origine del sistema di riferimento (K-spazio) poiché ogni colonna risulterà avere una media pari a zero e permette inoltre di rendere la variabilità di ciascuna variabile ugualmente importante nella costruzione del modello PCA (Wise *et al.,* 1996).
L'idea su cui si basa la PCA è che, se due o più variabili originali della matrice **X** sono correlate (ovvero il rango della matrice **X** è minore di K), allora è possibile individuare una direzione comune di variabilità che può essere descritta da una singola PC. In questo caso la matrice **X** può essere rappresentata con un numero A di PC, tale che A «min (N, K) (Valle *et al.,* 1999). A viene selezionato in modo tale da descrivere un'adeguata percentuale di variabilità del set di dati come descritto in §1.1.1.2 (Valle *et al.,* 1999). La ricerca delle direzioni del nuovo sistema di coordinate può essere indicato come un problema di ottimizzazione, la cui soluzione analitica è data dalla decomposizione degli autovettori della matrice di covarianza di **X** (Wise *et al.,* 1996, Burnham *et al.,* 1996):

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p}_1 = \lambda_1\mathbf{p}_1, \tag{1.1}$$

dove $\mathbf{p}_1$ è il vettore [$K \times 1$] dei coefficienti (chiamati *loadings*) per la prima componente principale. Il vettore dei *loadings* $\mathbf{p}_1$ rappresenta i coseni direttori della prima PC, ed è l'autovettore della matrice di covarianza di **X** (ossia, di $\mathbf{X}^{\mathrm{T}}\mathbf{X}$) corrispondente al più grande autovalore λ_1 di $\mathbf{X}\mathbf{X}^{\mathrm{T}}$, che è una misura della varianza spiegata dalla prima PC (Eriksson *et al.,* 2001). Quanto appena descritto per l'autovettore 1 può essere iterato per determinare tutte le S = rango (**X**) PC del modello PCA; vale a dire, tutti i *loadings* $\mathbf{p}_s$ (s = 1, 2, ..., S) del modello PCA, che sono ortonormali, tramite l'Eq. 1.1. Il vettore degli *scores* $\mathbf{t}_s$, ossia la proiezione dei dati originali lungo la direzione PC, è dato da:

$$\mathbf{t}_s = \mathbf{X}\mathbf{p}_s\,. \tag{1.2}$$

Si noti che i vettori degli *scores* sono ortonormali. Il set di dati **X** può essere rappresentato come il prodotto scalare degli S-vettori *lodings* e *scores*:

$$\mathbf{X} = \sum_{s=1}^{S} \mathbf{t}_s \mathbf{p}_s^{\mathrm{T}} \,. \tag{1.3}$$

Quindi, assumendo che solo le prime *A* PC vengano mantenute, e definendo la matrice degli *scores* $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_A]$ e la matrice dei *loadings* $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_A]$, la 1.3 può essere riscritta:

$$\mathbf{X} = \sum_{s=1}^{A} \mathbf{t}_s \mathbf{p}_s^{\mathrm{T}} + \sum_{s=A+1}^{S} \mathbf{t}_s \mathbf{p}_s^{\mathrm{T}} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E}\,, \tag{1.4}$$

dove $\mathbf{E}$ è la $[N \times K]$ matrice dei residui generati dalle $(S - A)$ PC scartate del modello PCA, quando $\mathbf{X}$ è ricostruita (cioè, approssimata) utilizzando solo le prime *A* PC (cioè, $\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}}$). La matrice dei residui riflette la variabilità dei dati che non viene catturata dal modello. Se gli elementi $\mathbf{e}_{n,k}$ del vettore $\mathbf{e}_k$ (ovvero la *k*-esima colonna di **E)** seguono una distribuzione normale, la variabilità non descritta dal modello è considerata non deterministica,

1.1.1.1 Interpretazione grafica della PCA

Di seguito in Figura 1.1, viene fornita una rappresentazione grafica del significato geometrico dell'analisi alle componenti principali, considerando, per semplicità, di considerare 7 campioni e di prendere in esame due variabili di processo, x_1 e x_2.

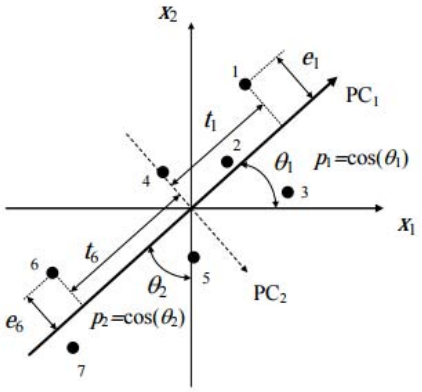


Figura 1.1. *Interpretazione geometrica di scores e loading della PCA di un set di dati con sette campioni e due variabili. Da: Tomba, (2013).*

Come si può notare, i dati seguono un andamento definito nello spazio (bidimensionale) delle variabili originali (x_k). Se si applica un modello PCA, la direzione di massima variabilità dei dati è identificata da PC_1. I *loadings* del modello ($\mathbf{p}_1$, $\mathbf{p}_2$) rappresentano i coseni direttori di PC_1, ovvero i coseni degli angoli tra le direzioni latenti e gli assi dello spazio delle variabili originali. Gli *scores* rappresentano le coordinate dei campioni di dati della matrice $\mathbf{X}$ nel nuovo sistema di riferimento rappresentato da PC_1. La mancanza di rappresentatività dei dati da parte del modello viene quantificata dai residui, rappresentati dalle distanze perpendicolari dei punti

dalla linea che rappresenta la direzione della PC_1. Nella Figura 1.1, la seconda componente principale che può essere valutata dai dati (PC_2) viene riportata come una linea tratteggiata. Come si può osservare la PC_2 è ortogonale alla PC_1, e rappresenta una variabilità molto limitata dei dati rispetto a quanto riportato dalla PC_1. In questo caso, può essere concluso che PC_1 è sufficiente per descrivere adeguatamente **X**.

1.1.1.2 Selezione del numero di PC

Il numero di PC per la costruzione del modello può essere selezionato in base a diversi criteri. In questa Tesi è stata considerata la regola dell'autovalore maggiore di 1 (Mardia et al., 1979). Tale regola prevede che le componenti principali i cui corrispondenti autovalori siano minori di 1 non vengano incluse nel modello. Questo criterio si basa sul fatto che i dati sono sottoposti a riduzione di scala e quindi si può assumere che l'autovalore associato a ogni PC rappresenti il numero di variabili la cui variabilità è catturata dalla quella componente principale. Dunque, se una PC non rappresenta almeno una variabile, non è necessaria alla costruzione del modello. Sebbene tale criterio sia molto semplice da implementare, è necessario considerare con cautela lo scarto di una PC il cui autovalore è molto vicino a 1, che può portare a non considerare la variabilità da esso spiegata, la quale potrebbe essere non trascurabile.

1.1.1.3 Indici diagnostici per PCA

In generale sono disponibili alcuni indici diagnostici per valutare le prestazioni di un modello PCA, o l'adeguatezza del set di dati e variabili considerati (Eriksson *et al.*, 2001). Per valutare le prestazioni del modello, è importante considerare la quantità di variabilità dei dati originali spiegati dal modello tramite l'indice R^2 (per dati autoscalati):

$$R^2 = 1 - \frac{\sum_{n=1}^{N}\sum_{k=1}^{K}(x_{n,k} - \hat{x}_{n,k})^2}{\sum_{n=1}^{N}\sum_{k=1}^{K}(x_{n,k})^2} = 1 - \frac{ESS}{TSS}, \tag{1.5}$$

dove *ESS* (*sum of square errors*) e *TSS* (*total sum of squares*) si distinguono rispettivamente per la somma dei quadrati degli errori e somma totale dei quadrati. Nella (1.5) $\hat{x}_{n,k}$ rappresenta l'elemento della riga *n*-esima e della colonna *k*-esima della matrice **X** ricostruita attraverso il modello PCA. R^2 è pertanto calcolato per un diverso numero di PC incluse nel modello e viene riportato anche come valore cumulativo (R^2_{CUM}):

$$R^2_{CUM} = \sum_{a=1}^{A} R^2_a , \tag{1.6}$$

È inoltre possibile valutare i dati utilizzati per la calibrazione di un modello PCA, al fine di rilevare eventuali valori anomali. Due statistiche vengono utilizzate per questo scopo: il T^2 di Hotelling e l'errore quadratico in predizione (SPE). Il T^2 di Hotelling (Hotelling, 1933) misura la distanza complessiva delle proiezioni di una osservazione (cioè un campione) del set di dati **X** dall'origine del nuovo spazio del modello PCA. Poiché ogni PC del modello cattura una diversa percentuale di varianza dei dati, si utilizza la distanza di Mahalanobis (Mardia et al., 1979):

$$T^2{}_i = \mathbf{t}_i^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} \mathbf{t}_i = \sum_{a=1}^{A} \frac{t_{a,i}^2}{\lambda_a}, \tag{1.7}$$

dove $\mathbf{t}_i$ è il vettore [$A \times 1$] tra le proiezioni t_{ai}, dell'i-esima osservazione sulle A PC utilizzate per creare il modello, mentre $\boldsymbol{\Lambda}$ è la matrice diagonale degli autovalori [$A \times A$]. In generale, il T^2 di Hotelling è utilizzato per valutare la deviazione di un'osservazione dalle condizioni medie rappresentate dall'insieme di dati. Graficamente, è possibile identificare i limiti di confidenza relativi a questo indice grazie alla determinazione di un'ellissi di confidenza, in cui la lunghezza del semiasse maggiore a dell'ellisse è calcolato secondo la (1.8) dove il pedice 1 si riferisce alla prima componente principale:

$$a = \sqrt{\lambda_1 T_\alpha^2}, \tag{1.8}$$

mentre il semiasse minore b secondo la (1.9) dove il pedice 2 si riferisce alla seconda componente principale:

$$b = \sqrt{\lambda_2 T_\alpha^2}, \tag{1.9}$$

La capacità rappresentativa di ogni campione da parte del modello PCA viene invece valutata attraverso la statistica *SPE*:

$$SPE_i = (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\mathrm{T}} (\mathbf{x}_i - \hat{\mathbf{x}}_i) = \mathbf{e}_i^{\mathrm{T}} \mathbf{e}_i, \tag{1.10}$$

Dove $\mathbf{e}_i$ [$N \times 1$] è il vettore dei residui nella ricostruzione dell'i-esima osservazione in $\mathbf{x}_i$. L'SPE_i misura la distanza ortogonale dell'i-esima osservazione dallo spazio latente identificato dal modello. In altre parole, rappresenta la mancata corrispondenza tra la ricostruzione di $\mathbf{x}_i$ del modello e l'effettivo valore di tale osservazione. Campioni con un valore elevato di SPE sono caratterizzati da una diversa struttura di correlazione rispetto a quella descritta dal modello PCA e, di conseguenza, non sono ben rappresentati dal modello. Tale incongruenza può essere

sfruttata nel controllo statistico di processo in quanto permette di identificare la presenza di campioni misurati che si discostano sensibilmente dai valori previsti dal modello. In tal caso può essere utile identificare le variabili che sono maggiormente responsabili della sua distanza dall'origine dello spazio PC o dallo spazio PC stesso. Questo può essere fatto analizzando i contributi di ciascuna variabile nel set di dati **X** per le statistiche T^2 e *SPE* del campione. In particolare, i contributi al T^2 possono essere calcolati come segue:

$$\mathbf{t}_{CONT,i}^2 = \mathbf{t}_i^{\mathrm{T}} \Lambda^{-1/2} \mathbf{P}^{\mathrm{T}} . \tag{1.11}$$

$\mathbf{t}_{CONT,i}$ è un [N × 1] vettore dei contributi di ogni variabile della statistica T^2 e può essere considerato una versione in scala dei dati all'interno del modello PCA. La formulazione nella (1.11) ha la proprietà che la somma degli elementi al quadrato di $\mathbf{t}_{CON,i}$ dà T^2_i per l'osservazione i-esima. Il contributo di ciascuna variabile alla statistica SPE_i per l'i-esimo campione coincide invece con i residui nella ricostruzione del campione attraverso il modello (ossia ogni singolo elemento $\boldsymbol{e}_{i,n}$ della riga i-esima della matrice dei residui **E**).

$$SPE_{CONT,i,n} = \mathbf{e}_{i,n_i} . \tag{1.12}$$

L'analisi dei contributi delle variabili possono rivelare quali variabili determinano principalmente la posizione di un campione nello spazio degli *scores* o fuori di esso. Questo, combinato con la conoscenza fisica del sistema, può essere utile soprattutto quando vengono individuati valori anomali, per capire la causa principale del problema.

1.1.2 Proiezione su strutture latenti (PLS)

La proiezione su strutture latenti (*partial least squares regression*; Wold et al, 1983; Höskuldsson 1988) è una tecnica di regressione lineare che correla un set di regressori **X** ad un set di variabili di risposta **Y** individuandone le direzioni di variabilità comuni. la PLS esegue una trasformazione dei dati in **X** in modo da massimizzare la covarianza delle sue variabili latenti (LV) con le variabili del dataset **Y**. I set di dati **X** e **Y** sono modellati secondo la decomposizione degli autovettori della matrice $\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}$ (Wold, 1976):

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{w}_1 = \lambda_1 \mathbf{w}_1 , \tag{1.13}$$

dove $\mathbf{w}_1$ [N × 1] è il vettore dei pesi (*weight*) per la prima LV, che rappresenta i coefficienti della combinazione lineare delle variabili in **X** che determinano gli *scores* PLS $\mathbf{t}_1$:

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1 . \tag{1.14}$$

Per ottenere i vettori dei pesi per le ulteriori LV, il problema in (1.14) può essere risolto iterativamente usando le matrici deflazionate $\mathbf{X}_a$ e $\mathbf{Y}_a$. Il processo di deflazione, per a = 1, ..., A-1 essendo A il numero di LV da prendere in considerazione, è definito come:

$$\mathbf{X}_{a+1} = \left(\mathbf{I}_i - \frac{\mathbf{t}_a \mathbf{t}_a^{\mathrm{T}}}{\mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a} \right) \mathbf{X}_a , \tag{1.15}$$

$$\mathbf{Y}_{a+1} = \left(\mathbf{I}_i - \frac{\mathbf{t}_a \mathbf{t}_a^{\mathrm{T}}}{\mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a} \right) \mathbf{Y}_a , \tag{1.16}$$

dove $\mathbf{I}_i$ è la [$I \times I$] matrice identità. Vale a dire, al passaggio a-esimo le ricostruzioni di ogni set di dati dalla a-esima LV stimata sono sottratte ai set di dati stessi. In particolare, dal secondo termine delle equazioni (1.15) e (1.16) risulta che:

$$\mathbf{p}_a^{\mathrm{T}} = \frac{\mathbf{t}_a^{\mathrm{T}} \mathbf{X}_a}{\mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a} , \tag{1.17}$$

$$\mathbf{q}_a^{\mathrm{T}} = \frac{\mathbf{t}_a^{\mathrm{T}} \mathbf{Y}_a}{\mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a} , \tag{1.18}$$

dove $\mathbf{p}_a$ e $\mathbf{q}_a$ rappresentano rispettivamente i vettori [$N \times 1$] e [$K \times 1$] dei *loadings* nella ricostruzione delle matrici $\mathbf{X}_a$ e $\mathbf{Y}_a$. Alla fine, i set di dati sono decomposti e correlati attraverso le loro strutture latenti:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} , \tag{1.19}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} , \tag{1.20}$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}^{*} . \tag{1.21}$$

Nelle equazioni (1.19) - (1.21), $\mathbf{T}$ è la matrice [$I \times A$] degli *scores*, $\mathbf{P}$ e $\mathbf{Q}$ sono le matrici [$N \times A$] e [$M \times A$] dei *loadings*, mentre $\mathbf{E}$ ed $\mathbf{F}$ sono le matrici [$I \times N$] e [$I \times K$] residue rappresentano la mancata corrispondenza del modello. Nella (1.22), $\mathbf{W}^{*}$ è la matrice [$N \times A$] dei pesi, che è calcolata dalla matrice dei pesi $\mathbf{W}$, per fornire informazioni su come le variabili della matrice $\mathbf{X}$ si combinano per formare gli scores $\mathbf{t}$ (Eriksson 2001).

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

Il vantaggio nell'uso di un modello PLS è che fornisce un modello per la struttura di correlazione di $\mathbf{X}$, un modello per la struttura di correlazione di $\mathbf{Y}$ e un modello della loro relazione reciproca. Pertanto la PLS è più adatta per gestire insiemi di dati, in cui sono inclusi dati fortemente correlati e, eventualmente, rumorosi.

1.1.2.1 Indici diagnostici per PLS

Le diagnostiche per valutare le prestazioni di un modello PLS sono le stesse utilizzate per il modello PCA: Eq. (1.6) - (1.12). Inoltre, per una corretta taratura del modello PLS e per agevolare l'interpretazione del modello, può essere utile identificare i regressori che influenzano maggiormente le variabili risposta. A questo scopo è stato introdotto l'indice *VIP* (*variable influence on projection*, influenza delle variabili sulle proiezioni; Chong e Jun, 2005), che è definito come:

$$VIP_m = \sqrt{M \sum_{i=1}^{A} R_{y,a}^2 (\mathbf{w}_{m,a})^2 \Big/ \sum_{i=1}^{A} R_{y,a}^2} \quad , \tag{1.23}$$

dove M è il numero totale di variabili considerate, $R_{y,a}^2$, è la varianza di **Y** spiegata dalla a-esima LV del modello, mentre $\mathbf{w}_{m,a}$, è il peso della variabile m-esima sull'a-esima LV calcolata dal modello PLS. Dal momento che la somma dei quadrati di tutti gli M *VIP* è uguale al numero di termini nel modello, il *VIP* medio sarebbe pari a 1. Le variabili con $VIP_m \geq 1$ sono pertanto considerati fattori di primaria importanza rispetto agli altri nella predizione delle variabili in **Y** (Eriksson *et al.*, 2001).

1.1.3 Interpretazione dei diagrammi degli scores e dei loadings

L'analisi grafica degli *scores* e dei *loadings* del modello è di fondamentale importanza per estrarre informazioni riguardo alla relazione tra le variabili e tra i campioni considerati. Secondo la pratica comune, gli *scores* sono riportati come grafici a dispersione, in cui gli *scores* su una PC sono riportati contro gli *scores* su un'altra PC (Figura 1.2*b*). Dal momento che generalmente le prime componenti principali spiegano la maggior parte della variabilità dei dati, è spesso sufficiente considerare solo da 2 a 4 PC per estrarre l'informazione ricercata.

I *loadings* di solito sono riportati come diagrammi a barre o come grafici a dispersione. Nel primo caso (che è stato adottato in questa Tesi) viene riportato un grafico a barre dei *loadings* delle variabili originali su ciascuna PC, come in Figura 1.2*a*. Nel secondo caso, come nei diagrammi degli *scores*, i *loadings* delle variabili sono tracciati su una PC rispetto a i *loadings* delle stesse variabili su una PC diversa. In generale, i grafici dei *loadings* sono utili per due motivi: *i*) capire quali sono le variabili legate alla variabilità dei dati e quali no; *ii*) capire se ci sono correlazioni tra le variabili. Ricordando il significato dei *loadings* in PCA, una variabile misurata che mostra un elevato *loading* ha un'importanza rilevante sulla corrispondente PC, ovvero è responsabile di una parte significativa della variabilità dei dati.

Pertanto, i *loadings* aiutano a identificare le variabili "più importanti" per il sistema in fase di studio. Se questa informazione è combinata con la conoscenza fisica del sistema, è possibile identificare quali siano le forze motrici dei fenomeni fisici alla base del sistema in esame.

Quando due variabili presentano *loadings* simili su una PC, allora tali variabili sono correlate come per esempio le variabili x_1 e x_3 in Figura 1.2*a* lungo la prima PC. Se i valori assoluti di *loadings* sono simili, ma i valori sono opposti, allora le variabili sono inversamente correlate come per esempio la variabile x_3 con le variabili x_1 e x_2 in Figura 1.2*a* lungo la seconda PC Si noti che i *loadings* in PCA su ogni PC sono indipendenti.

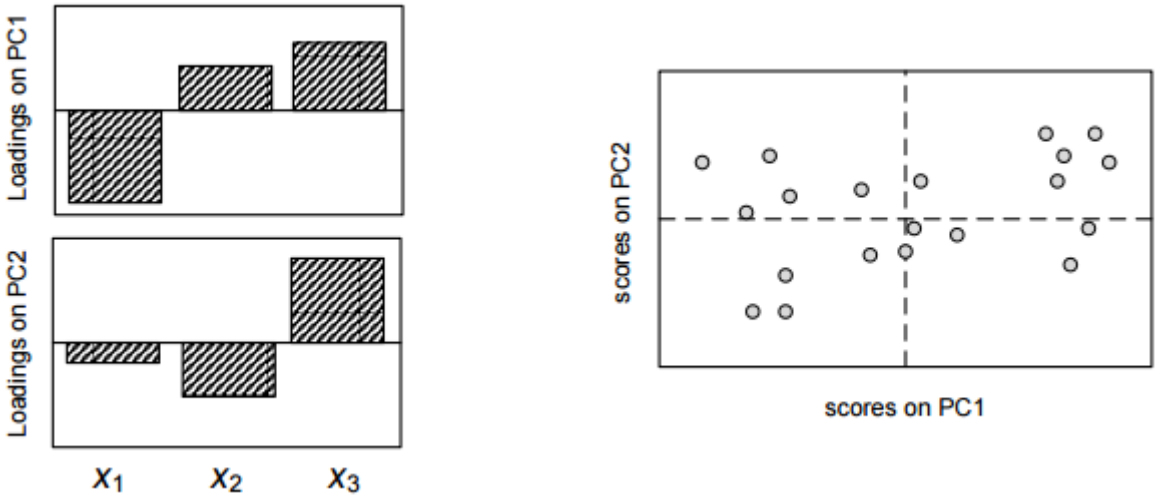


Figura 1.2. *Esempio di* (a) *grafico dei loadings a barre e* (b) *grafico degli scores per un modello con 2 PC. Da: Emanuele Tomba, (2013).*

Diagrammi degli *scores* come quello riportato in Figura 1.2*b* sono utili per identificare analogie tra i campioni. Ciò significa che i campioni con caratteristiche simili rientrano nella stessa regione del diagramma. Inoltre, il modello osservato in un diagramma degli *scores* riflette la struttura di correlazione individuata dai *loadings* delle variabili. Ad esempio, in Figura 1.2*b* si possono osservare tre gruppi principali lungo PC_1. I campioni sono quindi raggruppati secondo le somiglianze o differenze nelle variabili che hanno il *loading* più alto lungo PC_1 e analizzando il grafico dei *loadings*, si può identificare quali siano tali variabili (x_1 e x_3 in questo caso). Pertanto, campioni aventi un alto *score* positivo sul PC_1 avranno valori in x_3 elevati e valori inferiori in x_1, perché x_3 ha un *score* positivo e x_1 negativo sul PC_1.

1.2 Tecniche di identificazione delle possibili cause del disallineamento tra modello e processo

In questa Tesi vengono confrontate e migliorate le due metodologie proposte per l'identificazione delle possibili cause del disallineamento tra modello e processo (PMM, *process/model mismatch*). Di seguito vengono presentate in dettaglio entrambe le soluzioni: la

metodologia basata sull'analisi dell'indice MRLR (*mean residuals-to-limit ratio,* Meneghetti *et al.,* 2014) e quella basata sull'analisi dei coefficienti di correlazione.

1.2.1 Analisi dell'indice MRLR

La metodologia proposta da Meneghetti *et al.,* (2014) per la diagnosi di un PMM si basa sull'analisi dell'incongruenze tra la struttura di correlazione di dati storici e di dati ricavati utilizzando il modello in esame indagate grazie all'utilizzo di un modello PCA.

1. Generazione della matrice di modello e di processo. In generale, per l'analisi di un PMM sono disponibili due matrici, indicate rispettivamente dai pedici Π e M, in cui la prima matrice rappresenta il set di N campioni (o batch) registrati durante il normale monitoraggio di processo o durante la fase di modellazione del processo stesso, mentre la seconda rappresenta il set di stime per le stesse variabili, ottenute grazie all'utilizzo del modello a principi primi in analisi, a partire dagli stessi ingressi.

   Per ogni campione, le uscite simulate, le misure storiche e i parametri del modello a principi primi sono opportunamente combinati per ottenere due insiemi di variabili ausiliarie V ciascuno: un insieme si riferisce alle combinazioni delle misure simulate e parametri del modello, e l'altro si riferisce alle stesse combinazioni, ma utilizzando le misure storiche. Come verrà chiarito in seguito, le variabili vengono combinate in base alla struttura del modello a principi primi. Si noti che ogni variabile ausiliaria deve includere almeno un ingresso o una variabile di uscita, ovvero, nessuna variabile ausiliaria è ottenuta solo mediante combinazione di parametri del modello, a meno che i parametri del modello non siano essi stessi soggetti a variazione per ogni campione campioni (ad esempio, se i parametri dipendono dalle proprietà del materiale, e il materiale trattato cambia in tutti i campioni).

   Le due serie di variabili ausiliarie sono organizzate nelle colonne di due matrici, $\mathbf{X}_M$ [$N \times V$] e $\mathbf{X}_\Pi$ [$N \times V$], chiamate matrice di modello e matrice di processo, rispettivamente. A causa dell'esistenza di PMM, la struttura di correlazione di $\mathbf{X}_\Pi$ dovrebbe essere diversa da quella di $\mathbf{X}_M$.
2. Sviluppo di un modello PCA per la matrice del modello. Sia $\mathbf{X}_M$ che $\mathbf{X}_\Pi$ sono centrate sulle medie scalate sulle deviazioni standard delle colonne di $\mathbf{X}_M$ e. Un modello PCA è poi costruito su $\mathbf{X}_M$ e la matrice di residui $\mathbf{E}_M$ è calcolata come:

$$\hat{\mathbf{X}}_M = \mathbf{T}_M \mathbf{P}_M^T, \tag{1.24}$$

$$\mathbf{E}_M = \mathbf{X}_M - \hat{\mathbf{X}}_M. \tag{1.25}$$

Il modello PCA descrive la struttura di correlazione dei dati inclusi in $\mathbf{X}_M$. Il numero di PC da conservare nel modello PCA è determinato con la regola dell'autovalore maggiore di 1 (Mardia *et al.*, 1979).

3. Proiezione della matrice del processo sul modello PCA. $\mathbf{X}_\Pi$ è proiettata sullo spazio del modello PCA e la matrice dei residui $\mathbf{E}_\Pi$ è calcolata come:

$$\mathbf{T}_\Pi = \mathbf{X}_\Pi \mathbf{P}_M \,, \tag{1.26}$$

$$\hat{\mathbf{X}}_\Pi = \mathbf{T}_\Pi \mathbf{P}_\Pi^{\mathrm{T}} \,, \tag{1.27}$$

$$\mathbf{E}_\Pi = \mathbf{X}_\Pi - \hat{\mathbf{X}}_\Pi \,. \tag{1.28}$$

4. Analisi delle matrici dei residui e diagnosi del PMM. Le due matrici dei residui, $\mathbf{E}_\Pi$ e $\mathbf{E}_M$, vengono analizzate per identificare le variabili ausiliarie che contribuiscono maggiormente alla discrepanza delle strutture di correlazione di $\mathbf{X}_\Pi$ e $\mathbf{X}_M$. Queste variabili ausiliarie, insieme a una valutazione ingegneristica, vengono poi utilizzate per individuare le equazioni del modello a principi primi o i parametri che più contribuiscono al PMM osservato. Si noti che $\mathbf{E}_\Pi$ rappresenta sia la mancata corrispondenza tra $\mathbf{X}_\Pi$ e $\mathbf{X}_M$, che la frazione di variabilità di $\mathbf{X}_\Pi$ non descritta dal modello PCA creato sui dati di $\mathbf{X}_M$. Per tenere conto solo del contributo dovuto al PMM, viene rimosso il contributo relativo alla variabilità non modellata di $\mathbf{X}_\Pi$ dalla $\mathbf{E}_\Pi$. Quindi, per ogni colonna *v* della matrice $\mathbf{E}_\Pi$ si esegue l'analisi dei residui in termini di rapporto medio tra residui e limiti di confidenza (*mean residuals-to-limit ratio* MRLR_v):

$$\mathrm{MRLR}_v = \frac{\sum_{n=1}^{N} \sqrt{\left(\frac{(e_{\Pi n,v})^2}{CL_{95\%\, e_{Mv}}} \right)}}{N} \,. \tag{1.29}$$

I limiti di confidenza possono essere valutati per $\mathbf{e}_k$ nella forma (Montgomery *et al.*, 2005):

$$CL_{\alpha,\mathbf{e}_k} = z_{\alpha/2} \cdot \sigma(\mathbf{e}_k) \,, \tag{1.30}$$

dove α è il livello di significatività e in genere si assume un valore di 0.01 o 0.05, $z_{\alpha/2}$ è il corrispondente valore della statistica *z*, e $\sigma(\mathbf{e}_k)$ è la deviazione standard di $\mathbf{e}_k$. In questo lavoro viene utilizzato $\alpha = 0{,}05$ (vale a dire, il 95% di confidenza), e $z_{\alpha/2}$ assume il valore approssimato di 1.96. L'indice MRLR*v* esprime quindi la media dei rapporti tra i residui di ciascuna colonna della matrice $\mathbf{E}_\Pi$ e il limite di confidenza corrispondente 95%, calcolato considerando una distribuzione normale dei residui, si veda (1.5). Si noti che, se i residui non sono distribuiti normalmente, l'intervallo di confidenza non può essere calcolato dalla

(1.5). In tal caso è necessario ricorrere a espressioni alternative per la stima dei limiti di confidenza (Martin *et al.*, 1996; Doymaz *et al*, 2001).

1.2.2 Analisi dei coefficienti di correlazione

L'idea alla base dell'utilizzo dei coefficienti di correlazione per l'identificazione del PMM è di mettere in evidenza la differenza tra la rete di correlazione delle variabili del processo e la rete di correlazione delle variabili del modello in seguito all'insorgenza del PMM. A tal scopo è necessario valutare le correlazioni tra le variabili considerate rimuovendo le relazioni dovute a variabili terze (Melissa *et al.*, 2009; Rao *et al.*, 2007a; Rao *et al.*, 2007b). Questo può essere fatto condizionando (cioè, controllando, o tenendo costante) una o più altre variabili (*k*) prima di verificare l'associazione tra le due variabili designate (*i* e *j*). Questa correlazione è nota come correlazione di grado 1 ($r_{i,j\cdot k}$) delle variabili originali ed è una misura del livello di associazione diretta tra le componenti di *i* e *j* che sono correlate con *k*. In particolare, i coefficienti di correlazione di grado 0, 1 e 2, dove l'ordine di un coefficiente di correlazione è dato dal numero di variabili condizionate considerate (Rato e Reis, 2014) sono calcolati come:

$$r_{i,j} = \frac{\mathrm{cov}(i,\, j)}{\sqrt{\mathrm{var}(i) \cdot \mathrm{var}(j)}}\,. \tag{1.31}$$

$$r_{i,j\cdot k} = \frac{r_{i,j} - r_{i,k} \cdot r_{j,k}}{(1-r_{i,k})^2 (1-r_{j,k})^2}\,. \tag{1.32}$$

$$r_{i,j\cdot k,l} = \frac{r_{i,j\cdot k} - r_{i,l\cdot k} \cdot r_{j,l\cdot k}}{(1-r_{i,l\cdot k})^2 (1-r_{j,l\cdot k})^2}\,. \tag{1.33}$$

Se si verifica un cambiamento nelle relazioni tra le variabili, i coefficienti di correlazione in cui appare come variabile controllata una variabile associata alla causa principale del PMM, dovrebbero rimanere o vicini ai valori normali poiché in tali circostanze, la fonte di variabilità viene rimossa (Rato e Reis, 2015). Di seguito è riportata la procedura di diagnosi del PMM adattata a quella sviluppata da Rato e Reis (2015) per il monitoraggio di processo.

1. Generazione delle matrici di modello e di processo. Per poter applicare la procedura di diagnosi del PMM basata sull'analisi dei coefficienti di correlazione è necessario adattare il set di dati introdotto in §1.2.1 ossia le matrici $\mathbf{X}_{\mathrm{M}}$ e $\mathbf{X}_{\Pi}$. Per le matrici $\mathbf{X}$, costituite da *N* campioni (ovvero *N* batch che differiscono in base alle condizioni iniziali) di *K* variabili (da analizzare per stabilire quali siano maggiormente legate alla presenza del PMM), è possibile calcolare solo un vettore di coefficienti di correlazione. Per lo sviluppo della metodologia diagnostica, per ogni coefficiente deve essere disponibile un vettore di elementi, di cui ne è valutata la distribuzione. Per questo motivo per ognuno degli *N* campioni del set di dati vengono eseguite *B* simulazioni con i medesimi ingressi ottenendo dei dataset, chiamati

ripetizioni, che si differenziano per rumore bianco. In questo modo si procede con la costruzione delle strutture di modello $\underline{\mathbf{X}}_{BM}$ e di processo $\underline{\mathbf{X}}_{B\Pi}$ tramite l'assemblamento in una matrice tridimensionale rispettivamente di B matrici di modello $\mathbf{X}_M$ [$N \times V$] e di B matrici di processo $\mathbf{X}_\Pi$ [$N \times V$] costruite secondo quanto riportato in §1.2.1.

2. Calcolo dei coefficienti di correlazione di grado 1 normalizzati per il modello. Per ogni campione della matrice di modello $\mathbf{X}_M$ di dimensioni [$V \times B$] sono calcolati i coefficienti di correlazione di grado 1 considerando tutte le combinazioni di variabili associate e controllate secondo la (1.32). I coefficienti di correlazione sono quindi normalizzati rispetto al valore medio di popolazione ρ del coefficiente di correlazione (la media di ciascuna distribuzione degli N coefficienti di correlazione) tramite una delle seguenti equazioni:

$$\mathbf{w}_{i,j\cdot k} = \frac{\sqrt{N-2}\cdot(r_{i,j\cdot k}-\rho)}{1-\rho^2}, \tag{1.34}$$

$$\mathbf{w}_{i,j\cdot k} = \frac{\sqrt{N-2}}{1-\rho^2} ln\left[\left(\frac{1+r_{i,j\cdot k}}{1-r_{i,j\cdot k}}\right) - ln\left(\frac{1+\rho}{1-\rho}\right)\right]. \tag{1.35}$$

3. Calcolo dei coefficienti di correlazione di grado 1 normalizzati per il processo. Si ripete quanto descritto nel punto precedente per ogni campione della matrice di processo $\mathbf{X}_\Pi$ effettuando però la normalizzazione rispetto al vettore media ρ dei coefficienti di correlazione di grado 1 di $\mathbf{X}_M$. Un vettore di coefficienti di correlazione risulta quindi di dimensione [$1 \times V \cdot (V\text{-}1) \cdot (V\text{-}2)/2$].
4. Definizione dei limiti di confidenza per i coefficienti di correlazione di grado. Assumendo che i coefficienti di correlazione di grado 1, calcolati dalla matrice di modello, siano distribuiti normalmente, vengono definiti dei limiti di confidenza, per ogni coefficiente normalizzato in base alla percentuale di confidenza α pari al 95%:

$$CL_{i,j\cdot k} = z_{\alpha/2}\cdot\sigma(\boldsymbol{w}_{i,j\cdot k}) \tag{1.36}$$

5. Costruzione della matrice di diagnosi **D**. Per ognuna delle ripetizioni, viene calcolata una matrice **D** di dimensioni [$V \times V$], riportata in Figura 1.4. Si ricorda che un coefficiente di correlazione di grado 1 ($r_{i,j\cdot k}$) è una misura di associazione tra la coppia di variabili (i, j) quando k è controllata. Di conseguenza, definiamo "variabili in coppia" una qualsiasi coppia di variabili per le quali la correlazione viene misurata e "variabile in controllo" la corrispondente variabile controllata. La prima fase della procedura proposta prevede la realizzazione di una matrice **D** [$V \times V$] (le righe corrispondono alle variabili in controllo e le colonne alle variabili in coppia). Ogni elemento $d_{k,i}$ della matrice rappresenta il numero di

volte che la variabile *i*-esima in coppia e in controllo alla variabile *k* si trova al di fuori dei limiti di confidenza calcolati secondo la (1.36) per ogni *j* diversa da *i* e *k*, ossia:

$$d_{k,j} = \sum_{k\neq i, k\neq j} f(\boldsymbol{w}_{i,j,k}) \quad dove \quad f(\boldsymbol{w}_{i,j,k}) = \begin{cases} 1 \;\; se \;\; \boldsymbol{w}_{i,j,k} > CL_{i,j\cdot k} \\ 0 \;\; altrimenti \end{cases} \tag{1.37}$$

In questo modo, si è possibile valutare quali variabili in controllo riportino il minor numero di correlazioni parziali oltre la soglia considerando la norma di ogni riga della matrice **D**. Analogamente, è possibile valutare quale variabile in coppia presenta più frequentemente valori di correlazione anomali, attraverso la norma di ogni colonna di **D**. In particolare queste quantità sono chiamate rispettivamente "distanza di controllo" e "distanza di coppia". Di seguito in Figura 1.3 si riporta una rappresentazione grafica della matrice **D**.

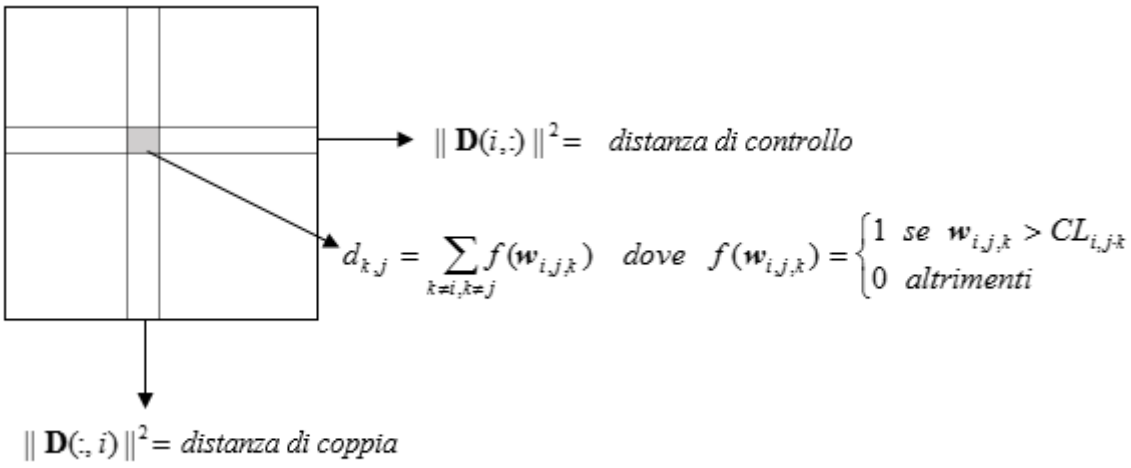


Figura 1.3. *Rappresentazione della matrice di diagnosi **D**. Adattata da Ibrahim (2016)*

Pertanto, una variabile che presenta una piccola distanza di controllo e una grande distanza di coppia è in linea di principio maggiormente riconducibile alla causa principale del PMM, dato che quando questa variabile è in controllo le correlazioni parziali tendono ad essere all'interno dei limiti di confidenza, mentre quando è in coppia, ne vengono evidenziati i cambiamenti di correlazione con le altre variabili. Le variabili che presentano solo la minima distanza di controllo sono ancora riconducibili alla causa del PMM, ma gerarchicamente ad un livello inferiore delle precedenti, come infine, le variabili con elevate distanze di coppia. Tali considerazioni sono riassunte nella classificazione descritta di seguito che indica la priorità per analizzare le variabili durante la fase di diagnosi:

a) Le variabili con distanza di controllo minima e distanza di coppia massima sono contrassegnate con l'indice 'ROSSO'. Questi sono i candidati più forti per l'origine del PMM, in quanto soddisfano entrambi i criteri;

b) Le variabili con distanza di controllo minima, ma con distanza di coppia inferiore al valore massimo, sono contrassegnate con l'indice 'ARANCIONE'. Quando queste variabili sono controllate si ottengono, per un piccolo numero di correlazioni parziali, dei valori superiori alla soglia ma la loro correlazione con le altre variabili non cambia in modo significativo;
c) Le variabili con distanza di coppia massima, ma con distanza di controllo superiore al valore minimo sono contrassegnate con l'indice 'GIALLO'. Queste variabili sono correlate con correlazioni parziali al di sopra della soglia, ma quando sono controllate, non eliminano i contributi relativi al PMM.

Queste tre regole di classificazione forniscono informazioni utili per restringere l'elenco di potenziali candidati per l'origine del PMM. Si prevede che il più delle volte, la regola 1 (cioè, variabili rosse) identificano l'insieme di variabili in relazione con la causa principale del PMM.

Tabella 1.1. *Regole di assegnazione delle etichette di diagnosi (Rato e Reis, 2015).*

	$\| \mathbf{D}(i,:) \|^2 = \min\left(\| \mathbf{D}(i,:) \|^2\right)$	$\| \mathbf{D}(:,i) \|^2 = \max\left(\| \mathbf{D}(:,i) \|^2\right)$
ROSSO	SI	SI
ARANCIONE	SI	NO
GIALLO	NO	SI

CAPITOLO 2

Caso studio: un modello di fermentazione

In questo Capitolo viene presentato il modello a principi primi sviluppato per la descrizione di un processo di fermentazione per la produzione di penicillina utilizzato come modello di riferimento per l'analisi del disallineamento tra modello e processo (*process/model mismatch*, PMM). Viene inoltre descritta la procedura utilizzata per generare i dati su cui applicare le procedure di diagnosi del modello a principi primi in esame implementate.

2.1 Caso studio

La necessità di disporre di un modello a principi primi per descrivere l'evoluzione di processi batch nell'industria chimica, biotecnologica o farmaceutica è sensibilmente aumentata negli ultimi anni per rispondere a diverse esigenze industriali, come il controllo di processo e l'ottimizzazione di processo e di prodotto. Per questo motivo, in questa Tesi è stato studiato il modello a principi primi sviluppato da Birol *et al.* (2002) per rappresentare un processo batch di produzione di penicillina.

La produzione di metaboliti secondari come la penicillina tramite microrganismi filamentosi è stato oggetto di molti studi a causa della sua importanza accademica e industriale (Atkinson e Mavituna, 1991). La formazione del prodotto di destinazione, l'antibiotico, non è di solito associata alla crescita cellulare. Per questo motivo, è pratica comune far crescere le cellule in un brodo di coltura e successivamente eseguire un'operazione fed-batch per promuovere la sintesi dell'antibiotico. Vi è una vasta letteratura sulla modellazione della produzione di penicillina, con vari gradi di complessità (Dussap *et al,* 1985 *et al.*, 1970; Heijnen *et al.*, 1979; Bajpai e Reuss 1980; Nestaas e Wang, 1983; Menezes *et al.*, 1994). I modelli riportati possono essere raggruppati come modelli *structured* e *unstructured*. Nei modelli *unstructured*, tutte le informazioni di fisiologia cellulare sono raccolte in un unico termine di biomassa in modo che non vi siano esplicite informazioni strutturali sull'attività cellulare, risultando quindi in un modello piuttosto semplice. I modelli *structured* invece comprendono gli effetti della fisiologia cellulare sulla produzione di metaboliti, tenendo conto della fisiologia e differenziazione delle cellule lungo la lunghezza delle ife e durante la fermentazione. Il modello meccanicistico di Bajpai e Reuss (1980), il modello di segregazione di Nestaas e Wang (1983) e il modello di Heijnen *et al.* (1979) sono alcuni dei modelli *structured* frequentemente citati. Alcuni di questi modelli non considerano gli effetti di variabili operative quali pH, temperatura, tasso di

aerazione, potenza di agitazione, portata di alimentazione di substrato per la crescita della biomassa e la produzione di metaboliti. Altri non considerano tutta la crescita della biomassa, la produzione di CO_2 e di metaboliti, il consumo di substrato (sia fonte di carbonio e ossigeno), e le condizioni di generazione del calore. Nel modello sviluppato da Birol *et al.* (2002) sono stati utilizzati i dati sperimentali disponibili in letteratura (Pirt e Righoletto, 1967; Metz *et al.*, 1981) per migliorare la simulazione della produzione di penicillina, estendendo i modelli matematici esistenti. Il modello meccanicistico di Bajpai e Reuss (1980) è stato utilizzato come base per la modellazione e gli effetti di variabili ambientali quali pH e temperatura e variabili di input come velocità di aerazione, potenza di agitazione, portata di alimentazione del substrato sulla formazione biomassa sono stati inclusi nel modello.
Di seguito in Figura 2.1 è riportato il diagramma di flusso per il processo di produzione della penicillina.

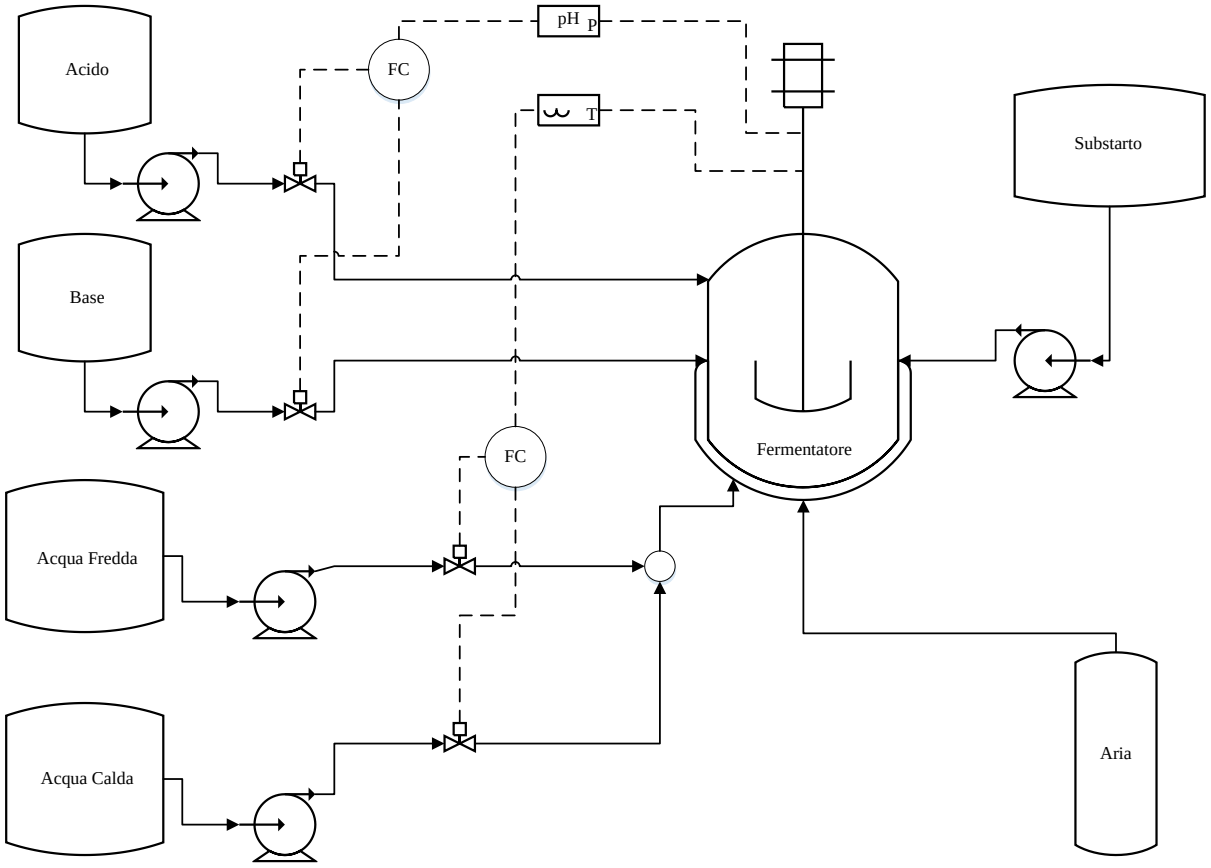


Figura 2.1. *Diagramma di flusso per il processo di produzione della penicillina. Adattato da Cinar et al., (2000).*

2.1.1 Equazioni e parametri del modello

Di seguito sono riportate le equazioni e i parametri implementati nel modello per descrivere le grandezze coinvolte nel processo di produzione di penicillina secondo il modello di Birol *et al.* (2002). Le equazioni differenziali descritte (Eq 2.1-13) vengono risolte simultaneamente.

2.1.1.1 Crescita della biomassa

La crescita della biomassa è descritta secondo l'equazione (2.1):

$$\frac{\mathrm{d}C_x}{\mathrm{d}t} = \mu \cdot C_x - C_x \frac{\mathrm{d}V}{\mathrm{d}T}, \tag{2.1}$$

Dove C_x è la concentrazione di biomassa (g/ l), V il volume (l), T la temperatura (K) e μ è la velocità specifica di crescita della biomassa (mol/ h). Quest'ultimo parametro esprime la dipendenza della crescita di biomassa dai substrati fonte di carbonio (glucosio), dalla concentrazione di ossigeno, dalla quantità di penicillina in fermentazione e dalla temperatura. Tali dipendenze sono descritte assumendo una cinetica di Contois (Bajpai e Reuss, 1980), come riportato in Eq. 2.2:

$$\mu = \mu_x \frac{C_s}{(K_x C_x + C_s)} \frac{C_L}{(K_{OX} C_x + C_L)} \left[\frac{1}{1 + K_1 / H^+ + H^+ / K_2} \right] k_g \exp\left(\frac{-E_g}{RT} \right) - k_d \exp\left(\frac{-E_d}{RT} \right), \tag{2.2}$$

dove C_s è la concentrazione di substrato (g/ l), C_L di ossigeno (g/ l), H^+ di idrogeno (g/ l). μ_x è la velocità specifica di crescita della biomassa per il microorganismo considerato (mol/ h), K_x la costante di saturazione di Contois (g / l), K_{ox} la costante di limitazione dell'ossigeno (-) K_1 K_2 le costanti empiriche (-), R è la costante universale dei gas.

È bene tener presente che la temperatura riveste un ruolo fondamentale nella crescita della biomassa. Infatti la velocità di crescita specifica di un microrganismo μ è proporzionale ad un aumento della temperatura fino ad un valore critico specifico per ogni microrganismo. Oltre questo valore, invece, si osserva una rapida diminuzione della crescita cellulare. Questa diminuzione può essere modellata considerando un corrispondente tasso di mortalità (Shuler e Kargi, 2002). Nella formulazione presentata da Birol *et al.,* (2002) in Eq. (2.2) l'effetto della temperatura sulla velocità di crescita specifica è stata introdotta tramite il modello cinetico di Arrhenius nei termini k_g (-) e E_g (cal/ mol), ossia: la costante e l'energia di attivazione per la crescita cellulare, e k_d (-) e E_d (cal/ mol): la costante e l'energia di attivazione per la morte cellulare. Valori tipici per questi parametri sono disponibili in letteratura (Shuler e Kargi, 2002). Durante il processo biologico di produzione della penicillina la temperatura del mezzo di coltura è mantenuta costante a 25 °C in conformità con gli esempi disponibili in letteratura

(Atkinson e Mavituna, 1991; Mou e Cooney, 1983). Quest'obbiettivo viene raggiunto tramite l'introduzione di un regolatore PID progettato per controllare la temperatura della coltura manipolando la portata acqua di riscaldamento / raffreddamento (Marlin, 1995).

2.1.1.2 Effetto e controllo del pH

È pratica comune, per mantenere costante il pH durante la fermentazione della penicillina, aggiungere una soluzione di NH_4OH (Atkinson e Mavituna, 1991; Mou e Cooney, 1983). Infatti, il decorrere delle reazioni coinvolte tende ad aumentare l'acidità del mezzo di coltura, per cui, all'aumentare della concentrazione di biomassa, il pH viene mantenuto costante grazie all'aggiunta di questa soluzione. Sulla base di questa osservazione, la concentrazione di ioni idrogeno $[H^+]$ è legata alla formazione di biomassa secondo l'Eq. (2.3):

$$\frac{\mathrm{d}H^+}{\mathrm{d}t} = \gamma\left(\mu \cdot C_x - \frac{F \cdot C_x}{V}\right) + \left[\frac{-B + B^2 + 4\times 10^{-14}}{2} - H^+\right]\frac{1}{\Delta t}, \tag{2.3}$$

dove F (l/ h) è la portata di substrato alimentata e γ una costante di proporzionalità che viene stimata come 10^{-5} mol $[H^+]$/g di biomassa, sulla base dei dati sperimentali di Mou e Cooney (1983). Il pH è mantenuto costante ad un valore di 5 (-), per simulare la produzione di penicillina, utilizzando una un *controller* (PID) che regola le portate di acido e base. Il parametro B in Eq. (2.3) viene espresso come:

$$B = \frac{[10^{-10}/H^+ - H^+]V - C_{a/b}(F_a + F_b)\Delta t}{V + (F_a + F_b)\Delta t}. \tag{2.4}$$

In Eq. 2.4, F_a e F_b rappresentano le portate di acido e di base (l/h), in cui, le concentrazioni in entrambe le soluzioni sono assunte pari a $C_{a/b}$ = 3 M. La concentrazione di ioni idrogeno viene calcolata prendendo in considerazione la dissociazione acido/ base dell'acqua e la produzione di idrogeno da biomassa.

2.1.1.4 Produzione della penicillina

Nel modello di Birol *et al*. (2002) produzione di penicillina è descritta in funzione della concentrazione della biomassa, della variazione del volume del brodo di coltura nel tempo e del tasso di produzione specifico di penicillina (Bajpai e Reuss, 1980):

$$\frac{\mathrm{d}C_p}{\mathrm{d}t} = \mu_{pp} \cdot C_x - K \cdot C_p - \frac{C_p}{V}\frac{\mathrm{d}V}{\mathrm{d}t}, \tag{2.5}$$

dove, C_p è la concentrazione di penicillina e μ_{pp} è il tasso di produzione specifico di penicillina definito come:

$$\mu_{pp} = \mu_p \frac{C_s}{(K_p + C_s + C_s^2 / K_I)} \frac{C_L^P}{(K_{OP} \cdot C_x + C_L^P)}. \tag{2.6}$$

In Eq 2.6, μ_p è il tasso di produzione di penicillina specifico per il microrganismo in esame, K_p (-) e K_{OP} (-) sono le costanti di inibizione dell'ossigeno e p (-) è un parametro empirico.

2.1.1.5 Utilizzo del substrato

I termini che descrivono l'utilizzo del substrato comprendono la crescita della biomassa, la formazione del prodotto e il mantenimento del microorganismo come suggerito da Bajpai e Reuss (1980):

$$\frac{dC_s}{dt} = \frac{\mu}{Y_{x/s}} C_x - \frac{\mu_{pp}}{Y_{p/s}} C_x - m_x C_x + \frac{Fs_f}{V} - C_x \frac{dV}{dt}. \tag{2.7}$$

La variazione di ossigeno disciolto, è espressa come:

$$\frac{dC_L}{dt} = \frac{\mu}{Y_{x/o}} C_x - \frac{\mu_{pp}}{Y_{p/o}} C_x - m_o C_x + K_l a(C_L^* - C_L) - \frac{C_L}{V} \frac{dV}{dt}, \tag{2.8}$$

dove i parametri Y (-) e m (-) sono le costanti di resa e di mantenimento, il parametro s_f (g/l) è la concentrazione di substrato in alimentazione e $K_l a$ il coefficiente di trasferimento di massa complessiva (1/ h). Nel modello originale di Bajpai e Reuss, $K_l a$ è costante (Bajpai e Reuss, 1980). Nel modello di Birol *et al.* (2002), invece, $K_l a$ è considerato in funzione della potenza di agitazione applicata P_w (W) e della portata di ossigeno f_g (g/ l) come suggerito da Bailey e Ollis (1986).

$$K_l a = \alpha \sqrt{f_g} \left(\frac{P_w}{V} \right)^{\beta}. \tag{2.9}$$

I valori di α (-) e β (-) sono assegnati in modo che la dipendenza della concentrazione della penicillina su $K_l a$ mostrino un comportamento molto simile alle previsioni di Bajpai e Reuss (1980).

2.1.1.6 Evoluzione della concentrazione di CO_2

L'evoluzione della concentrazione della CO_2 dipende dalla crescita della biomassa, dalla sintesi della penicillina e dal mantenimento dei microrganismi come suggerito da Montague *et al.*,

(1986). Per semplificare le relazioni tra queste grandezze e l'evoluzione della CO_2 Birol *et al.,* descrivono il bilancio dell'anidride carbonica come:

$$\frac{dC_{CO_2}}{dt} = \alpha_1 \frac{dC_x}{dt} + \alpha_2 C_x + \alpha_3, \tag{2.10}$$

dove i valori di α_1, α_2, α_3, (-) sono stimati per in base alprofilo della C_{CO2} ricavati dalle previsioni di Montague *et al.* (1986). L'evoluzione della C_{CO2} è quasi la stessa dell'ossigeno per la produzione di penicillina e si annulla dopo il passaggio alla fase fed-batch.

2.1.1.7 Variazione del volume

L'operazione di processo fed-batch provoca una variazione di volume nel fermentatore. Questo è calcolato come:

$$\frac{dV}{dt} = F + F_{a/b} - F_{loss}. \tag{2.11}$$

Nel modello di Birol *et al.,* (2002) per tenere in considerazione l'effetto di addizione con acido/ base ma soprattutto delle perdite per evaporazione sul cambiamento totale del volume della coltura, sono stati inclusi i termini $F_{a/b}$ e F_{loss} (l/ h) rispettivamente portata di acido/base e perdita di volume per evaporazione. La perdita di volume per evaporazione è infatti più significativa nelle fermentazioni industriali rispetto al termine della aggiunta di base. Normalmente l'aria che entra nel fermentatore è abbastanza secca e raggiunge un circa umidità relativa del 90/100 % dopo gorgogliamento nella coltura.

2.1.1.8 Calore di reazione

Trascurando tutte le altre fonti di generazione di calore diverse da quella causata da reazioni microbiche, il tasso di produzione di calore volumetrico è dato come:

$$\frac{dQ_{rxn}}{dt} = r_{q_1} \frac{dC_x}{dt} V + r_{q_2} C_x V, \tag{2.12}$$

dove, Q_{rxn} (cal) è il calore di reazione, r_{q1} è la resa di generazione di calore: (cal/ g biomassa) considerata costante e può essere trattata come un coefficiente di rendimento (Nielsen e Villadsen, 1994). Durante la fase di sintesi del prodotto, quando il tasso di formazione di biomassa diventa molto piccolo è significativa la generazione di calore da attività di manutenzione metabolica. Pertanto, è stato incluso il secondo termine in (2.12) per tenere conto della produzione di calore durante la manutenzione. La generazione di calore e l'evoluzione di C_{CO2} mostrano profili simili quindi il loro tasso di produzione dovrebbe avere lo stesso rapporto

in relazione alla crescita e alla quantità di biomassa. r_{q2}, la costante di generazione di calore (cal/ g biomassa · h) viene calcolato sulla base di questa osservazione. Il bilancio energetico è descritto da uno scambiatore di calore di tipo a spirale che è adatto per un fermentatore di scala da laboratorio (Nielsen, 1997):

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{F}{s_f}(T_f - T) + \frac{1}{V\rho c_p}\left[Q_{rnx} - a\frac{F_c^{b+1}}{F_c + (aF_c^b / 2\rho_c c_{pc})}\right], \tag{2.13}$$

dove *Tf* (K) è la temperatura di mandata di substrato, *a* (cal/ h·°C) il coefficiente di scambio termico del liquido raffreddamento/riscaldamento, *Fc* (l/h) la portata di acqua di raffreddamento, $\rho_c \cdot C_{pc}$ la densità (g/l) per la capacità termica del liquido di raffreddamento (J/(K·g)), $\rho \cdot C_p$ la densità per la capacità termica del mezzo (J/(K·g)).

2.1.2 Simulazione del processo e risultati

Per generare i dati analizzati in questa Tesi, è stato utilizzato il simulatore Pensim (di cui è disponibile una versione compatibile con Matlab®) che fornisce una soluzione numerica delle equazioni differenziali e algebriche che compongono il modello a principi primi di Birol *et al.* (2002).

2.1.2.1 Caratteristiche del simulatore

Il simulatore permette la scelta di un set di condizioni iniziali che costituiscono le condizioni al contorno per risolvere le equazioni differenziali considerate (2.1-2.13). Inoltre, possono essere modificati anche i valori di alcune delle variabili operative. In Tabella 2.1 sono riportati gli intervalli suggeriti dagli sviluppatori del simulatore per le variabili in ingresso e per i parametri di processo.

Tabella 2.1 *Variabili in ingresso e parametri e relativi intervalli di validità .*

Variabile	**Valore iniziale**
Concentrazione iniziale di substrato: C_s	5 – 50 (g/ l)
Concentrazione iniziale di biomassa: C_x	0 - 0.2 (g/ l)
Volume coltura iniziale: V	100 – 200 (l)
Concentrazione iniziale di anidride carbonica: C_{CO2}	0.5 -1.0 (mmol/ l)
Concentrazione iniziale ioni idrogeno: H^+	10^{-4} – 10^{-6} (mol/ l)
Temperatura iniziale del fermentatore: T	298 – 300 (K)
Portata di aria: f_g	3 - 10 (l/ h)
Potenza di agitazione: P_w	20 - 50 (W)
Portata di alimentazione del substrato: F	0.035 – 0.045 (l/ h)
Temperatura di mandata di substrato: T_f	296 – 298 (K)
Set point pH	5 – 6 (-)
Set point temperatura	298 – 300 (K)

Le simulazioni vengono eseguite sotto il controllo ad anello chiuso di pH e temperatura, e in anello aperto per il glucosio. Nel simulatore, allo scopo di riprodurre delle condizioni di processo verosimili, vengono introdotte delle deviazioni dai valori nominali delle variabili operative, definite come PRBS (*pseudo random binary signals*), che producono piccole fluttuazioni nei profili delle concentrazioni. Rumore bianco viene aggiunto anche al profilo di concentrazione dell'ossigeno disciolto, alla portata di alimentazione di substrato e alla potenza di agitazione. Nelle Tabelle 2.2 - 2.3 -2.4 sono riportati i valori delle condizioni iniziali nel caso base del simulatore, i parametri del modello e quelli dei sistemi di regolazione.

Tabella 2.2 *Valori delle condizioni iniziali delle concentrazioni di substrato, ossigeno, biomassa, prodotto; volume iniziale.*

Variabile/ parametro	**Valore**
Concentrazione di substrato: C_s	15 (g / l)
Concentrazione di ossigeno disciolto: C_L (=C_L^* alla saturazione)	1.16 (g / l)
Concentrazione di biomassa: C_x	0.1 (g / l)
Concentrazione di penicillina: C_p	0 (g / l)
Volume coltura: V	100 (l)
Concentrazione di anidride carbonica: C_{CO2}	0.5 (mmol / l)
Concentrazione ioni idrogeno: H^+	$10^{-5.1}$ (mol / l)
Temperatura: T	297 (K)
Generazione di calore: Q_{rxn}	0 (cal)

Tabella 2.3 *Parametri cinetici e costanti di resa (Birol et al., 2002).*

Variabile/ parametro	**Valore**
Portata di aria: f_g	8.6 (l/h)
Potenza di agitazione: P_w	29.9 (W)
Concentrazione del substrato di alimentazione: s_f	600 (g / l)
Portata di alimentazione del substrato: F	0.0426 (l / h)
Temperatura di mandata di substrato: T_f	298 (K)
Costante di resa: $Y_{x/s}$ (g biomassa/ g di glucosio)	0.45 (-)
Costante di resa: $Y_{x/o}$ (g biomassa/ g di ossigeno)	0.04 (-)
Costante di resa: $Y_{p/s}$ (g penicillina/ g di glucosio)	0.90 (-)
Costante di Resa: $Y_{p/o}$ (g penicillina/ g di ossigeno)	0.20 (-)
Costante: K_1, K_2	10^{-10}, 7×10^{-5} (mol / l)
Coefficiente di mantenimento substrato, ossigeno: m_x (per h), m_o (per h)	0.014, 0.467 (-)
Costante della CO_2 relativa alla crescita: α_1 (mmol CO_2 /g biomassa)	0.143 (mmol/ g)
Costante della CO_2 dell''energia mantenimento: α_2 (mmol CO_2/ g biomassa h)	4×10^{-7} (mmol/ g h)
Costante della CO_2 relativa alla produzione di penicillina α_3 (mmol CO2/ l h)	10^{-4} (mmol/ l h)
Massima velocità specifica di crescita: μ_x (per h)	0.092 (-)
Costante di saturazione di Contois: K_x	0.15 (g / l)
Costanti di limitazione dell'Ossigeno: K_{ox}, K_{op} (nessuna limitazione)	0 (-)
Costanti di limitazione dell'Ossigeno: K_{ox}, K_{op} (con limitazione)	2×10^{-2}, 5×10^{-4} (-)
Tasso specifico di produzione di penicillina: m_p (per h)	0.005 (-)
Costante di inibizione: K_p	0.0002 (g / l)
Costante di inibizione della formazione del prodotto: K_I	0.10 (g /l)
Constanti: p, b	3, 0.60 (-)
Tasso di idrolisi della penicillina: K (per h)	0.04 (-)
Costante di Arrhenius per la crescita delle cellule: k_g	7×10^{-3} (-)
Energia di attivazione per la crescita delle cellule: E_g	5100 (cal / mol)
Costante di Arrhenius per la morte delle cellule: k_d	1033 (-)
Energia di attivazione per la morte delle cellule: E_d	50000 (cal / mol)
Densità × capacità termica del mezzo: $\rho \cdot C_p$	1/1500 (g J/ K l)
Densità × capacità termica del liquido di raffreddamento: $\rho_c \cdot C_{pc}$	1/2000 (g J/ K l)
Resa di generazione di calore: r_{q1} (cal / g biomassa)	60 (cal / g)
Costante di generazione di calore: r_{q2} (cal / g biomassa h)	$1{,}6783\times10^{-4}$ (cal / g h)
Coefficiente di scambio termico del liquido raffreddamento a	1000 (cal/h°C)
Costanti in $K_l a$: α, β	70, 0.4 (-)
Costante in F_{loss}: λ	2.5×10^{-4} (1/h)
Costante di proporzionalità: γ (mol H^+ / g biomassa)	10^{-5} (mol/ g)

Tabella 2.4 *Parametri dei controller.*

Variabile/ parametro	**Valore**
pH: (base) K_c, τ_I, τ_d:	8×10^{-4}(-), 4.2 (h), 0.2625 (h)
pH: (acid) K_c, τ_I, τ_d	1×10^{-4}(-), 8.4 (h), 0.125 (h)
Temperatura: (raffreddamento) K_c, τ_I, τ_d	70 (-), 0.5 (h), 1.6 (h)
Temperatura: (riscaldamento) K_c, τ_I, τ_d	5 (-), 0.8 (h), 0.05 (h)

2.2 Generazione dei dati per la procedura di diagnosi del PMM

La procedura di diagnosi di un PMM prevede di comparare le misure storiche del processo in esame con le stime ottenute dal modello utilizzato per descrivere il processo stesso. In accordo con le definizioni fornite in § 1.2.1, il primo set corrisponde ad un insieme di misure delle uscite dal processo. Il secondo set invece, è ottenuto utilizzando gli stessi valori degli ingressi del primo set di dati utilizzando un modello a principi primi. In questa Tesi anche le misure storiche del processo sono generate tramite il simulatore in cui è implementato il modello a principi primi in esame. Un modello alterato invece, rappresenta il modello utilizzato per ottenere le misure simulate, in modo da forzare la presenza di *mismatch* modificando opportunamente il modello implementato nel simulatore.

2.2.1 Selezione delle variabili incluse nel set di dati

In Tabella 2.1 sono state riportate tutte le variabili di ingresso che è possibile fissare nel simulatore Pensim. Per semplificare il sistema in esame, è stato selezionato un sottoinsieme di variabili tra quelle disponibili, in base al loro effetto sulla concentrazione finale delle variabili in uscita. Seguendo la stessa procedura adottata da Ibrahim (2016), per selezionare tali variabili è stato utilizzato un modello PLS (*partial least square regression*, regressione parziale ai minimi quadrati; §1.1.2). A tal scopo, vengono simulati 100 batch i cui dati vengono selezionati per realizzate due matrici: la matrice degli ingressi (predittori) **X** e la matrice delle uscite del processo (risposte) **Y**. Ogni riga della matrice **X** corrisponde ai valori iniziali delle variabili di ingresso per un particolare batch mentre ogni riga della matrice **Y** corrisponde ai valori finali delle variabili di uscita per il medesimo batch.
I 100 batch sono simulati in seguito alla selezione casuale all'interno degli intervalli riportati in Tabella 2.1 di 100 valori iniziali per le variabili di ingresso. In Tabella 2.5 sono riportate le variabili incluse nella matrice **X** e nella matrice **Y**.

Tabella 2.5 *Matrice degli ingressi* **X** *e matrice delle uscite* **Y** *costruite a partire dalle variabili disponibili nel simulatore Pensim.*

Matrice X	**Matrice Y**
Concentrazione iniziale di substrato	Concentrazione finale di substrato
Concentrazione iniziale di biomassa	Concentrazione finale di biomassa
Portata di aria	Concentrazione finale di penicillina
Potenza di agitazione	Concentrazione finale di ossigeno disciolto
Portata di substrato	Concentrazione finale di anidride carbonica disciolta
Temperatura dell'alimentazione	Volume della coltura

Con riferimento al §1.1.2.1 dal modello PLS vengono calcolati gli indici *VIP* (*variable influence on projection*, influenza delle variabili sulle proiezioni; §1.1.2.1). Tali indici secondo la regola proposta da Eriksson *et al.*, (1999) che prevede di scartare i predittori il cui *VIP* è minore di 1 permettono di individuare quali siano le variabili che hanno maggiore influenza sulle risposte. In Figura 2.2 sono riportati gli indici *VIP* ottenuti per i 100 batch simulati.

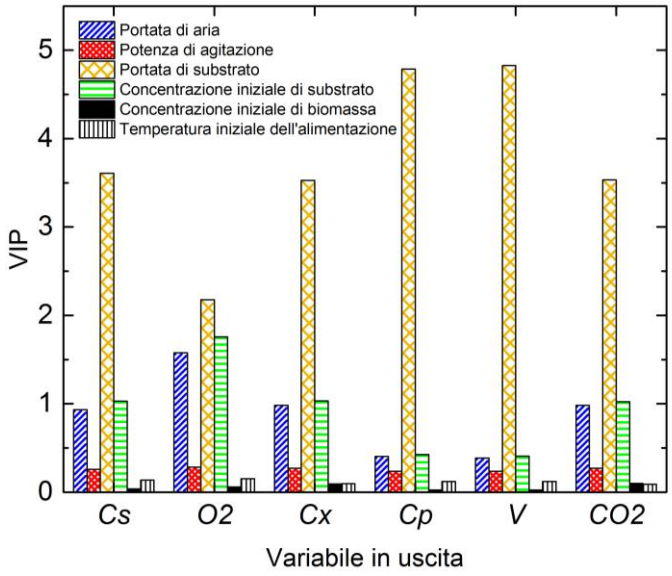


Figura 2.2 *Indici VIP delle variabili in entrata. Le variabili in uscita dal processo sono (da sinistra): concentrazione finale di substrato C_s, concentrazione finale di ossigeno disciolto O_2, Concentrazione finale di biomassa C_x, concentrazione finale di penicillina C_p, volume finale V, Concentrazione finale di anidride disciolta CO_2.*

Come si può osservare da Figura 2.2 gli indici *VIP* calcolati per la temperatura dell'alimentazione (per la presenza del sistema di regolazione) e la concentrazione iniziale di biomassa sono poco significativi. Sono quindi reputati poco rilevanti nel determinare le risposte del processo e per questo motivo vengono mantenuti ai valori nominali (riportati nel caso base del simulatore). La potenza di agitazione sebbene riporti indici *VIP* bassi viene comunque considerata nel set di ingressi iniziale, in quanto per come si può osservare in (2.9) tale variabile

ha un ruolo determinante sul parametro $K_l a$, il quale, come verrà poi descritto in 2.3, sarà modificato per forzare la presenza di un *mismatch*. La portata di aria e la concentrazione iniziale di substrato invece hanno indici *VIP* significativi e sono quindi rilevanti nel determinare le risposte del processo. Osserviamo infine come la portata di substrato sia la variabile con indici *VIP* più elevato e sia quindi la variabile con il maggior impatto sulle uscite del processo. Di seguito in Tabella 2.6 vengono riportate le variabili in ingresso e i relativi intervalli di validità all'interno dei quali vengono selezionati i valori iniziali per le simulazioni dei diversi batch.

Tabella 2.6 *Variabili in ingresso e relativi intervalli di validità selezionati per la generazione dei diversi batch.*

Variabile	**Valore iniziale**
Concentrazione iniziale di substrato: C_s	5 – 50 (g/ l)
Portata di aria: f_g	3 - 10 (l/ h)
Potenza di agitazione: P_w	20 - 50 (W)
Portata di alimentazione del substrato: F	0.035 – 0.045 (l/ h)

Da Tabella 2.1 si procede poi con la selezione delle variabili che permettono la costruzione delle matrici $\mathbf{X}_{\mathrm{Ms}}$ e $\mathbf{X}_{\Pi \mathrm{s}}$. Tale selezione comprende le variabili indagate nell'applicazione delle procedure diagnostiche e le variabili utilizzate per la definizione delle variabili ausiliarie (§3.2). Da questa selezione si ha l'esclusione della variabile di uscita concentrazione di anidride carbonica. Questo dipende da due osservazioni: la sovrapponibilità con la concentrazione di ossigeno degli indici *VIP* e la scarsa interazione della variabile concentrazione di anidride carbonica con le restanti variabili del modello come si può osservare dalla (2.10). Ogni colonna delle matrici $\mathbf{X}_{\mathrm{Ms}}$ e $\mathbf{X}_{\Pi \mathrm{s}}$ corrisponde a una variabile del processo, le righe invece rappresentano i campioni, ognuno dei quali generato con una differente combinazione di variabili in entrata. Le variabili incluse in questa selezione sono riportate in Tabella 2.7.

Tabella 2.7 *Variabili che compaiono nelle matrici* $\mathbf{X}_{\mathrm{Ms}}$ *e* $\mathbf{X}_{\Pi \mathrm{s}}$.

Variabile	**Simbolo**
Portata di aria	f_g
Portata di substrato	F
Potenza di agitazione	P_w
Concentrazione di substrato	Cs
Concentrazione di biomassa	C_x
Concentrazione di penicillina	C_p
Concentrazione di ossigeno disciolto	C_L
Volume della coltura	V

2.3 Analisi delle cause di un PMM: casi studio considerati

Un modello è costituito da equazioni e parametri. In un modello a principi primi (FP, *first-principle*) le equazioni rappresentano le conoscenze disponibili sui meccanismi alla base del processo, mentre i valori dei parametri vengono appositamente stimati per rappresentare in modo adeguato il comportamento del sistema. Quando le stime di un modello a principi primi di un processo sono confrontate con un set di dati storici, le uscite del modello potrebbero non corrispondere con le misure storiche di processo: in questo caso si parla di disallineamento tra modello e processo o *process model mismatch* (PMM). Ciò può essere dovuto a diversi motivi: (i) la conoscenza del processo di base è limitata, e quindi le equazioni del modello sono inadeguate; (ii) la complessità dei fenomeni fisici coinvolti nel processo è stata semplificata eccessivamente, ad esempio, perché il modello deve essere utilizzato in linea; (iii) ad alcuni dei parametri del modello sono stati assegnati valori non appropriati (ricavati per esempio da letteratura o da studi semi teorici) (Meneghetti *et al.,* 2014). Il verificarsi del PMM può essere critico quando il modello è utilizzato per scopi di progettazione, ottimizzazione, o di controllo. Attualmente, le tecniche per correggere un eventuale PMM e migliorare quindi l'aderenza del modello alla realtà si basano sulle tecniche di progettazione di esperimenti basate su modelli (*model-based design of experiment,* MBDoE) (Franceschini *et al.,* 2008, Marquardt *et al.,* 2005). Anche se efficaci, tali tecniche possono essere molto dispendiose, se non si conosce in anticipo quali equazioni o parametri sono maggiormente responsabili del PMM osservato. L'identificazione di tali termini senza dover ricorrere a nuove campagne di sperimentazione permetterebbe un'applicazione più efficace di tecniche come il MBDoE (o addirittura di evitarle) con notevoli riduzioni dei tempi e dei costi di intervento (Meneghetti *et al.,* 2014).

2.3.1 Caso studio 1: forzatura di un PMM tramite modifica del valore di $K_l a$

Un parametro critico del modello di fermentazione considerato è il coefficiente volumetrico di trasporto di massa dell'ossigeno. Infatti, nei processi biologici aerobici, l'ossigeno è il substrato fondamentale impiegato per la crescita, il mantenimento e altre vie metaboliche, tra cui la sintesi del prodotto. A causa della sua scarsa solubilità nel brodo di coltura, solitamente soluzioni acquose, l'ossigeno deve essere continuamente fornito in fase gassosa, e quindi è necessaria la conoscenza della velocità di trasferimento dell'ossigeno (*oxygen transfer rate,* OTR). L'OTR dipende dalla concentrazione di ossigeno disciolto nel volume di coltura e dal suo tasso di consumo da parte del microrganismo (*oxygen uptake rate,* OUR).

Sia il trasferimento di massa di ossigeno dal gas alla fase liquida che il suo consumo da parte del microrganismo hanno una importanza decisiva, poiché in molti processi il trasferimento di ossigeno rappresenta lo stadio di controllo per la crescita microbica, il quale a sua volta può influenzare l'evoluzione dei processi biologici (Aiba, *et al.,* 1973)

Durante un processo biologico aerobico, l'ossigeno viene trasferito da una bolla di gas a una fase liquida ed infine al sito di fosforilazione ossidativa all'interno della cellula, che può essere considerato come una particella solida. Il trasporto di ossigeno dalla bolla d'aria verso le cellule può essere rappresentato da una serie di passaggi intermedi come schematizzato in Figura 2.3, in cui la resistenza del film liquido intorno alle bolle di solito controlla la velocità di trasferimento complessiva. La teoria più semplice per il trasferimento di massa gas-liquido è il modello dei due film (Whitman, 1923) e la velocità di trasferimento di massa gas-liquido nel modello di Birol *et al.,* (2002) è modellata secondo questa teoria.

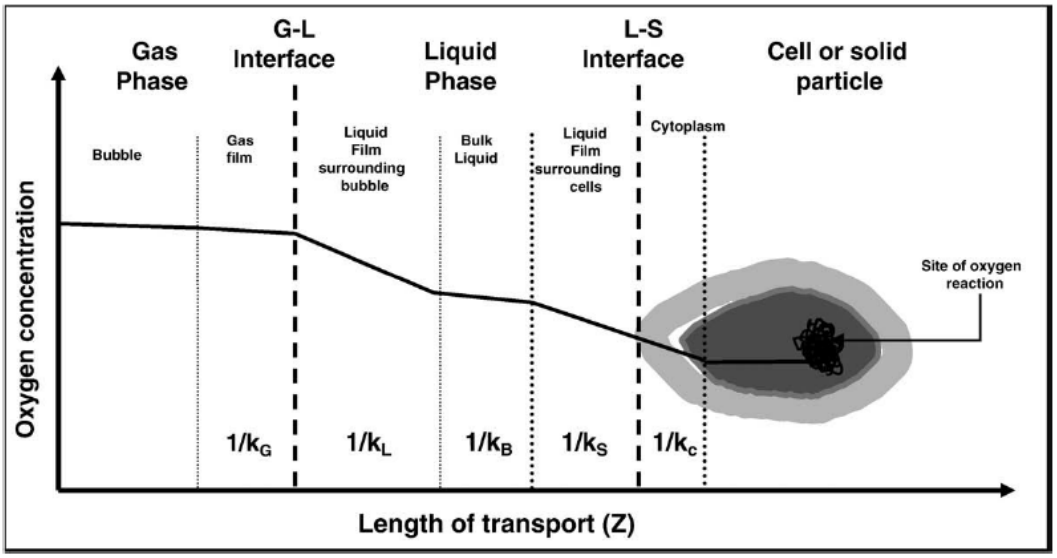


Figura. 2.3 *stadi e resistenze per il trasferimento di ossigeno da bolla di gas alla cellula. (i) trasferimento dall'interno della bolla al film; (ii) movimento attraverso l'interfaccia gas-liquido; (iii) diffusione attraverso il film liquido stagnante che circonda la bolla; (iv) trasporto attraverso il bulk del liquido; (v) diffusione attraverso il film liquido stagnante che circonda le cellule; (vi) movimento attraverso l'interfaccia liquido-cellule; e (vii) trasporto attraverso il citoplasma al sito di reazione biochimica. Da Garchia-Ochoa (2009).*

La forza motrice è il gradiente tra la concentrazione di ossigeno alla superficie e nel *bulk* del liquido (concentrazione media). Fattori che influenzano questo gradiente includono la solubilità e l'attività metabolica. La solubilità del gas, C^*, in soluzioni elettrolitiche è generalmente inferiore alla solubilità del gas in acqua pura (effetto "*salting-out*"). La solubilità del gas dipende principalmente dalla temperatura, la pressione, concentrazione tipo di sali presenti e le reazioni chimiche (Linek e Vacek, 1981; Hermann *et al*, 1995; Weissenborn e Pugh, 1996).

2.3.1.1 Determinazione del parametro K_la

La determinazione di K_la. in bioreattori è essenziale per stabilire l'efficienza di aerazione e per quantificare gli effetti delle variabili operative sulla fornitura di ossigeno disciolto. Sono stati sviluppati diversi metodi per determinare la velocità di trasferimento dell'ossigeno nei

bioreattori (van't Riet, 1979). Nella seleziona del metodo più appropriato per il sistema in uso, diversi fattori devono essere presi in considerazione (Novak e Klekner, 1988) quali:

- i sistemi di aerazione e di omogeneizzazione utilizzati;
- il tipo di bioreattore e la sua progettazione meccanica;
- la composizione del mezzo di fermentazione;
- il possibile effetto della presenza del microrganismo.

Il bilancio di massa per l'ossigeno disciolto nella fase liquida ben miscelato può essere stabilito come:

$$\frac{\mathrm{d}C}{\mathrm{d}t} = OTR - OUR, \tag{2.14}$$

dove d*C*/d*t* è il tasso di accumulo di ossigeno in fase liquida. Il termine *OUR* può essere espresso dal prodotto $q_{O_2} \cdot C_x$, essendo q_{O_2} il tasso di assorbimento di ossigeno specifico del microrganismo impiegato e C_x la concentrazione della biomassa.

La determinazione del parametro $K_l a$ può avvenire per via indiretta assumendo l'assenza di consumo biologico di ossigeno: in assenza di biomassa o con le cellule non respirano, le reazioni biochimiche non hanno luogo, OUR = 0. In questo caso, la (2.14) può essere semplificata per:

$$\frac{\mathrm{d}C}{\mathrm{d}t} = K_l a \cdot (C^* - C), \tag{2.15}$$

Alcuni metodi di misurazione sono basati sulla (2.15) :

- Metodi chimici.
- Metodi fisici: impiegano la risposta della sonda di ossigeno alle variazioni di concentrazione nel gas disperso nel mezzo, in condizioni non stazionarie. Questi metodi sono oggi i più comunemente usati per la stima trasferimento di ossigeno, poiché si basano sulla misura della concentrazione di ossigeno disciolto nel liquido durante l'assorbimento o il desorbimento di ossigeno nella soluzione (Dussap *et al*, 1985; Baird *et al*, 1993; Nocentini *et al*, 1993; Merchuk *et al*, 1994; Garcia-Ochoa e Gomez, 1998; Sanchez *et al*, 2000; Puthli *et al*, 2005; Clarke *et al*, 2006; Zhan *et al*, 2006).
- Metodo dinamico indiretto

Altri metodi per determinazione del parametro $K_l a$ sono basati sulla misurazione diretta dell'*OTR*:

- Analisi della fase gas, van't Riet, 1979).
- Metodo dinamico diretto, Taguchi e Humphrey (1966)

In Figura 2.7 si riportano valori $K_l a$ ottenuti con metodi diversi in funzione della velocità di agitazione in soluzioni non-Newtoniane in reattore da 20 l ($V_S = 2 \cdot 10^{-3}$ m·s^{-1} e μ_a = da $8 \cdot 10^{-3}$ a $30 \cdot 10^{-3}$ Pa·s).

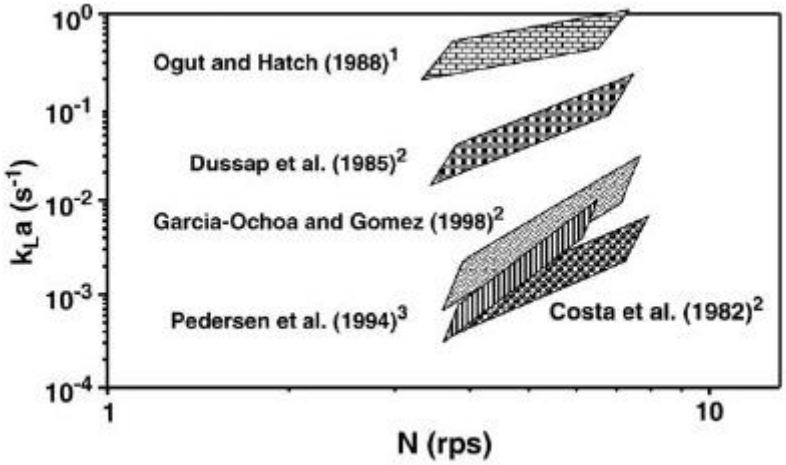


Figura 2.4 *Confronto dei valori di Kıa ottenuti con differenti metodi, in funzione della velocità dell'agitatore in soluzioni a comportamento non newtoniano. Legenda:1: metodo chimico;2: metodo fisico;3: metodo Kr (misurazione di radioattività nella portata in uscita a seguito dell'iniezione di un tracciante nel mezzo di coltura). Da: Garcia-Ochoa e Gomez, 1998.*

Come si può osservare da Figura 2.4 i valori del parametro $K_l a$ possono variare in modo consistente (anche di tre ordini di grandezza) a seconda del metodo sperimentale utilizzato. Per questo motivo l'introduzione di un errore su tale parametro è stato scelto come causa di *mismatch* nel primo caso studio considerato, in modo tale da simulare condizioni simili a quelle che possono accadere durante la normale attività industriale.

Per sistemi non viscosi, solitamente, il parametro $K_l a$ è calcolato direttamente utilizzando la correlazione empirica (anche nel caso del modello di Birol *et al.* 2002 implementato nel simulatore Pensim) di van't Riet (1979), (2.16).

$$K_l a = \mathrm{C} \cdot V_s^{a} \cdot (P/V)^{b} \mu_a^{c}, \tag{2.16}$$

dove i valori di $K_l a$, sono in funzione della velocità dell'agitatore, N, velocità superficiale del gas, V_S, la viscosità efficace del liquido, μ_a, la costante C che dipende dai parametri geometrici del reattore e dell'agitatore impiegato e la potenza media per volume, P/V. Questa correlazione deriva dal fatto che in bioreattori agitati meccanicamente, l'agitatore è lo strumento di dispersione del gas mentre la velocità e il design hanno entrambi un effetto pronunciato sul trasferimento di massa.

2.3.1.2 Perturbazione del parametro $K_l a$

Per verificare l'effetto di eventuali errori di stima su $K_l a$, è stata riprodotta l'analisi di sensitività sulla riduzione del parametro $K_l a$ della resa di penicillina condotta da Ibrahim (2016). In Figura 2.5 si può osservare che per variazioni del -90% di $K_l a$, la concentrazione del prodotto al termine della simulazione diminuisce del 10%. Considerando che, a seconda dei metodi utilizzati per stimare $K_l a$ sperimentalmente, si ottengono valori negli intervalli di 10^{-1} h e 10^{3} h, una variazione negativa del 90% è considerata accettabile, dato che, per il valore iniziale di 100 L del volume il valore del coefficiente di trasporto dell'ossigeno è 123 h^{-1}.

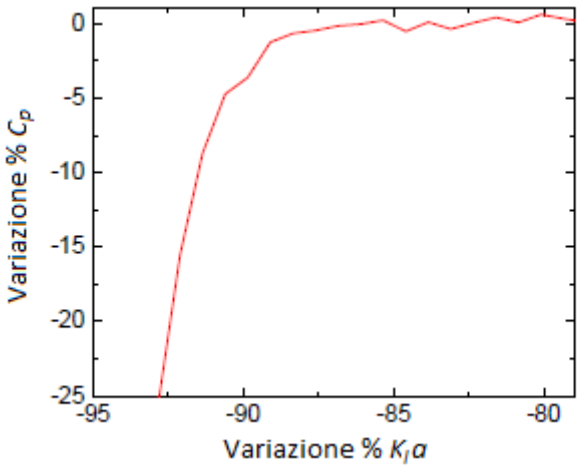


Figura 2.5. *Variazione percentuale sulla concentrazione finale di penicillina in funzione della variazione percentuale sul coefficiente volumetrico di trasporto di massa dell'ossigeno*

Nell'analisi di sensitività non sono state studiate variazioni della resa per variazioni positive del parametro in quanto un incremento positivo del coefficiente volumetrico di trasporto di massa produce variazioni estremamente ridotte sulla concentrazione di prodotto.

Tale andamento può essere giustificato osservando l'equazione di variazione della concentrazione di ossigeno nel tempo (2.8), infatti all'aumentare di $K_l a$, la forza motrice (C_L^*-C_L) diminuisce come effetto dell'aumento del trasporto di ossigeno nella fase liquida, creando così una compensazione che impedisce un impatto apprezzabile sulla resa del processo. Nella diagnosi del PMM parametrico, il set di dati storico viene generato utilizzando il valore originale del parametro α che compare nella (2.9) mentre il set di dati di modello riceve gli stessi valori delle variabili in ingresso delle misure storiche, ma α è diminuito del 92%, in modo da garantire una variazione apprezzabile della resa di prodotto.

2.3.2 Caso studio 2 e 3: forzatura di un PMM tramite modifica del valore di $Y_{p/s}$ e modifica del valore di $Y_{x/s}$

Il secondo e terzo caso studio in esame in questa Tesi, sono basati sulla modifica dei coefficienti di resa inclusi nel modello a principi primi di Birol *et al*. (2002). Tali parametri sono spesso

critici da stimare, e quindi maggiormente soggetti ad errore. Questo è dovuto al fatto che tali parametri spesso vengono stimati sfruttando il bilancio elettronico con conseguenti sovrastime del valore assunto addirittura dal 27 al 36% (Khandran e Smets, 1999). Inoltre, la scelta di forzare il secondo e terzo *mismatch* modificando questi parametri si basa sull'idea di studiare la robustezza della procedura diagnostica nel caso in cui l'errore coinvolga termini del modello oggetto di studio fortemente correlati tra loro.

Nel modello in esame, al fine di correlare il substrato consumato, le nuove cellule formate, e il prodotto generato, vengono definiti i coefficienti di resa riportati in Eq. 2.17- 2.20. La stechiometria delle reazioni coinvolte è molto complessa e varia con microorganismo/ sistema nutriente e condizioni ambientali, quali pH, temperatura e potenziale redox.

Il coefficiente di rendimento per cellule e substrato è espresso come:

$$Y_{x/s} = \frac{massa\ di\ nuove\ cellule}{massa\ di\ substrato\ consumato} = -\frac{\Delta C_x}{\Delta C_s}, \tag{2.17}$$

$$Y_{x/s} = \frac{1}{Y_{s/x}}. \tag{2.18}$$

Il coefficiente di rendimento per il prodotto e cellule come:

$$Y_{p/x} = \frac{massa\ di\ prodotto}{massa\ di\ cellule} = -\frac{\Delta C_p}{\Delta C_x}, \tag{2.19}$$

Il coefficiente di resa stechiometrica che riguarda la quantità di prodotto formato per massa di substrato è consumata è definito come:

$$Y_{p/s} = \frac{massa\ di\ prodotto\ formato}{massa\ di\ substrato\ consumato} = -\frac{\Delta C_p}{\Delta C_s}. \tag{2.20}$$

2.3.2.1 Determinazione dei parametri di resa

I parametri di resa posso essere determinati sperimentalmente tramite la misurazione diretta delle grandezze riportate in (2.20) in un reattore batch per il processo fermentazione di interesse. Tramite una sonda si misura la concentrazione iniziale e finale del prodotto, della biomassa e del substrato nel brodo di coltura. Noti tali valori dalle (2.17) e (2.20) si ricavano i parametri di resa: $Y_{x/s}$ e $Y_{p/s}$.

Non di rado tuttavia la determinazione dei parametri di resa possono essere determinati anche tramite un metodo predittivo basato sul bilancio elettronico. Per utilizzare tale metodo è necessario distinguere tra fermentazioni aerobiche e anaerobiche. In fermentazioni aerobiche, la resa di crescita per elettroni disponibili in molecole di ossigeno è di circa 3.14 ± 0.11 (gdw

di cellule/ elettroni) quando l'ammoniaca viene usata come fonte di azoto. Il numero di elettroni disponibili per molecola di ossigeno (O_2) è quattro. Se è noto il numero di molecole di ossigeno per mole di substrato consumato, il coefficiente di rendimento di crescita, $Y_{x/s}$, può essere facilmente calcolato. Si considera il catabolismo aerobico del glucosio:

$$C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O \tag{2.21}$$

Il numero totale di elettroni disponibili a 1 mole di glucosio è 24. La resa cellulare per elettrone disponibile è $Y_{x/s} = 24 \cdot (3.14) = 76$ (gdw di cellule/mol). Il coefficiente di rendimento di crescita previsto è $Y_{x/s} = 76/180 = 0.4$ (gdw di cellule/g di glucosio).
Si definisce la resa ATP ($Y_{x/ATP}$) la quantità di biomassa sintetizzata per mole di ATP generato dal microrganismo. In molte fermentazioni anaerobiche è pari a circa 10.5 ± 2 (gdw di cellule/ mol di ATP). In fermentazioni aerobiche, tale resa varia tra 6 e 29. Quando il rendimento energetico di una via metabolica è noto (*N* moli di ATP prodotti per grammo di substrato consumato), la resa di crescita $Y_{x/s}$ può essere calcolata con la seguente equazione:

$$Y_{x/s} = Y_{x/ATP} N \,. \tag{2.22}$$

Successivamente si calcola il coefficiente di resa massima teorico dalla stechiometria della reazione metabolica considerata secondo l'equazione:

$$Y_{p/s} = \frac{\nu_p PM_p}{\nu_s PM_s} \,. \tag{2.23}$$

In pratica, queste rese massime non vengono mai raggiunte. Le rese di prodotto sono circa il 90% al 95% dei valori massimi, perché il substrato viene convertito in biomassa e altri prodotti metabolici (Shuler e Kargi, 2002).

2.3.2.2 Perturbazione del parametro $Y_{p/s}$

Il secondo parametro sul quale viene introdotto un errore è la costante di resa di prodotto su consumo di substrato $Y_{p/s}$. Anche in questo caso viene condotta una analisi di sensitività per valutare gli effetti della variazione del parametro $Y_{p/s}$ sulla resa finale di penicillina, Figura 2.6.

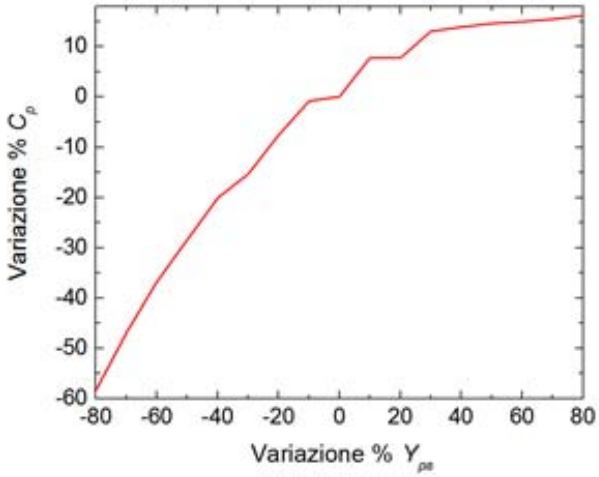


Figura 2.6 *Variazione percentuale della concentrazione finale di penicillina in funzione della variazione percentuale del coefficiente di resa di penicillina su substrato.*

Come possiamo osservare da Figura 2.6 l'effetto della riduzione del parametro di resa $Y_{p/s}$ sulla concentrazione finale di penicillina risulta più marcato dell'effetto ottenuto tramite la riduzione del coefficiente volumetrico di trasporto di massa dell'ossigeno $K_l a$. Da notare inoltre come l'andamento delle due analisi di sensitività sia simile. Infatti se a una riduzione del parametro $Y_{p/s}$ si assiste a una riduzione della concentrazione finale di penicillina, ad un aumento del medesimo parametro non si osserva una sensibile variazione della concentrazione di penicillina. Tale andamento è dovuto al fatto che un aumento del coefficiente di resa di formazione del prodotto su consumo di substrato corrisponde a una minore disponibilità di quest'ultimo per la formazione di biomassa. I due effetti infatti si compensano in quanto a una miglior efficienza da parte delle cellule di formare prodotto corrisponde un più rapido consumo di substrato e quindi a una minor disponibilità dello stesso per la crescita della biomassa e quindi una minor capacità produttiva. In questa Tesi il coefficiente di resa di prodotto su consumo di substrato $Y_{p/s}$ viene ridotto per un valore percentuale del 45% per forzare la presenza di un *mismatch*.

2.3.2.3 Perturbazione del parametro $Y_{x/s}$

Il terzo parametro sul quale viene introdotto un errore è la costante di resa di biomassa su consumo di substrato $Y_{p/s}$.

Per verificare l'effetto di eventuali errori di stima su $K_l a$, è stata riprodotta l'analisi di sensitività sulla riduzione del parametro $K_l a$ della resa di penicillina condotta da Ibrahim (2016), Figura 2.6.

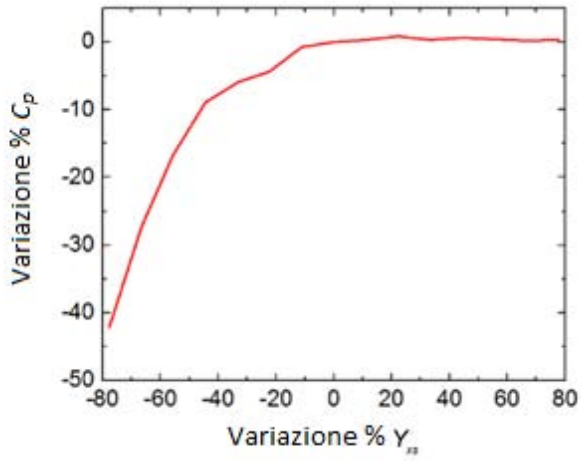


Figura 2.7 *Variazione percentuale della concentrazione finale di penicillina in funzione della variazione percentuale del coefficiente di resa di biomassa su substrato.*

Da Figura 2.7 si osserva come l'effetto della riduzione del parametro di resa $Y_{x/s}$ sulla concentrazione finale di penicillina risulti meno marcato dell'effetto ottenuto tramite la riduzione del parametro di resa $Y_{p/s}$. Tuttavia, anche in questo caso se a una riduzione del parametro $Y_{p/s}$ si assiste a una riduzione della concentrazione finale di penicillina, ad un aumento del medesimo parametro non si osserva una sensibile variazione della concentrazione di penicillina. Tale andamento è dovuto al fatto che un aumento del parametro di resa di formazione di biomassa su consumo di substrato corrisponde a una minore disponibilità di quest'ultimo per la produzione di penicillina. I due effetti quindi si compensano. Per forzare la presenza del *mismatch* il coefficiente di resa di biomassa su consumo di substrato $Y_{x/s}$viene ridotto per un valore percentuale del 65%.

CAPITOLO 3

Diagnosi tramite l'analisi dell'indice MRLR

In questo Capitolo viene descritta l'applicazione della procedura diagnostica per l'identificazione delle possibili cause del disallineamento tra modello e processo (*process/model mismatch*, PMM) sviluppata da Meneghetti *et al.* (2014) per tre casi di studio. Il modello considerato è il modello a principi primi sviluppato per un processo di fermentazione di penicillina (Birol *et al.*, 2002) mentre i casi di studio considerati coinvolgono tre diversi tipi di PMM parametrico.

3.1 Generazione dei dati

Come anticipato in §1.2 in questa Tesi vengono utilizzati dei set di dati simulati con l'ausilio del simulatore di processo sviluppato da Birol *et al.,* (2012) per un processo di fermentazione di penicillina. Con tale simulatore, in cui non è stata introdotta alcuna modifica al modello di base, viene generato un set di dati che rappresenta le misure storiche di processo (§ 1.2.1). Il set di dati viene realizzato generando 100 combinazioni dei valori delle variabili di ingresso per il simulatore selezionate secondo quanto riportato in § 2.2.1. Queste combinazioni sono ottenute dalla selezione casuale secondo una distribuzione normale dei valori degli ingressi all'interno degli intervalli di validità del simulatore Pensim. Tali combinazioni costituiscono le righe della matrice degli ingressi $\mathbf{X}_i$ [$N \times I$] dove N sono i campioni (o combinazioni) e I le variabili di ingresso. Per ogni riga della matrice $\mathbf{X}_i$ viene eseguita una simulazione realizzando la matrice tridimensionale $\underline{\mathbf{X}}_\Pi$ di dimensioni [$N \times K \times T$], dove N è il numero di campioni, K il numero di variabili di variabili selezionate per le analisi e T è il numero di istanti di campionamento.
In questo lavoro sono stati considerati solo i valori ottenuti al tempo t=T, ottenendo una matrice bidimensionale $\mathbf{X}_\Pi$ [$N \times K$] in cui ogni colonna corrisponde a una variabile del processo e ogni riga alle uscite finali di simulazione di un dato campione. In Figura 3.1 è riportata una rappresentazione grafica della procedura descritta e in Tabella 3.1 le variabili selezionate per la costruzione della matrice $\mathbf{X}_\Pi$.

Tabella 3.1 *Variabili selezionate per la generazione della matrice di processo* $\boldsymbol{X_{\Pi}}$.

Variabili incluse nella matrice di processo $\mathbf{X}_{\Pi}$			
f_g	Portata di aria	V	Volume finale della coltura
P_w	Potenza di agitazione	C_{CO2}	Concentrazione di anidride carbonica disciolta
F	Portata di substrato	pH	pH
T_f	Temperatura dell'alimentazione	Q_{rxn}	Calore generato
C_s	Concentrazione di substrato	F_a	Portata di acido
C_L	Concentrazione di ossigeno disciolto	F_b	Portata di base
C_x	Concentrazione di biomassa	F_c	Portata di acqua di raffreddamento
C_p	Concentrazione di penicillina	T	Temperatura

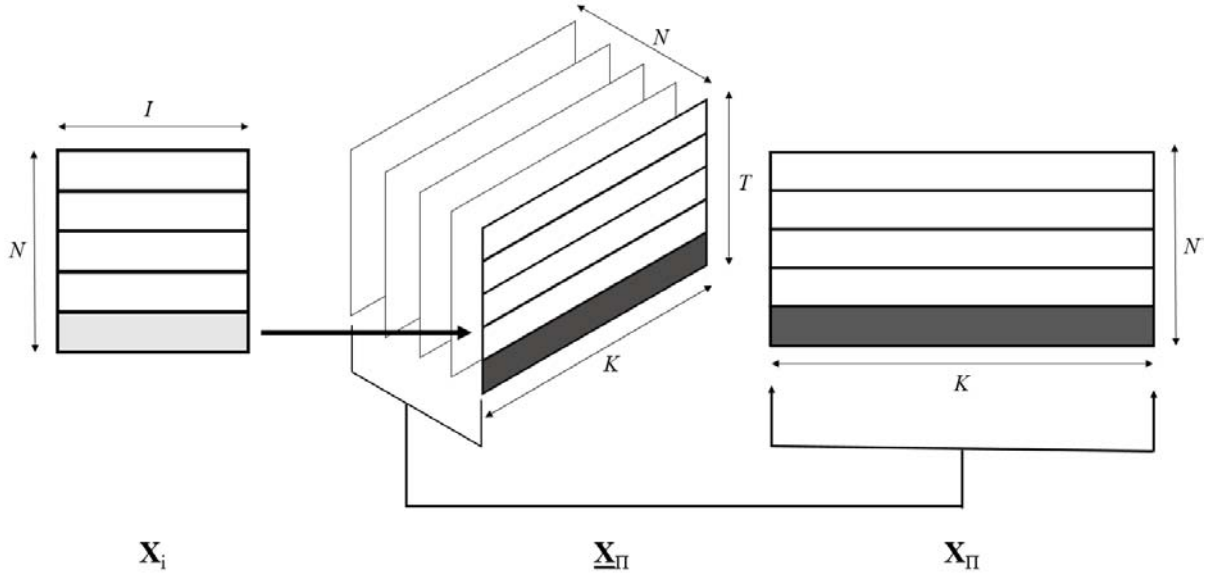


Figura 3.1 *Procedura per la costruzione della matrice di processo* $\mathbf{X}_{\Pi}$

La procedura appena descritta viene ripetuta per la costruzione della matrice di modello $\mathbf{X}_M$ tramite la preventiva modifica nel modello a principi primi di Birol *et al.*, (2002) dei parametri K_la, $Y_{p/s}$ e $Y_{x/s}$ secondo quanto descritto in §2.3. La riduzione del parametro K_la viene fissata a un valore percentuale del 92% con conseguente sottostima della resa in penicillina in media del 10%. Mentre le riduzioni dei parametri $Y_{p/s}$ e $Y_{x/s}$ vengono fissate a un valore percentuale del 45% e del 35% con conseguente sottostima delle rese in penicillina in media del 50% e del 40%.

3.1.1 Analisi della distribuzione dei dati generati

Prima di applicare la procedura di diagnosi, è importante analizzare la distribuzione dei dati della matrice di processo $\mathbf{X}_{\Pi}$ e della matrice di modello $\mathbf{X}_M$. In Figura 3.2 sono riportate le

distribuzioni dei valori finali della concentrazione di penicillina, una delle uscite del simulatore per la matrice del processo (Figura 3.2*a*) e le matrici del modello sia nel caso della modifica del parametro $K_l a$ (Figura 3.2*b*) che dei parametri $Y_{p/s}$ (Figura 3.2*c*) e $Y_{x/s}$ (Figura 3.2*d*).

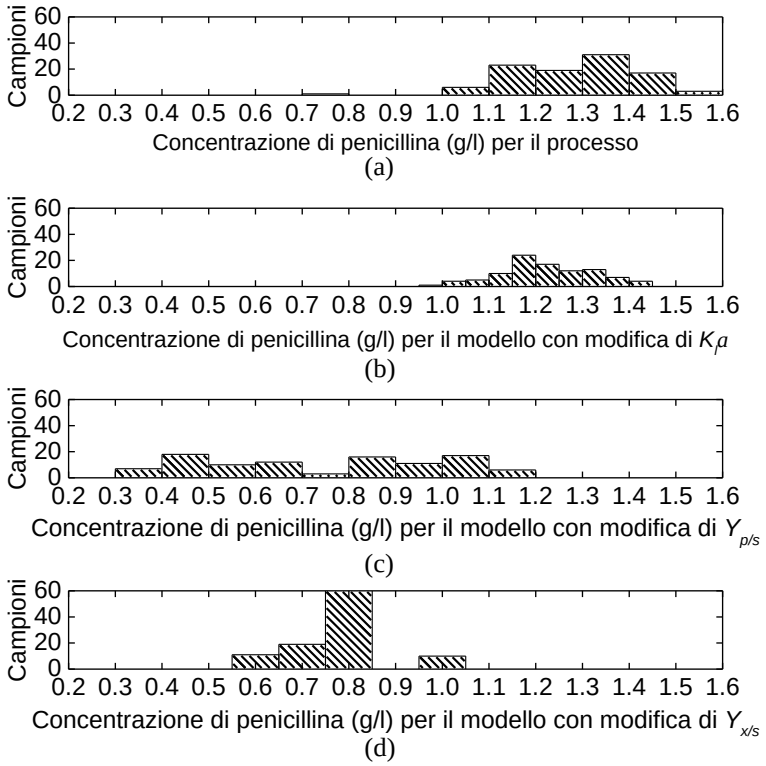


Figura 3.2 *Distribuzioni dei valori finali della concentrazione di penicillina considerando i diversi campioni.* (a) *Dati di processo,* (b) *dati di modello in cui è modificato il parametro* $K_l a$, (c) *dati di modello in cui è modificato il parametro* $Y_{p/s}$ (d) *dati di modello in cui è modificato il parametro* $Y_{x/s}$.

La distribuzione dei valori in uscita della concentrazione di penicillina finale nei dati storici (ottenuti utilizzando un modello in cui non è stata introdotta alcuna alterazione) presenta un andamento tendente alla normalità (Figura3.2*a*): lo stesso vale per i dati in uscita ottenuti dal modello con modifica del parametro $K_l a$ (Figura 3.2*b*), mentre nel caso della modifica del parametro $Y_{p/s}$ (Figura 3.2*c*) le concentrazioni finali di penicillina si presentano con una distribuzione bimodale. Infine nel caso della matrice del modello con modifica del parametro $Y_{x/s}$ (Figura 3.2*d*) i valori finali di penicillina ottenuti si assestano principalmente in uno stretto intorno di 0.8 (g/l). Questa analisi dimostra quindi, il diverso impatto dei PMM parametrici sui dati di uscita del simulatore, sia per quanto riguarda le deviazioni standard che la media. Notiamo infatti come le medie per le distribuzioni dei dati di concentrazione finale di penicillina si riducano notevolmente considerando la matrice del processo, la matrice del modello con modifica di $K_l a$ e le matrici del modello con modifica di $Y_{p/s}$ e di $Y_{x/s}$.

3.2 Definizione delle variabili ausiliarie

In base alla metodologia riportata in §1.2.1, vengono definite *V* variabili ausiliarie a partire dalla combinazione delle variabili originali (sia di $\mathbf{X}_{\Pi}$ che di $\mathbf{X}_{M}$) con i parametri del modello. In questo lavoro vengono sfruttate le stesse combinazioni analizzate da Ibrahim (2016), ottenute in seguito all'analisi delle equazioni del modello di Birol *et al.,* (2002) riportate in §2.1.1. In particolare, le variabili ausiliarie (Eq. 3.1) vengono definite su alcuni addendi delle equazioni differenziali, i quali sono già combinazioni non lineari di variabili del processo e di parametri. In questo modo le variabili ausiliarie assumono un preciso significato fisico, in quanto ognuna di esse rappresenta un termine presente in bilanci di materia, energia e fattori cinetici.

$$
\begin{aligned}
x_1 &= K_l a(C_L^{\ *} - C_L) & x_4 &= KC_p \\
x_2 &= C_x \mu & x_5 &= \frac{x_2}{Y_{x/s}} + \frac{x_3}{Y_{p/s}} + C_x m_x \\
x_3 &= \mu_{pp} C_x & x_6 &= \frac{x_2}{Y_{x/o}} + \frac{x_3}{Y_{p/o}} + C_x m_o
\end{aligned}
\qquad (3.1)
$$

Ogni variabile ausiliaria viene calcolata utilizzando i dati di $\mathbf{X}_{\Pi}$ e $\mathbf{X}_{M}$ per generare rispettivamente le matrici $\mathbf{X}_{\Pi v}$ e $\mathbf{X}_{Mv}$ ovvero la matrice di processo e la matrice di modello di dimensioni [$N \times V$] dove *N* sono i campioni mentre *V* le variabili ausiliarie.

3.3 Caso studio 1: errore introdotto sul parametro $K_l a$

Con l'obbiettivo di identificare il termine del modello responsabile della presenza del *mismatch* nel primo caso analizzato si procede con un'analisi preliminare dei diagrammi degli *scores* e dei *loadings* ottenuti da un modello PCA sui dati di modello. Successivamente si esegue la procedura di diagnosi tramite l'indice MRLR (Meneghetti *et al.*, 2014).

3.3.1 Comparazione delle matrici $\mathbf{X}_{Mv}$ e $\mathbf{X}_{\Pi v}$ analisi delle componenti principali

La procedura suggerita per l'individuazione del PMM basata sull'analisi dell'indice MRLR prevede la costruzione di un modello a componenti principali (PCA, *principal component analysis*; Jackson, 1990), sulla matrice di modello $\mathbf{X}_{Mv}$. Nel caso in esame il modello è stato costruito considerando solo le prime due componenti principali, scelte in base alla regola dell'autovalore maggiore di 1 (Mardia et al., 1979) riportato in §1.1.1.2. Queste descrivono l'87% della variabilità dei dati (Tabella 3.2). La matrice $\mathbf{X}_{\Pi v}$ viene poi scalata sulla media e la deviazione standard di $\mathbf{X}_{Mv}$, e viene proiettata nello spazio del modello calibrato su $\mathbf{X}_{Mv}$.

Tabella 3.2 *Variabilità dei dati catturata da ciascuna componente principale del modello PCA costruito su* $\mathbf{X}_{Mv}$.

Numero PC	Autovalore	R^2	R^2 cumulato
1	5.40	60.01	60.01
2	2.49	27.63	87.64

In Figura 3.3 viene riportato il grafico degli *scores* della matrice di modello (simboli rossi) sul quale vengono proiettati gli *scores* del processo (simboli blu) e l'ellisse di confidenza (linea nera) costruita secondo quanto riportato in §1.1.1.3.

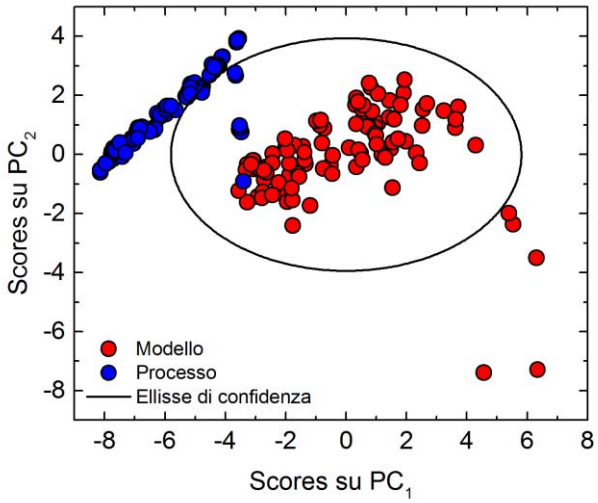


Figura 3.3 *Grafico degli scores per il modello PCA costruito su* $\mathbf{X}_{M,v}$ *(simboli rossi) lungo le due componenti principali e proiezione dei dati di processo (simboli blu).*

La distribuzione degli *scores* per i due dataset offre chiare indicazioni della presenza di un PMM. In Figura 3.3 si può osservare come gli *scores* per i campioni del processo si scostino sensibilmente dagli *scores* del modello con un numero non trascurabile di campioni al di fuori dell'ellisse di confidenza §1.1.1.3. È inoltre interessante notare come gli *scores* del processo non siano semplicemente traslati rispetto a quelli del processo, ma assumano una distribuzione completamente diversa nel piano identificato dalle prime due PC. Questo indica che il PMM incide probabilmente su una variabile del modello in cui è particolarmente forte l'effetto della non linearità delle equazioni considerate.

Infine la traslazione degli *scores* del processo avviene preferenzialmente lungo la prima componente principale, indicando che la variabile su cui incide il *mismatch* è molto probabilmente descritta da questa direzione del piano. Per identificare quali siano le variabili più importanti lungo questa direzione, vengono riportati in Figura 3.4 i *loadings* delle variabili ausiliarie rispetto alle due componenti principali del modello PCA.

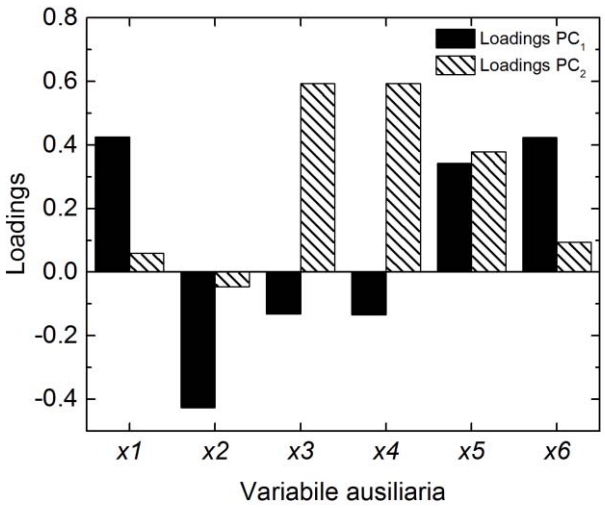


Figura 3.4 *Grafico dei loadings per il modello PCA costruito su* $\mathbf{X}_{Mv}$.

Come possiamo osservare in Figura 3.4 vi sono quattro variabili ausiliarie fortemente correlate con PC_1 e quindi verosimilmente responsabili del PMM. Tali variabili sono: x_1 x_2 x_5 x_6 che presentano tutte *loadings* tra loro confrontabili, impedendo quindi l'identificazione della vera causa del *mismatch*.

3.3.2 Analisi dell'indice MRLR

Secondo quanto riportato in §1.2.1, le due matrici dei residui $\mathbf{E}_\Pi$ e $\mathbf{E}_M$ vengono analizzate per determinare quali siano le variabili ausiliarie che contribuiscono maggiormente alla diversa struttura di correlazione tra le matrici $\mathbf{X}_{\Pi v}$ e $\mathbf{X}_{Mv}$ e, quindi alla formazione del PMM. Per tenere conto solo dalla discrepanza tra i dati storici e quelli simulati rimuovendo il contributo legato alla variabilità di $\mathbf{X}_{\Pi v}$ che non viene identificata dal modello viene calcolato l'indice MRLR (1.31). I risultati dell'analisi sono riportati in Figura 3.5, in cui si può osservare che le variabili con indice MRLR più elevato e quindi maggiormente riconducibili alla causa del *mismatch* sono le variabili x_1, x_2, x_6, tra cui la variabile x_1 presenta indice MRLR nettamente superiore alle altre. Questo risultato suggerisce che x_1 è la variabile maggiormente responsabile del PMM, mentre, le variabili che sono direttamente correlate a quest'ultimo. Quindi come riporta Ibrahim (2016) l'indice MRLR permette in questo caso di indentificare con precisione la principale variabile responsabile del PMM e discriminare quest'ultima sia dalle variabili direttamente correlate al PMM che dalle variabili non correlate al PMM.

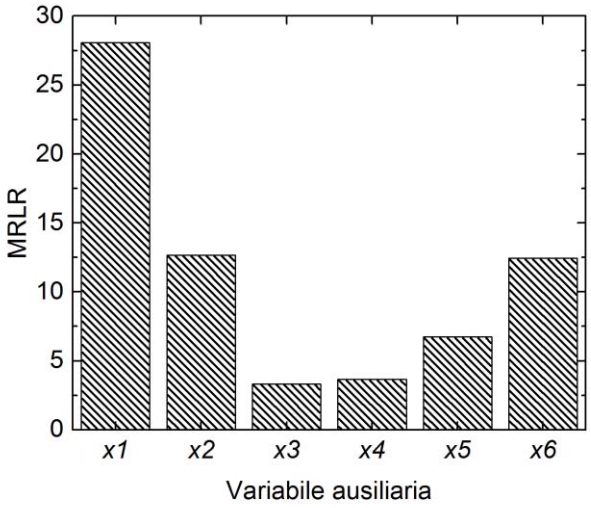


Figura 3.5 *Indici MRLR calcolati a partire dalle proiezioni di* $\mathbf{X}_{\Pi}$ *sul modello PCA costruito su* $\mathbf{X}_{M,v}$

3.4 Caso studio 2: errore introdotto sul parametro $Y_{p/s}$

Anche in questo secondo esempio, con l'obbiettivo di identificare la variabile ausiliaria responsabile del PMM dovuto alla presenza di un *mismatch* sul parametro $Y_{p/s}$ si esegue un'analisi preliminare dei diagrammi degli *scores* e dei *loadings* ottenuti dal modello PCA costruito sui dati di modello. Successivamente si procede con l'implementazione della procedura di diagnosi tramite l'indice MRLR (Meneghetti *et al.*, 2014).

3.4.1 Comparazione delle matrici $\boldsymbol{X}_{Mv}$ e $\boldsymbol{X}_{\Pi v}$ analisi delle componenti principali

Il modello PCA viene costruito sulla matrice di modello $\mathbf{X}_{Mv}$ considerando solo le prime due componenti principali, scelte in base alla regola dell'autovalore maggiore di 1 (Mardia et al., 1979) riportato in §1.1.1.2.

Queste descrivono l'87% della variabilità dei dati (Tabella 3.3). La matrice $\mathbf{X}_{\Pi v}$ viene poi scalata sulla media e la deviazione standard di $\mathbf{X}_{Mv}$, e viene proiettata nello spazio del modello calibrato su $\mathbf{X}_{Mv}$.

Tabella 3.3 *Variabilità dei dati catturata da ciascuna componente principale del modello PCA costruito su* $\mathbf{X}_{M,v}$.

Numero PC	Autovalore	R^2	R^2 cumulato
1	4.51	75.21	75.21
2	1.43	23.85	99.06

In Figura 3.6 viene riportato il grafico degli *scores* della matrice di modello (simboli rossi) sul quale vengono proiettati gli *scores* del processo (simboli blu) nel piano identificato dalle prime due componenti principali e l'ellisse di confidenza (linea nera).

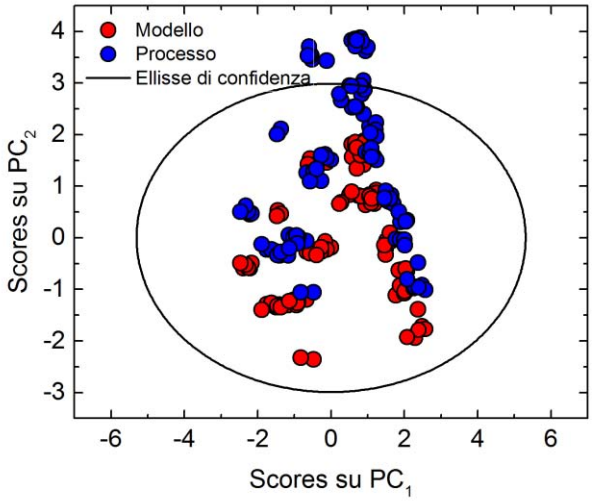


Figura 3.6 *Grafico degli scores per il modello PCA costruito su* $\mathbf{X}_{M,v}$ *(simboli rossi) lungo le due componenti principali e proiezione dei dati di processo (simboli blu).*

Il diagramma degli *scores* evidenzia come la distribuzione dei dati sia nel caso del modello che nel caso del processo presenti un certo raggruppamento in accordo con la distribuzione della concentrazione di penicillina osservata in Figura 3.2. Si osserva infatti come i dati sia del modello che del processo si distribuiscano sul piano in piccoli raggruppamenti caratterizzati da valori di scores molto simili tra loro.

Anche in questo caso la distribuzione degli *scores* per i due dataset offre alcune prime indicazioni della presenza di PMM. In Figura 3.6 si può osservare come gli *scores* per i campioni del processo si scostino sensibilmente dagli *scores* del modello, presentando un certo numero di campioni al di fuori dell'ellisse di confidenza, ma in modo meno marcato rispetto al caso precedente.

Infine osservando come la traslazione degli *scores* del processo avvenga preferenzialmente lungo la seconda componente principale possiamo ipotizzare che la variabile su cui incide il *mismatch* sia maggiormente descritta da quest'ultima componente.

A tal scopo, in Figura 3.7 vengono riportati i *loadings* delle variabili ausiliarie rispetto alle due componenti principali del modello PCA.

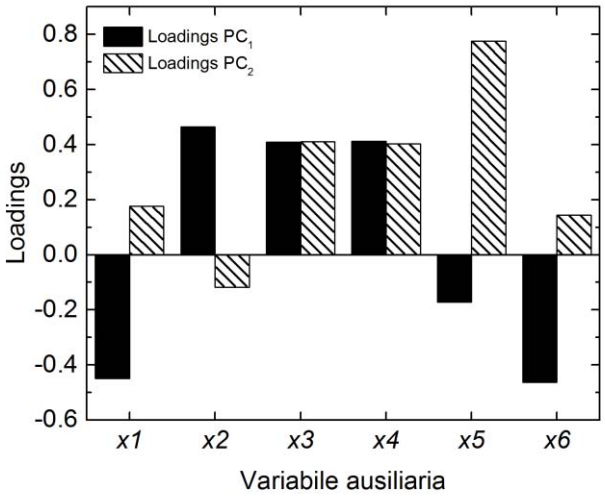


Figura 3.7 *Grafico dei loadings per il modello PCA costruito su* $\mathbf{X}_{Mv}$.

L'analisi del diagramma dei *loadings* permette di identificare quali siano le variabili ausiliare che presentano un peso maggiore lungo PC_2 e quindi probabilmente legate alla causa del PMM. Come possiamo osservare dalla figura vi sono tre variabili ausiliarie fortemente correlate con la PC_2 e quindi verosimilmente responsabili del PMM. Tali variabili sono: x_3, x_4, x_5. È interessate osservare come queste tre variabili siano effettivamente correlate all'errore introdotto sul parametro $Y_{p/s}$. Infatti la variabile x_3 moltiplica tale parametro della definizione in base alla variabile ausiliaria x_5 e la variabile ausiliaria x_4 è espressione diretta della concentrazione di penicillina la quale a sua volta è correlata alla definizione del parametro $Y_{p/s}$ (resa di penicillina su substrato). Tuttavia tale analisi offre un risultato interessante, in quanto la variabile x_5, responsabile del PMM, viene riportata con il *loading* più elevato lungo PC_2.

3.4.2 Analisi dell'indice MRLR

Per confermare i risultati ottenuti tramite l'analisi dei parametri del modello PCA, si esegue la metodologia proposta da Meneghetti *et al.* (2014) secondo quanto riportato in §1.2.1. tramite il calcolo dell'indice MRLR (1.31), i cui valori per ogni variabile sono riportati in Figura 3.8.
Come si può osservare, le variabili con indice MRLR più elevato e quindi responsabili del PMM sono le variabili x_4, x_5 e in particolare x_3, la quale sebbene sia influenzata indirettamente dalla presenza del *mismatch* non ne è la causa principale. Infatti, la causa principale del PMM è presente nella variabile x_5 che, sebbene presenti errori elevati, sono molto simili a x_4 e inferiori a x_3. In questo caso, oltre all'effetto della stretta correlazione tra le tre variabili in oggetto, è da considerare anche che la variazione della resa in penicillina dei dati del modello rispetto a quelli di processo causata dalla modifica della costante di resa è elevata (il 50% circa). È inevitabile quindi che più variabili risentano in modo marcato di quella variazione, e l'indice MRLR non è in grado in questo caso di eliminare tale effetto.

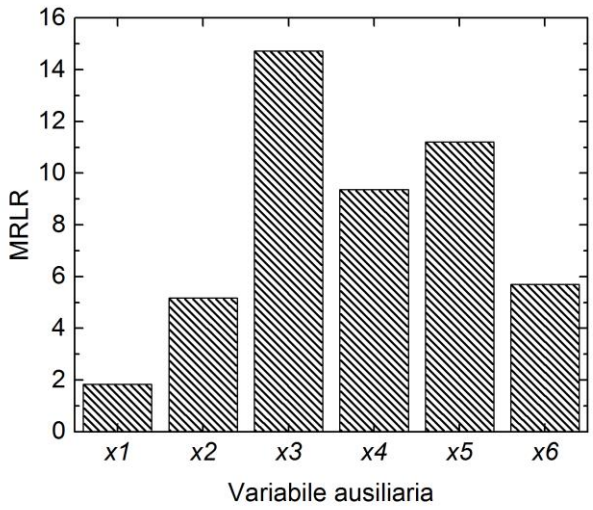


Figura 3.8 *Indici* MRLR *calcolati a partire dalle proiezioni di* $\mathbf{X}_{\Pi v}$ *sul modello PCA costruito su* $\mathbf{X}_{Mv}$.

In questo caso, la metodologia applicata per l'identificazione del PMM permetta di individuare le variabili direttamente influenzate dal PMM sebbene non sia in grado di discriminare tra le variabili maggiormente collegate al *mismatch* e la variabile causa del *mismatch*. La combinazione di questa analisi con lo studio preliminare condotto sul diagramma dei *loadings* si potrebbe superare il limite della procedura evidenziato in questo caso studio e procedere con l'identificazione della variabile responsabile del PMM nella variabile x_5.

3.5 Caso studio 3: errore introdotto sul parametro $Y_{x/s}$

In questo terzo esempio, viene modificato parametro $Y_{x/s}$ per forzare la presenza di un *mismatch*. La scelta di introdurre un errore su tale parametro viene fatta con l'obbiettivo di verificare se metodologia diagnostica applicata è in grado di discriminare tra due errori incidenti sulla medesima variabile ausiliaria. Si osserva infatti da (3.1) come entrambi i parametri figurino della definizione della variabile ausiliaria x_5 ma $Y_{p/s}$ moltiplica la variabile x_3 mentre $Y_{x/s}$ moltiplica la variabile x_2.
L'obiettivo di questo caso studio è quindi di testare la capacità diagnostica della procedura implementata nel fornire risultati diversi e accettabili rispetto al caso precedente, in cui il parametro modificato coinvolge termini simili del modello.

3.5.1 Comparazione delle matrici $\mathbf{X}_{Mv}$ e $\mathbf{X}_{\Pi v}$ analisi delle componenti principali

Il modello PCA viene costruito sulla matrice di modello $\mathbf{X}_{Mv}$ considerando solo le prime due componenti principali, scelte in base alla regola dell'autovalore maggiore di 1 (Mardia et al.,

1979) riportato in §1.1.1.2. Queste descrivono il 99.35% della variabilità dei dati (Tabella 3.4). La matrice $\mathbf{X}_{\Pi v}$ viene poi scalata sulla media e la deviazione standard di $\mathbf{X}_{Mv}$, e viene proiettata nello spazio del modello calibrato su $\mathbf{X}_{Mv}$.

Tabella 3.4 *Variabilità dei dati catturata da ciascuna componente principale del modello PCA costruito su* $\mathbf{X}_{M,v}$.

Numero PC	Autovalore	R^2	R^2 cumulato
1	5.76	94.94	95.94
2	0.21	3.42	99.35

In Figura 3.9 viene riportato il grafico degli *scores* della matrice di modello (simboli rossi) sul quale vengono proiettati gli *scores* del processo (simboli blu) nel piano identificato dalle prime due componenti principali e l'ellisse di confidenza (linea nera).

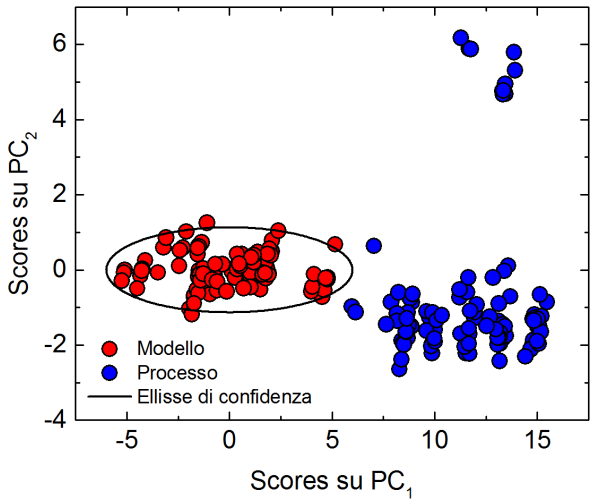


Figura 3.9 *Grafico degli scores per il modello PCA costruito su* $\mathbf{X}_{M,v}$ *(simboli rossi) lungo le due componenti principali e proiezione dei dati di processo (simboli blu).*

Anche in questo caso la distribuzione degli *scores* per i due dataset offre chiare indicazioni della presenza di un PMM come riscontrato in precedenza. In Figura 3.9, infatti si può osservare come gli *scores* per tutti i campioni del processo si scostino sensibilmente dagli *scores* del modello al di fuori dell'ellisse di confidenza §1.1.1.3.

Si nota come la traslazione degli *scores* del processo avvenga preferenzialmente lungo la prima componente principale, indicando che la variabile su cui incide il PMM è con maggiore probabilità descritta da questa direzione del piano.

Confrontando Figura 3.9 con Figura 3.6 si può osservare facilmente come i diagrammi degli *scores* per i due casi studio risultino nettamente differenti sia per quanto riguarda la distribuzione degli scores sul piano, sia per quanto riguarda la traslazione dei dati tra modello

e processo. In Figura 3.6 la traslazione dei dati di processo è lungo PC_2 mentre in Figura 3.9 la traslazione è lungo PC_1.
In conclusione l'analisi del diagramma degli *scores* per questo terzo caso studio permette di identificare la presenza di un *mismatch* e qualitativamente permette di distinguerlo dal *mismatch* riportato in §3.4.
In Figura 3.10 vengono riportati i *loadings* delle variabili ausiliarie rispetto alle due componenti principali del modello PCA.

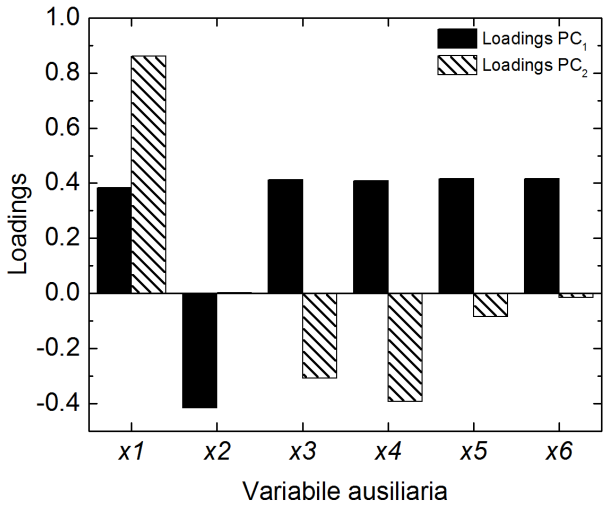


Figura 3.10 *Grafico dei loadings per il modello PCA costruito su* $\mathbf{X}_{Mv}$.

In questo caso a differenza di quanto riportato in §3.4, 1'analisi del diagramma dei *loadings* non permette di identificare quali siano le variabili ausiliare maggiormente legate al PMM. Come possiamo osservare dalla figura infatti tutte le variabili ausiliarie presentano pesi simili lungo PC_1.

3.5.2 Analisi dell'indice MRLR

I risultati ottenuti dall'applicazione della metodologia proposta da Meneghetti *et al.* (2014) sono riportati in Figura 3.11, in cui si osserva che le variabili con indice MRLR più elevato sono le variabili x_2, x_3, x_5, x_6, tra le quali la variabile x_5 presenta indice MRLR maggiore mentre la variabile x_3 quello minore. Confrontando Figura 3.11 con Figura 3.8 è possibile osservare che le variabili indicate come direttamente influenzate o responsabili del *mismatch* differiscono nei due casi studio.
È bene sottolineare tuttavia, che in entrambi i casi studio gli indici MRLR per le variabili correlate al *mismatch* sono molto simili, impedendo un'identificazione univoca della variabile responsabile del PMM come nel caso studio in cui è stato forzato la presenza del *mismatch* modificando il parlamento $K_l a$.

Tuttavia, in questo caso al contrario di quanto osservato in Figura 3.8 per l'errore introdotto sul parametro $Y_{p/s}$, l'indice MRLR permette almeno indicativamente di indentificare correttamente la principale variabile responsabile del PMM, ovvero la variabile x_5. Si ricorda infatti che il parametro $Y_{x/s}$ in cui è stato introdotto un errore figura esplicitamente in quella variabile ausiliaria. In questo caso i risultati ottenuti differiscono marcatamente da quanto ottenuto da Ibrahim (2016) per il medesimo caso studio. Si ritiene che ciò sia dovuto al fatto che i dataset considerati nei due casi sono diversi a causa della generazione dei dati tramite selezione casuale dei valori di ingresso al simulatore Pensim.

Si nota inoltre che in questo caso la modifica del parametro $Y_{x/s}$ causa una riduzione più contenuta della resa di penicillina di quanto avvenga in §3.4.2. Quello è probabilmente il motivo per cui la variabile x_4 non presenta un indice MRLR confrontabile con le variabili direttamente influenzate dal PMM in aggiunta al fatto che, in questo caso, tale variabile non è direttamente influenzata dal *mismatch*.

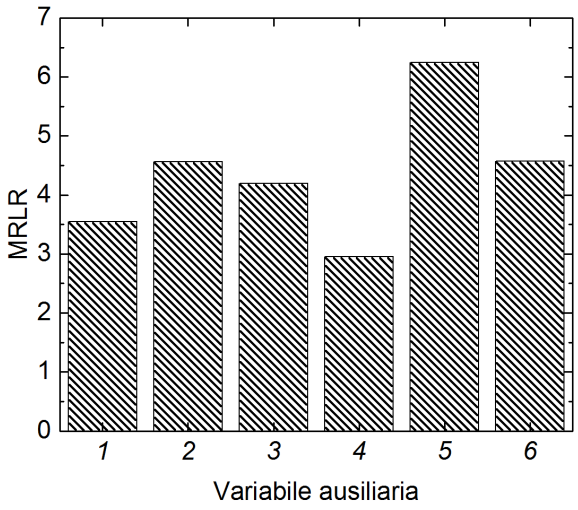


Figura 3.11 *Indici* MRLR *calcolati a partire dalle proiezioni di* $\mathbf{X}_{\Pi}$ *sul modello PCA costruito su* $\mathbf{X}_{M,v}$

3.6 Conclusioni

In questo Capitolo è stata testata la metodologia sviluppata da Meneghetti *et al.,* (2014) e basata sulla analisi delle componenti principali (PCA) per l'identificazione della variabile responsabile di un PMM su tre casi studio, i quali hanno previsto l'introduzione di un errore rispettivamente sui parametri: $K_l a$, $Y_{p/s}$ e $Y_{x/s}$.

Nei tre casi studio affrontati, l'analisi preliminare dei diagrammi degli *scores* e dei *loadings* del modello PCA costruito per la determinazione delle variabili responsabili del PMM ha permesso di ottenere buone informazioni preliminari per identificare le variabili maggiormente responsabili del *mismatch*. E bene tuttavia sottolineare che la qualità e la quantità di

informazioni che è stato possibile estrarre da tali analisi sono risultate strettamente dipendenti dal caso studio indagato in quanto sensibili all'effetto del PMM sulle uscite del modello e alle correlazioni che intercorro tra le variabili indagate.

Si è infatti osservato come l'individuazione della variabile ausiliaria su cui è stato introdotto l'errore è risultata univoca nel caso della modifica del coefficiente volumetrico di trasporto di massa dell'ossigeno $K_l a$ in quanto tale variabile risulta debolmente correlata con le altre variabili ausiliarie e l'effetto del PMM è contenuta.

Per quanto riguarda invece gli errori introdotti sui parametri $Y_{p/s}$ e $Y_{x/s}$ l'analisi dei residui di un modello PCA tramite l'indice MRLR non ha fornito un'indicazione chiara della variabile responsabile del PMM. Questo a causa delle forti correlazioni che intercorrono tra le variabili associate al PMM e del notevole effetto del *mismatch* su queste ultime. Tuttavia è risultato interessante osservare come l'analisi dell'indice MRLR sia stata in grado di fornire comunque una diagnosi distinta tra questi due PMM sebbene incidenti, in modo analogo ma non uguale sulla medesima variabile.

Complessivamente si evince quindi la potenzialità della metodologia diagnostica basata sull'indice MRLR se combinata con un analisi preliminare del relativo modello PCA. Tuttavia si individuano alcuni limiti della tecnica nell'identificare con precisione la variabile responsabile del PMM nel caso di variabili fortemente correlate e di PMM con forte impatto sulle uscite del modello.

CAPITOLO 4

Diagnosi tramite analisi dei coefficienti di correlazione di variabili originali

In questo Capitolo viene testata la seconda metodologia in esame implementata per l'identificazione delle possibili cause del disallineamento tra modello e processo (*process/model mismatch,* PMM). Tale soluzione, che si basa sull'analisi dei coefficienti di correlazione, è stata sviluppata partendo dalla metodologia proposta da Rato e Reis (2015) nell'ambito del monitoraggio di processo.
Il modello considerato è il modello a principi primi sviluppato per rappresentare un processo di fermentazione della penicillina (Birol *el al.*, 2002) descritto nel Capitolo 2, mentre i due casi di studio considerati coinvolgono entrambi un PMM parametrico (§3.4-3.5). La procedura è stata sviluppata per fornire una soluzione ai limiti rilevati dall'applicazione della procedura diagnostica sviluppata da Meneghetti (2014), Capitolo 3.

4.1 Generazione dei dati

La procedura di diagnosi del PMM, viene implementata adattando la metodologia proposta da Rato e Reis (2015) per il monitoraggio di sistemi continui. Tale procedura prevede il calcolo dei coefficienti di correlazione di grado 1 (Eq. 1.31) per diverse variabili costituite da un certo numero di osservazioni successive. Allo scopo di riprodurre condizioni simili a quelle in cui la procedura è stata sviluppata, è necessario avere a disposizione una distribuzione di coefficienti di correlazione per ogni variabile. Quindi per ognuno degli N campioni della matrice di processo $\mathbf{X}_{\Pi}$ (§3.1) [$N \times K$] vengono eseguite B simulazioni, chiamate ripetizioni, che si differenziano tra loro per rumore bianco. Il risultato ottenuto è una matrice tridimensionale $\underline{\mathbf{X}}_{B\Pi}$ di dimensioni [$N \times K \times B$], dove N *è* il numero di campioni, K il numero di variabili e B il numero di ripetizioni. Lo scopo è di testare l'efficacia della procedura in condizioni ottimali, e in un secondo momento, riadattarla in base alle caratteristiche dei dati disponibili. La procedura appena descritta viene ripetuta per la costruzione della matrice di modello $\underline{\mathbf{X}}_{BM}$ tramite la preventiva modifica nel modello a principi primi di Birol *et al.,* (2002) dei parametri $K_l a$ e $Y_{p/s}$ secondo quanto descritto in §2.3, (casi studio no1 e 2 Capitolo 3).
Sia per la matrice del processo $\underline{\mathbf{X}}_{B\Pi}$ che per la matrice del modello $\underline{\mathbf{X}}_{BM}$ vengono quindi calcolati $J = 1 \times K \cdot (K\text{-}1) \cdot (K\text{-}2) / 2$ coefficienti di correlazione di grado 1 secondo la (1.31) relativi al

processo e al modello. Ogni coefficiente di correlazione è calcolato considerando gli N campioni disponibili per ogni K-esima variabile di ogni B-esima ripetizione delle matrici tridimensionali $\underline{\mathbf{X}}_{BП}$ e $\underline{\mathbf{X}}_{BM}$. In questo modo si ottiene una distribuzione di B coefficienti di correlazione di grado 1 per ognuna delle K variabili considerate. I coefficienti di correlazione del modello sono normalizzati rispetto al valore medio di popolazione ρ del coefficiente di correlazione (la media di ciascuna distribuzione dei J coefficienti di correlazione) secondo la (1.34). La procedura di normalizzazione viene ripetuta per i coefficienti ottenuti considerando la matrice di processo ma effettuando la normalizzazione rispetto al vettore media ρ dei coefficienti di correlazione della matrice di modello. Infine, per ogni distribuzione di coefficienti di correlazione normalizzati generati dalla matrice di modello, vengono calcolati i limiti di confidenza (secondo l'Eq.1.30) necessari per applicare la procedura di diagnosi descritta in §1.2.1, secondo la quale ad ogni variabile viene assegnato l'indice 'GIALLO', 'ARANCIONE' o 'ROSSO' a seconda del grado di probabilità di rappresentare la causa del PMM indagato.

4.1.1 Analisi della distribuzione dei dati generati

Di seguito viene riportata l'analisi della distribuzione dei valori della concentrazione di penicillina generati dal simulatore per i diversi dataset utilizzati per le indagini descritte in questo Capitolo. L'obbiettivo di questa analisi è duplice: (i) verificare se le B ripetizione degli N campioni delle matrici di modello e di processo $\underline{\mathbf{X}}_{BM}$ $\underline{\mathbf{X}}_{BП}$ si distribuiscono normalmente con un intorno ristretto; (ii) selezionare dei sottogruppi di dati distribuiti normalmente su cui applicare le procedure diagnostiche e garantire quindi delle condizioni di analisi ottimali, ovvero più simili possibile a quelle utilizzate da Rato e Reis (2015) nei diversi casi studio riportati.

In Figura 4.1 viene riportata la distribuzione dei valori finali di concentrazione di penicillina nelle diverse ripetizioni per un campione selezionato casualmente sia per i dati definiti storici, ovvero generati senza alcuna alterazione del modello che per i dati simulati ottenuti modificando il parametro $K_l a$ o il parametro $Y_{p/s.}$ Come si può osservare i dati della matrice di processo si distribuiscono normalmente (le diverse ripetizioni differiscono solo a causa di rumore bianco) con media 1.38 (g/l) e deviazione standard 0.016 (g/l). La distribuzione normale viene riscontrata anche in seguito alla forzatura di un errore sui parametri $K_l a$ e $Y_{p/s}$. Nel primo caso i dati presentano una media pari a 1.19 (g/l) e una deviazione standard pari a 0.031 (g/l) nel secondo caso una media paria a 0.6 (g/l) e una deviazione standard pari a 0.012 (g/l).

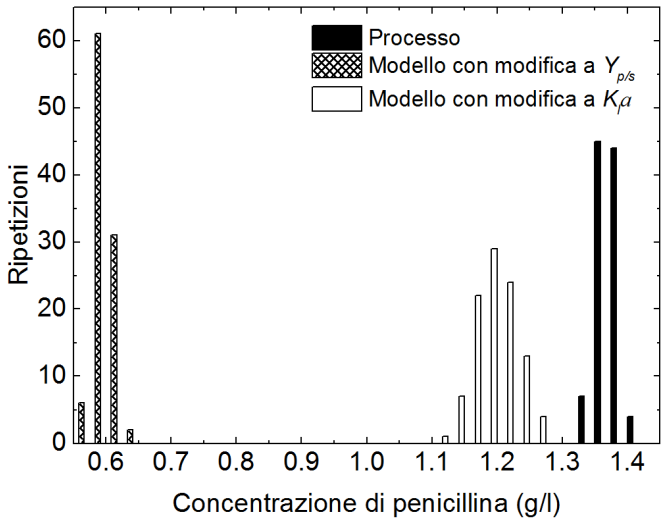


Figura 4.1 *Distribuzioni dei valori finali della concentrazione di penicillina nelle diverse ripetizioni per: il processo, il modello con modifica del parametro $K_l a$, il modello con modifica del parametro $Y_{p/s}$.*

Infine, per garantire delle condizioni di analisi ottimali, si procede con la selezione di alcuni sottogruppi di dati che presentano con una distribuzione normale ristretta. Tale selezione viene condotta tenendo in considerazione il fatto che l'eventuale presenza di sottoinsiemi di dati con diverse caratteristiche è dovuta ad un diverso effetto del PMM sulle variabili analizzate. Quest'ultima evidenza potrebbe quindi limitare l'efficacia della procedura di diagnosi del PMM che in questo caso si basa proprio sull'analisi dettagliata delle diverse correlazioni tra le variabili analizzate.

Come già osservato in Figura 3.2*b* e come riportato in Figura 4.2*a*, l'insieme dei valori finali di penicillina nei campioni della matrice di modello in cui è stato modificato il parametro $K_l a$ non presenta sottoinsiemi significativi. Si osserva infatti come tutti i campioni si distribuiscano normalmente intorno a un valore medio pari a 1.19 (g/l) eccetto alcuni *outlier* prossimi a 0.3 (g/l) e 0.65 (g/l). Questi ultimi vengono rimossi, mentre i restanti costituiscono il dataset $\underline{\mathbf{X}}_{Bs}$ di dimensioni [$H \times K \times B$] dove $H = 96$ è il numero di campioni, K il numero delle variabili e B il numero delle ripetizioni (barre in grigio scuro in Figura 4.2*a*).

Per quanto riguarda invece l'insieme dei valori finali di penicillina dei campioni della matrice di modello in cui è stato modificato il parametro $Y_{p/s}$ da Figura 4.2*b* è possibile identificare la presenza di tre sottoinsiemi con medie rispettivamente pari a: 0.4, 0.7, 1 (g/l).

Per questo motivo, viene selezionato un sottoinsieme di campioni della matrice $\underline{\mathbf{X}}_{BM}$ con media 0.7 (g/l) e deviazione standard inferiore a 0.1 (g/l) per definire il dataset $\underline{\mathbf{X}}_{Bs}$ di dimensioni [$G \times K \times B$] dove $G = 25$ è il numero di campioni, K il numero delle variabili e B il numero delle ripetizioni (barre in grigio scuro in Figura 4.2*b*).

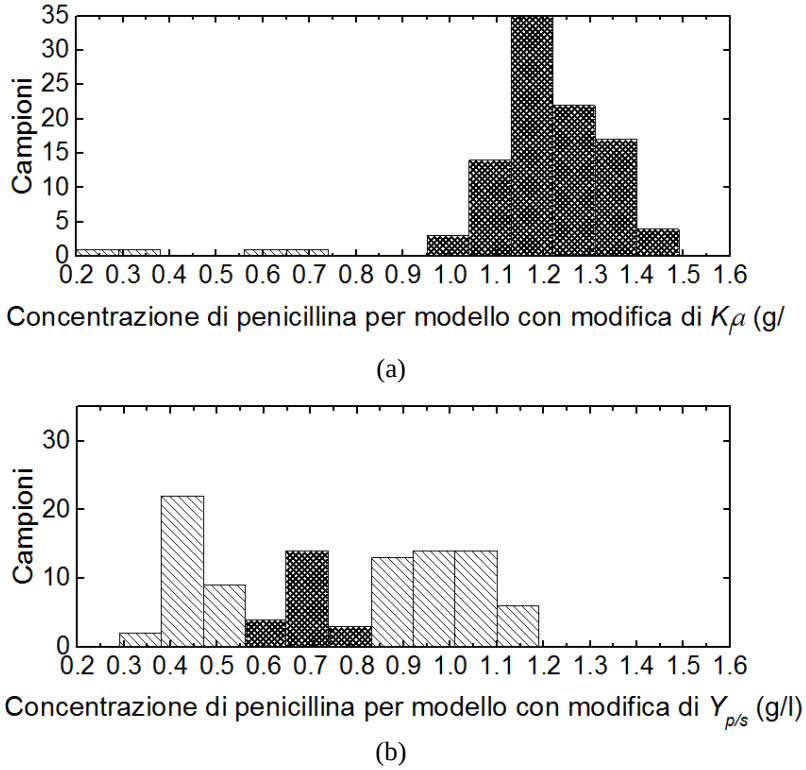


Figura 4.2 *Distribuzione dei valori finali della concentrazione di penicillina nei diversi campioni per:* (a) *il modello con modifica del parametro* $K_l a$, (b) *il modello con modifica del parametro* $Y_{p/s}$. *Le barre in grigio scuro riportano i campioni selezionati per l'applicazione della procedura diagnostica.*

4.2 Caso studio 1: errore introdotto sul parametro $K_l a$

Come primo esempio, viene riportata l'applicazione della procedura diagnostica adattata dalla metodologia di Rato e Reis (2015) per l'identificazione della variabile maggiormente responsabile del PMM causato dalla modifica del coefficiente volumetrico di trasporto di massa dell'ossigeno $K_l a$.

In seguito, viene riportata l'analisi effettuata per valutare il ruolo rivestito dalla scelta delle variabili considerate nelle matrici analizzate e dalla presenza del rumore nella classificazione delle variabili responsabili del PMM. Le diagnosi sono eseguite sui dataset $\underline{\mathbf{X}}_{Bs}$ di dimensioni $[H \times K \times B]$.

4.2.1 Applicazione della procedura diagnostica

Per applicare la procedura diagnostica in condizioni di analisi ottimali, non è sufficiente ampliare il set di dati considerato ma è necessario considerare le variabili originali. Di conseguenza, l'obbiettivo non è quello di identificare il termine del modello maggiormente responsabile del PMM, che si può ottenere tramite l'analisi delle variabili ausiliarie ma, la

variabile di processo che ne è maggiormente influenzata. Di queste ne vengono selezionate le principali: la concentrazione di substrato, C_s, la concentrazione di ossigeno disciolto, C_L, la concentrazione di biomassa, C_x e la concentrazione di penicillina, C_p.

Definito tale set di variabili, si esegue la procedura di diagnosi basata sull'assegnazione delle classi definite in §1.2.2, ovvero per ogni ripetizione, si verifica quali variabili abbiano la maggior distanza di coppia e la minore distanza di controllo. Si ricorda che la distanza di coppia per la *k*-esima variabile quantifica la frequenza con cui i coefficienti di correlazione di grado 1, in cui tale variabile rappresenta una delle due variabile in coppia, si trovano al di fuori dei limiti di confidenza nelle diverse ripetizioni. Viceversa la distanza di controllo per la *k*-esima variabile rappresenta la frequenza con cui i coefficienti di correlazione di grado 1, in cui tale variabile figura come variabile in controllo, si trovano al di fuori dei limiti di confidenza nelle diverse ripetizioni.

In Figura 4.3 sono riportati i risultati della procedura di classificazione per la diagnosi della variabile principalmente responsabile del PMM, da cui risulta evidente che la concentrazione di substrato *Cs* venga indicata come la principale variabile responsabile del PMM (indice 'ROSSO', per il 100% delle ripetizioni), mentre la concentrazione di ossigeno disciolto C_L (indice 'GIALLO' per il 50% delle ripetizioni), appare solo parzialmente riconducibile alla causa del PMM.

Tale risultato non è conforme alle aspettative, in quanto ci si aspetta che la concentrazione di ossigeno, direttamente collegata al parametro $K_l a$, venga classificata con una maggiore percentuale di indice 'ROSSO'.

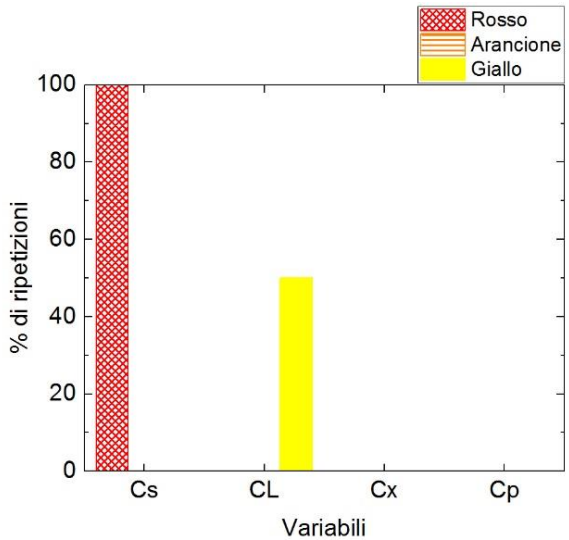


Figura 4.3 *Percentuale di ripetizioni in cui ogni variabile è stata classificata con indice 'ROSSO', 'ARANCIONE' o 'GIALLO'.*

4.2.1.1 Analisi dei PCC

Per verificare che l'analisi riportata in §4.2.1, sebbene non riconduca all'identificazione della variabile responsabile del PMM, sia effettivamente in grado di identificare modifiche nella struttura di correlazione delle matrici in esame, si procede con l'analisi dei singoli coefficienti di correlazione di grado 1 per il modello e per il processo, in base ai quali è stata definita la classificazione di ogni variabile riportata in Figura 4.3. A tal scopo, in Figura 4.4, sono riportati i valori medi delle distribuzioni per i coefficienti di correlazione normalizzati secondo l'Eq.1.34, mentre in Tabella 4.1 sono riportate le corrispondenze tra il numero di ogni coefficiente di correlazione riportato in Figura 4.4 e le relative variabili rappresentate.

Tabella 4.1 *Corrispondenza tra numero identificativo di ogni coefficienti di correlazione di grado 1 e rispettive variabili rappresentate.*

Coefficienti di correlazione di grado 1							
1	2	3	4	5	6	7	8
$C_sC_xC_L$	$C_sC_LC_x$	$C_sC_LC_p$	$C_LC_xC_s$	$C_LC_sC_x$	$C_LC_sC_p$	$C_xC_LC_s$	$C_xC_sC_L$
9	10	11	12	13	14	15	16
$C_xC_sC_p$	$C_pC_LC_s$	$C_pC_sC_L$	$C_pC_sC_x$	$C_sC_pC_L$	$C_sC_pC_x$	$C_sC_xC_p$	$C_LC_pC_s$
17	18	19	20	21	22	23	24
$C_LC_pC_x$	$C_LC_xC_p$	$C_xC_pC_s$	$C_xC_pC_L$	$C_xC_LC_p$	$C_pC_xC_s$	$C_pC_xC_L$	$C_pC_LC_x$

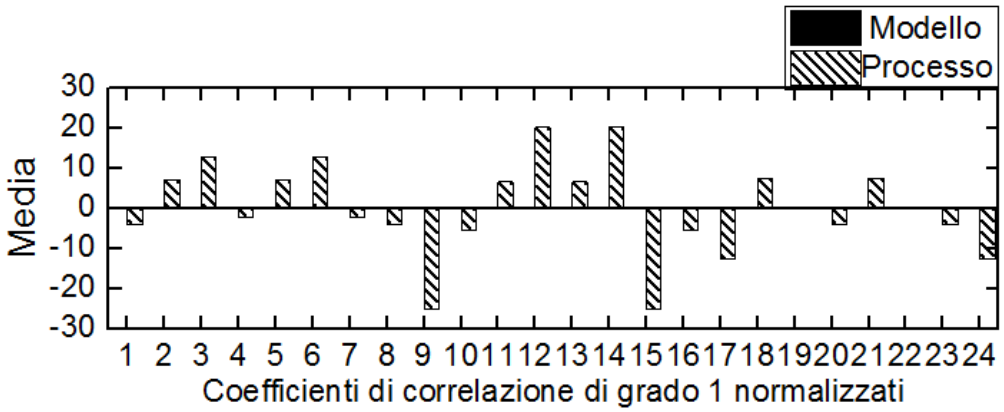


Figura 4.4 *Valori medi dei coefficienti di correlazione di grado 1 normalizzati secondo la (1.34).*

Come è possibile osservare in Figura 4.4 la maggior parte dei valori medi dei coefficienti di correlazione relativi al processo si discostano sensibilmente dai valori medi relativi al modello (che dopo la normalizzazione risultano pari a zero), indicando chiaramente la presenza di differenze nella struttura di correlazione tra processo e modello dovuta all'insorgenza del PMM. In particolare, i coefficienti che si discostano in modo più marcato rispetto alla media sono i coefficienti: 3, 6, 9, 12, 14, 15, 17, 24. Si osserva (Tabella 4.1) che solo in alcuni di questi

coefficienti (coefficienti 3, 6, 17 e 24) compare come variabile di coppia la concentrazione di ossigeno disciolta (C_L) variabile che dovrebbe essere maggiormente influenzata dalla modifica del parametro $K_l a$, mentre nella maggior parte di questi compare come variabile in coppia la concentrazione di substrato (C_s). I coefficienti che si discostano in misura trascurabile rispetto alla media, sono invece i coefficienti 4, 7, 10, 16, 19 e 22. Tra questi coefficienti, la variabile in controllo è ancora la concentrazione di substrato. In entrambi i casi comunque lo scostamento è ben oltre il valore di soglia considerato.
Inoltre, in generale, la distanza di controllo per la variabile concentrazione di substrato è minore di quella relativa alla concentrazione di ossigeno e questo fa sì che la prima variabile sia classificata con indice 'ROSSO' e la seconda 'GIALLO'. Quindi, sebbene l'analisi sia in grado di verificare la presenza di una discrepanza nella struttura di correlazione delle due matrici, non è in grado, almeno per questo set di campioni, di identificare la causa del *mismatch*.
Lo studio dettagliato dei coefficienti di correlazione di grado 1 è stata ripetuta sfruttando anche l'equazione di normalizzazione (1.35). Tuttavia non si sono evidenziate particolari modifiche nella rappresentazione grafica dei valori medi e delle deviazioni standard per i coefficienti di correlazione tali da alterare l'analisi condotta precedentemente.

4.2.1.2 Analisi dell'effetto del rumore sulla procedura diagnostica

In Figura 4.1 in cui è riportata la distribuzione della concentrazione di penicillina finale ottenuta nelle diverse ripetizioni di uno stesso campione, è possibile può osservare che l'effetto del rumore tra le diverse ripetizioni in alcuni casi è confrontabile con l'effetto dell'errore introdotto sul parametro $K_l a$. Si può infatti verificare che per 10 ripetizioni del campione considerato in Figura 4.1 la concentrazione di penicillina per il processo è prossima a 1.35 (g/ l) come per 5 ripetizioni del modello in cui è stato introdotto l'errore su $K_l a$. Per questo motivo l'analisi diagnostica viene ripetuta per un nuovo set di campioni per i quali l'effetto del *mismatch* non è confrontabile con quello del rumore. A tal scopo, per ogni campione si calcola la differenza tra il valore finale della concentrazione di penicillina per le misure storiche e il valore finale della concentrazione di penicillina per le misure simulate. Il calcolo viene ripetuto per tutte le B ripetizioni delle matrici, ottenendo così la matrice $\mathbf{Z}$ di dimensioni $[H \times B]$ (Step 1). Successivamente per ogni riga della matrice $\mathbf{Z}$ si calcola la differenza tra il valore massimo e il valore minimo all'interno della relativa distribuzione lungo le B ripetizioni, le quali, si ricorda differiscono le une dalle altre solo per rumore bianco. In questo modo si costruisce il vettore $\mathbf{a}$ $[H \times 1]$ (Step 2). Delle matrici $\underline{\mathbf{X}}_{\mathrm{BsП}}$, $\underline{\mathbf{X}}_{\mathrm{BsM}}$ si mantengono solo le ripetizioni di ogni campione in cui la differenza tra il relativo elemento della matrice $\mathbf{Z}$ e il relativo elemento del vettore $\mathbf{a}$ risulta positiva (Step 3). Infine vengono rimossi tutti i campioni delle matrici $\underline{\mathbf{X}}_{\mathrm{BsП}}$, $\underline{\mathbf{X}}_{\mathrm{BsM}}$ che presentano un numero di ripetizioni che rispettano le condizioni imposte nello stadio precedente inferiore al 20% del numero iniziale (Step 4).
In Figura 4.5 è riportato lo schema dell'algoritmo di selezione dei campioni descritto.

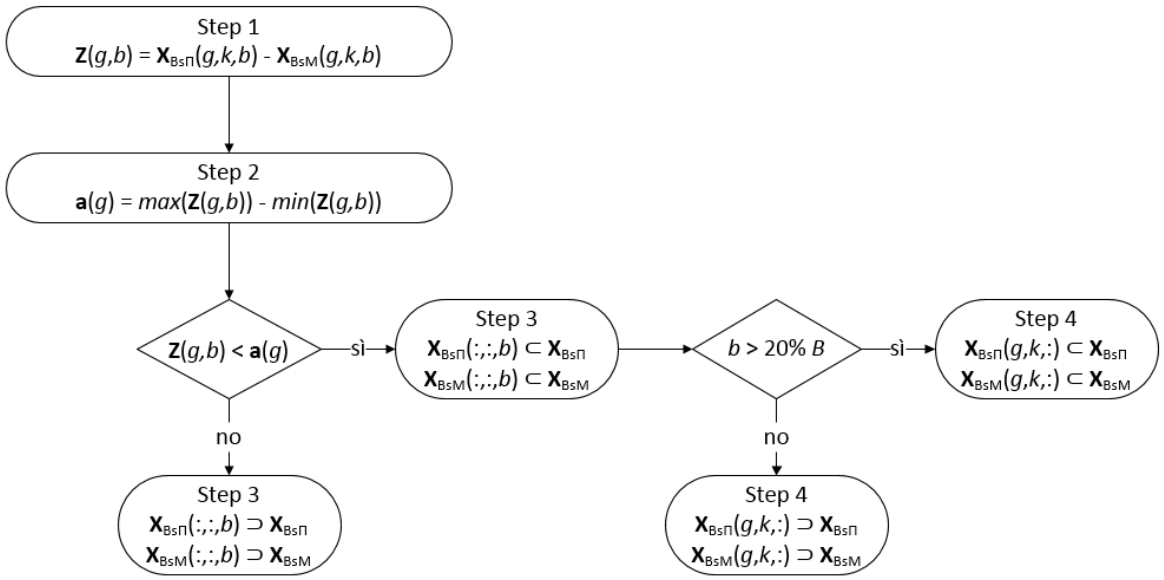


Figura 4.5. *Algoritmo per selezionare i campioni adatti alla procedura diagnostica*

Solo 13 dei 100 campioni originali soddisfano le condizioni imposte dalla procedura di selezione di Figura 4.7.e per ognuno di questi campioni sono state selezionate solo 20 ripetizioni.

In Figura 4.6 si riportano i risultati della procedura di classificazione iterata considerando il nuovo dataset ottenuto, in cui la deviazione della concentrazione di penicillina dovuta al rumore è minore della deviazione dovuta alla modifica del parametro $K_l a$ e quindi in conseguenza al PMM. Come si può osservare la concentrazione di ossigeno disciolto, C_L, presenta una classificazione con indice 'ROSSO' per ognuna delle 20 ripetizioni mentre le restanti quattro variabili riportano una classificazione con indice 'ARANCIONE' per il 5% delle ripetizioni considerate. Tale risultato soddisfacente evidenzia il ruolo determinante svolto del rumore nel modificare l'efficacia della procedura diagnostica basata sui coefficienti di correlazione nel caso in cui la variabilità dei dati di processo dovuta al PMM sia confrontabile con la variabilità introdotta dal rumore.

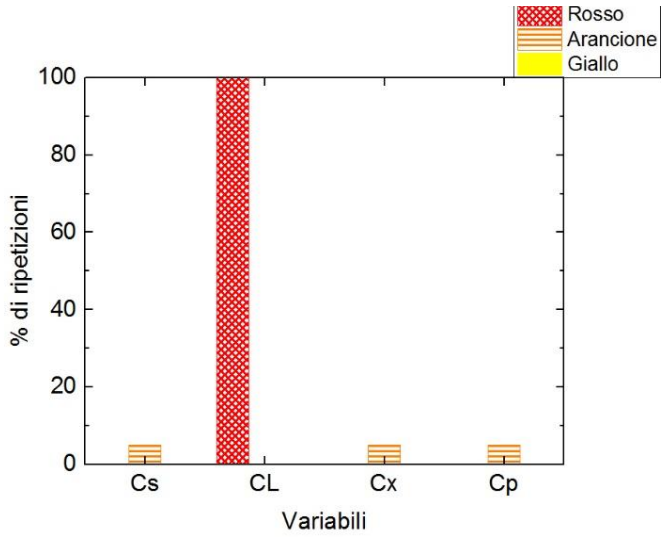


Figura 4.6 *Percentuale di ripetizioni in cui ogni variabile è stata classificata con indice 'ROSSO', 'ARANCIONE' o 'GIALLO'.*

4.2.1.3 Valutazione della robustezza della procedura diagnostica

Oltre all'effetto del rumore, è interessante valutare la robustezza della procedura proposta al variare del dataset disponibile per l'analisi. A tal scopo, vengono costruiti 10 diversi dataset estraendo in modo random 50 campioni del dataset $\underline{\mathbf{X}}_{BsM}$ e $\underline{\mathbf{X}}_{Bs\Pi}$ (con le relative 100 ripetizioni). Le matrici risultanti, sono rispettivamente $\underline{\mathbf{X}}_{BM50}$ e $\underline{\mathbf{X}}_{B\Pi 50}$ di dimensioni $[M \times K \times B]$ dove M sono i campioni, K le variabili e B le ripetizioni.

I risultati della procedura di classificazione delle variabili selezionate per la diagnosi della variabile principalmente responsabile del PMM, sono riportati in Figura 4.7 per ognuno dei 10 dataset considerati. In particolare lungo l'asse x sono riportati i dieci dataset considerati, lungo l'asse y le variabili selezionate per ognuno dei dieci dataset e lungo l'asse z le percentuali di ripetizioni per cui una certa variabile è classificata con un certo indice.

Come è possibile osservare il *trend* della classificazione delle variabili nei 10 dataset è molto simile. Infatti in ogni dataset la concentrazione di substrato è l'unica variabile che presenta classificazione con indice 'ROSSO' per circa il 100% delle ripetizioni.

È quindi evidente che la tecnica diagnostica sia robusta al variare del dataset considerato nell'identificare una variabile come variabile principalmente responsabile del PMM sebbene in questo caso tale variabile non coincida con l'effettiva causa PMM.

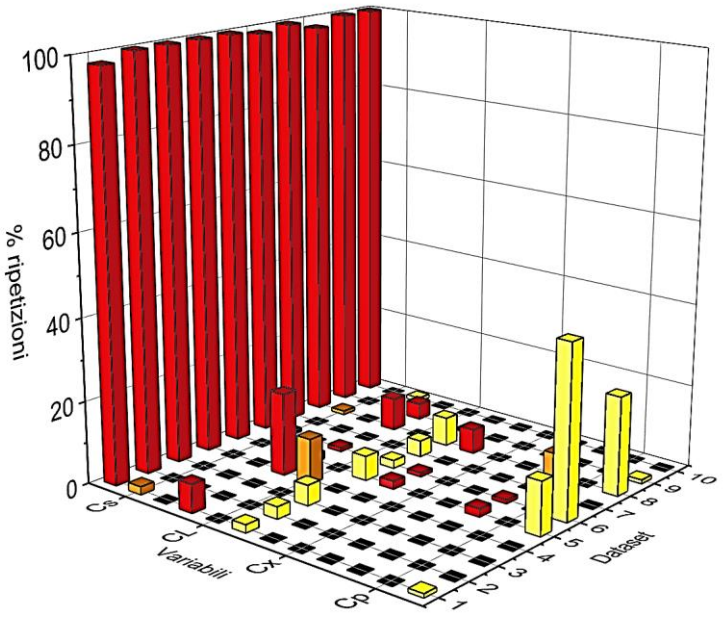


Figura 4.7 *Percentuale di ripetizioni in cui ogni variabile è stata classificata come 'ROSSO', 'ARANCIONE' o 'GIALLO' per i 10 dataset* $\underline{\mathbf{X}}_{BM50}$ e $\underline{\mathbf{X}}_{BП50}$.

4.3 Caso studio 2: errore introdotto sul parametro $Y_{p/s}$

In questo secondo esempio, viene forzata la presenza di un *mismatch* introducendo un errore sul coefficiente di resa di produzione di penicillina su consumo di substrato $Y_{p/s}$. La procedura di diagnosi in esame viene testata considerando un set di 4 variabili in uscita dal processo, investigando in particolare il ruolo rivestito dalle correlazioni tra le variabili e l'incidenza del rumore sulla procedura diagnostica. Infine viene sviluppata una tecnica decorrelativa di pretrattamento dei dati per migliorare l'efficacia diagnostica della procedura. Le diagnosi sono eseguite sui dataset $\underline{\mathbf{X}}_{Bs}$ di dimensioni [$25 \times 4 \times 100$].

4.3.1 Applicazione della procedura diagnostica

In Figura 4.8 si riportano i risultati della procedura di classificazione per la diagnosi della variabile principalmente responsabile del PMM ottenuta secondo quanto descritto in §4.2.1. Le variabili di processo considerate sono le stesse del caso precedente: concentrazione di substrato (C_s), concentrazione di ossigeno disciolto (C_L) concentrazione di biomassa (C_x), concentrazione di penicillina (C_p).

.

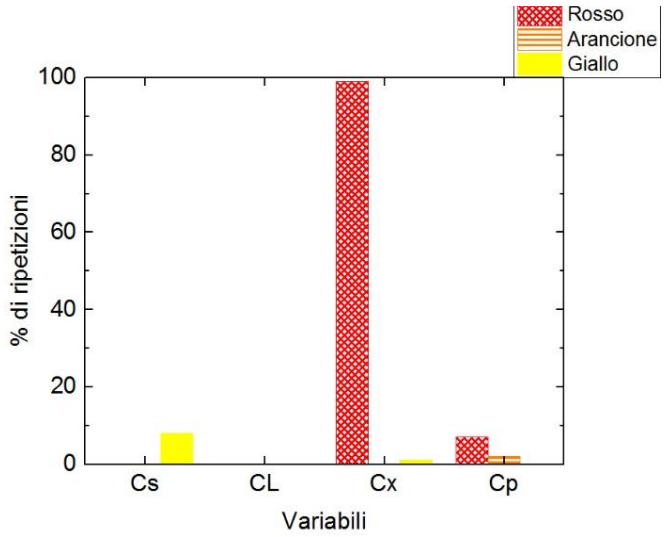


Figura 4.8 *Percentuale di ripetizioni in cui ogni variabile è stata classificata con indice 'ROSSO', 'ARANCIONE' o 'GIALLO'.*

Come si può osservare dalla Figura 4.8 la concentrazione di biomassa presenta una classificazione con indice 'ROSSO' per tutte le 100 ripetizioni considerate. Purtroppo, sebbene questa variabile sia altamente correlata con la concentrazione di penicillina e substrato, non è la causa principale del PMM. Infatti, come si può osservare dalle equazioni costitutive (2.1) – (2.11) del modello a principi primi di Birol *et al.*, (2002), queste ultime (ed in particolare C_s) sono direttamente interessate dall'errore introdotto, ma vengono classificate rispettivamente con indice 'ROSSO' e con indice 'GIALLO' solo per il 10% delle ripetizioni.
Dall'analisi del modello di Birol *et al.,* (2002) tale risultato viene preliminarmente imputato alle forti correlazioni che si stabiliscono tra le variabili del modello e che impediscono un efficace diagnosi da parte della metodologia utilizzata.

4.3.1.1 Analisi dei PCC

Si procede con l'analisi dei coefficienti di correlazione di grado 1 calcolati per la matrice di modello e di processo. Tale analisi viene effettuata per verificare se la mancata identificazione della variabile responsabile del PMM riportata in §4.3.1, sia dovuta al fatto che le deviazioni dei coefficienti di correlazione tra matrice di modello e matrice di processo per tale variabile sono confrontabili con le deviazioni tra i coefficienti di correlazione di variabili indipendenti dal PMM. Ovvero per verificare se i coefficienti di correlazione per le variabili indicate come responsabili del PMM si discostino notevolmente dal valore di soglia pari a 0.
A tal scopo, in Figura 4.9, sono riportati i valori medi delle distribuzioni dei coefficienti di correlazione normalizzati secondo la (1.34). In Tabella 4.2 sono riportate le corrispondenze tra il numero identificativo di ogni coefficiente di correlazione riportato in Figura 4.9 e le variabili considerate nel calcolarlo.

Tabella 4.2 *Corrispondenza tra numero identificativo di ogni coefficienti di correlazione di grado 1 e rispettive variabili rappresentate.*

Coefficienti di correlazione di grado 1							
1	2	3	4	5	6	7	8
$C_sC_xC_L$	$C_sC_LC_x$	$C_sC_LC_p$	$C_LC_xC_s$	$C_LC_sC_x$	$C_LC_sC_p$	$C_xC_LC_s$	$C_xC_sC_L$
9	10	11	12	13	14	15	16
$C_xC_sC_p$	$C_pC_LC_s$	$C_pC_sC_L$	$C_pC_sC_x$	$C_sC_pC_L$	$C_sC_pC_x$	$C_sC_xC_p$	$C_LC_pC_s$
17	18	19	20	21	22	23	24
$C_LC_pC_x$	$C_LC_xC_p$	$C_xC_pC_s$	$C_xC_pC_L$	$C_xC_LC_p$	$C_pC_xC_s$	$C_pC_xC_L$	$C_pC_LC_x$

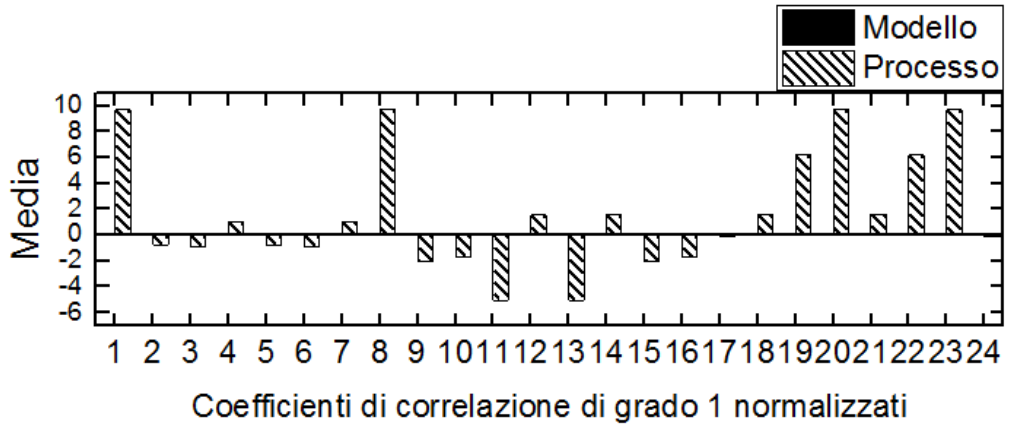


Figura 4.9 *Valori medi dei coefficienti di correlazione di grado 1 normalizzati secondo la (1.34).*

Come è possibile osservare in Figura 4.9 la maggior parte dei valori medi dei coefficienti di correlazione relativi al processo si discostano sensibilmente dai valori medi dei coefficienti relativi al modello (soglia), indicando chiaramente la presenza di differenze nella struttura di correlazione tra processo e modello dovuta all'insorgenza del PMM. In particolare, i coefficienti che si discostano in modo più marcato rispetto alla soglia sono i coefficienti: 1, 8, 11, 13, 19, 22, 23. Mentre i coefficienti che si discostano in misura trascurabile rispetto alla soglia sono i coefficienti 2, 3, 4, 5, 6, 7, 17, 24.

Nella prima serie di coefficienti da Tabella 4.2 possiamo osservare come per i coefficienti 1, 8, 22, 23 nelle variabili di coppia figurino o la concentrazione di substrato e la concentrazione di biomassa o la concentrazione di penicillina e la concentrazione di biomassa.

Per quanto riguarda la seconda serie di coefficienti si rileva come nella maggioranza dei casi la variabile in controllo sia la concentrazione di biomassa, in particolar modo nei casi in cui la deviazione del coefficiente tende a zero. Questo risultato giustifica la classificazione con indice 'ROSSO' per tale variabile nella diagnosi effettuata in §4.3.1.

Per questo caso studio non viene eseguita la valutazione della robustezza della procedura, in quanto a causa del raggruppamento dei dataset indagati il numero di campioni disponibili non è sufficiente per eseguire tale analisi.

4.3.2 Decorrelazione delle variabili

L'applicazione della procedura diagnostica nel caso studio in cui è stato introdotto un errore sul parametro $Y_{p/s}$ ha evidenziato i propri limiti nell'identificazione della variabile responsabile del PMM nel caso in cui tra termine del modello contenete l'errore e le altre variabili considerate siano presenti delle forti correlazioni. Si osserva infatti che la discrepanza tra i coefficienti di correlazione parziale calcolati considerando i dati della matrice di modello e quelli calcolati considerando i dati della matrice di processo è fortemente dipendente dalle correlazioni che intercorrono tra le variabili considerate. Per esempio, quando il rapporto tra due variabili altamente correlate subisce una piccola deviazione, la corrispondente variazione nel coefficiente di correlazione è minima. Al contrario, ogni volta che due variabili inizialmente indipendenti diventano correlate in qualche misura, la loro correlazione cambia in modo evidente. Pertanto, al fine di rilevare piccole variazioni nella struttura, è vantaggioso elaborare i dati in modo da ottenere variabili non correlate. Un principio simile è stato applicato da Hawkins (1993) con variabili *adjusted-regressed.* Tuttavia, nel suo lavoro, la trasformazione delle variabili è volta al monitoraggio di processo tramite l'identificazione delle deviazioni dal valore medio delle stesse e non a migliorare la capacità di rilevamento di variazioni delle correlazioni coinvolte. L'uso di variabili non correlate è stata applicata anche per monitorare la covarianza marginale a fini di semplificazione dei metodi di monitoraggio di processo (Hawkins e Maboudou-Tchao 2008; Huwang *et al.,* 2007). Per l'identificazione di un PMM, invece, Ibrahim (2016) ha utilizzato delle tecniche decorrelative basate sulla decomposizione di Cholesky (Press *et al.,* 2007), evidenziando i limiti delle tecniche applicate ai casi studio in esame. Per questo motivo viene testata una tecnica decorrelativa alternativa basata sull'identificazione della rete di causalità esistente tra le variabili considerate come proposto da Rato e Reis (2014).

La trasformazione applicata, necessita dell'identificazione dei legami rilevanti tra variabili considerate, in base ai quali viene ricostruita una rete di causalità (e.g., Figura 4.10*a* e *b*). Ad ogni variabile, viene quindi sottratto il contributo di ogni variabile riconosciuta come 'genitore', attraverso la determinazione di un modello di regressione Figura 4.10*c*. Infine, il calcolo dei coefficienti di correlazione e la conseguente diagnosi, vengono effettuati sulle nuove variabili ottenute da tale regressione. L'obiettivo è quello di rompere i legami primari tra le variabili, in modo da ottenere dei nuovi coefficienti di correlazione con un valore molto basso, prossimo a zero.

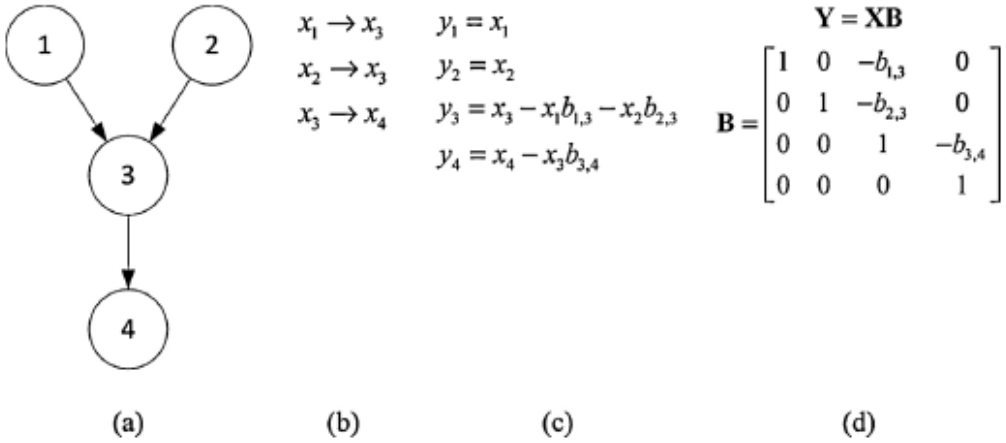


Figura 4.10 *Implementazione della procedura di trasformazione proposta:* (a) *rete di causalità esistente tra 4 variabili prese in esempio,* (b) *individuazione dei legami rilevanti,* (c) *implementazione di regressioni lineari basate sulla rete di causalità identificata* (d) *modello finale. Da Rato e Reis 2014.*

4.3.2.1 Metodo di decorrelazione 1

L'applicazione della procedura di decorrelazione delle variabili necessita della determinazione dei legami causali tra le variabili del modello ossia della cosiddetta ricostruzione della rete di causalità. Il punto di partenza della procedura sono i coefficienti di correlazione, i quali rappresentano una misura di associazione tra variabili continue. Tramite tali coefficienti è possibile la costruzione di una rete di legami tra le variabili e da questa successivamente lo sviluppo di una rete di causalità tra le stesse. Di tale approccio sono presenti diversi esempi in letteratura, per esempio Fuente *et al.,* (2007) e Pellet e Elisseeff (2004), utilizzano coefficienti di correlazione fino al 2° grado al fine di recuperare i legami significativi tra variabili. La procedura implementata in questa Tesi per la ricostruzione della rete di legami rappresenta di fatto un adattamento della procedura sviluppata da Rato e Reis (2014) per il monitoraggio di processo. Le modifiche apportate alla procedura consistono nella rimozione degli step di costruzione delle matrici *time-shifted,* in quanto la procedura è implementata considerando dei dataset che contengono solo i valori finali delle simulazioni, e non dataset dinamici. La procedura include 5 step principali (Rato e Reis 2014):

1. calcolo dei legami per tutte le possibili coppie di variabili;
2. determinazione per tutte le possibili coppie di variabili del coefficiente di correlazione di grado 0 tra ogni coppia (x, y):
   a. costruzione di un modello di regressione:

$$y = x\mathrm{b}. \tag{4.1}$$

b. determinazione dell'associazione tra variabili attraverso l'analisi della significatività statistica di *a* (Eq 4.2). Se *a* è inferiore alla soglia di significatività ossia presenta *P*-value inferiore a 0.05, non viene considerato il legame per la coppia di variabili (*x*, *y*).

$$a = corr(y, x\,\mathrm{b})\,. \tag{4.2}$$

3. determinazione per tutti i legami rimanenti del coefficiente di correlazione di grado 1 per la coppia (*x*, *y*) controllata da *z*:
   a. costruzione dei seguenti modelli di regressione:

$$x = z\,\mathrm{b}\,, \tag{4.3}$$
$$y = z\,\mathrm{b}\,. \tag{4.4}$$

   b. calcolo dei residui:

$$e = x - z\,\mathrm{b}\,. \tag{4.5}$$
$$f = y - z\,\mathrm{b}\,. \tag{4.6}$$

   c. ripetizione del passaggio 2 per la coppia di variabili (*e*, *f*);
4. per tutti i legami rimanenti determinazione del coefficiente di correlazione di grado 2 per la coppia (*x*, *y*) controllata da *z* e *q*:
   a. costruzione di modelli con la seguente struttura:

$$x = [z\,q]\,\mathrm{b}\,, \tag{4.7}$$
$$y = [z\,q]\,\mathrm{b}\,. \tag{4.8}$$

   b. calcolo dei residui:

$$e = x - [z\,q]\,\mathrm{b}\,, \tag{4.9}$$
$$f = y - [z\,q]\,\mathrm{b}\,. \tag{4.10}$$

   c. ripetizione del passaggio 2 per la coppia di variabili (*e*, *f*);
5. determinazione dei legami significativi tra le variabili, per stimare la struttura di rete risultante $\mathbf{E}_0$.

Una volta ricostruita la rete di casualità, è necessario determinare le relazioni di causalità tra le variabili, ovvero i versi dei legami identificati. Anche questa procedura consiste nell'adattamento della procedura sviluppata da Rato e Reis (2014). In questo caso le modifiche apportate alla metodologia riguardano la ricostruzione delle gerarchie che rimangono

indeterminate nel primo step della procedura tramite la determinazione del miglior modello di regressione per la rete di legami ricostruita tramite la minimizzazione della somma dei quadrati degli errori tra il modello regredito e le variabili originali.
Le direzioni dei legami possono essere determinate osservando che, quando una variabile (figlio) ha più di una causa (padre), se si imposta la direzione di causalità scorretta si provoca la comparsa di un legame precedentemente non identificato che coinvolge altri genitori. La procedura implementata prevede i seguenti step:

1. per ogni coppia di direzioni delle variabili (x, y):
   a. considerata l'ipotesi che x causi y ($x \rightarrow y$):
      i. rimozione dell'effetto di x su y costruendo un modello con struttura:

$$y = x\mathrm{b}, \tag{4.11}$$

      ii. calcolo dei residui:

$$e = y - x\mathrm{b}, \tag{4.12}$$

      iii. sul set di dati originale, sostituzione di y e determinazione della nuova rete di legami, $\mathbf{E}_{\mathrm{NEW}}$;
      iv. determinazione del numero di nuovi legami $N_{x \rightarrow y}$ (i legami in $\mathbf{E}_{\mathrm{NEW}}$ che non sono stati identificati in $\mathbf{E}_0$);
      v. determinazione del numero di nuovi legami che coinvolgono variabile y, $L_{x \rightarrow y}$.
   a. ripetizione del punto 1.a considerando l'ipotesi che y causi x ($y \rightarrow x$), ottenendo $N_{y \rightarrow x}$ e $L_{y \rightarrow x}$.
   b. determinazione della direzione in base alla seguente regola: nei casi in cui il figlio (y) presenti più di un genitore (x), quando x è considerato come il "figlio di tentativo", la sostituzione di x con i residui della regressione provoca la comparsa di legami tra i residui (e) e gli altri genitori di y nella stima di una nuova rete. In caso contrario, se y è considerato come il "figlio di tentativo", non sono attesi nuovi legami tra le variabili. Pertanto, la direzione che porta al minor numero di nuovi legami è assunta come la direzione corretta (o almeno, predominante):
      i. se $L_{x \rightarrow y} < L_{y \rightarrow x}$, allora $x \rightarrow y$;
      ii. altrimenti, se $L_{y \rightarrow x} < L_{x \rightarrow y}$, allora $y \rightarrow x$;
      iii. altrimenti, se $N_{x \rightarrow y} < N_{y \rightarrow x}$, allora $x \rightarrow y$;
      iv. altrimenti, se $N_{y \rightarrow x} < N_{x \rightarrow y}$, allora $y \rightarrow x$.
2. per le restanti coppie di direzioni per le variabili (x, y):
   a. verifica della rilevanza del legame condizionando i genitori precedentemente identificati. Se il legame non è più rilevante, eliminazione del legame;

b. altrimenti, determinazione della direzione in modo tale da ottenere il miglior modello di regressione:
   i. ipotesi 1: $x \rightarrow y$, regressione di y su x seguita dal calcolo della somma dei quadrati degli errori calcolati sulla differenza tra y e y regredita.
   ii. ipotesi 2: $y \rightarrow x$, regressione di x su y a seguita dal calcolo della somma dei quadrati degli errori calcolati sulla differenza tra x e x regredita.
   iii. L'ipotesi considerata è stabilita in base al minor valore della somma dei quadrati degli errori del modello di regressione.

Una volta ottenuta la rete di causalità per le variabili considerate si procede con la regressione multilineare di ogni variabile figlio sulle variabili genitori in modo tale da ottenere il modello di regressione con cui procedere alla decorrelazione delle variabili, secondo quanto descritto in Figura 4.10.

4.3.2.2 Risultati dell'applicazione del metodo di decorrelazione 1

In questo esempio, le variabili prese in esame sono la concentrazione di substrato, C_s, la concentrazione di ossigeno disciolto, C_L, la concentrazione di biomassa, C_x, e la concentrazione di penicillina, C_p, ovvero le stesse variabili selezionate per l'applicazione della procedura diagnostica riportata in §4.3.2.1. A queste quattro variabili viene aggiunta una quinta variabile la potenza di agitazione, P_w. Tale scelta deriva dal fatto che quest'ultima variabile è direttamente in relazione con la concentrazione di ossigeno disciolto (C_L) come si può osservare dalle equazioni (2.9) e (2.10). Di conseguenza, tale relazione viene utilizzata come verifica della ricostruzione della rete di causalità. La ricostruzione della rete di causalità viene eseguita per ogni ripetizione della matrice $\underline{\mathbf{X}}_{Bs}$ di modello dove si è considerato un particolare raggruppamento dei dati, come riportato in §4.1.1.

In Figura 4.11 si riporta la rete di legami ottenuta al termine della procedura di ricostruzione in cui sequenzialmente si sono considerati i coefficienti di correlazione di grado 0, 1 e 2. Le reti di legami nei tre stadi della procedura non hanno subito modifiche ad indicazione che i legami individuati dall'analisi dei coefficienti di correlazione di grado 0 sono diretti. È bene specificare che le reti di casualità riportate sono specifiche per una particolare ripetizione della matrice $\underline{\mathbf{X}}_{Bs}$ di modello (selezionata casualmente).

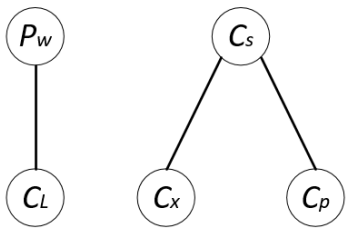
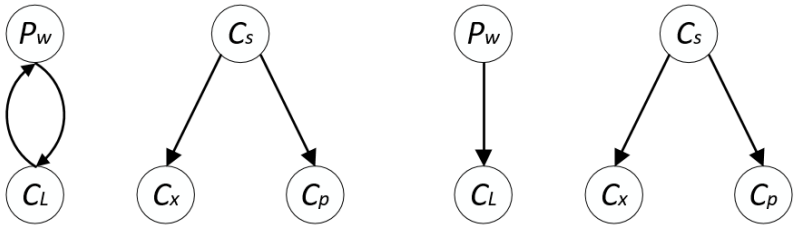


Figura 4.11 *Ricostruzione sui coefficienti di correlazione di grado 0, 1 e 2.*

Come si può osservare da Figura 4.11 la ricostruzione della rete di legami risulta consistente con la struttura del modello di Birol *et al.* (2002). Confrontando tale ricostruzione con le equazioni: (2.1), (2.2) e (2.6) – (2.10) possiamo infatti riconoscere tutte i legami identificati dalla procedura. Nello specifico il legame diretto tra C_s, C_x e C_p, l'assenza di una relazione significativa di queste tre variabili con C_L e il legame tra quest'ultima e la variabile P_w.
In Figura 4.12 si riportano le reti di causalità per la stessa ripetizione della matrice $\underline{\mathbf{X}}_{Bs}$ considerata in Figura 4.11. Nello specifico in Figura 4.12*a* si riporta la rete di causalità ottenuta tramite l'applicazione del primo stadio della procedura di identificazione dei versi dei legami (2.a), mentre in Figura 4.12*b* è riportata la rete di causalità dopo l'esecuzione del secondo stadio della stessa procedura (2.b).



(a) (b)

Figura 4.12 *Implementazione della procedura di ricostruzione della rete di causalità:* (a) *ricostruzione sul il primo stadio della procedura,* (b) *ricostruzione sul secondo stadio della procedura.*

L'identificazione dei versi dei legami appare consistente con il modello di Birol *et al.*, (2002), almeno per quanto riguarda la relazione di causalità diretta tra le variabili C_L e P_w, facilmente deducibile dalle equazioni (2.9) e (2.10) del modello.
La procedura di ricostruzione delle reti di causalità viene eseguita per ogni ripetizione della matrice $\underline{\mathbf{X}}_{Bs}$ di modello (le quali, si ricorda, differiscono gli uni dagli altri solo per del rumore bianco). Di seguito si riporta in Figura 4.13 la percentuale di ripetizioni in cui si presentano i legami di causalità tra le diverse variabili del modello nei dataset della matrice. Lungo l'asse *x* e lungo l'asse *y* sono riportate le variabili considerate mentre lungo *z* è riportata la frequenza (ovvero il numero di ripetizioni) con cui un legame si presenta per ogni coppia di variabili.

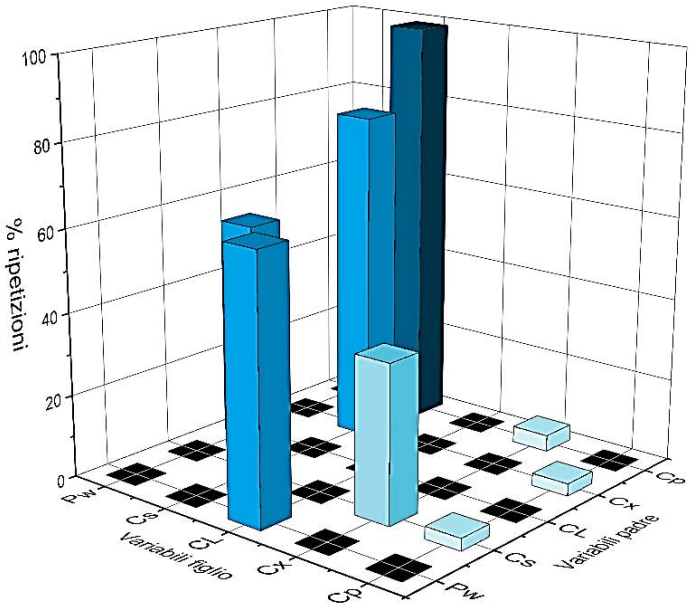


Figura 4.13 *Percentuale di ripetizioni in cui si presentano i legami di causalità tra le variabili nella matrice* $\underline{\mathbf{X}}_{Bsubset}$.

In Figura 4.13 si osserva come la maggior parte dei legami identificati dalla metodologia proposta è presente in una percentuale superiore al 50% nelle ripetizioni della matrice $\underline{\mathbf{X}}_{Bsubset}$. Tale risultato permette di considerare la rete di causalità ricostruita e riportata nelle Figure 4.11 e 4.12 una buona rappresentazione nella rete di legami media della matrice $\underline{\mathbf{X}}_{Bsubset}$. Si deve tuttavia notare che il legame tra le variabili C_L e P_w si presenta con una percentuale del 40%. Tale percentuale è legata al fatto che per il 60% delle ripetizioni l'analisi dei coefficienti di correlazione individuano il legame tra le variabili C_L e P_w non come una relazione diretta ma dipendente dalla variabile C_p.

Tale incongruenza è dovuta al fatto che il coefficiente di correlazione in cui le variabili C_L e P_w sono in coppia e la variabile C_p è in controllo è molto vicino alla soglia di significatività (*P*-value = 0.05). Conseguentemente sulle 100 ripetizioni della matrice $\underline{\mathbf{X}}_{Bs}$ vi è un'alta probabilità che il legame tra C_L e P_w possa essere considerato non diretto. Questo risultato rivela il ruolo fondamentale del valore di soglia fissato per stabilire se un legame è diretto o indiretto.

Al termine di questa fase si determina il modello di regressione multilineare in cui ogni variabile figlio viene regredito rispetto alle variabili genitori definite dalla rete di causalità media la quale è definita come l'insieme dei legami che si presentano in più del 40% delle ripetizioni della matrice $\underline{\mathbf{X}}_{Bs}$. In questo modo, per quanto visto in Figura 4.13 si procede a decorrelare le variabili considerate come rappresentato in Figura 4.12*b*. L'operazione viene eseguita per ogni ripetizione della matrice $\underline{\mathbf{X}}_{Bs}$ di modello.

Sulle variabili ottenute è stata eseguita l'analisi dettagliata dei coefficienti di correlazione di grado 1. Tale analisi, ha evidenziato la presenza di una correlazione residua non trascurabile tra

le variabili C_x e *Cp*. Come è possibile osservare da Figura 4.12*b* tale relazione non è stata identificata dalla metodologia decorrelativa per la ricostruzione della rete di causalità e conseguentemente non è stata rimossa durante la generazione delle variabili trasformate. Tuttavia il coefficiente di correlazione per tali variabili, sebbene sia al di sotto del valore di soglia, non risulta trascurabile. Tale legame è inoltre deducibile da un'analisi qualitativa delle equazioni del modello di Birol *et al.*, (2002) . Pertanto è stata modificata la rete di causalità di Figura 4.12*b* introducendo un legame diretto tra le variabili C_x e *Cp*, Figura 4.14:

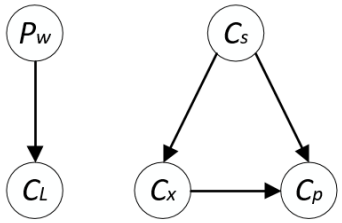


Figura 4.14 *Rete di correlazione causale modificata.*

Al termine di questa fase si determina il modello di regressione multilineare in cui ogni variabile figlio viene regredito rispetto alle variabili genitori definite dalla rete di causalità media modificata, Figura 4.14. L'operazione viene eseguita per ogni ripetizione della matrice $\underline{\mathbf{X}}_{Bs}$ di modello.

4.3.2.3 Analisi dei PCC in seguito all'applicazione del metodo di decorrelazione 1

Per valutare il grado di decorrelazione ottenuto vengono analizzati in dettaglio le medie delle distribuzioni di tutti i coefficienti di correlazione di grado 1 calcolati per il modello e per il processo (Figura 4.15) In Tabella 4.3 sono riportate le corrispondenze tra il numero identificativo di ogni coefficiente di correlazione riportato in Figura 4.15 e le variabili considerate nel calcolarlo.

In Figura 4.17 si osserva come le correlazioni tra le variabili nel modello C_x e *Cp* in seguito alla procedura di decorrelazione (coefficienti 46, 47, 48) siano tutte tendenti a zero. In questo caso quindi la combinazione della conoscenza ingegneristica del processo alla procedura di ricostruzione della rete di causalità Ha permesso di ottenere risultati molto soddisfacenti.

Dal confronto tra i coefficienti di correlazione calcolati per il modello per le cinque variabili in esame prima di applicare la procedura decorrelativa (In Figura 4.15*a*) e dopo l'applicazione della procedura (In Figura 4.15*b*) è possibile verificare una notevole riduzione dei valori medi per tutti i coefficienti di correlazione, a conferma dell'efficacia della procedura utilizzata per decorrelare le variabili coinvolte. In particolare, si può osservare che anche per le coppie di variabili fortemente correlate in cui i coefficienti di correlazione tendono a |1|, ossia C_s, C_p e C_s, C_x (si vedano Eq.2.1-2.15) grazie alla procedura decorrelativa si ottengo dei coefficienti di

correlazione tendenti a 0. Vi sono due coppie di variabili per cui permane una leggera correlazione, ossia per le coppie P_w, C_x e P_w, C_s. Pertanto, è necessario considerare che queste correlazioni residue potrebbero avere un effetto negativo sulla procedura di diagnosi implementata.

Infine, dal confronto tra i coefficienti di correlazione per il processo e i coefficienti di correlazione per il modello in seguito all'applicazione della procedura decorrelativa (Figura 4.15*b*) è possibile ricavare un'indicazione preliminare delle variabili che con maggior probabilità rappresentano la causa di *mismatch*, come dimostrato in seguito.

Tabella 4.3 *Corrispondenza tra numero identificativo di ogni coefficienti di correlazione di grado 1 e rispettive variabili rappresentate.*

Coefficienti di correlazione di grado 1							
1	2	3	4	5	6	7	8
$P_wC_sC_L$	$P_wC_sC_x$	$P_wC_sC_p$	$P_wC_LC_s$	$P_wC_LC_x$	$P_wC_LC_p$	$P_wC_xC_s$	$P_wC_xC_L$
9	10	11	12	13	14	15	16
$P_wC_xC_p$	$P_wC_pC_s$	$P_wC_pC_L$	$P_wC_pC_x$	$C_sP_wC_L$	$C_sP_wC_x$	$C_sP_wC_p$	$C_sC_LP_w$
17	18	19	20	21	22	23	24
$C_sC_LC_x$	$C_sC_LC_p$	$C_sC_xC_s$	$C_sC_xC_L$	$C_sC_xC_p$	$C_sC_pP_w$	$C_sC_pC_L$	$C_sC_pC_x$
25	26	27	28	29	30	31	32
$C_LP_wC_s$	$C_LP_wC_x$	$C_LP_wC_p$	$C_LC_sP_w$	$C_LC_sC_x$	$C_LC_sC_p$	$C_LC_xP_w$	$C_LC_xC_s$
33	34	35	36	37	38	39	40
$C_LC_xC_p$	$C_LC_pP_w$	$C_LC_pC_s$	$C_LC_pC_x$	$C_xP_wC_s$	$C_xP_wC_L$	$C_xP_wC_p$	$C_xC_sP_w$
41	42	43	44	45	46	47	48
$C_xC_sC_L$	$C_xC_sC_p$	$C_xC_LP_w$	$C_xC_LC_s$	$C_xC_LC_p$	$C_xC_pP_w$	$C_xC_pC_s$	$C_xC_pC_L$
49	50	51	52	53	54	55	56
$C_pP_wC_s$	$C_pP_wC_L$	$C_pP_wC_x$	$C_pC_sP_w$	$C_pC_sC_L$	$C_pC_sC_x$	$C_pC_LP_w$	$C_pC_LC_s$
57	58	59	60				
$C_pC_LC_x$	$C_pC_xP_w$	$C_pC_xC_L$	$C_pC_xC_s$				

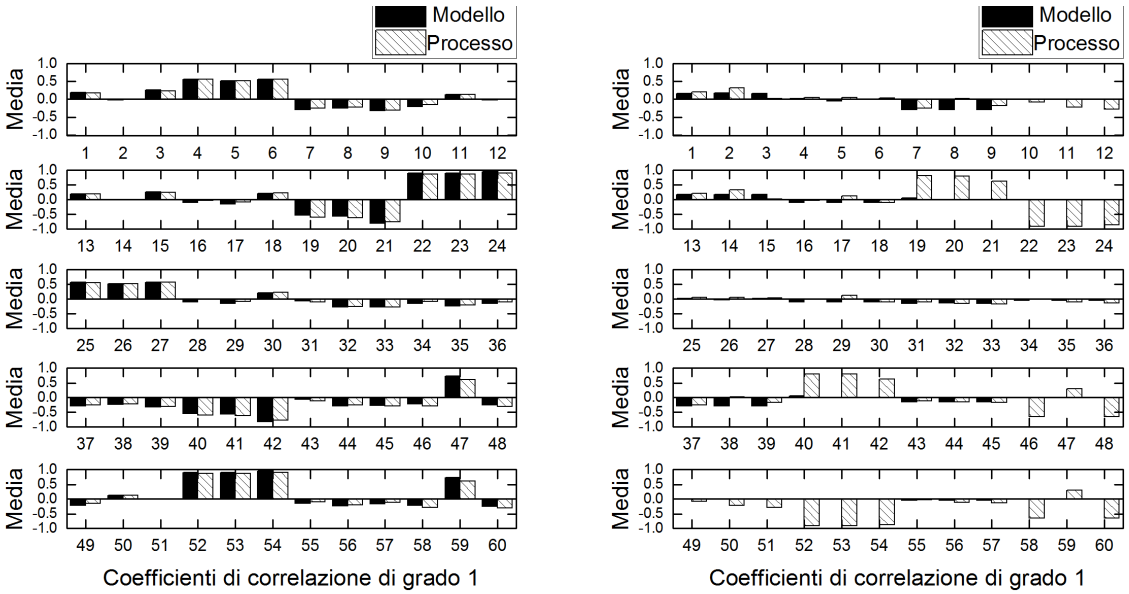


Figura 4.15 (a) *Media dei coefficienti di correlazione di grado 1 prima di applicare la procedura decorrelativa.* (b) *Media dei coefficienti di correlazione di grado 1 dopo aver applicato la procedura decorrelativa, per la matrice di processo (barra chiare) per la matrice di modello (barre nere).*

4.3.2.4 Applicazione della procedura diagnostica in seguito all'applicazione del metodo di decorrelazione 1

In Figura 4.16 si riportano i risultati della procedura di classificazione implementata per la diagnosi delle variabili principalmente responsabili del PMM ottenuta secondo quanto descritto in §4.4.2. Le variabili in esame sono: la potenza di agitazione, P_w, la concentrazione di substrato, C_s, la concentrazione di ossigeno disciolto, C_L, la concentrazione di biomassa, C_x, e la concentrazione di penicillina, C_p. I valori considerati per il calcolo dei coefficienti di correlazione sono stati ottenuti in seguito all'applicazione della procedura decorrelativa descritta in §4.3.2.1.

Come si può osservare dalla Figura 4.16 l'applicazione della procedura diagnostica permette di identificare con successo la variabile maggiormente responsabile del PMM ossia la concentrazione di substrato, C_s alla quale è attribuita una classificazione con indice 'ROSSO' per il 98% delle ripetizioni indagate. Tale risultato, se confrontato con quanto ottenuto eseguendo la stessa procedura diagnostica ma in assenza di un'adeguata decorrelazione delle variabili (Figura 4.8) dimostra l'efficacia dell'operazione di decorrelazione dei dati nel rendere efficace l'identificazione della variabile su cui incide il PMM.

Si osserva inoltre come la diagnosi identifichi con classificazione con indice 'ROSSO' anche la variabile potenza di agitazione, P_w, ma solo per un numero poco significativo di ripetizioni (2%), e quindi non rilevante ai fini diagnostici.

Quest'ultimo risultato può essere giustificato considerando la presenza di una correlazione residua tra le variabili P_w, C_s al termine della procedura decorrelativa.

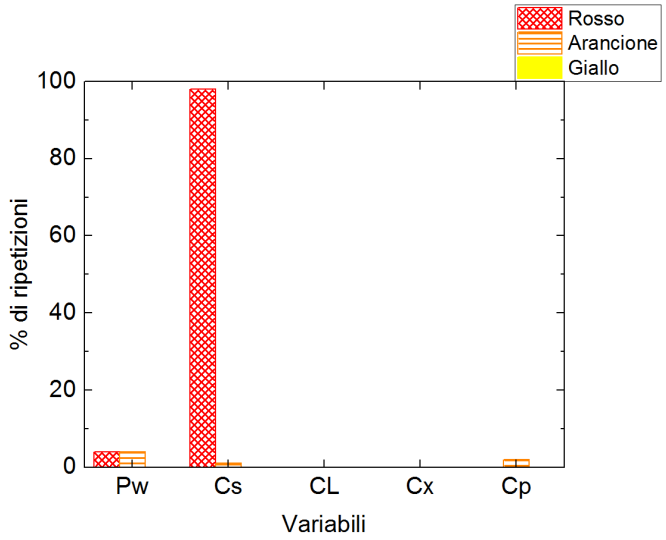


Figura 4.16 *Percentuale di ripetizioni in cui ogni variabile è stata classificata con indice 'ROSSO', 'ARANCIONE' o 'GIALLO'.*

4.4 Conclusioni

In questo Capitolo è stata testata la metodologia adattata da Rato e Reis (2015) per l'identificazione della variabile responsabile di un PMM su due casi studio. La procedura è stata applicata sulle variabili originali e ampliando il set di dati in modo tale da porsi nelle condizioni di analisi ottimali, ovvero quanto più simili a quelle utilizzate da Rato e Reis (2015) nella loro diagnosi.

I casi studi indagati hanno coinvolto l'introduzione di un errore sul coefficiente volumetrico di trasporto di massa dell'ossigeno K_la e sul parametro di resa di produzione di penicillina su consumo di substrato $Y_{p/s}$.

In questo modo è stato possibile valutare l'efficienza della metodologia implementata per un *mismatch* la cui presenza è stata forzata su una variabile poco correlata alle altre variabile (caso studio K_la) e su una variabile fortemente correlata alle altre (caso studio $Y_{p/s}$).

Nel caso dell'errore introdotto sul parametro K_la è stato osservato che il rumore casuale che differenzia le ripetizioni dei campioni nelle matrici di modello e di processo, provoca una variazione dei valori in uscita al simulatore Pensim paragonabile a quella dovuta all'incidenza del PMM. Tale effetto impedisce l'esatto riconoscimento della variabile maggiormente legata al PMM.

In seguito ad una selezione accurata dei dati con l'obbiettivo di selezionare i campioni in cui il rumore incide sono marginalmente sulla distribuzione dei dati, la procedura diagnostica

applicata al nuovo set di dati, è in grado di individuare con successo la variabile su cui è stato introdotto l'errore.

La procedura di diagnosi, inoltre, risulta poco sensibile nel perseguire il proprio scopo di identificazione della variabile responsabile del disallineamento tra modello e processo ai diversi set di dati indagati.

Nel caso dell'errore introdotto sul parametro $Y_{p/s}$ la diagnosi su variabili originali si rivela inefficace nell'individuarne la principale responsabile, ma in questo caso non è possibile imputare il parziale insuccesso della diagnosi all'effetto del rumore, dato che l'analisi della distribuzione dei dati evidenzia come la variazione dei valori in uscita al simulatore Pensim a causa del rumore casuale non sia paragonabile a quella dovuta all'incidenza del PMM.

Dall'analisi della struttura del modello di Birol *et al.,* (2002) si riscontra come il parametro $Y_{p/s}$ sul quale è stato introdotto l'errore incida su una variabile che risulta particolarmente correlata a tutte le variabili del modello. Per questo motivo è stata applicata una procedura di decorrelazione dei dati per ottenere variabili non correlate ad evidenziare maggiormente eventuali cambiamenti nella struttura di correlazione delle matrici di modello e di processo. Tale procedura, basata sulla ricostruzione della rete di causalità tramite l'analisi del grado di correlazione delle variabili considerate permette di decorrelare efficacemente le variabili originali. Il risultato diagnostico ottenuto in seguito alla decorrelazione risulta soddisfacente in quanto identifica correttamente la variabile su cui è stato introdotto l'errore. Quest'ultimo risultato conferma il ruolo di primo piano rivestito dalle correlazioni tra le variabili su cui è forzata la presenza del *mismatch* nel determinare l'efficacia della procedura diagnostica sviluppata.

CAPITOLO 5

Diagnosi tramite analisi dei coefficienti di correlazione di variabili ausiliarie

L'analisi effettuata considerando un set di variabili in uscita dal processo, ha rivelato l'efficacia della metodologia proposta da Rato e Reis (2015) nell'identificare la/e variabile/i riconducibili all'origine di un certo PMM (Capitolo 4). In base a questi risultati promettenti, la stessa procedura diagnostica è stata implementata considerando un set di variabili ausiliarie, allo scopo di determinare i termini del modello maggiormente responsabili del PMM. In particolare sono stati presi in esame il secondo e terzo caso di studio considerati in questa tesi, in cui l'errore è stato introdotto sui parametri di resa: $Y_{p/s}$ e $Y_{x/s}$.

5.1 Caso studio 2: errore introdotto sul parametro $Y_{p/s}$

Come primo esempio, viene riportata l'applicazione della procedura diagnostica adattata dalla metodologia di Rato e Reis (2015) per l'identificazione della variabile ausiliaria maggiormente responsabile del PMM in seguito alla modifica del parametro di resa $Y_{p/s}$. Tale parametro presenta delle forti correlazioni tra il termine del modello contenete l'errore e le altre variabili ausiliarie considerate per le quali l'analisi dell'indice MRLR (Meneghetti *et al.*, 2014) non si è dimostrata efficace nell'identificare il responsabile del PMM (§3.4). Successivamente vengono implementate due tecniche di pretrattamento dei dati per migliorare l'efficacia diagnostica della procedura. Le diagnosi sono eseguite sui dataset $\underline{\mathbf{X}}_{Bsv}$ di dimensioni [$G \times V \times B$].

5.1.1 Applicazione della procedura diagnostica

Per garantire delle condizioni di analisi ottimali, l'analisi è stata condotta considerando come in §4.1 un sottoinsieme di dati della matrice $\underline{\mathbf{X}}_{BM}$ con media 0.7 (g/l) e deviazione standard inferiore a 0.1 (g/l). In questo modo è stato definito il dataset $\underline{\mathbf{X}}_{Bs}$ di dimensioni [$G \times K \times B$] dove G = 25 è il numero di campioni, K = 4 il numero delle variabili e B=100 il numero delle ripetizioni. I dataset $\underline{\mathbf{X}}_{BΠs}$ e $\underline{\mathbf{X}}_{BMs}$ così ottenuti, sono utilizzati per definire rispettivamente la matrice di modello e di processo, in base al set di variabili ausiliarie definite in §3.1. Per tali matrici tridimensionali, di dimensioni [$G \times V \times B$] dove G = 25 sono i campioni mentre V = 6 le variabili ausiliarie e B = 100 le ripetizioni, vengono calcolati $J = 1 \times V \cdot (V\text{-}1) \cdot (V\text{-}2) / 2$ coefficienti di correlazione di grado 1 secondo quanto descritto in §4.1.

$$
\begin{aligned}
x_1 &= K_l a(C_L^{*} - C_L) & x_4 &= KC_p \\
x_2 &= C_x \mu & x_5 &= \frac{x_2}{Y_{x/s}} + \frac{x_3}{Y_{p/s}} + C_x m_x \qquad (5.1) \\
x_3 &= \mu_{pp} C_x & x_6 &= \frac{x_2}{Y_{x/o}} + \frac{x_3}{Y_{p/o}} + C_x m_o
\end{aligned}
$$

In Figura 5.1 si riportano i risultati della procedura di diagnosi delle variabili principalmente responsabili del PMM ottenuta secondo quanto descritto in §4.2.1.

Nessuna delle variabili indagate mostra una percentuale significativa di classificazioni con indice 'ROSSO', necessaria per determinare la variabile maggiormente responsabile del PMM. Tuttavia per la seconda variabile tale percentuale raggiunge il 20% mentre la quarta variabile raggiunge il 90% di classificazione con indice 'ARANCIONE'. Questo risultato, sebbene non permetta di identificare la variabile responsabile del PMM, permette di individuare una delle variabili maggiormente influenzata da quest'ultimo. Infatti, è possibile assumere che la variabile x_4 rappresenti la concentrazione di penicillina, C_p, (5.1) la quale è direttamente correlata al parametro $Y_{p/s}$ come si può verificare dalla definizione (2.20).

È bene tuttavia sottolineare che in Figura 5.1 la variabile x_2 viene classificata con indice 'ROSSO' sebbene quest'ultima non sia influenzata dall'errore introdotto. Questo risultato può essere tuttavia giustificato considerando il fatto che la variabile x_2 è strettamente correlata alle variabili x_5 e x_3 le quali sono effettivamente influenzate dal PMM. In questo modo si evidenzia ulteriormente la sensibilità della diagnosi alle correlazioni tra le variabili indagate e quindi la necessità di un pretrattamento dei dati tramite una tecnica decorrelativa.

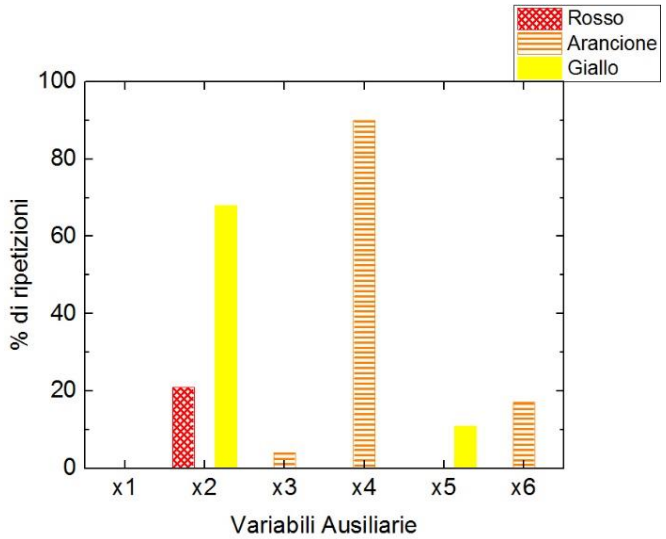


Figura 5.1 *Percentuale di ripetizioni in cui ogni variabile è stata classificata con indice 'ROSSO', 'ARANCIONE' o 'GIALLO'.*

5.1.2 Decorrelazione delle variabili ausiliarie

La variabile ausiliaria responsabile del *mismatch* è fortemente correlata con la maggior parte delle altre variabili ausiliarie. Pertanto, allo scopo di migliorare la diagnosi riportata al paragrafo precedente e in base ai risultati ottenuti applicando una procedura decorrelativa all'analisi delle variabili originali (§4.3) la stessa procedura viene applicata (descritta in §4.3.2.1) anche alle variabili ausiliarie considerate in questo studio.

5.1.2.1 Metodo di decorrelazione 1

La ricostruzione della rete di causalità per le variabili ausiliarie considerate viene eseguita per ogni ripetizione della matrice $\underline{\mathbf{X}}_{Bsv}$ tramite la procedura descritta in §4.3.2.1. In Figura 5.2 sono riportate in sequenza le reti di causalità ottenute al termine della procedura di ricostruzione in cui si sono considerati rispettivamente i coefficienti di correlazione di grado 0 (Figura 5.2*a*), di grado 1 (Figura 5.2*b*) e di grado 2 (Figura 5.2*c*).
La progressiva riduzione del numero di legami negli stadi sequenziali di ricostruzione della rete indica come la maggior parte dei legami tra le variabili non sia diretto ma dovuto a variabili intermedie. La rete di legami risultante collega le variabili x_2, x_5 e x_6, mentre le variabili x_1, x_3 e x_4 risultano completamente indipendenti dalle altre.

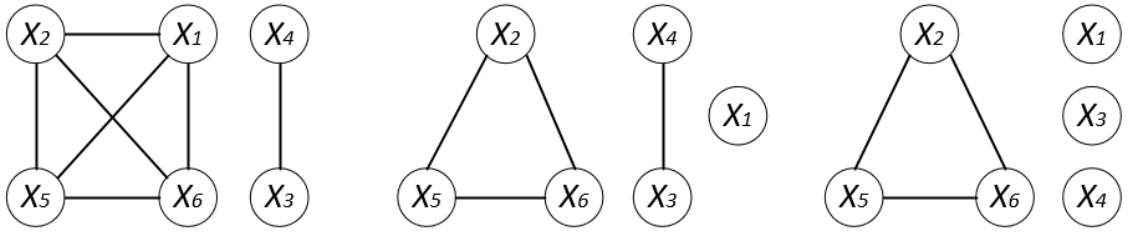


Figura 5.2 *Rete di legami risultante ottenuta da:* (a) *coefficienti di correlazione di grado 0,* (b) *coefficienti di correlazione di grado 1 e* (c) *coefficienti di correlazione di grado 2.*

Confrontando la rete di legami individuata dalla procedura applicata con la definizione delle variabili ausiliarie riportata in (5.1) si può osservare come alcuni legami che logicamente appaiono espliciti vengano efficacemente individuati mentre altri non vengano rilevati. In particolare, viene identificato il legame tra le variabili x_5 e x_6 con la variabile x_2, la quale risulta inclusa nella definizione di entrambe le variabili ausiliarie, mentre le correlazioni di tali variabili con la variabile x_3, anch'essa inclusa nella definizione di entrambe le variabili ausiliarie, non viene rilevata. Tuttavia l'analisi dei relativi coefficienti di correlazione evidenzia che per le variabili x_5 e x_3 il coefficiente di correlazione sia molto prossimo al valore di soglia utilizzato per valutare la significatività del legame. In generale, nella ricostruzione di una rete di causalità è necessario prestare particolare attenzione ai legami che presentano un coefficiente

di correlazione vicino al valore di soglia calcolato per determinarne la significatività. Tali legami infatti, possono risultare o meno nella rete finale di casualità in base al valore di soglia, che dipende anche dal numero di campioni (§5.2.1.4).
In Figura 5.3 si riportano le reti di causalità per la stessa ripetizione della matrice di modello $\underline{\mathbf{X}}_{Bsv}$ considerata in Figura 5.2. Nello specifico in Figura 5.3*a* si riporta la rete di causalità ottenuta tramite l'applicazione del primo stadio della procedura di identificazione dei versi dei legami, mentre in Figura 5.3*b* è riportata la rete di causalità dopo l'esecuzione del secondo stadio della stessa procedura.

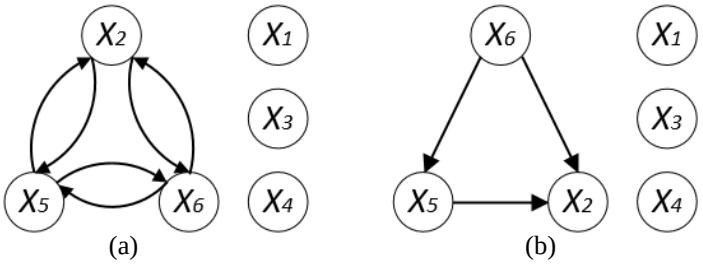


Figura 5.3 *Rete di causalità risultante per le variabili ausiliarie della matrice* $\underline{\mathbf{X}}_{Bsv}$*:* (a) *ricostruzione in seguito al primo stadio della procedura*, (b) *in seguito al secondo stadio della procedura.*

La procedura di ricostruzione delle reti di causalità viene eseguita per ogni ripetizione della matrice $\underline{\mathbf{X}}_{Bsv}$ le quali, si ricorda, differiscono le une dalle altre solo per del rumore bianco. Le relazioni identificate dalla metodologia proposta sono le stesse per tutte le ripetizioni della matrice $\underline{\mathbf{X}}_{Bsv}$. Tale risultato dimostra che la rete di causalità ricostruita è una buona rappresentazione nella rete di legami media della matrice $\underline{\mathbf{X}}_{Bsv}$.
Al termine di questa fase si procede con la regressione multilineare di ogni variabile figlio sulle variabili genitori definite dalla rete di causalità media in modo tale da ottenere il modello di regressione con cui procedere alla decorrelazione delle variabili, secondo quanto descritto in Figura 4.12. L'operazione viene eseguita su ogni ripetizione della matrice $\underline{\mathbf{X}}_{Bsv}$.

5.1.2.2 Analisi dei PCC in seguito all'applicazione del metodo di decorrelazione 1

Per valutare il livello di decorrelazione ottenuto a seguito della regressione di ogni variabile ausiliaria rispetto alle variabili identificate come 'genitori', viene effettuata un'analisi dettagliata dei coefficienti di correlazione di grado 1 per il modello e per il processo. A tal scopo in Figura 5.4, 5.5 sono riportate i valori medi delle distribuzioni per i coefficienti di correlazione di grado 1, di cui in Tabella 5.1 sono riportate le corrispondenze con le relative variabili rappresentate.

Tabella 5.1 *Corrispondenza tra numero identificativo di ogni coefficienti di correlazione di grado 1 e rispettive variabili rappresentate.*

Variabili ausiliarie nelle terne dei coefficienti di correlazione di grado 1	
pedice 1	variabile ausiliaria y_1
pedice 2	variabile ausiliaria y_2
pedice 3	variabile ausiliaria y_3
pedice 4	variabile ausiliaria y_4
pedice 5	variabile ausiliaria y_5
pedice 6	variabile ausiliaria y_6

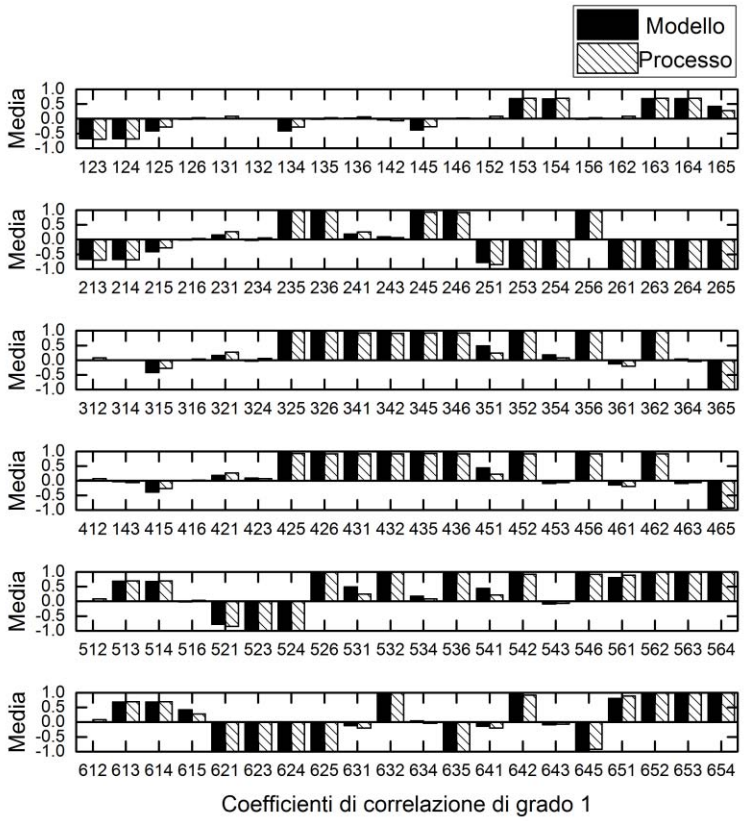


Figura 5.4 *Media dei coefficienti di correlazione di grado 1 prima dell'applicazione della procedura decorrelativa, per la matrice di processo (barra chiare) per la matrice di modello (barre nere).*

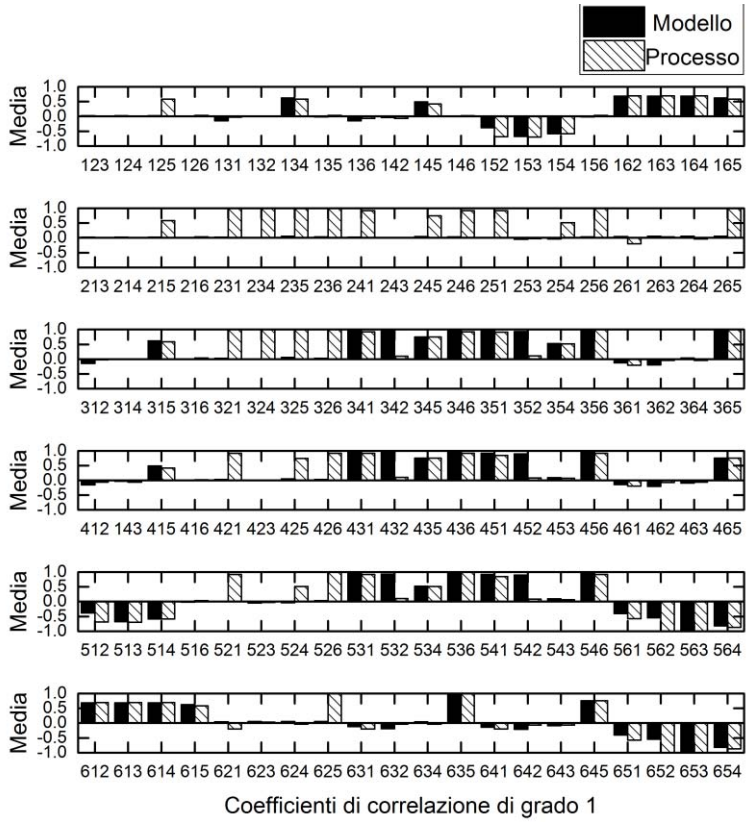


Figura 5.5 *Media dei coefficienti di correlazione di grado 1 dopo l'applicazione della procedura decorrelativa, per la matrice di processo (barra chiare) per la matrice di modello (barre nere).*

Dal confronto tra i coefficienti di correlazione per il modello per le sei variabili ausiliarie in esame prima dell'applicazione della procedura decorrelativa (Figura 5.4) e dopo l'applicazione della procedura (Figura 5.5) si osserva che la decorrelazione è stata ottenuta solo per alcune coppie di variabili considerate. In particolare, si osserva una netta riduzione dei valori medi dei i coefficienti di correlazione in cui compare la variabile x_2 come variabile di coppia, ad indicazione di una buona decorrelazione di tale variabile. Per i restanti coefficienti di correlazione invece, i valori medi si assestano in prossimità del valore massimo (± 1). Da quest'analisi si evidenzia il successo solo parziale della procedura di ricostruzione della rete di causalità attuata sulle variabili ausiliarie e della successiva procedura decorrelativa. A differenza del set di variabili originali analizzato nel Capitolo precedente, in questo caso, alcune variabili ausiliarie sono incluse nella definizione di altre variabili introducendo probabilmente relazioni fortemente non lineari nel set di variabili analizzato.

Conseguentemente la procedura di decorrelazione, basata sulla regressione di un modello lineare tra i termini del modello non è in grado di identificare e rimuovere completamente le relazioni tra questi ultimi portando così a un successo soltanto parziale della tecnica decorrelativa.

5.1.2.3 Applicazione della procedura diagnostica in seguito all'applicazione del metodo di decorrelazione 1

In Figura 5.6 si riportano i risultati della procedura di classificazione per la diagnosi della variabile principalmente responsabile del PMM ottenuta secondo quanto descritto in §4.2.1. Le variabili ausiliarie in esame sono le variabili ausiliarie x_1, x_2, x_3, x_4, x_5, x_6 a seguito all'applicazione della procedura decorrelativa previa ricostruzione della rete di causalità.

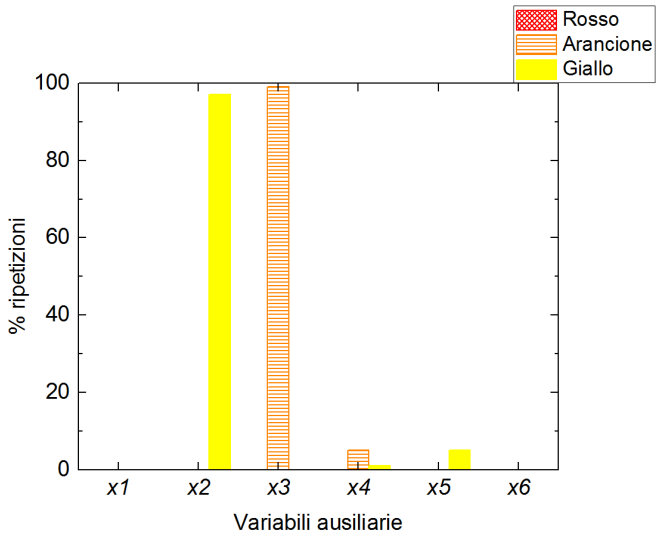


Figura 5.6 *Percentuale di ripetizioni in cui ogni variabile è stata classificata con indice 'ROSSO', 'ARANCIONE' o 'GIALLO'.*

Anche a seguito della procedura decorrelativa, la variabile x_5 in cui è contenuto il parametro $Y_{p/s}$ responsabile del PMM viene classificata erroneamente con indice 'GIALLO' e solo per il 5% delle ripetizioni. La variabile x_3 invece, riporta una classificazione con indice 'ARANCIONE' per il 100% delle ripetizioni per e la variabile x_2 viene classificata con indice 'GIALLO' per il 95% dei campioni. Entrambe queste variabili sono correlate al parametro $Y_{p/s}$ (specialmente x_3) ma non lo contengono. Si riporta inoltre una classificazione con indice 'ARANCIONE' del 5% per la variabile x_4 la quale per quanto riportato in §5.1.2.1 risulta correlata indirettamente al PMM. Dal confronto del risultato diagnostico prima dell'applicazione della procedura decorrelativa di Figura 5.1 si osserva un leggero miglioramento nel risultato ottenuto in quanto si evidenziano come principali responsabili del PMM soltanto le variabili direttamente correlate alla variabile responsabile del PMM (x_5). Purtroppo però la procedura non è in grado di identificare questa variabile come variabile responsabile del PMM. Dal confronto dei risultati riportati in Figura 5.6 e il risultato ottenuto in Figura 4.18 per le variabili originali, è possibile dedurre la necessità di una decorrelazione completa delle variabili affinché la procedura diagnostica risulti efficiente.

5.1.2.4 Analisi dell'effetto della dimensione del dataset sul metodo decorrelativo 1

La tecnica di ricostruzione della rete di causalità utilizzata in §4.3.2.1 e 5.1.2 permette di determinare i legami diretti tra le variabili indagate, individuando i coefficienti di correlazione che presentano valore medio al di sopra di una soglia di significatività pari a *P*-value = 0.05. Tale tale soglia dipende dal numero di campioni disponibili per ogni variabile daci che il *P*-value è basato sul t-test, il quale, come riportato in (5.2), dipende dal numero di campioni (Montgomery (2009):

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}, \tag{5.2}$$

dove $\bar{x}$ è la media dei campioni, μ è la media di riferimento (μ=0), *s* è la deviazione standard dei campioni, e n è il numero dei campioni.

Al variare del numero di campioni, è quindi possibile che la rete di causalità ricostruita possa variare. Per questo motivo vengono generati due dataset analoghi al dataset $\underline{\mathbf{X}}_{Bsv}$ su cui è stata fino ad ora condotta l'analisi, ma costituiti da 207 e 343 campioni con concentrazione finale di penicillina media paria a 0.5 (g/l) e una deviazione standard pari a 0.11 (g/l) definiti rispettivamente $\underline{\mathbf{X}}_{B207v}$ [$F \times V \times B$] e $\underline{\mathbf{X}}_{B343v}$ [$D \times V \times B$] dove F = 207, D =343 sono i numeri di campioni, V il numero delle variabili ausiliarie e B il numero delle ripetizioni. La procedura di ricostruzione della rete di causalità descritta in §5.1.2.1 è ripetuta per entrambi i dataset $\underline{\mathbf{X}}_{B207v}$, $\underline{\mathbf{X}}_{B343v}$, ottenendo la stessa rete di causalità riportata in Figura 5.7 per entrambi i dataset.

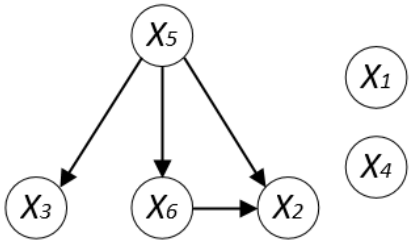


Figura 5.7 *Rete di causalità risultante per le variabili ausiliarie delle matrici* $\underline{\mathbf{X}}_{B207v}$, $\underline{\mathbf{X}}_{B343v}$.

Come si può osservare da Figura 5.7 le reti di causalità ricostruite per i dataset $\underline{\mathbf{X}}_{B207v}$, $\underline{\mathbf{X}}_{B343v}$ differiscono dalla rete di causalità ricostruita per il dataset $\underline{\mathbf{X}}_{Bsv}$ (Figura 5.3*b*) per la presenza di un legame causale tra la variabile x_5 e la variabile x_3. Tale risultato dimostra come effettivamente la procedura di ricostruzione della rete di causalità dipenda delle dimensioni dei dataset indagati. È bene sottolineare, inoltre, che sebbene la procedura impiegata identifichi il legame tra le variabili x_5 e x_3 non è in grado di individuare il legame tra le variabili x_3 e x_6, nonostante anche in questo caso la variabile x_3 sia inclusa nella definizione della variabile x_6 (5.1).

Per verificare se la rete di causalità ricostruita sui dataset $\underline{\mathbf{X}}_{B207v}$ e $\underline{\mathbf{X}}_{B343v}$ permetta una decorrelazione migliore di quella ottenuta tramite la rete di causalità ricostruita sul dataset $\underline{\mathbf{X}}_{Bsv}$ si procede con la regressione multilineare di ogni variabile figlio sulle variabili genitori in modo tale da ottenere il modello di regressione da utilizzare per decorrelare le variabili ausiliarie in esame. La decorrelazione ottenuta, sebbene si differenzi da quella ottenuta sul dataset $\underline{\mathbf{X}}_{Bsv}$ non risulta migliore di quanto riportato in §5.1.2.2 per entrambi i dataset indagati, indicando come ad un aumento della dimensione del dataset disponibile non corrisponda necessariamente ad una migliore capacità decorrelativa della procedura. Infine, per verificare se tale decorrelazione sia sufficiente per migliorare le prestazioni della procedura di diagnosi, tale procedura viene eseguita considerando i due nuovi dataset. In Figura 5.8 si riporta la classificazione delle variabili ausiliarie indagate ottenuta per il dataset $\underline{\mathbf{X}}_{B207v}$. Per quanto riguarda il dataset $\underline{\mathbf{X}}_{B343v}$ la classificazione ottenuta presenta lo stesso andamento riportato i Figura 5.8.

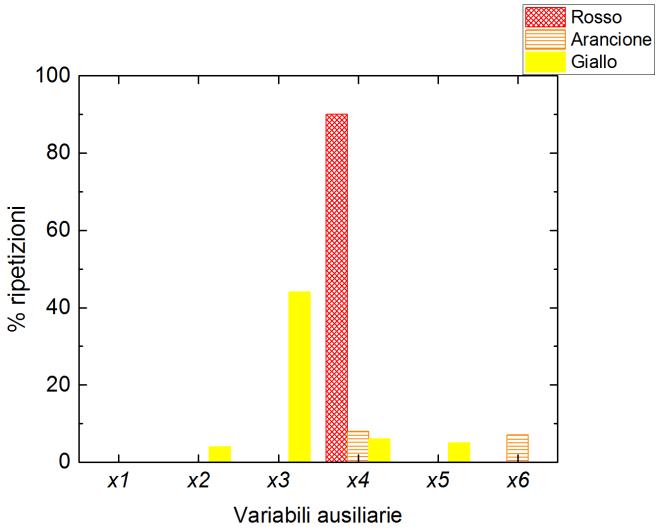


Figura 5.8 *Percentuale di ripetizioni in cui ogni variabile è stata classificata con indice 'ROSSO', 'ARANCIONE' o 'GIALLO'.*

I risultati ottenuti, riportati in Figura 5.8, evidenziano una classificazione del 90% delle ripetizioni con indice 'ROSSO' per la variabile x_4, e una classificazione con indice 'GIALLO' e per il 40% delle ripetizioni per la variabile x_3.

Questo risultato, sebbene non permetta di identificare la variabile responsabile del PMM, (x_5) permette di individuare due variabili influenzate da quest'ultimo. Infatti la variabile x_3 e il parametro su cui è stato introdotto l'errore ossia $Y_{p/s}$, rappresentano rispettivamente il dividendo e il divisore del rapporto incluso nel calcolo della variazione di C_s (Eq.5.1). Mentre la variabile x_4 rappresenta la concentrazione di penicillina, C_p, (5.1) la quale è direttamente correlate al parametro $Y_{p/s}$ come si può verificare dalla definizione (2.20).

Quanto riportato conferma la necessità di ottenere una decorrelazione completa delle variabili affinché sia possibile giudicare le prestazioni della procedura diagnostica implementata. Per

questo motivo è stata valutata una tecnica decorrelativa alternativa, i cui risultati sono riportati nel seguente paragrafo.

5.1.2.5 Metodo di decorrelazione 2

La procedura decorrelativa adattata dalla metodologia di Rato e Reis (2014) non è risultata efficace nella decorrelazione delle variabili ausiliarie. Pertanto è stata implementata una tecnica decorrelativa alternativa. Questa tecnica prevede di ricostruire direttamente un modello di regressione lineare per ogni variabile considerata e identificare a posteriori tramite i *P*-value la rete di legami tra le variabili necessaria alla decorrelazione di queste ultime.

La regressione lineare per ogni variabile è basata sulla funzione *fitglm* disponibile in Matlab®, v.2015b. La tecnica decorrelativa viene eseguita per ogni ripetizione della matrice $\underline{\mathbf{X}}_{Bsv}$. La procedura prevede di definire un modello di regressione lineare in cui la risposta coincide con ogni *i*-esima variabile mentre i regressori consistono di volta in volta nelle *V-i* variabili ausiliarie rimanenti. Per ogni regressore la funzione *fitglm* restituisce il relativo coefficiente di regressione, mentre il relativo *P*-value basato sull'F-test è calcolato tramite la funzione *coefTest* (Matlab®, v.2015b). Quando il *P*-value di un regressore rispetto auna certa variabile risposta è minore di 0.05 (valore di soglia considerato in questo studio) allora la correlazione tra le due variabili è considerata rilevante e viene quindi considerato il legame esistente tra le due variabili. In caso contrario le due variabili vengono considerate prive di legame. Di seguito si riportano i coefficienti di regressione lineare stimati per ogni variabile ausiliaria e le relative statistiche di qualità del fitting, quali: l'errore standard (*Standard Error*, SE) e il *P*-value.

Tabella 5.2 *Coefficienti di regressione lineare generalizzati e statistiche di fitting per la variabile* x_1.

	Coefficiente	**SE**	***P*-value**
x_2	$-7.48\cdot10^4$	$4.68\cdot10^4$	0.87
x_3	0	0	--
x_4	$1.70\cdot10^2$	$1.44\cdot10^2$	0.25
x_5	$2.09\cdot10$	$4.68\cdot10^2$	0.96
x_6	$-3.41\cdot10$	$2.29\cdot10^2$	0.88

Tabella 5.3 *Coefficienti di regressione lineare generalizzati e statistiche di fitting per la variabile x_2.*

	Coefficiente	**SE**	***P*-value**
x_1	$8.90 \cdot 10^{-18}$	$2.04 \cdot 10^{-17}$	0.66
x_3	-1.59	$6.83 \cdot 10^{-11}$	$2.51 \cdot 10^{-186}$
x_4	$2.89 \cdot 10^{-15}$	$1.36 \cdot 10^{-14}$	0.83
x_5	0.68	$2.90 \cdot 10^{-11}$	$2.44 \cdot 10^{-186}$
x_6	-0.02	$8.50 \cdot 10^{-13}$	$1.58 \cdot 10^{-186}$

Tabella 5.4 *Coefficienti di regressione lineare generalizzati e statistiche di fitting per la variabile x_3.*

	Coefficiente	**SE**	***P*-value**
x_1	$-8.18 \cdot 10^{-20}$	$1.54 \cdot 10^{-17}$	0.99
x_2	-0.62	$3.24 \cdot 10^{-11}$	$9.50 \cdot 10^{-185}$
x_4	$4.36 \cdot 10^{-15}$	$1.03 \cdot 10^{-14}$	0.67
x_5	0.42	$3.25 \cdot 10^{-14}$	$1.52 \cdot 10^{-238}$
x_6	-0.01	$1.58 \cdot 10^{-14}$	$1.64 \cdot 10^{-215}$

Tabella 5.5 *Coefficienti di regressione lineare generalizzati e statistiche di fitting per la variabile x_4.*

	Coefficiente	**SE**	***P*-value**
x_1	$3.83 \cdot 10^{-4}$	$3.23 \cdot 10^{-4}$	0.25
x_2	$-4.86 \cdot 10^{2}$	$6.93 \cdot 10^{2}$	0.49
x_3	0	0	--
x_5	0.87	0.67	0.21
x_6	-0.24	0.33	0.47

Tabella 5.6 *Coefficienti di regressione lineare generalizzati e statistiche di fitting per la variabile x_5.*

	Coefficiente	**SE**	***P*-value**
x_1	$5.30 \cdot 10^{-18}$	$3.14 \cdot 10^{-17}$	0.86
x_2	1.47	$6.57 \cdot 10^{-11}$	$5.60 \cdot 10^{-186}$
x_3	2.35	$1.54 \cdot 10^{-13}$	$9.22 \cdot 10^{-240}$
x_4	$2.74 \cdot 10^{-15}$	$2.09 \cdot 10^{-14}$	0.89
x_6	0.03	$3.02 \cdot 10^{-14}$	$3.01 \cdot 10^{-217}$

Tabella 5.7 *Coefficienti di regressione lineare generalizzati e statistiche di fitting per la variabile x_6.*

	Coefficiente	**SE**	***P*-value**
Intercetta	$-2.67\cdot10^{-15}$	$5.65\cdot10^{-15}$	0.64
x_1	$-6.12\cdot10^{-16}$	$7.51\cdot10^{-16}$	0.42
x_2	$-4.91\cdot10$	$1.53\cdot10^{-9}$	$6.54\cdot10^{-189}$
x_3	$-7.83\cdot10$	$6.04\cdot10^{-11}$	$1.79\cdot10^{-219}$
x_4	$5.10\cdot10^{-13}$	$5.01\cdot10^{-13}$	0.32
x_5	$3.33\cdot10$	$2.41\cdot10^{-11}$	$5.44\cdot10^{-220}$

La ricostruzione grafica della rete di legami risultante dai legami individuati a posteriori con questa tecnica sono riportati in Figura 5.9.

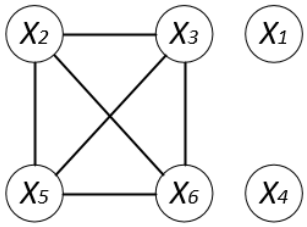


Figura 5.9 *Modello di regressione lineare generalizzato.*

Confrontando la rete di causalità ottenuta in Figura 5.9 con la rete di causalità ottenuta tramite la tecnica descritta in §5.1.2.1 possono rilevare delle somiglianze con la rete di causalità ottenuta dall'analisi dei coefficienti di correlazione di grado 0 di Figura 5.2*a*, se non per l'assenza di legami tra le variabili x_4 e x_3.

5.1.2.6 Analisi puntuale dei PCC in seguito all'applicazione del metodo di decorrelazione 2

Per valutare le prestazioni del metodo di decorrelazione 2 si procede con l'analisi dettagliata dei coefficienti di correlazione di grado 1 per il modello e per il processo, riportati in
Figura 5.10. In Tabella 5.1 sono riportate le corrispondenze tra il numero del coefficiente di correlazione riportato in Figura 5.10 e le relative variabili rappresentate.
Dal confronto dei valori medi dei coefficienti di correlazione prima dell'applicazione del metodo di decorrelazione 2 (Figura 5.4) e dopo (Figura 5.10) si può osservare una notevole decorrelazione per la maggior parte delle variabili. Infatti il valore medio di tali coefficienti non supera il valore di 0.5 per nessuna terna di variabili, risultato significativamente migliore rispetto a quello ottenuto tramite la ricostruzione della rete di causalità. In particolar modo le variabili x_1 e x_4 risultano completamente decorrelate. Mentre per le terne che riportano le variabili x_2 e x_3 come variabili di coppia, si nota un un aumento del valore medio.

Il secondo metodo di decorrelazione comporta quindi una migliore decorrelazione delle variabili rispetto a quanto ottenuto tramite la ricostruzione della rete di causalità tuttavia non viene raggiunta una completa decorrelazione.

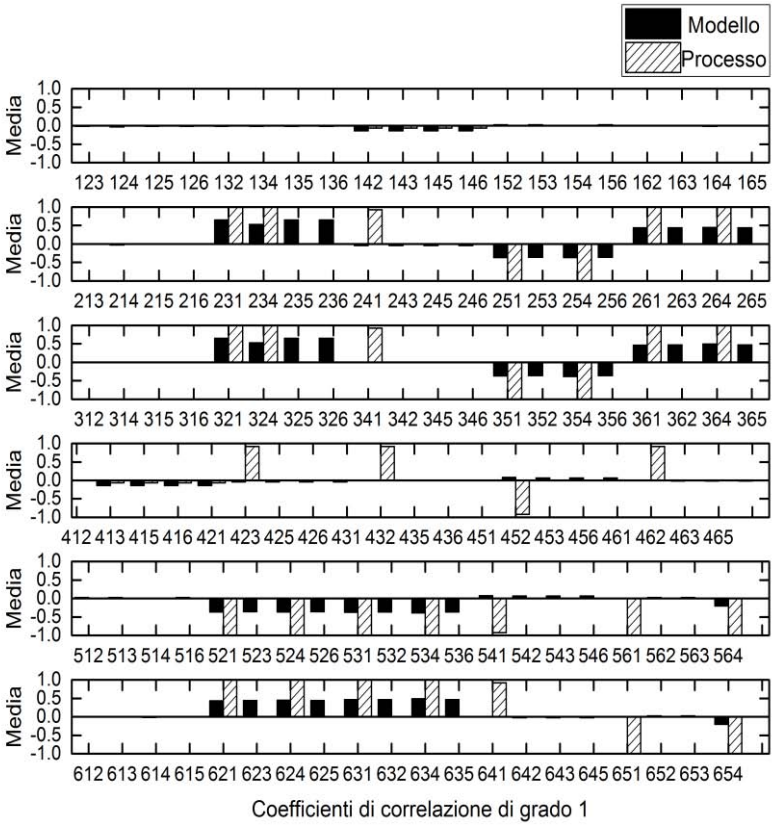


Figura 5.10 *Media dei coefficienti di correlazione di grado 1 dopo l'applicazione della procedura decorrelativa.*

5.1.2.7 Applicazione della procedura diagnostica in seguito all'applicazione del metodo di decorrelazione 2

In Figura 5.11 si riportano i risultati della procedura di classificazione della variabile principalmente responsabile nel PMM ottenuta secondo quanto descritto in §4.2.1, considerando le variabili ausiliarie x_1, x_2, x_3, x_4, x_5, x_6 a ottenute in seguito all'applicazione del secondo metodo di decorrelazione.

Il risultato della procedura diagnostica riportato in Figura 5.11 mostra che per il 98% delle ripetizioni alle variabili x_2, x_3, x_5, x_6 è assegnata la classificazione con indice 'ROSSO' e per il 100% delle ripetizioni per la variabile x_1 una classificazione con indice 'ARANCIONE'.

Tali classificazioni evidenziano come un miglioramento parziale nella decorrelazione non comporta un miglioramento nella diagnostica del PMM. Confrontando Figura 5.11 con Figura 5.8 relativa alla diagnostica successiva alla decorrelazione tramite ricostruzione della rete di causalità, si rileva infatti un peggioramento delle prestazioni dato che anche la variabile x_6 non direttamente correlata al parametro $Y_{p/s}$ viene classificata con indice 'ROSSO' e la variabile x_1

anch'essa non direttamente correlata a tale parametro viene classificata con indice 'ARANCIONE'.

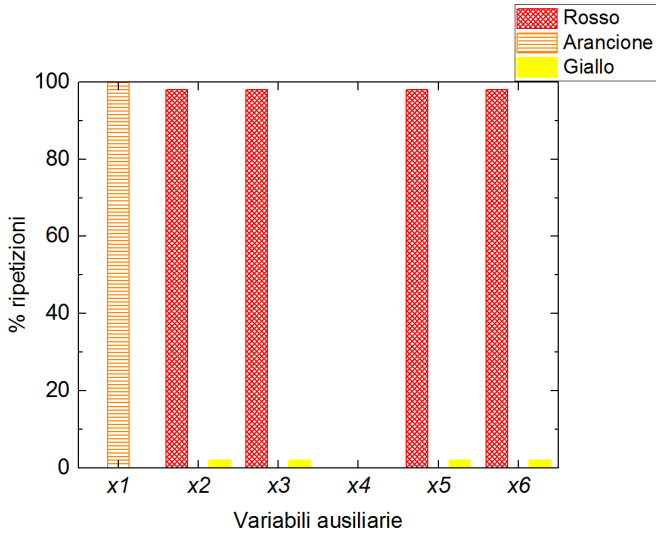


Figura 5.11 *Percentuale di ripetizioni in cui ogni variabile è stata classificata con indice 'ROSSO', 'ARANCIONE' o 'GIALLO'.*

5.2 Caso studio 3: errore introdotto sul parametro $Y_{x/s}$

In questo secondo esempio, viene forzata la presenza di un *mismatch* introducendo un errore sul coefficiente di resa di produzione di biomassa su consumo di substrato $Y_{p/s}$.
Come nel caso studio precedente il parametro presenta delle forti correlazioni tra il termine del modello contenete l'errore e le altre variabili ausiliarie considerate ma in questo caso l'analisi dell'indice MRLR (Meneghetti *et al.*, 2014) si è dimostrata efficace nell'identificare il responsabile del PMM (§3.5). La procedura di diagnosi in esame viene testata considerando il set di variabili ausiliarie (5.1). Vengono quindi applicate le tecniche di pretrattamento dei dati per migliorare l'efficacia diagnostica della procedura sviluppate in 5.1.2 e ne viene indagata l'efficienza decorrelativa.

5.2.1 Applicazione della procedura diagnostica

Anche in questo secondo caso studio viene definito un sottoinsieme di dati, $\underline{\mathbf{X}}_{Bs}$ [$G \times K \times B$], per la matrice di modello in cui è stato introdotto un errore sul parametro $Y_{x/s}$ selezionando G = 25 campioni con concentrazione finale di penicillina media pari a 0.7 (g/l) e deviazione standard inferiore a 0.01 (g/l). Tramite le definizioni riportate in (5.1) si procede alla costruzione delle matrici tridimensionali $\underline{\mathbf{X}}_{B\Pi sv}$ e $\underline{\mathbf{X}}_{BMsv}$ rispettivamente matrice di processo e matrice di modello in variabili ausiliarie di dimensioni [$G \times V \times B$] dove G = 25 sono i campioni mentre V = 6 le variabili ausiliarie e B = 100 le ripetizioni.

Su tali matrici si esegue la procedura diagnostica descritta in §4.2.1. di cui si riportano i risultati in Figura 5.12. In questo caso nessuna delle variabili indagate mostra una classificazione con indice 'ROSSO' necessaria per determinare la variabile maggiormente responsabile del PMM la quale si ricorda coincide con la variabile ausiliaria x_5. Tuttavia la prima e la seconda variabile presentano una classificazione con indice 'ARANCIONE' per il 10 e 90% delle ripetizioni. Le variabili x_4 e x_5, invece, presentano una classificazione con indice 'GIALLO' per il 70 e 50% delle ripetizioni. Questa classificazione, sebbene non permetta di identificare la variabile responsabile del PMM, permette di individuare una delle variabili maggiormente influenzate da quest'ultimo. Infatti come possiamo osservare dalla definizione delle variabili ausiliarie (5.1) la variabile x_2 moltiplica il parametro su cui è stato introdotto l'errore ossia $Y_{x/s}$. Tale risultato è stato ottenuto anche da Ibrahim (2016) conducendo la medesima analisi di identificazione della variabile responsabile del PMM sullo stesso caso studio qui indagato.
Confrontando questo risultato con quanto ottenuto in §5.1.1 per la diagnosi dell'errore sul parametro $Y_{p/s}$ si possono rilevare delle somiglianze, poiché le variabili indicate come principali responsabili del PMM siano le medesime, ovvero le variabili x_2 e x_4. Questo risultato mette in evidenza, quindi, l'incapacità della procedura diagnostica implementata, di discriminare errori incidenti, in modo analogo ma non uguale sulla stessa variabile ausiliaria.

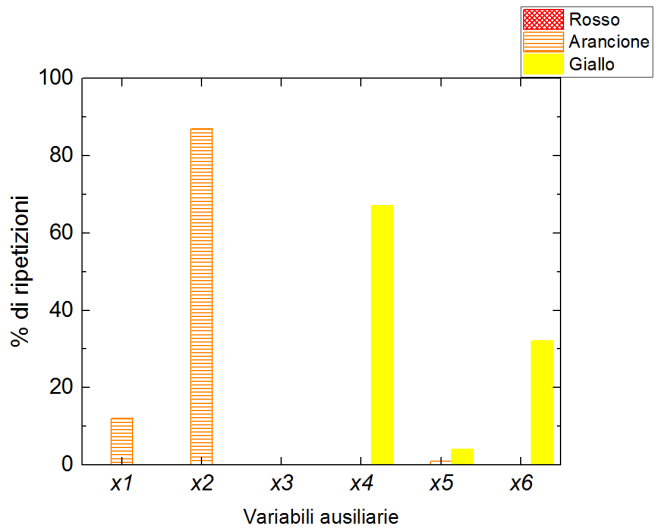


Figura 5.12 *Percentuale di ripetizioni in cui ogni variabile è stata classificata con indice 'ROSSO', 'ARANCIONE' o 'GIALLO'.*

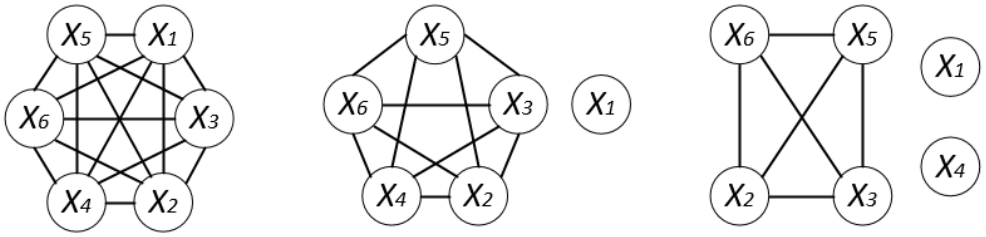
5.2.2 Decorrelazione delle variabili

Seguendo la procedura di analisi utilizzata per il primo esempio riportato in questo Capitolo, sono state valutate le prestazioni della procedura decorrelativa descritta in §4.2.1 in questo nuovo caso studio e gli eventuali miglioramenti ottenuti in fase diagnostica.

5.2.2.1 Metodo di decorrelazione 1

La ricostruzione della rete di causalità viene eseguita per ogni ripetizione della matrice $\underline{\mathbf{X}}_{Bsv}$ in cui le variabili prese in esame sono le sei variabili ausiliarie x_1, x_2, x_3, x_4, x_5, x_6.

In Figura 5.11 sono riportate in sequenza le reti di causalità ottenute al termine della procedura di ricostruzione (descritta in §4.2.1) in cui si sono considerati rispettivamente i coefficienti di correlazione di grado zero (Figura 5.13*a*), di grado 1 (Figura 5.13*b*) e di grado 2 (Figura 5.13*c*). La rete di legami risultante collega le variabili x_2, x_3, x_5 e x_6, mentre le variabili x_1 e x_4 risultano completamente indipendenti dalle altre.

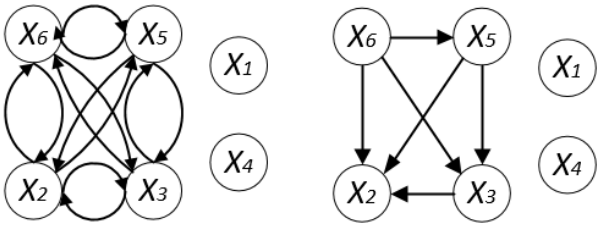


(a) (b) (c)

Figura 5.13 *Rete di legami risultante ottenuta da:* (a) *coefficienti di correlazione di grado 0,* (b) *coefficienti di correlazione di grado 1 e* (c) *coefficienti di correlazione di grado 2.*

La rete di legami individuata risulta consistente con le definizioni delle variabili ausiliarie in (5.1). Si osserva infatti come le variabili x_2 e x_3 che nelle equazioni (5.1) siano incluse nella definizione delle variabili x_5 e x_6 risultino legate nella rete di Figura 5.11a.

In Figura 5.14 si riportano le reti di causalità per la stessa ripetizione della matrice di modello $\underline{\mathbf{X}}_{Bsv}$ considerata in Figura 5.11. Nello specifico in Figura 5.12*a* si riporta la rete di causalità ottenuta tramite l'applicazione del primo stadio della procedura di identificazione dei versi dei legami, mentre in Figura 5.12*b* è riportata la rete di causalità dopo l'esecuzione del secondo stadio della stessa procedura.



(a) (b)

Figura 5.14 *Rete di causalità risultante per le variabili ausiliarie della matrice* $\underline{\mathbf{X}}_{Bsv}$*:* (a) *ricostruzione in seguito al primo stadio della procedura*, (b) *in seguito al secondo stadio della procedura.*

La procedura di ricostruzione delle reti di causalità viene eseguita per ogni ripetizione della matrice $\underline{\mathbf{X}}_{Bsv}$ le quali, si ricorda, differiscono le une dalle altre solo per del rumore bianco. Le relazioni identificate dalla metodologia proposta sono le stesse per tutte le ripetizioni della matrice $\underline{\mathbf{X}}_{Bsv}$. Tale risultato dimostra che la rete di causalità ricostruita è una buona rappresentazione nella rete di legami media della matrice $\underline{\mathbf{X}}_{Bsv}$.
Confrontando Figura 5.14*b* con Figura 5.3*b* relativa alla ricostruzione della rete di legami eseguita sulla matrice di modello in cui è stato forzato un errore sul parametro $Y_{p/s}$ si evidenzia come le strutture determinate differiscano notevolmente tra loro. Questo risultato mette in luce le modifiche strutturali apportate alla struttura di correlazione delle variabili da parte di *mismatch* causati da diversi errori, come nel caso dei parametri di resa $Y_{p/s}$ e $Y_{x/s.}$
Al termine di questa fase si procede con la regressione multilineare di ogni variabile figlio sulle variabili genitori definite dalla rete di causalità media in modo tale da ottenere il modello di regressione con cui procedere alla decorrelazione delle variabili, secondo quanto descritto in Figura 4.12. L'operazione viene eseguita su ogni ripetizione della matrice $\underline{\mathbf{X}}_{Bsv}$.

5.2.2.2 Analisi dei PCC in seguito all'applicazione del metodo di decorrelazione 1

Con l'obbiettivo di valutare il livello di decorrelazione ottenuto a seguito della regressione di ogni variabile ausiliaria rispetto alle variabili identificate come 'genitori', viene effettuata un'analisi dettagliata dei coefficienti di correlazione di grado 1 per il modello e per il processo. A tal scopo in Figura 5.15, 5.16 sono riportate i valori medi delle distribuzioni per i coefficienti di correlazione di grado 1, di cui in Tabella 4.3, sono riportate le corrispondenze con le relative variabili rappresentate.
Dal confronto tra i coefficienti di correlazione per il modello per le sei variabili in esame prima dell'applicazione della procedura decorrelativa (Figura 5.15) e dopo l'applicazione della procedura (Figura 5.16) si osserva che la decorrelazione è stata ottenuta solo per alcune coppie di variabili considerate. In particolare, si osserva una netta riduzione dei valori medi dei coefficienti di correlazione in cui compaiono le variabili x_2 e x_3 come variabili di coppia ad indicazione di una buona decorrelazione di tali variabile. Per i restanti coefficienti di correlazione invece, i valori medi si assestano tra ±0.5 e ±1.
Anche in questo caso studio quindi la procedura decorrelativa risulta efficace solo marginalmente nel caso si considerino le variabili ausiliarie. Confrontando Figura 5.16 con Figura 5.5 (relativa ai coefficienti di correlazione in seguito all'applicazione della procedura decorrelativa sul caso studio $Y_{p/s}$) si osserva che la decorrelazione delle variabili ausiliarie in questo caso studio risulta maggiore di quella ottenuta per il caso studio precedente. Questo risultato evidenzia la sensibilità della tecnica decorrelativa ai dataset considerati.

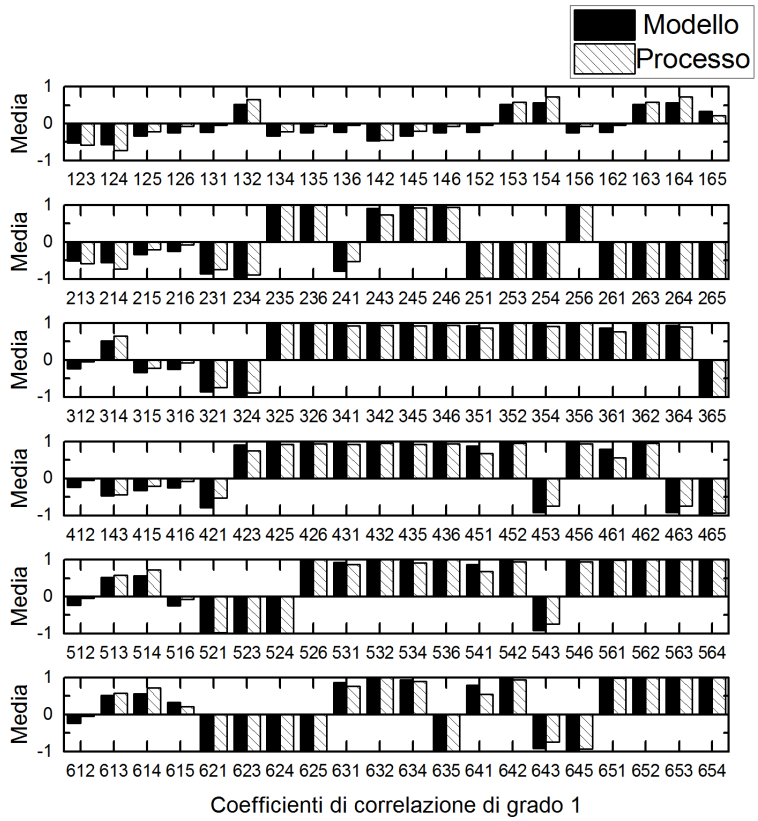


Figura 5.15 *Media dei coefficienti di correlazione di grado 1 prima dell'applicazione della procedura decorrelativa.*

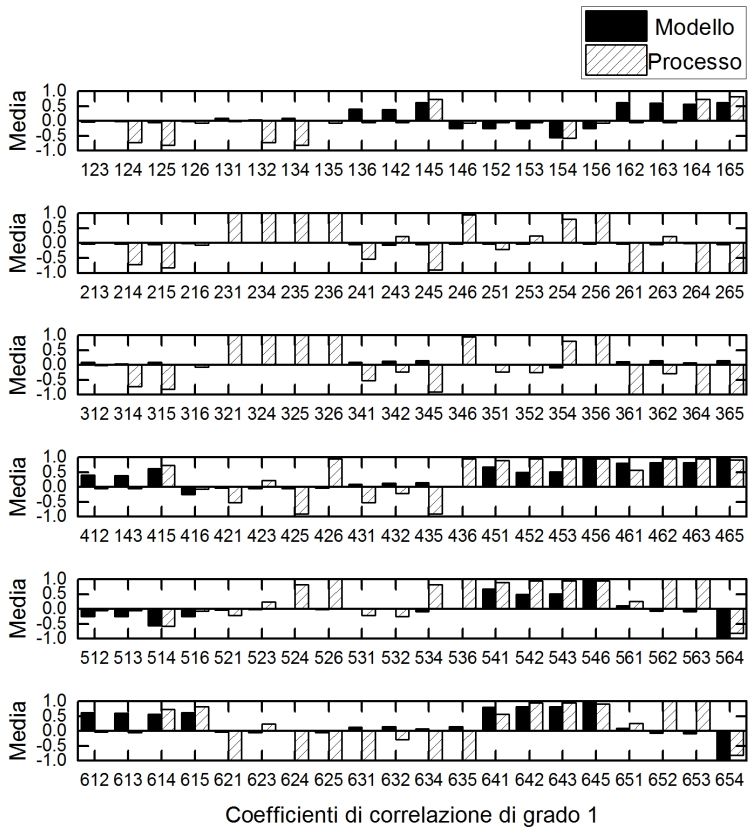


Figura 5.16 *Media dei coefficienti di correlazione di grado 1 dopo l'applicazione della procedura decorrelativa.*

5.2.2.3 Applicazione della procedura diagnostica in seguito all'applicazione del metodo di decorrelazione 1

In Figura 5.17 si riportano i risultati della procedura di classificazione per la diagnosi della variabile principalmente responsabile del PMM ottenuta secondo quanto descritto in §4.2.1. Le variabili in esame sono le variabili ausiliarie x_1, x_2, x_3, x_4, x_5, x_6 a seguito all'applicazione della procedura decorrelativa via ricostruzione della rete di causalità.

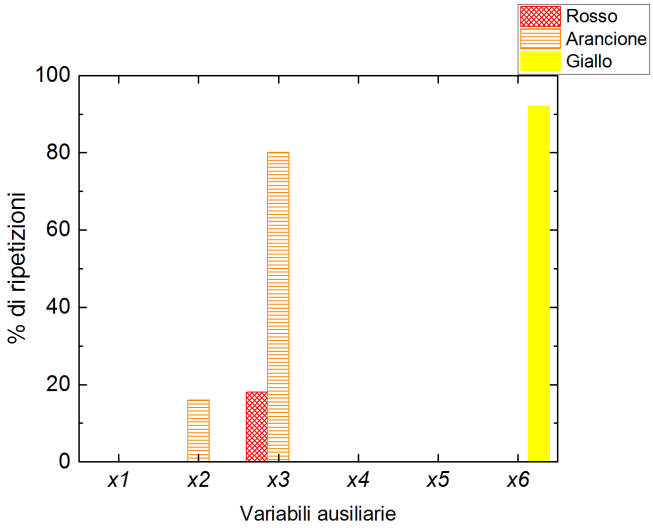


Figura 5.17 *Percentuale di ripetizioni in cui ogni variabile è stata classificata con indice 'ROSSO', 'ARANCIONE' o 'GIALLO'.*

L'applicazione della procedura decorrelativa non porta in effetti a nessun evidente miglioramento dei risultati, dato che la variabile x_5 in cui è contenuto il parametro $Y_{x/s}$, non è mai riconosciuta come variabile in relazione al PMM.

Si riporta invece una classificazione con indice 'ROSSO' per il 20% delle ripetizioni e con indice 'ARANCIONE' per l'80% delle ripetizioni per la variabile x_3, una classificazione del 20% 'ARANCIONE' per la variabile x_2 e una classificazione del 90% 'GIALLO' per la variabile x_6. Queste variabili sono correlate al parametro $Y_{x/s}$ ma non lo contengono.

5.2.2.4 Metodo di decorrelazione 2

La procedura decorrelativa basata sulla ricostruzione della rete di causalità non si è dimostrata efficace nella decorrelazione completa delle variabili (§5.2.2). Viene quindi nuovamente implementato il metodo di decorrelazione 2 descritto in §5.1.2.5. I modelli di regressione nonché i risultati in termini di decorrelazione ottenuti sono analoghi ai risultati ottenuti per il caso studio precedente (§5.1.2.6). Tali risultati dimostrano quindi sia l'inefficacia di questa metodologia nella decorrelazione delle variabili sia la bassa sensibilità nell'identificare variazioni della rete di correlazione indagata a seconda del *mismatch* considerato.

5.3 Set di variabili ausiliarie alternativi

In questo Capitolo allo scopo di determinare i termini del modello maggiormente responsabili del PMM è stata implementata la procedura diagnostica considerando un set di variabili ausiliarie. Tuttavia come riportato in §5.1 e §5.2 in seguito all'introduzione di tali variabili la procedura diagnostica non si è rivelata in grado di determinare la variabile responsabile del PMM al contrario di quanto ottenuto in §4 considerando le variabili originali. Tale insuccesso diagnostico si ritiene sia dovuto alla particolare selezione di variabili ausiliarie considerate. Per tanto, in mancanza della conoscenza della reale causa del *mismatch*, è necessario definire uno o più indici in grado guidare la selezione di un adeguato set di variabili ausiliarie.
Nello specifico si considerano cinque set di variabili ausiliarie ottenuti tramite cinque diverse combinazioni dei termini di modello di Birol *et al.*, (2002) presenti nelle equazioni: (2.5), (2.7) e (2.8). Per ogni set di variabili ausiliarie si esegue la decorrelazione delle variabili tramite ricostruzione della rete di causalità (4.3.2§). Per valutare quale set di variabili permetta di ottenere la migliore decorrelazione, è stato calcolato un indice di correlazione dato dalla somma delle medie quadratiche dei coefficienti di correlazione di grado 1 per ogni variabile del set di variabili ausiliarie calcolate per la matrice di modello.
In (5.3) si riporta il set di variabili ausiliarie per il quale è stato calcolato l'indice di correlazione più basso. Confrontando questo set di variabili ausiliarie con il set (5.1) si può osservare come in questo caso non siano presenti variabili incluse nella definizione di altre variabili impedendo in questo modo la generazione di relazioni fortemente non lineari tra i termini del modello. Rimuovendo tali variabili però, non è possibile differenziare i casi in cui i parametri modificati per causare il PMM sono contenuti nella stessa variabile, come nel caso di $Y_{p/s}$ e $Y_{x/s}$.

$$x_1 = K_l a(C_L^{\ *} - C_L) \qquad x_3 = \frac{\mu C_x}{Y_{x/o}} + \frac{\mu_{pp} C_x}{Y_{p/o}} + C_x m_o$$
$$x_2 = \frac{\mu C_x}{Y_{x/s}} + \frac{\mu_{pp} C_x}{Y_{p/s}} + m_x C_x \qquad x_4 = K C_p \tag{5.3}$$

In Figura 5.18 si riportano i risultati della procedura di classificazione per la diagnosi della variabile principalmente responsabile del PMM ottenuta secondo quanto descritto in §4.2.1. Le variabili in esame sono le variabili ausiliarie x_1, x_2, x_3, x_4 definite in (5.3) a seguito all'applicazione della procedura decorrelativa via ricostruzione della rete di causalità (§4.3.2). Come sé possibile dalla Figura 5.18 si riporta una classificazione con indice 'ROSSO' per il 100% delle ripetizioni per la variabile x_2 e una classificazione con indice 'ARANCIONE' per il 95% delle ripetizioni per la variabile x_4.

Questo risultato comporta un notevole miglioramento della diagnosi rispetto a quanto ottenuto con il set di variabili ausiliarie (5.1) riportato in §5.1.1 in cui la variabile responsabile del PMM, x_5, veniva classificata con indice 'GIALLO' per il 5% delle ripetizioni.
Si osserva infatti che la variabile ausiliaria responsabile del PMM, x_2, vieni identificata come tale dalla procedura diagnostica. Questo risultato conferma quindi il ruolo di primo piano rivestito dalla definizione delle variabili ausiliarie nel permettere una sufficiente decorrelazione necessaria ad una diagnosi efficace del PMM.

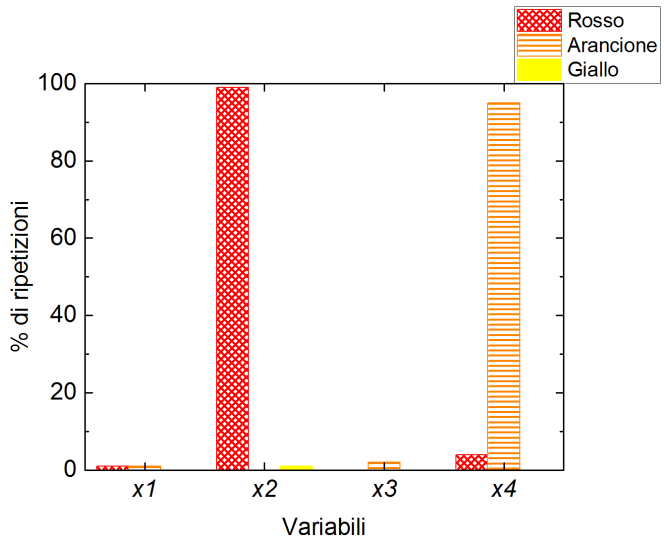


Figura 5.18 *Percentuale di ripetizioni in cui ogni variabile è stata classificata con indice 'ROSSO', 'ARANCIONE' o 'GIALLO'.*

5.4 Conclusioni

L'analisi delle variabili ausiliarie per identificare i termini del modello causa di *mismatch*, ovvero i parametri di resa $Y_{p/s}$ e $Y_{x/s}$, a causa delle forti correlazioni presenti in entrambi i casi tra le variabili coinvolte, ha dimostrato la necessità di implementare il pretrattamento dei dati tramite decorrelazione delle variabili. Tuttavia, il primo metodo di decorrelazione adottato, basato sulla ricostruzione della rete di causalità, non risulta efficace. Infatti sebbene vengano identificate variabili correlate al *mismatch* non si è in grado di identificare la variabile effettivamente responsabile del PMM, probabilmente a causa delle relazioni fortemente non lineari che si instaurano in conseguenza alla definizione delle variabili ausiliarie.
L'applicazione di una procedura decorrelativa alternativa, basato sulla ricostruzione della rete di legami successiva alla regressione di un modello lineare per ogni variabile, permette una maggior decorrelazione delle variabili di quanto ottenuto tramite la tecnica basata sulla ricostruzione della rete di causalità. Tuttavia, anche in questo caso, le correlazioni parziali che non vengono rimosse non permettono un'adeguata diagnosi del modello in esame. Infine

l’analisi di diversi set di variabili ausiliarie ha identificato nella definizione delle variabili ausiliarie il principale limite di generalizzazione della procedura diagnostica, nel caso di variabili fortemente correlate tra loro. Si evidenzia infatti come un primo tentativo di selezione di un set di variabili ausiliarie alternativo in cui non siano presenti variabili incluse nella definizione di altre variabili impedendo in questo modo la generazione di relazioni fortemente non lineari tra i termini del modello consenta una diagnosi efficace del PMM.

Conclusioni

In questa Tesi sono state valutate delle possibili soluzioni per la diagnosi di un modello a principi primi per il quale è stato rilevato un disallineamento (*process/model mismatch*, PMM) tra i dati storici del processo descritto e le corrispondenti stime calcolate tramite il modello stesso. A tal scopo sono state analizzate le prestazioni e le limitazioni di due diverse metodologie sviluppate per accelerare i tempi solitamente necessari al miglioramento di un modello a principi primi. In particolare, le metodologie indagate sono: la metodologia sviluppata da Meneghetti *et al.* (2014), basata sull'analisi delle componenti principali (PCA), e la metodologia adattata dalla tecnica sviluppata da Rato e Reis (2015) per il monitoraggio di processo, basata sull'analisi di coefficienti di correlazione parziale.

Entrambe le tecniche diagnostiche sono state impiegate per l'identificazione di *mismatch* parametrici opportunamente generati grazie all'utilizzo di un simulatore di processo sviluppato per descrivere un processo a due stadi (batch e semi-batch) di fermentazione per la produzione di penicillina. Utilizzando tale simulatore infatti, è stato generato un set di dati che rappresenta le misure 'storiche' del processo, e tre set di dati generati utilizzando il medesimo modello, ma forzando la presenza di un PMM tramite la modifica di tre diversi parametri presenti nel modello. In tutti i dataset sono stati consideranti solo i valori di fine batch.

I tre parametri in cui è stato introdotto un errore sono rispettivamente: il coefficiente volumetrico di trasporto di massa dell'ossigeno $K_l a$, il coefficiente di resa di produzione di penicillina su consumo di substrato $Y_{p/s}$ e quello di resa di produzione di biomassa su consumo di substrato $Y_{x/s}$. Il primo errore incide principalmente su una variabile debolmente correlata con le altre variabili del modello. Il secondo e il terzo, invece, incidono su una variabile fortemente correlata alle altre e sono presenti sulla medesima equazione di modello.

Per agevolare l'analisi delle possibili cause del *mismatch* osservato, è utile verificare se l'errore presente abbia modificato in modo marcato la distribuzione dei dati in uscita dal modello rispetto a quella dei dati di processo. Nel caso in cui l'errore incida su relazioni altamente non lineari, è infatti possibile che i dati in uscita dal modello presentino dei raggruppamenti o delle distribuzioni molto più ampie rispetto ai dati di processo. Quando questo accade, si suggerisce di limitare il set di dati analizzato ad un gruppo ristretto per il quale l'effetto del PMM possa essere considerato comparabile, come effettuato in questa Tesi per entrambi i casi studio in cui è stato modificato il parametro di resa.

Nella prima parte della Tesi sono state valutate le prestazioni e i limiti della metodologia sviluppata da Meneghetti *et al.* (2014) per l'identificazione del PMM. Tale analisi, si basa sull'analisi tramite un modello PCA (analisi delle componenti principali) delle cosiddette variabili ausiliarie, ossia delle variabili ottenute dalla combinazione non lineare di variabili e

parametri del modello in modo tale da permettere l'identificazione del termine del modello responsabile del PMM. I risultati ottenuti si differenziano notevolmente a seconda del PMM considerato. Infatti nel caso dell'errore sul parametro $K_l a$ è possibile identificare correttamente il termine del modello responsabile del *mismatch*. Nel caso in cui gli errori incidano sui parametri $Y_{p/s}$ e $Y_{x/s}$, invece, il risultato non è soddisfacente, in quanto sebbene venga identificato un gruppo di termini direttamente influenzato dalla presenza del *mismatch*, la procedura non è in grado di individuare in modo chiaro i termini che contengono i parametri $Y_{p/s}$ e $Y_{x/s}$ e che ne sono quindi responsabili.

Nella seconda parte della Tesi invece, è stata testata la seconda metodologia adattata dalla procedura proposta da Rato e Reis (2015) per il monitoraggio di processi stazionari. Tale procedura è stata inizialmente applicata ampliando in modo opportuno il set di dati disponibile (ripetendo diversi batch le cui uscite differiscono solo a causa di rumore bianco introdotto artificialmente nel simulatore) e considerando le principali variabili di processo invece di analizzare variabili ausiliarie, ovvero: concentrazione di substrato, di ossigeno, di biomassa e di penicillina. Tali scelte sono state inizialmente necessarie per testare la metodologia nelle medesime condizioni di analisi che hanno permesso a Rato e Reis (2015) di ottenere ottimi risultati nelle applicazioni da loro riportate. La metodologia basata sull'identificazione dell'alterazione della struttura di correlazione del modello in seguito all'incidenza del PMM grazie alla comparazione dei coefficienti di correlazione parziale calcolati per un set di variabili di processo e di modello, viene applicata per identificare gli errori introdotti sui parametri $K_l a$, $Y_{p/s}$ e $Y_{x/s}$. In questo caso l'analisi dell'errore introdotto su $K_l a$, ha permesso di dimostrare che per la corretta identificazione della causa del *mismatch*, quando questo è poco marcato, è necessario verificare che per i dati considerati l'effetto del *mismatch* non sia comparabile con l'effetto dovuto al rumore di cui sono affette le variabili analizzate. La causa del secondo *mismatch* invece, (ovvero $Y_{p/s}$) incidente su una variabile fortemente correlata con le altre variabili del modello, non viene identificata correttamente dalla metodologia diagnostica. Si evidenzia quindi come anche questa procedura risenta sensibilmente delle correlazioni tra le variabili indagate. Una possibile soluzione a tale problema prevede l'implementazione di una procedura decorrelativa basata sulla ricostruzione della rete di legami causali tra le variabili analizzate (Rato e Reis, 2014). L'obiettivo della tecnica impiegata è di rompere i legami primari tra le variabili, in modo da ottenere dei nuovi coefficienti di correlazione con un valore molto basso, prossimo a zero e permettere così un efficace rilevazione di modifiche alla struttura di correlazione e quindi una corretta diagnosi del PMM. Grazie all'applicazione della procedura decorrelativa, è stato possibile individuare anche per il secondo caso in esame la variabile direttamente influenzata dall'errore introdotto nel modello. In base ai buoni risultati ottenuti, la stessa procedura è stata implementata considerando l'analisi delle
variabili ausiliarie per la diagnosi del secondo e del terzo *mismatch* parametrico (dovuti agli errori sui parametri $Y_{p/s}$ e $Y_{x/s}$). L'obbiettivo primario infatti, è di fornire una procedura che non

sia solo in grado di identificare la variabile maggiormente influenzata dal *mismatch* ma di identificare il termine del modello responsabile del *mismatch*. In questo caso le procedure sviluppate non si sono però rivelate efficaci nel decorrelare le variabili ausiliarie analizzate e di conseguenza nell'indentificare il termine responsabile del PMM. Si ritiene infatti che la presenza di relazioni fortemente non lineari tra i termini del modello considerati nelle variabili ausiliarie, non permetta la completa decorrelazione di tali variabili, dato che il modello di regressione utilizzato per individuare la rete di legami su cui si basa la decorrelazione delle variabili è un modello lineare, che probabilmente non è in grado di rappresentare correttamente le relazioni tra i termini del modello. Tale considerazione è supportata anche dai risultati ottenuti utilizzando una tecnica decorrelativa alternativa che prevede di ricostruire un modello di regressione lineare direttamente per ogni variabile considerata e identificare a posteriori tramite i valori dei *P*-value di ogni coefficiente calcolato la rete di legami tra le variabili necessaria alla decorrelazione di queste ultime. Infatti, sebbene tale alternativa dimostri migliori prestazioni nel decorrelare le variabili considerate rispetto alla tecnica decorrelativa basata sulla ricostruzione della rete di causalità, tale miglioramento non è sufficiente a riconoscere correttamente il termine maggiormente responsabile del PMM.

Si ritiene che una possibile soluzione a tale problema, che rappresenta anche un aspetto critico dell'applicazione di entrambe le procedure, sia la definizione di un appropriato set di variabili ausiliarie che permetta di ottenere una decorrelazione efficace e di conseguenza la corretta diagnosi del PMM. A questo scopo è stata proposta una procedura basata sulla valutazione del grado di decorrelazione ottenuta in seguito all'applicazione dalla metodologia implementata per diversi set di variabili ausiliarie. Un'analisi preliminare di tale procedura ha effettivamente permesso di identificare un nuovo set di variabili ausiliarie con cui è stato possibile ottenere un netto miglioramento nell'identificazione del termine maggiormente responsabile del PMM. Tuttavia, in possibili lavori futuri, sarebbe necessario valutare non solo il grado di correlazione, ma anche la quantità di informazione che è possibile ottenere considerando diversi set di variabili ausiliarie e considerare non solo i dati di fine batch, ma tutta l'evoluzione del processo in esame.

Nomenclatura

A = numero di componenti principali (-)
a = indicatore generico per il numero di variabili latenti (-)
b = costante empirica del modello di birol *et al.*, (2002) (-)
C = parametro geometrico del reattore e dell'agitatore (-)
C_{CO2} = concentrazione di anidride carbonica (mmol/ l)
C_L = concentrazione di ossigeno disciolto (g/ l)
CO_2 = concentrazione di anidride carbonica (g/ l)
C_p = concentrazione di penicillina (g/ l), capacità termica del mezzo (1/ °C)
C_s = concentrazione di substrato (g/ l)
C_x = concentrazione di biomassa (g/ l)
D = matrice di diagnosi (-)
$d_{k,j}$ = elemento della matrice **D** (-)
E = matrice degli errori statistici multivariati per la matrice **X** (-)
E_d = energia di attivazione per la morte delle cellule (cal/ mol)
E_g = energia di attivazione per la crescita delle cellule (cal/ mol)
$\mathbf{e}_i$ = generico vettore della matrice dei residui **E** (-)
$\mathbf{e}_{n,k}$ = errore di ricostruzione corrispondente a $\hat{x}_{n,k}$ (-)
F = matrice degli errori statistici multivariati per la matrice **Y** (-)
F = portata di alimentazione del substrato (l / h)
F_c = portata acqua di raffreddamento (l/ h)
f_g = portata di aria (l/h)
H^+ = concentrazione ioni idrogeno (mol/ l)
i = indicatore generico di un'osservazione o pedice generico (-)
$\mathbf{I}_i$ = matrice identità (-)
J = Numero di coefficienti di correlazione (-)
k = indicatore generico per le variabili (-)
K = tasso di idrolisi della penicillina (1/h)
K_1 = costante empirica del modello di Birol *et al.*, (2002) (mol/ l)
K_2 = costante empirica del modello di Birol *et al.*, (2002) (mol/ l)
K_c = *controller gain*
k_d = costante di Arrhenius per la morte delle cellule (-)
k_g = costante di Arrhenius per la crescita delle cellule (-),
K_I = costante di inibizione della formazione del prodotto (g /l)
$K_l a$ = coefficiente di trasporto dell'ossigeno (1/ h)
K_{op} = costante di limitazione dell'ossigeno (-)
K_{ox} = costante di limitazione dell'ossigeno (-)
K_p = costante di inibizione (g / l)

K_x = costante di saturazione di Contois (g / l)
M = numero di variabili considerate nella matrice delle risposte (-)
m_o = coefficiente di mantenimento ossigeno (1/h)
m_p = tasso specifico di produzione di penicillina (1/ h)
m_x = coefficiente di mantenimento substrato (1/h)
N = numero di campioni considerati in un set di dati (-)
n = Indicatore generico per i campioni (-)
$\mathbf{P}$ = matrice dei *loadings* di **X** (-)
p = constante empirica del modello di Birol *et al.,* (2002) (-)
$\mathbf{p}_i$ = generico vettore della matrice dei *loadings* **P** (-)
P_w = potenza di agitazione (W)
$\mathbf{Q}$ = matrice dei *loadings* di **Y** (-)
$\mathbf{q}_i$ = generico vettore della matrice dei *loadings* **Q** (-)
Q_{rxn} = generazione di calore (cal)
R^2 = variabilità dei dati originali spiegati dal modello PCA (-)
R^2_{CUM} = variabilità cumulata dei dati originali spiegati dal modello PCA (-)
r_{q1} = resa di generazione di calore (-)
r_{q2} = costante di generazione di calore (cal/ g biomassa h)
$r_{i,j}$ = coefficiente di correlazione di grado 0 tra le variabili *i, j* (-)
$r_{i,j\cdot k}$ = coefficiente di correlazione di grado 1 tra le *i, j, k* (-)
$r_{i,j\cdot k,l}$ = coefficiente di correlazione di grado 2 tra le *i, j, k, l* (-)
S = rango della matrice **X** (-)
s_f = concentrazione del substrato di alimentazione (g / l)
$\mathbf{T}$ = matrice degli *scores* di **X** (-)
T = temperatura (K)
T^2 = statistica di Hotelling (-)
$\mathbf{t}_{CONT,i}$ = generico vettore dei contributi delle variabili alla statistica T^2 del campione (-)
T_f = temperatura di mandata di substrato (K)
$\mathbf{t}_i$ = generico vettore della matrice degli *scores* **T** (-)
$\mathbf{U}$ = matrice degli *scores* di **Y** (-)
V = volume coltura (l)
$\mathbf{W}$ = matrice dei pesi (-)
$\mathbf{w}_i$ = generico vettore della matrice dei *pesi* **W** (-)
$\mathbf{w}_{m,a}$ = peso della variabile *m*-esima sull'*a*-esima LV calcolata dal modello PLS (-)
$\mathbf{w}_{i,j\cdot k}$ = coefficiente di correlazione parziale normalizzato tra le variabili *i, j, k* (-)
$\mathbf{X}$ = matrice bidimensionale delle variabili di processo misurate (-)
$\underline{\mathbf{X}}$ = matrice tridimensionale dei dati (-)
$\underline{\mathbf{X}}_{\text{Bsubset}}$ = matrice tridimensionale del sottoinsieme di dati di *B* simulazioni (-)
x_i = generica variabile ausiliaria
$\hat{\mathbf{X}}$ = matrice ricostruita dal modello PCA (-)
$\hat{x}_{n,k}$ = elemento della *n*-esima riga, *k*-esima colonna ricostruito con modello PCA (-)
$x_{n,k}$ = elemento della *n*-esima riga, *k*-esima colonna di **X** (-)

$\mathbf{x}_i$ = colonna *i*-esima della matrice **X** (-)
$\hat{\mathbf{x}}_i$ = colonna *i*-esima del modello PCA (-)
$\mathbf{X}_i$ = matrice bidimensionale degli ingressi del simulatore (-)
$\mathbf{Y}$ = matrice bidimensionale delle variabili di risposta (-)
$\mathbf{Y}$ = matrice bidimensionale delle variabili di risposta (-)
$Y_{p/o}$ = costante di resa (g penicillina/ g ossigeno)
$Y_{p/s}$ = costante di resa (g penicillina/ g glucosio)
$Y_{x/o}$ = costante di resa (g biomassa/ g ossigeno)
$Y_{x/s}$ = costante di resa (g biomassa/ g glucosio)
$Y_{x/ATP}$ = resa ATP (mol biomassa/ mol di ATP)
$\mathbf{Z}$ = matrice per determinazione della robustezza della procedura diagnostica (-)
$z_{\alpha/2}$ = statistica *z* (-)

Apici

T = matrice trasposta (-)
* = condizione di equilibrio termodinamico (-)

Pedici

M = indice generico dei dati di modello (-)
B = indice generico del numero di simulazioni dell'*n*-esimo campione (-)
T = indice generico di istanti di simulazione dell'*n*-esimo campione (-)
v = indice generico per le variabili ausiliarie (-)
Π = indice generico dei dati di processo (-)

Lettere greche

μ = massima velocità specifica di crescita della biomassa: (mol/ h)
μ_{max} = velocita specifica di crescita massima delle cellule (mol/s)
α = percentuale di confidenza (-)
α_1 = costante della CO_2 per la crescita (mmol CO_2 /g biomassa)
α_2 = costante della CO_2 per l'energia mantenimento (mmol CO_2/g biomassa h)
α_3 = costante della CO_2 relativa alla produzione di penicillina (mmol CO2/ l h)
β = costante in $K_l a$ (-)
γ = costante di proporzionalità: (mol H^+ / g biomassa)
Δ = differenza
θ = angolo tra le direzioni latenti e assi dello spazio delle variabili originali (rad)
λ = autovalore della matrice di covarianza **C** (-)
Λ = matrice degli autovalori (-)
μ_a = viscosità efficace del liquido (Pa s)
ρ = densità (g/ l)
σ = deviazione standard (-)
τ_d = *derivative time constant* (h)
τ_I = *integral time constant* (h)

Acronimi

CL	=	*confidence limits*
DD	=	*data driven*
ESS	=	*sum of square errors*
FP	=	*first principle*
LV	=	*latent variables*
MBDoE	=	*model-based design of experiment*
MRLR	=	*mean residuals-to-limit ratio*
OTR	=	*oxygen transfer rate*
OUR	=	*oxygen uptake rate*
PC	=	*principal component*
PCA	=	*principal component analysis*
PCC	=	*partial correlation coefficient*
PLS	=	*partial least squares regression*
PMM	=	*process/model mismatch*
SE	=	*standard error*
SET	=	*sensitivity enhancing transformations*
SPE	=	*sum prediction error*
TSS	=	*total sum of squares*
VIP	=	*variable influence on projection*

Riferimenti bibliografici

Aiba S., A.E. Humphrey (1973). *Biochemical Engineering*. Academic Press, New York (U.S.A.).

Atkinson, B. e F. Mavituna (1991). *Biochemical Enginnering and Biotechnology Handbook*. Stockton Press, New York (U.S.A.).

Baird M., N. Rama e Z. Shen (1993). Oxygen absorption in a baffled tank agitated by delta paddle impeller. *Can. J. Chem. Eng.*, **71**, 195-201.

Bajpai, R. e M. Reuss (1980). A mechanistic model for penicillin production. Journal of *Chemical Technology and Biotechnology*, **30**, 330-344.

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

Burnham A.J., R. Viveros, e J.F. MacGregor (1996). Frameworks for latent variable multivariate Regression. *J. Chemom.*, **10**, 31-45.

Chandran K. e B.F. Smets (2001). Estimating biomass yield coefficients for autotrophic ammonia and nitrite oxidation from batch respirograms. *Wat. Res*., **35**, 3153-3156.

Chew W. e P. Sharratt (2010). Trends in process analytical technology. *Anal. Methods*, **2**, 1412-1438.

Chong, I.G., e C.H. Jun (2005). Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.*, **78**, 103-112.

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

Clarke K.G., P.C. Williams, M.S. Smit e S.T.L Harrison (2006). Enhancement and repression of the volumetric oxygen transfer coefficient through hydrocarbon addition and its influence on oxygen transfer rate in stirred tank bioreactors. *Biochem. Eng. J.*, **28**, 237-242.

Constantinides, A., J. Spencer e E.J. Gaden (1970). Optimization of batch fermentation processes. Development of mathematical models for batch penicillin fermentations. *Biotechnology and Bioengineering*, **12**, 803.

Doymaz F., J. Chen, J.A. Romagnoli e A. Palazoglu (2001). Robust Strategy for Real-Time Process Monitoring. *J. Process Control*, **11**, 343-359.

Dussap C.G., J. Decorps e J.B. Gros (1985). Transfert d'oxygene en presence de polysaccharides exocellulaires dans un fermenteur agite aeré et dans un fermenteur de type gazosiphon. *Entropie*, **123**, 11-20.

Eriksson L., E. Johansson., N. Kettaneh-Wold e S. Wold (2001). *Multi and Megavariate Data Analysis: Principles and Applications*, Umetrics, Umea.

Franceschini G. e S. Macchietto (2008). Model-Based Design of Experiments for Parameter Precision: State of the Art. *Chem. Eng. Sci*., **63**, 4846-4872.

Fuente A.D.L., N. Bing, I. Hoeschele e P. Mendes (2004). Discovery of meaningful associations in genomic data usingpartial correlation coefficients. *Bioinformatics*, **20**, 3565–3574.

Gabrielsson J., N.O. Lindberg and T. Lundstedt (2002). Multivariate methods in pharmaceutical applications. *J. Chemom*., **16**, 141-160.

Garcia-Ochoa F. e E. Gomez (1998). Mass transfer coefficient in stirrer tank reactors for xanthan solutions. *Biochem. Eng. J.*, **1**, 1-10.

Garcia-Ochoa F., E. Gomez Victoria Santosa e J.C. Merchukb (2009). Oxygen uptake rate in microbial processes: An overview Biotechnology Advances. *Biochem. Eng. J.*, **27**, 153–176

Hawkins D.M. (1993). Regression adjustment for variables in multivariate quality control. *J. Qual. Technol.*, **25**, 170–182.

Hawkins D.M. e E.M. Maboudou-Tchao (2008). Multivariate exponentially weightedmoving covariance matrix, *Technometrics*, **50**, 155–166.

Heijnen J., J. Roels e A. Stouthamer (1979). Application of balancing methods in modeling the penicillin fermentation. *Biotechnology and Bioengineering*, **21**, 2175-2201.

Hermann C, I. Dewes e A. Shumpe (1995). The estimation of gas solubilities in salt solutions. *Chem Eng Sci.*, **50**, 1673–5.

Höskuldsson A. (1988). PLS regression methods. *J. Chemom.*, **2**, 211-228.

Hotelling H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psycol.*, **24**, 417-441.

Huwang L., A.B. Yeh e C.W. Wu, Monitoring multivariate process vari-ability for individual observations, *J. Qual. Technol,*. **39**, 258–278.

Ibrahim A. (2016). *Metodologie per la diagnosi di modelli a principi primi per sistemi dinamici*. Università degli Studi di Padova.

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

Linek V. e V. Vacek (1981). Volumetric mass transfer coefficient in stirred reactors. *Chem. Eng. Technol*.,**11**, 249–51.

Mardia K.V., J.T. Kent e J.M. Bibby (1979). *Multivariate analysis*. Academic Press Limited, London (U.K.).

Marlin T. (1995). *Process Control. McGraw Hill*, New York (U.S.A.).

Marquardt W. (2005). Model-Based Experimental Analysis of Kinetic Phenomena in Multiphase Reactive Systems. *Chem. Eng. Res. Des.*, **83**, 561−573.

Martin E. B. e A.J. Morris (1996). Non-parametric Confidence Bounds for Process Performance Monitoring Charts. *J. Process Control*, **6**, 349−358.

Melissa A.S., K.R. Raghuraj e S. Lakshminarayanan (2009). Partial correlation metric based classifier for food product characterization, *J. Food Eng.*, **90**, 146-152.

Meneghetti N., P. Facco, F. Bezzo e M. Barolo (2014). A Methodology to diagnose Process/Model Mismatch in First Principles Models. *Ind. Eng. Chem. Res.*, **53**, 14002-14013.

Menezes J., S. Alves, Lemos e S. Azevedo (1994). Mathematical modelling of industrial pilot-plant penicillin-G fed-batch fermentastions. *Journal of Chemical Technology and Biotechnology*, **61**, 123-138.

Merchuk J.C e Asenjo J. (1995). *Bioreactor System Design. Fundamentals of Bioreactor Design*. Marcel Dekker Inc., New York (U.S.A.).

Metz B., Bruijin E. e J. van Suijdam (1981). Method for quantitative representation of the morphology of molds. *Biotechnology and Bioengineering*, **23**, 149-162.

Montague G., A. Morris, A. Wright, M. Aynsley e A. Ward (1986). Growth monitoring and control through computer-aided on-line mass balancing in fed-batch penicillin fermentation. *Canadian Journal of Chemical Engineering*, **64**, 567-580.

Montgomery D. C. (2005). *Introduction to Statistical Quality Control*. John Wiley & Sons Inc., New York (U.S.A.).

Montgomery D. C. (2009). *Design and Analysis of Experiments.* John Wiley & Sons Inc., New York (U.S.A.).

Mou D. e C. Cooney (1983). Modeling and adaptive control of fedbatch penicillin production. *Biotechnology and Bioengineering*, **25**, 225-255.

Nestaas E. e D. Wang (1983). Computer control of the penicillin fermentation using the filtration probe in conjunction with a structured process model. *Biotechnology and Bioengineering*, **25**, 781-796.

Nielsen J. (1997). *Physiological Engineering Aspects of Penicillin chrysogenum*. World Scientific, Singapore.

Nielsen J. e J. Villadsen (1994). *Bioreaction Engineering Principles*. Plenum Press., New York (U.S.A.).

Nocentini M., D. Fajner, G. Pasquali e F. Magelli (1993). Gas–liquid mass transfer and holdup in vessels stirred with multiple rushton turbines: water and water-glycerol solutions. *Ind. Eng. Chem. Res*., **32**, 19–26.

Novak M. e V. Klekner (1988). Comparison of various methods of KLa estimation in cultures of filamentous microorganisms. *Biotechnol. Techn*., **2**, 243–8.

Pellet J.P. e A. Elisseeff (2007). A partial correlation-based algorithm for causal structurediscovery with continuous variables, in: Proceedings of the 7th International Conference on Intelligent Data Analysis. Springer-Verlag, Ljubljana, Slovenia 229–239.

Pirt S. e R. Righoletto (1967). Effect of growth rate on the synthesis of penicillin by Penicillium chrysogenum in batch and chemostat cultures. *Applied Microbiology*, **15**, 1284-1290.

Press W.H., S.A. Teukolsky, W.T. Vetterling e B.P. Flannery (2007). *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, Cambridge (U.K.).

Puthli M.S., V.K. Rathod e A.B. Pandit (2005). Gas–liquid mass transfer studies with triple impeller system on a laboratory scale bioreactor. *Biochem. Eng. J.*, **23**, 25–30.

Rao K.R. e S. Lakshminarayanan (2007). Variable interaction network based variableselection for multivariate calibration. *Anal. Chim. Acta*, **599**, 24–35.

Rao K.R. e S. Lakshminarayanan (2007). Partial correlation based variable selection approach for multivariate data classification methods. *Chemometr. Intell. Lab.Syst.*, **86**, 68-81.

Rato T. J. e M.S. Reis (2014). Sensitivity enhancing transformations for monitoring the process correlation structure. *J. Process Control*, **24**, 905-915.

Rato T.J. e M.S. Reis (2015). On-line process monitoring using local measures of association. Part II: Design issues and fault diagnosis. *Chemom. Intell. Lab. Syst*. **142**, 266-275.

Sanchez A, F. Garcia, A. Contreras, E. Molina e Y. Chisti (2000). Bubble-column and airlift photobioreactors for algal culture. *AIChE J*., **46**, 1872–87.

Seborg D. E., T.F. Edgar, A. Mellichamp, J. F. Doyle (2010). *Process Dynamics and Control*. John Wiley & Sons Inc., New York (U.S.A.).

Shuler M. e F. Kargi (2002). *Bioprocess Engineering Basic Concepts*. Prentice Hall. Saddle River, New York (U.S.A.).

Taguchi H. e A.E Humphrey (1966). Dynamic Measurement of the volumetric oxygen transfer coefficient in fermentation systems. *J. Ferment Technol*., **44**, 881–9.

Tomba E. (2013). *Latent variable modelling approaches to assist the implementation of quality-by-design paradigms in pharmaceutical development and manufacturing*. Università degli Studi di Padova.

Valle S., W. Li e S.J. Qin (1999). Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Ind. Eng. Chem. Res.*, **38**, 4389-4401.

Van't Riet K. (1979). Review of measuring methods and nonviscous gas–liquid mass transfer in stirred vessels. *Ind. Eng. Chem. Process. Des. Dev.*, **18**, 357–64.

Weissenborn P.K. e R.J. Pugh (1996). Surface tension of aqueous solutions of electrolytes: relationship with ion hydration, oxygen solubility, and bubble coalescence. *J. Colloid Interface Sci.*, **184**, 550–63.

Whitman W.G. (1923). Preliminary experimental confirmation of the two-film theory of gas absorption. *Chem. Metall. Eng*., **29**, 146–9.

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

Wold H. (1966). *Estimation of principal components and related models by iterative least squares. In Multivariate analysis*. Academic Press Limited, New York (U.S.A.).

Wold S., H. Martens e H. Wold (1983). The multivariate calibration problem in chemistry solved by the PLS method. *Lecture Notes in Math*., **973**, 286-293.

Zhan T., T. Wang T e J. Wang (2006). Analysis and measurement of mass transfer in airlift loop reactors. *Chin. J. Chem. Eng.*, **14**, 604–10.