**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**

# PREDIZIONE DEL TITOLO VIRALE IN UN PROCESSO INDUSTRIALE DI PRODUZIONE DI REOVIRUS IMPIEGATI PER LA FORMULAZIONE DI VACCINI AVIARI

***Relatore: Prof. Massimiliano Barolo***
***Correlatori: Dott.ssa Donatella Bernini***
***Ing. Martina Largoni***

***Laureando: RICCARDO VEDOLIN***

ANNO ACCADEMICO 2013 – 2014

# Riassunto

In questa Tesi è affrontato il problema del monitoraggio in tempo reale di un processo biologico-farmaceutico per la produzione di reovirus impiegati per la formulazione di vaccini aviari. Il processo in esame viene condotto mediante fermentazioni batch, al termine delle quali si ottiene un prodotto di qualità variabile. Nella Tesi vengono sviluppati dei modelli basati su dati, che permettono di stimare la qualità finale del prodotto a partire dai dati di processo disponibili. I modelli sono stati costruiti in modo da fornire la stima del titolo virale finale alla conclusione del batch, riducendo così il tempo di attesa rispetto a quanto ad oggi avviene grazie ai test del laboratorio interno, che rendono noto il valore del titolo 15 giorni dopo la conclusione del batch. I modelli sviluppati sono stati ottimizzati in modo da predire, con sufficiente accuratezza, la qualità finale sia di batch in specifica che di batch fuori specifica. I risultati ottenuti in termini di stima sono positivi e mostrano come il titolo possa essere predetto, in maniera affidabile, con un errore mediamente inferiore rispetto a quello compiuto dalle analisi di laboratorio, sia per i batch in specifica sia per i batch fuori specifica. In particolare, realizzando in tempo reale la stima del titolo virale finale, i modelli sviluppati permettono di stimare con precisione il titolo virale finale a partire dalla 30$^{a}$ ora di fermentazione, cioè da circa metà della durata di ciascun batch.

# Indice

# Introduzione

Nelle produzioni industriali è importante monitorare la qualità del prodotto finale. Essa, infatti, può essere soggetta a variabilità, per esempio dovuta alle materie prime o alle diverse condizioni alle quali viene esercito il processo. Accade piuttosto di frequente che, nonostante in ciascuna produzione si conducano le medesime azioni e si cerchi di garantire le stesse condizioni di processo (lavorazione secondo "ricetta"), la qualità del prodotto vari, comportando talvolta dei fuori specifica. Questo problema è particolarmente rilevante nell'industria farmaceutica, nella quale le specifiche di produzione sono molto restrittive (per effetto dei vincoli regolatori) e i prodotti hanno alto valore aggiunto.

In questa Tesi si studia un processo biologico-farmaceutico per la produzione di reovirus utilizzati nella formulazione di vaccini aviari. Il processo viene condotto nell'azienda Merial Italia S.p.A. di Noventa Padovana. Il problema affrontato è la variabilità della qualità del prodotto, rappresentata dal titolo virale finale del reovirus, che può comportare l'ottenimento di un prodotto fuori specifica. Il titolo virale finale viene determinato mediante analisi di laboratorio, il cui esito è noto solo 15 giorni dopo il termine del batch. L'obiettivo perseguito nella Tesi è sviluppare dei modelli che, a partire dai dati di processo disponibili, permettano di predire il titolo virale finale riducendo l'attesa dovuta ai tempi necessari per lo svolgimento delle analisi di laboratorio.

Nonostante prove per definire le condizioni sperimentali più appropriate per la crescita del reovirus siano note in letteratura (Robertson e Wilcox, 1986; Grande e Benavente, 2000), la produzione industriale di reovirus ha finora solo marginalmente beneficiato di queste informazioni e la maggior parte delle operazioni viene svolta sulla base dell'esperienza. La complessa gestione del processo rende necessario uno studio approfondito, basato sull'implementazione di opportuni metodi statistici per estrarre dai dati informazioni che permettano una migliore comprensione del processo produttivo.

Nella Tesi vengono condotte quattro principali attività:

- organizzazione razionale e trattamento dei dati dell'intero processo produttivo;
- elaborazione di una metodologia per individuare le attività che possono essere condotte per migliorare il monitoraggio di processo e, in particolare, sviluppo dell'attività legata alla stima della qualità del prodotto;
- analisi preliminare su tutte le variabili di processo per definire le operazioni rilevanti e significative per la definizione della qualità del prodotto;
- analisi dettagliata dello stadio di fermentazione, dove avviene la replicazione del virus.

Per legare i dati di processo disponibili alla variabile di qualità al fine di predirla, si utilizza il metodo statistico multivariato PLS (*partial least squares regression*, Geladi e Kowalski,

1986). Viene sviluppato un sensore virtuale (*soft sensor*) che, tramite modelli PLS, permette di predire il titolo virale finale del reovirus. Finora i metodi di *soft sensing* hanno trovato poche applicazioni in ambiti legati alla biotecnologia (Mandenius e Gustavsson, 2014), e in questa Tesi per la prima volta viene applicato il *soft sensing* ad un processo di produzione di reovirus. La predizione del titolo finale viene eseguita a batch concluso, al fine di predire il titolo dei batch in anticipo rispetto alle lunghe analisi di laboratorio.

Nella Tesi viene anche discussa la possibilità di predire in tempo reale il titolo finale, permettendo, nel caso di batch che stiano evolvendo verso condizioni di fuori specifica, di intervenire con azioni correttive.

La Tesi si sviluppa su 3 capitoli. Nel Capitolo 1 viene descritto il processo per la produzione di reovirus, presentando i dati di processo disponibili. Nel Capitolo 2 viene spiegata la tecnica statistica multivariata utilizzata per la predizione della qualità del prodotto. Il Capitolo 3 descrive la metodologia utilizzata per realizzare la stima del titolo finale; vengono riportate, dopo un'analisi preliminare sulle variabili iniziali di processo, le prestazioni di predizione ottenute dai modelli che utilizzano i dati relativi alla fermentazione, a seconda che il modello sia sviluppato su dati di batch in specifica o si considerino anche i dati dei batch fuori specifica. Una sezione finale riassume le conclusioni che possono essere tratte dal lavoro svolto.

# Capitolo 1

# Processo per la produzione di reovirus

In questo Capitolo viene descritto il processo per la produzione di reovirus dell'azienda Merial di Noventa Padovana. Il processo è l'oggetto di studio della Tesi al quale si applicano tecniche di analisi statistica multivariata al fine di monitorarne lo stato. Vengono descritti i dati a disposizione e particolare attenzione inoltre viene data alla descrizione del fermentatore e del sistema di regolazione.

## 1.1 Descrizione del processo

Un reovirus è l'antigene virale che viene utilizzato nei vaccini aviari contro le artriti. Il processo Merial ha lo scopo di produrre tale virus, con una ben determinata specifica di titolo infettante sul prodotto finale. Esso deve risultare infatti superiore ad un certo valore di soglia. Il titolo virale finale è espresso come concentrazione quindi e rappresenta la capacità infettante del virus, cioè il suo potere di moltiplicarsi all'interno delle cellule.
Il processo si sviluppa in tre stadi principali: raccolta e pretrattamento delle uova, digestione e fermentazione. La materia prima di partenza è l'uovo di pollo a 11 giorni di vita.
In Figura 1.1 viene riportato lo schema a blocchi del processo, articolato nei 3 stadi. Lo schema è strutturato secondo 3 livelli. In particolare vengono specificati all'interno di ogni stadio i materiali usati, le operazioni coinvolte e i dati raccolti. Questa tipologia di struttura viene utilizzata seguendo lo schema logico proposto da Tomba *et al.* (2013).

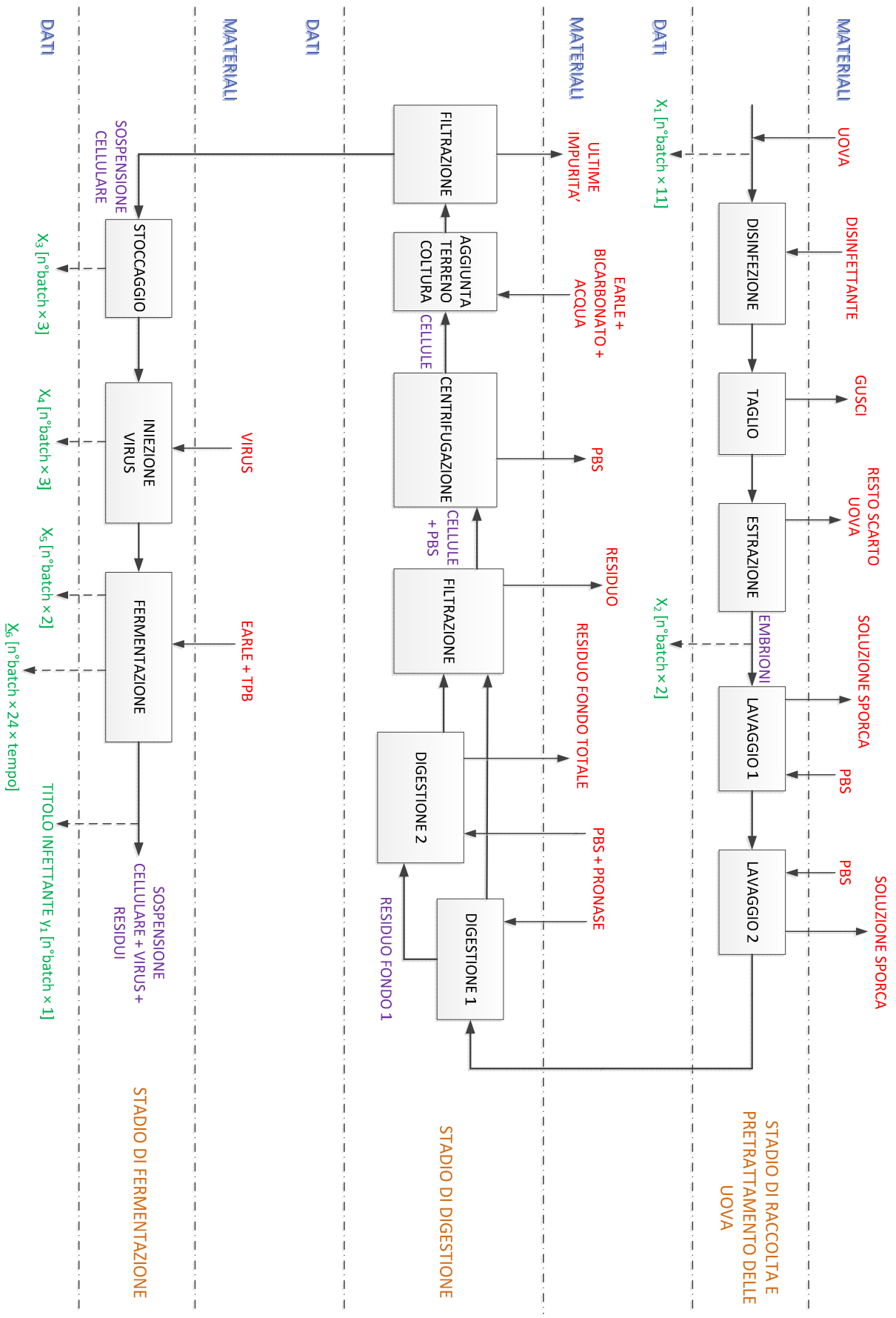



Figura 1.1.vsd

**Figura 1.1.** *Schema a blocchi del processo Merial per la produzione di reovirus.*

Nei paragrafi successivi è presente la descrizione dettagliata delle operazioni presenti all'interno di ciascuno dei 3 stadi.

### *1.1.1 Stadio di raccolta e pretrattamento delle uova*

Lo stadio di raccolta è strutturato secondo diverse fasi che qui vengono descritte facendo riferimento alla Figura 1.1.

Le uova vengono consegnate dal fornitore dopo aver subito una serie di trattamenti. In particolare, vengono tenute in incubazione per un periodo di tempo osservato, poi vengono disinfettate e sottoposte a speratura[1]. Successivamente, vengono caricate su camion e trasportate fino all'azienda.

Le uova consegnate subiscono una prima operazione di disinfezione in una camera in cui viene nebulizzato un disinfettante con lo scopo di abbassare la carica batterica che è normalmente presente nel guscio delle uova. Dopo il trattamento, le uova disinfettate subiscono l'operazione meccanica del taglio della calotta superiore del guscio, guidata da un operatore specializzato.

Eliminato il guscio, è possibile estrarre l'embrione contenuto all'interno dell'uovo. Questa operazione viene eseguita manualmente, con l'ausilio di pinzette, da parte di 2 addetti. L'operazione di estrazione è molto delicata perché si deve estrarre l'embrione separandolo nel miglior modo possibile dalla parte restante dell'uovo. Gli embrioni vengono poi posti in 14 apposite beute in vetro in modo che ciascuna beuta contenga circa 450-550 embrioni. In questa fase alcune uova vengono scartate, o per problemi di rottura del guscio durante il taglio o perché gli embrioni sono morti. La percentuale di scarto rappresenta un dato di processo. Quando ogni beuta è stata riempita, viene passata alla fase successiva attraverso un locale di disinfezione.

Per ogni beuta vengono effettuati 2 lavaggi a temperatura ambiente con una soluzione salina fisiologica (PBS, *phosphate buffered saline*) con lo scopo di eliminare le impurità delle uova presenti nelle beute, dovute alle operazioni precedenti. In particolare ogni beuta viene mantenuta in lenta agitazione per 2 minuti. Successivamente, si attende qualche istante affinché gli embrioni possano sedimentarsi sul fondo. Infine, si svuota manualmente la soluzione di lavaggio contenente la sporcizia avendo cura di non far fuoriuscire gli embrioni.

[1] La speratura è l'operazione che consiste nell'osservazione di un uovo controluce mediante una lampada sperauovo. Viene effettuata in incubatoio per verificare che le uova da cova siano fertili e con embrione vitale, al nono o decimo giorno di incubazione.

### *1.1.2 Stadio di digestione*

Dopo le operazioni iniziali, gli embrioni vengono trattati per estrarre le cellule (fibroblasti di pollo) che saranno utilizzate per la crescita del virus nel successivo stadio di fermentazione. Si passa, quindi, allo stadio di digestione vero e proprio.

In ogni beuta viene immessa una soluzione costituita da PBS e pronase, un enzima che serve a favorire la disgregazione delle cellule degli embrioni. Il sistema viene lasciato in agitazione a bassa intensità per 20 min a 41°C, temperatura alla quale l'enzima si attiva. Tale temperatura viene mantenuta grazie ad un bagno ad acqua nel quale è inserita la beuta. La digestione viene ripetuta per 2 volte. Già alla fine del primo ciclo, e ancor più chiaramente alla fine del secondo, è possibile osservare come nella beuta siano presenti 2 fasi nettamente separate: il liquido costituito da cellule disgregate e soluzione di PBS, e lo strato al fondo, costituito principalmente da impurità e parti di embrione non disgregate da eliminare.

Al termine di ciascuna delle digestioni, dopo aver atteso qualche secondo per far sedimentare lo strato di impurità, viene eseguito il prelievo della fase liquida. Più di metà del liquido contenuto in ogni beuta viene inviato in una tanica mediante un ago aspirante collegato ad una pompa, passando attraverso un sacchetto filtrante con setto poroso per trattenere le impurità residue e lasciar passare il prodotto desiderato. Il prodotto viene successivamente trasferito in barattoli per centrifuga da 1 L. Lo stadio di centrifugazione serve per separare lo strato di cellule, che si ritroveranno adese al fondo di ogni flacone, dal surnatante costituito da soluzione di PBS, che viene eliminato manualmente.

Dopo aver tolto il surnatante, ad ogni flacone viene aggiunto il terreno di coltura, costituito da *Earle* (un composto di sali di vario tipo), bicarbonato e acqua. Ogni flacone viene agitato in modo che le cellule vengano messe in sospensione col terreno di coltura, dove esse mantengono le loro funzioni vitali.

La sospensione poi viene inviata ad una tanica passando attraverso una garza filtrante per trattenere le ultime impurità. Si sottolinea che la fase di digestione determina variabilità.

### *1.1.3 Stadio di fermentazione*

La tanica che arriva dallo stadio di digestione viene mantenuta a temperatura ambiente e in lenta agitazione per evitare che le cellule aderiscano tra loro. In questa fase viene prelevato un campione di circa 5 mL, che viene utilizzato per la conta cellulare. La conta delle cellule è importante perché il fermentatore contiene una quantità di terreno che permette la sopravvivenza di $3 \div 6 \times 10^6$ cellule. Tramite una metodologia sperimentale con camera di *burker* viene fatto il conteggio delle cellule per mL. Poi viene utilizzata l'Equazione (1.1) al fine di determinare il numero di cellule per mL di terreno di fermentatore:

$$\frac{\text{n°cellule}}{\text{mLterreno}_{\text{fermentatore}}} = \frac{\text{n°cellule}_{\text{contate}} \times 1000 \times \text{volume}_{\text{tanica}}}{\text{volume}_{\text{fermentatore}}}. \quad (1.1)$$

In base al numero di cellule contenute nella tanica si verifica che tale numero sia adeguato al fine di permettere che la fermentazione possa avvenire nelle migliori condizioni. Il numero di cellule contenute nella tanica è un valore piuttosto indicativo, poiché la misura è poco precisa, ma l'intervallo di errore tollerato è piuttosto ampio.

Nella tanica, in agitazione e a temperatura ambiente, viene iniettata una piccola quantità di virus che andrà ad infettare le cellule, per poi moltiplicarsi all'interno del fermentatore. Il rapporto tra la quantità di virus introdotto (legato al titolo del virus, noto da prove di laboratorio) e il numero di cellule è legato ad un coefficiente adimensionale, detto MOI (*multiplicity of infection*). Esso in particolare viene definito biologicamente come:

$$\text{MOI} = \frac{\frac{\text{titolo}_{\text{virale}}}{1\text{mLmatrice}_{\text{virale}}} \times \text{volume}_{\text{matrice}}}{\frac{\text{n°cellule}}{1\text{mLterreno}_{\text{fermentatore}}} \times \text{volume}_{\text{fermentatore}}}. \quad (1.2)$$

Si utilizzano delle matrici di virus standard e si mantiene un MOI costante per ogni produzione.

Prima di alimentare al fermentatore le cellule con il virus vengono effettuati due cicli di sterilizzazione. Il primo è una sterilizzazione a vuoto con vapore, mirata principalmente a sterilizzare la valvola di fondo. Poi viene effettuato il carico di acqua di osmosi purificata nel reattore per la seconda sterilizzazione a pieno. Tra i due cicli vengono tarati i sensori di $O_2$ e pH dell'impianto. Terminati i cicli di sterilizzazione, nel fermentatore viene caricato un medium contenente *Earle* e TPB (anticoagulante acido e nutriente). Quindi, viene trasferito al bioreattore l'inoculo (sospensione cellulare e virus), grazie ad una pompa peristaltica e attraverso una linea sterilizzata. Durante la fermentazione si mantengono controllati i parametri di processo, cioè il pH, la temperatura e la percentuale di ossigeno. In queste condizioni, ottimali, la fermentazione continua per 61 h circa. I profili tipici della percentuale di ossigeno disciolto e del pH sono riportati in Figura 1.2.

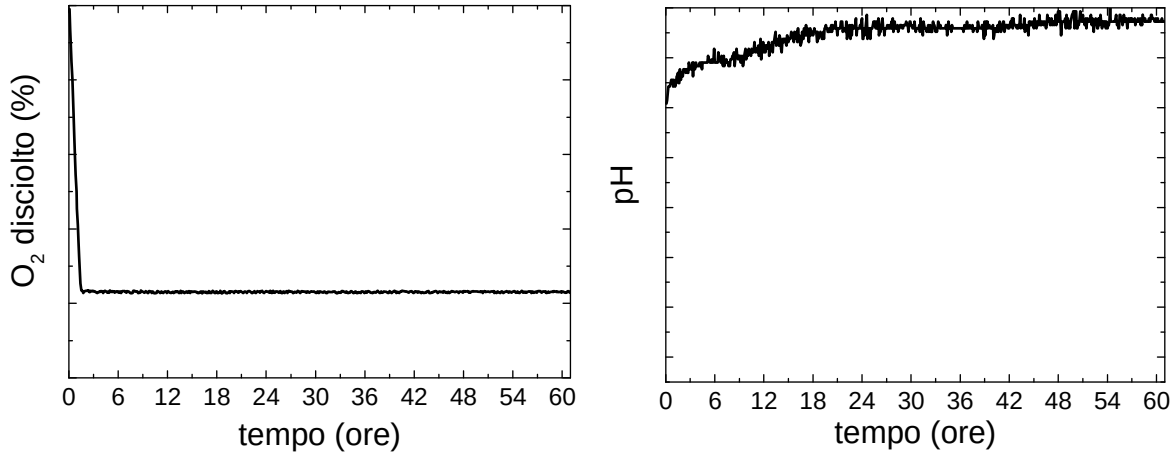



Figura 1.2a e Figura 1.2b.opj

(a) (b)

**Figura 1.2.** *Andamento tipico (a) della percentuale di ossigeno disciolto e (b) del pH in funzione del tempo di fermentazione.*

Nei primi istanti la percentuale di ossigeno disciolto è quasi pari a 100%. Poi, come si può osservare dalla Figura 1.2a, essa cala drasticamente a causa della respirazione cellulare fino ad attestarsi attorno al *set point*, tipicamente fino alla fine del batch. Il periodo di tempo in cui le cellule respirano è denominata fase aerobia.

In una prima fase della fermentazione viene a crearsi un ambiente acido, che deve essere regolato con l'immissione di bicarbonato. In una seconda fase, le cellule cominciano a morire a causa del moltiplicarsi del virus al loro interno. La loro morte crea un ambiente ossidante che fa aumentare il pH. L'ambiente a questo punto diventa basico e deve essere regolato impiegando $CO_2$. A prova di questo fenomeno è presentato l'andamento del pH in Figura 1.2b: il pH cresce fino alla dodicesima ora circa e poi si attesta attorno al valore di *set point*.

Durante la seconda fase l'aria serve sempre meno e meno frequentemente poiché le cellule stanno morendo a causa del moltiplicarsi del virus infettante al loro interno. L'andamento tipico della portata d'aria è pertanto quello riportato in Figura 1.3.

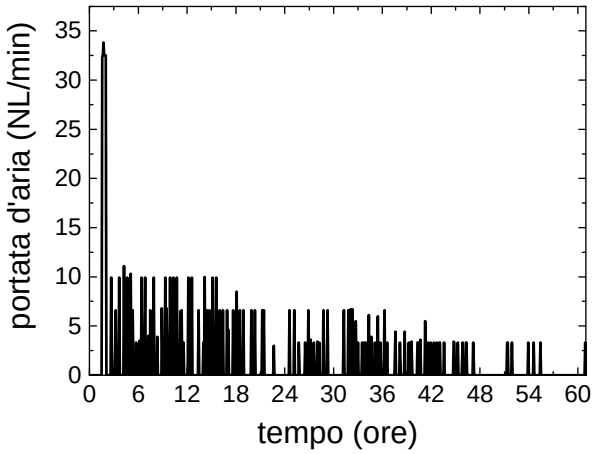



Figura 1.3.opj

**Figura 1.3.** *Andamento tipico della portata d'aria in funzione del tempo di fermentazione.*

Al termine dell'ultima ora, inizia un raffreddamento automatico del reattore fino a 15°C per bloccare la reazione di moltiplicazione del virus. A questo punto la crescita del virus si può ritenere conclusa e viene effettuato un prelievo del campione da cui verrà valutato il titolo.

## 1.2 Dati disponibili

Facendo riferimento alla Figura 1.1 si prendono in considerazione i dati disponibili nei diversi stadi del processo. Per ogni batch si hanno a disposizione diversi tipi di dati che caratterizzano il lotto. Le matrici $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$, $\mathbf{X}_4$, in particolare, contengono le caratteristiche iniziali del processo, e cioè diverse variabili che vengono misurate per ogni batch al tempo $t$=0.
La matrice $\mathbf{X}_1$ contiene le informazioni relative ai pretrattamenti che le uova subiscono prima di entrare nell'impianto. Essi sono dati riportati dal fornitore delle uova al momento della consegna e vengono definiti in Tabella 1.1.

**Tabella 1.1.** *Variabili di processo per la matrice* $\mathbf{X}_1$.

| N° Variabile | Descrizione |
|---|---|
| 1 | Età embrioni (giorni) |
| 2 | N° uova iniziali |
| 3 | Fornitore |
| 4 | Gruppo di incubazione |
| 5 | Età pollo (settimane) |
| 6 | Durata incubazione (h) |
| 7 | Tipologia disinfettante |
| 8 | Tempo da speratura a carico camion (h) |
| 9 | Durata trasporto (h) |
| 10 | N° uova incubate |
| 11 | N° uova non fertili |

Si precisa che, per quanto riguarda la matrice $\mathbf{X}_1$, le variabili fornitore (3), gruppo (4) e tipologia di disinfettante (7) sono variabili categoriali, cioè non definite da un numero.

Nella Tabella 1.2 invece sono riportate variabili che vengono definite allo stadio di pretrattamento e raccolta delle uova e che appartengono alla matrice $\mathbf{X}_2$.

**Tabella 1.2.** *Variabili di processo per la matrice* $\mathbf{X}_2$.

| N° Variabile | Descrizione |
|---|---|
| 1 | N° uova scartate |
| 2 | N° uova usate |
| 3 | N° uova rotte |
| 4 | N° uova morte |

La matrice $\mathbf{X}_3$ è composta da variabili che vengono misurate prima dello stadio di fermentazione, in particolare durante lo stoccaggio, e sono quelle proposte in Tabella 1.3. Nella matrice $\mathbf{X}_4$ compaiono variabili che entrano in gioco sempre prima dello stadio fermentativo e in particolare durante l'iniezione del virus. Esse sono presentate in Tabella 1.4.

**Tabella 1.3.** *Variabili di processo per la matrice* $\mathbf{X}_3$.

| N° Variabile | Descrizione |
|---|---|
| 1 | N° cellule / 1 mL terreno |
| 2 | N° cellule / embrione |
| 3 | N° cellule / mL terreno fermentatore |

**Tabella 1.4.** *Variabili di processo per la matrice* $\mathbf{X}_4$.

| N° Variabile | Descrizione |
|---|---|
| 1 | Volume matrice virus infettante (mL) |
| 2 | Titolo virus ($TCID_{50}$) / mL matrice |
| 3 | MOI |

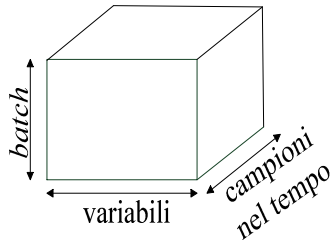
La matrice $\mathbf{X}_5$ contiene le variabili che appartengono allo stadio di fermentazione e che per ogni batch vengono misurate a $t$=0 o a $t$=$t_{finale}$. Esse vengono presentate in Tabella 1.5. La matrice tridimensionale $\underline{\mathbf{X}}_6$ contiene l'andamento temporale su 7320 istanti temporali (2 campionamenti al minuto) delle variabili di fermentazione misurate in linea, per ciascun batch. Queste vengono presentate in Tabella 1.6. Il vettore $\mathbf{y}_1$ contiene la specifica di qualità del virus, cioè il suo titolo infettante.

**Tabella 1.5.** *Variabili del processo fermentativo per la matrice* $\mathbf{X}_5$.

| N° Variabile | Descrizione |
|---|---|
| 1 | Volume fermentatore (L) |
| 2 | Durata fase aerobia (h) |

**Tabella 1.6.** *Variabili del processo fermentativo misurate in linea per la matrice* $\underline{\mathbf{X}}_6$.

| N° Variabile | Descrizione |
|---|---|
| 1 | Temperatura fermentatore (°C) |
| 2 | Temperatura camicia (°C) |
| 3 | pH (-) |
| 4 | Ossigeno disciolto (%) |
| 5 | Peso carica fermentatore (kg) |
| 6 | Pressione fermentatore (bar) |
| 7 | Velocità agitatore (rpm) |
| 8 | Portata d'aria (NL/min) |
| 9 | Set point temperatura fermentatore (°C) |
| 10 | Set point pH (-) |
| 11 | Set point ossigeno disciolto (%) |
| 12 | Set point pressione (bar) |
| 13 | Set point portata d'aria (NL/min) |
| 14 | Set point velocità agitatore (rpm) |
| 15 | Set point temperatura camicia (°C) |
| 16 | Apertura valvola di scarico dei gas esausti (%) |
| 17 | Output controllore temperatura reattore (%) |
| 18 | Output controllore portata di base (%) |
| 19 | Output controllore portata di acido (%) |
| 20 | Output controllore pressione (%) |
| 21 | Output controllore aria (%) |
| 22 | Output controllore temperatura reattore durante sterilizzazione a vuoto (%) |
| 23 | Output controllore temperatura reattore durante sterilizzazione a pieno (%) |
| 24 | Output controllore camicia (%) |




Durante la fermentazione, nel bioreattore vengono misurate in linea 24 variabili. Di queste, le prime 8 sono le misure delle variabili di processo. Le restanti sono variabili legate al sistema di controllo: i *set point* delle variabili di processo e i *controller output* (CO), ovvero i segnali di uscita dei regolatori. In particolare, le variabili 22 e 23 si riferiscono ai CO dei regolatori di temperatura durante i cicli di sterilizzazione che avvengono prima del carico del fermentatore e, come tali, non assumono significato durante il processo di fermentazione vero e proprio.
I batch a disposizione presenti nella matrice $\underline{\mathbf{X}}_6$ sono costituiti da:

- batch relativi al fermentatore da 300 L con un titolo virale finale a specifica, ossia superiore al valore di soglia;
- batch relativi al fermentatore da 600 L con un titolo virale finale a specifica;
- batch relativi al fermentatore da 300 L con un titolo virale finale fuori specifica, ossia inferiore al valore di soglia;
- batch relativi al fermentatore da 600 L conclusisi con un titolo virale finale fuori specifica.

Il numero totale di batch storici a disposizione è pari a 75.

## 1.3 Descrizione del fermentatore da 300 L

Lo stadio di fermentazione è il più importante del processo perché durante la fermentazione il virus si moltiplica e la qualità del prodotto finale viene "costruita". Si vuole quindi a questo punto studiare con maggiore attenzione a livello impiantistico il reattore in cui avviene il processo di fermentazione. In primo luogo viene considerato il reattore da 300 L, analizzando dettagliatamente il suo sistema di controllo.

In Figura 1.4 viene proposto lo schema di processo del reattore in esame e vengono anche rappresentati i vari *loop* di controllo, secondo il P&I fornito dalla ditta costruttrice del reattore.

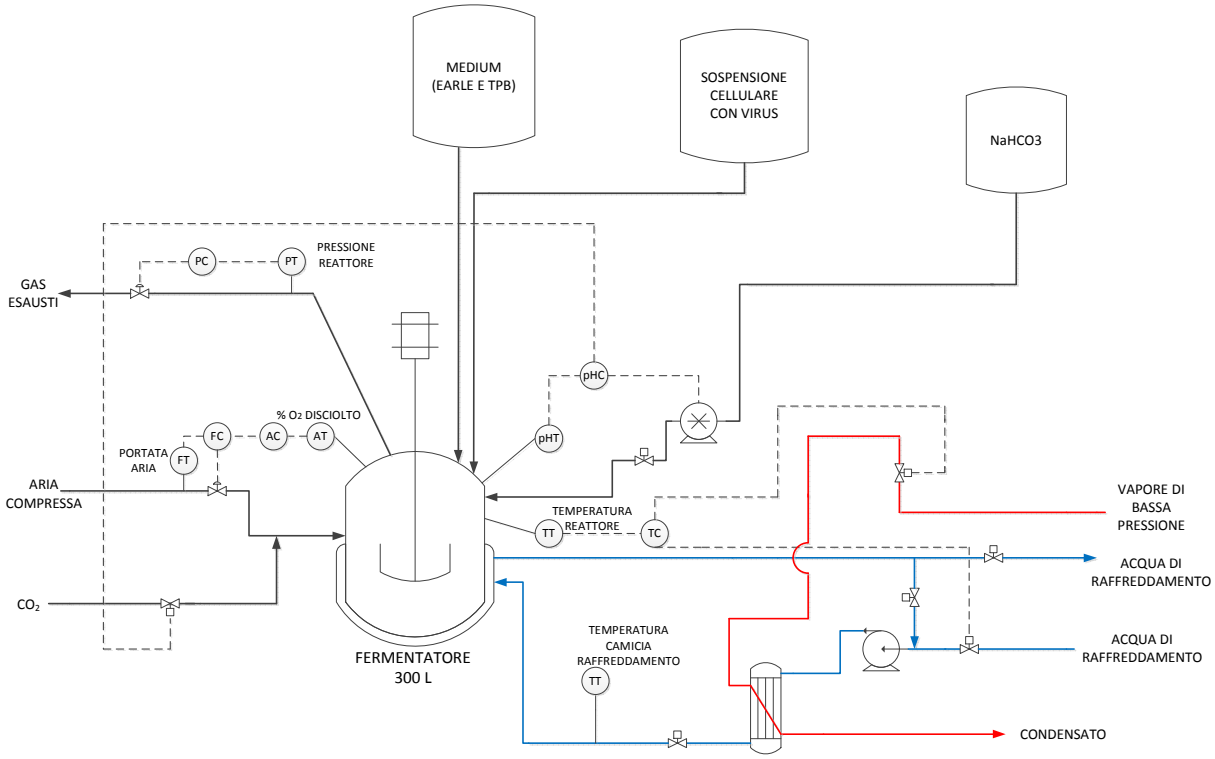



Figura 1.4.vsd

**Figura 1.4.** *Schema impiantistico del reattore da 300 L con relativo sistema di regolazione.*

In Figura 1.4, si può osservare come sia opportunamente evidenziata la corrente riservata al passaggio del vapore di bassa pressione, e la corrente riservata all'acqua di raffreddamento.

### *1.3.1 Analisi del sistema di controllo*

Si possono identificare 5 *loop* di controllo riguardanti:

- velocità d'agitazione;
- pressione del reattore;
- concentrazione di $O_2$ disciolto e della portata d'aria;
- pH;
- temperatura del reattore.

#### 1.3.1.1 Controllo della velocità d'agitazione

L'agitatore in questione è caratterizzato da un sistema a doppia elica corotante a spinta verso l'alto, installato dal fondo. Il senso di rotazione è orario se visto dalla cupola del reattore. Il misuratore coinvolto misura la velocità di rotazione dell'agitatore, che risulta costantemente pari a 71 rpm. Non è presente tuttavia un reale sistema di controllo; è previsto solamente un "trasferimento del *set point*" e ciò che agisce sull'agitatore è un *inverter*.

#### 1.3.1.2 Controllo della pressione del reattore

La pressione di testa del fermentatore viene misurata da un trasduttore di pressione installato sulla corrente dei gas in uscita e ha andamento oscillante fra 0.07 e 0.08 bar. Essa viene regolata tramite la portata uscente dei gas che si sviluppano dal processo di fermentazione. Il *controller output* del regolatore, come la percentuale di apertura della valvola (aria-apre) dei gas esausti, assume valore pari a 100%.

#### 1.3.1.3 Controllo della concentrazione di $O_2$ disciolto e della portata d'aria

In questo caso è presente un controllo in cascata, con due anelli *feedback* in cui lo *slave loop*, quello di portata d'aria, è "annidato" nel *master loop*, quello della concentrazione di $O_2$.
La percentuale di $O_2$ disciolto viene misurata e viene mantenuta dalla regolazione per gran parte della durata del batch (durata della fase aerobia) sul valore di *set point* locale. Dal regolatore PI dell'$O_2$ viene inviato il *set point* remoto (funzione del tempo) al regolatore PI della portata d'aria compressa, che quindi lo confronta con la misura di portata d'aria fornita da un flussimetro termico. Il *controller output* va poi a comandare la valvola modulatrice (aria-apre) dell'aria.

#### 1.3.1.4 Controllo del pH

Il pH deve rimanere attorno al valore di *set point* e la sua regolazione avviene mediante l'immissione di una portata di base ($NaHCO_3$) o una portata di acido ($CO_2$). Dal regolatore PI esce 1 *controller output*, che va a 2 elementi finali di controllo; uno per il controllo della base (bicarbonato) e uno per il controllo dell'acido ($CO_2$). Per quanto riguarda il controllo dell'acido, il segnale uscente dal regolatore è espresso in percentuale, ma viene convertito in segnale pulsato, secondo lo schema presentato in Figura 1.5.

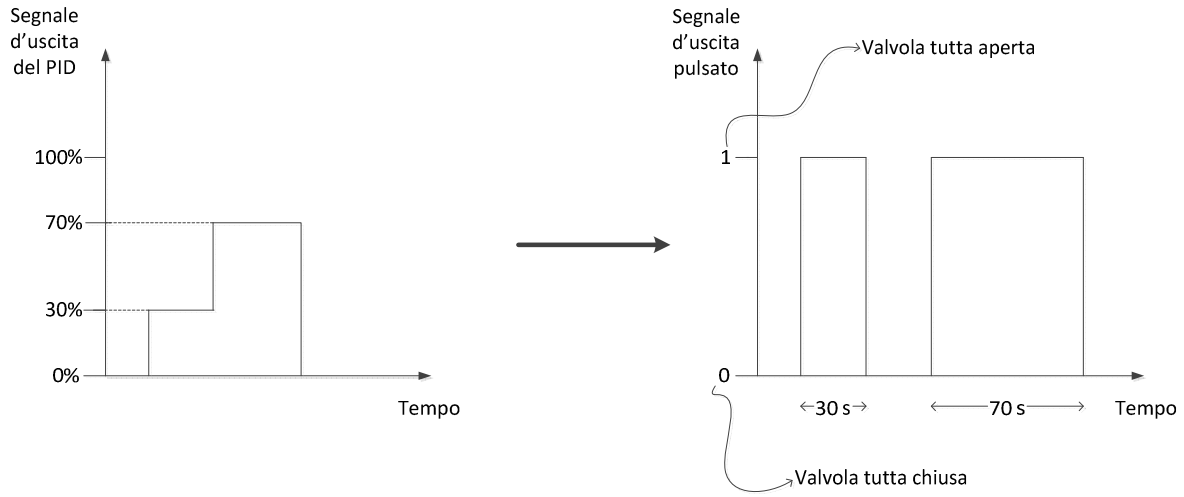



Figura 1.5.vsd

**Figura 1.5.** *Rappresentazione esemplificativa del principio di creazione del segnale pulsato.*

C'è bisogno di creare un segnale pulsato nel caso della regolazione dell'acido perché la valvola è di tipo on-off e non è modulatrice.
Il segnale poi viene convertito a segnale digitale e inviato all'elemento finale di controllo. Questo meccanismo vale per tutte le valvole regolatrici dei regolatori descritti in seguito.

### 1.3.1.5 Controllo della temperatura

La regolazione della temperatura del reattore viene fatta mediante una camicia in cui viene fatta passare acqua di raffreddamento. L'acqua però può servire, a seconda delle necessità, fredda o meno fredda e quindi è presente uno scambiatore di calore a piastre che ha lo scopo di riscaldarla opportunamente al bisogno con vapore. Il regolatore PI fornisce 1 *controller output*, secondo un ragionamento di tipo *split-range*. Nello specifico, esso è un *hot output* (*range* 0-100%), che, dopo essere stato convertito in segnale pulsato, va ad agire sulla valvola (ad angolo, aria-apre) del vapore per regolare la portata; oppure è un *cold output* (*range* -100-0%) che, allo stesso modo, va ad agire sulla valvola (ad angolo, aria-apre) dell'acqua di raffreddamento in entrata. La presenza di questo tipo di sistema di controllo implica obbligatoriamente che ci sia di base un ricircolo di acqua di raffreddamento (con annesso spurgo e reintegro) onde evitare che possa essere presente solo vapore nello scambiatore.
Infine, la temperatura della camicia di raffreddamento viene misurata da un termometro posizionato sulla corrente dell'acqua prima dell'entrata in camicia. Il suo andamento nel tempo mostra come essa oscilli all'interno dell'intervallo 33-43°C.

### 1.3.1.6 Legge di controllo e sintonizzazione del regolatore PI

I regolatori presenti nell'impianto per la regolazione delle varie grandezze fisiche sono tutti regolatori PID nei quali però è sempre disattivata l'azione derivativa. Essi sono quindi

regolatori PI e la regolazione di cui si parla è del tipo *feedback*. La variabile di processo viene misurata e il segnale che ne deriva viene convertito in analogico, linearizzato e quindi trasmesso al comparatore. La comparazione di *set point* e variabile di processo per ogni *loop* crea un errore che costituisce il segnale d'ingresso al regolatore. L'errore però prima di entrare viene convertito in errore normalizzato, al fine di ottenere poi un guadagno del regolatore ($k_C$) adimensionale. I guadagni dei regolatori sono tutti positivi. La logica di implementazione è quella in parallelo, secondo la:

$$\mathrm{CO}_{\mathrm{PI}}(t) = k_{\mathrm{C}}(\mathrm{P} + \mathrm{I}), \tag{1.3}$$

in cui $CO_{PI}$ è il *controller output* del regolatore, funzione del tempo, P rappresenta il termine relativo all'azione proporzionale e I quello dell'integrale.
La legge di controllo del regolatore PI è:

$$\mathrm{CO}_{\mathrm{PI}}(t) = k_{\mathrm{C}}(\varepsilon + \frac{1}{\tau_{\mathrm{I}}}\int_0^t \varepsilon \, \mathrm{d}t), \tag{1.4}$$

in cui $\varepsilon(t)$ è l'errore normalizzato che entra al regolatore e $\tau_I$ è il tempo dell'azione integrale.
Per ciascun regolatore possono essere attive entrambe le azioni o solo l'azione proporzionale, a seconda del *loop* di controllo. In Tabella 1.7 è riportata la sintonizzazione di ogni regolatore presente.

**Tabella 1.7.** *Valori dei parametri di sintonizzazione di ogni regolatore PI presente in riferimento al reattore da 300 L.*

| | | **Parametri di sintonizzazione** | |
|---|---|---|---|
| | | $k_C$ | $\tau_I$ |
| ***Loop* di controllo** | **Pressione fermentatore** | 150 | 80 s |
| | **pH base** | 20 | / |
| | **pH acido** | 5 | / |
| | **Concentrazione ossigeno disciolto** | 150 | / |
| | **Portata aria** | / | / |
| | **Temperatura reattore** | 300 | 25 s |
| | **Temperatura camicia raffreddamento** | 10 | 20 s |

Come si può notare dalla Tabella 1.7, solo per pressione del fermentatore, temperatura del fermentatore e temperatura della camicia di raffreddamento è prevista una regolazione PI.

## 1.4 Descrizione del fermentatore da 600 L

Si analizza ora il reattore da 600 L cercando di capire quali siano le differenze rispetto a quello da 300 L appena considerato. Inoltre si considera il sistema di controllo. In Figura 1.6 viene proposto lo schema di processo del reattore in esame.

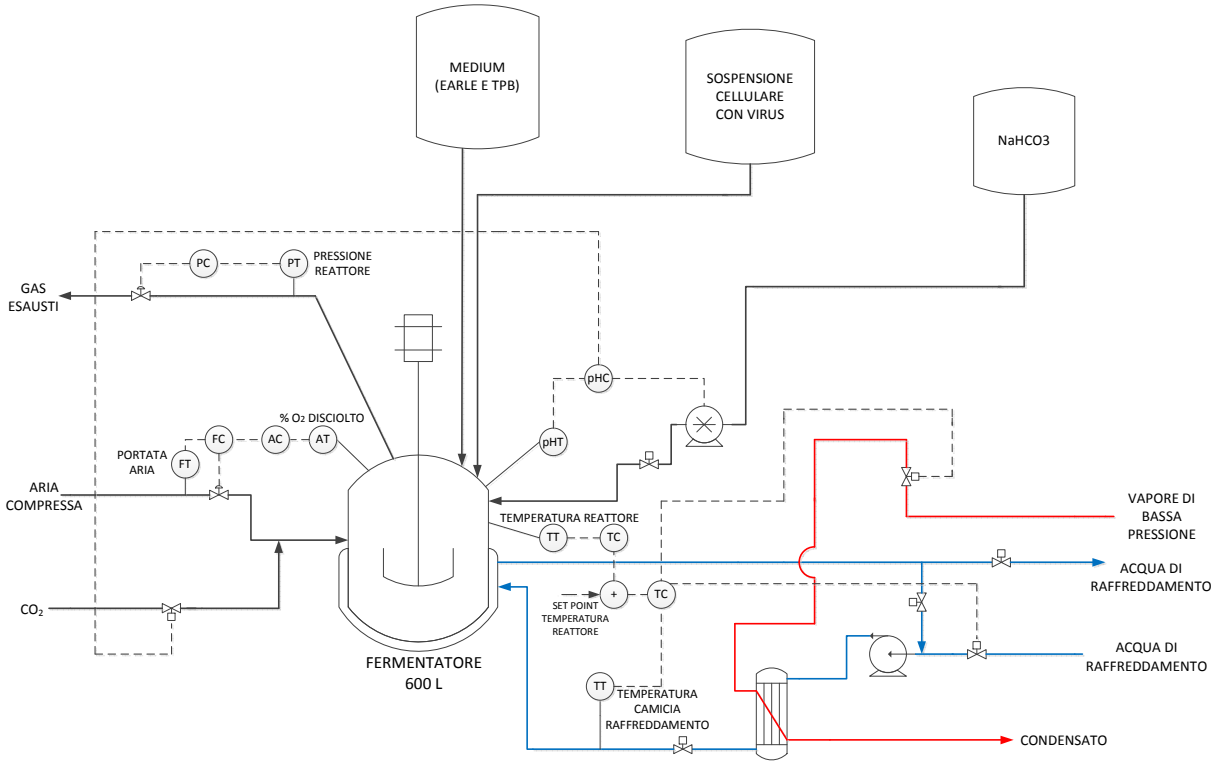



Figura 1.6.vsd

**Figura 1.6.** *Schema impiantistico del reattore da 600 L con relativo sistema di regolazione.*

In Figura 1.6, si può osservare come sia opportunamente evidenziata la corrente riservata al passaggio del vapore di bassa pressione, e la corrente riservata all'acqua di raffreddamento.

### *1.4.1 Analisi del sistema di controllo*

Si possono identificare concettualmente anche in questo caso 5 *loop* di controllo, analoghi a quelli del fermentatore da 300 L. Sono presenti però alcune differenze:

- la velocità d'agitazione misurata si attesta pari al valore di *set point*;
- la regolazione della pressione è ottenuta ancora tramite un'azione sulla valvola che modula la portata dei gas esausti in uscita, però la pressione di testa del fermentatore ha andamento oscillante fra 0.09 e 0.11 bar e il *set point* locale è diverso;
- è presente un sistema di regolazione in cascata in cui dal regolatore PI della temperatura del reattore esce un segnale che, una volta linearizzato, va a sommarsi con il *set point* locale dello stesso regolatore. Il segnale risultante costituisce il *set point* remoto per il regolatore PI della camicia di raffreddamento. Da quest'ultimo regolatore parte poi il

*controller output* per la modulazione della portata di vapore di bassa pressione o di acqua di raffreddamento. Il *controller output* del regolatore della camicia mostra un andamento che si attesta fra -5 e 5%. Il suo *set point* è fissato, e oscilla, quindi è variabile nel tempo.

La legge di controllo e il principio di funzionamento sottostante sono gli stessi descritti in precedenza nelle (1.3) e (1.4). I regolatori sono infatti tutti di tipo PI, e possono essere presenti entrambe le azioni o solo quella proporzionale. A tal proposito viene riportata in Tabella 1.8 la sintonizzazione di ogni regolatore presente.

**Tabella 1.8.** *Valori dei parametri di sintonizzazione di ogni regolatore PI presente in riferimento al reattore da 600 L.*

| | | **Parametri di sintonizzazione** | |
|---|---|---|---|
| | | $k_C$ | $\tau_I$ |
| ***Loop* di controllo** | **Pressione fermentatore** | 35 | 15 s |
| | **pH base** | 20 | / |
| | **pH acido** | 5 | / |
| | **Concentrazione ossigeno disciolto** | 150 | / |
| | **Portata aria** | / | / |
| | **Temperatura reattore** | 12 | / |
| | **Temperatura camicia raffreddamento** | 6 | 60 s |

Come si può dedurre dalle Tabella 1.8, solo per la pressione del fermentatore e per la temperatura della camicia di raffreddamento è prevista una regolazione di tipo PI.

# Capitolo 2

# Richiami statistici sul metodo PLS

Il Capitolo illustra la tecnica statistica multivariata che si utilizza in questa Tesi per la predizione della qualità finale di un prodotto. Si considera il metodo statistico della proiezione su strutture latenti, o *partial least-squares regression* (PLS), e le sue applicazioni in relazione ad un processo batch di tipo farmaceutico-biologico.

## 2.1 La qualità nei processi batch

La qualità dei prodotti è un aspetto essenziale nelle produzioni industriali. In un processo batch, dell'industria sia chimica che biologico-farmaceutica, la variabile "qualità" può presentarsi in diverse forme: composizione di un distillato, viscosità di un polimero, brillantezza di una superficie, carica batterica in un prodotto alimentare, ecc. La qualità di un prodotto viene spesso determinata da test di laboratorio che vengono eseguiti previo campionamento al termine del batch. Essa, però, rappresenta spesso una caratteristica difficilmente misurabile nel tempo perché la misura della variabile può essere ottenuta solo con un certo ritardo (non trascurabile) dal momento dell'effettivo campionamento. L'esito pertanto è noto solamente dopo un certo periodo di tempo, anche medio-lungo, dalla conclusione del batch. La bassa frequenza con la quale la qualità viene determinata comporta degli inevitabili ritardi nelle procedure imposte dalla ricetta di produzione, nonché nelle operazioni di correzione in caso di prodotto non in specifica.

Esistono però tecniche che permettono di creare un sensore *software* "virtuale" (*soft sensor*) al fine di predire la qualità sia dopo la conclusione del batch sia in tempo reale. Lo scopo è quello di poter conoscere la stima della qualità in largo anticipo e quindi capire se un batch si è concluso con un prodotto in specifica o fuori specifica. Inoltre è possibile informare istantaneamente gli operatori sullo stato di conduzione del batch attraverso la stima del parametro di qualità. Allo stesso tempo il sensore è in grado di attestare l'attendibilità della stima fornita al fine di evitare l'utilizzo di informazioni generate da condizioni di processo considerate anomale.

Una delle tecniche che permette di stimare la qualità di un prodotto a partire dalle variabili di processo è PLS (Geladi e Kowalski, 1986).

## 2.2 Metodo della proiezione su strutture latenti

Il metodo PLS utilizza una matrice di dati di processo **X**(*I*×*V*), in cui *I* è il numero dei batch e *V* è il numero delle variabili (di processo). Si utilizza anche una matrice di variabili di qualità **Y**(*I*×*M*) in cui *M* è il numero di variabili di qualità del prodotto. Il metodo PLS si focalizza sulla variabilità di **X** che è più predittiva per **Y** (Nomikos e MacGregor, 1995).

### *2.2.1 Teoria del metodo PLS*

La proiezione su strutture latenti, nota anche come metodo della regressione parziale ai minimi quadrati, è un metodo di regressione utilizzato per correlare due matrici di dati fra loro, con lo scopo in genere predittivo. Essa cerca cioè di trovare la relazione presente fra i dati contenuti in una matrice **X** e le variabili risposta della matrice **Y**, attraverso la costruzione di un modello che, noto il valore delle variabili in **X** (i regressori), ritorni il valore di un certo numero di variabili predette in **Y**. In generale il metodo PLS è di fondamentale importanza nella predizione della qualità di un prodotto soprattutto in quei casi in cui si devono trattare molti dati, spesso altamente correlati fra di loro, non solo per quanto riguarda le variabili predittrici, ma anche nel caso delle variabili predette. In questo senso, l'analisi PLS trova i fattori che catturano la parte della varianza nelle variabili in **X** maggiormente correlata alle variabili latenti che descrivono la variabilità delle variabili in **Y**.
Il metodo consiste di due relazioni esterne ed una interna. La prima relazione esterna è sulla matrice delle variabili di processo. La matrice di dati viene pretrattata secondo *autoscaling*, che viene descritto dettagliatamente al §2.2.3.
Il metodo suddivide la matrice **X**(*I*×*V*) di rango *R* in una somma di *R* matrici $\mathbf{M}_r$ di rango 1, con $r = 1, ..., R$:

$$\mathbf{X} = \mathbf{M}_1 + \mathbf{M}_2 + \mathbf{M}_3 + ... + \mathbf{M}_r + ... + \mathbf{M}_R . \tag{2.1}$$

La generica matrice $\mathbf{M}_r$ può essere rappresentata con il prodotto esterno di due vettori $\mathbf{t}_r$ e $\mathbf{p}_r$, rispettivamente *score* e *loading*. Riscrivendo la (2.1) si ottiene:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \mathbf{t}_3\mathbf{p}_3^{\mathrm{T}} + ... + \mathbf{t}_r\mathbf{p}_r^{\mathrm{T}} + ... + \mathbf{t}_R\mathbf{p}_R^{\mathrm{T}}, \tag{2.2}$$

dove l'apice T indica la trasposizione del vettore.
PLS esegue l'operazione algebrica di approssimazione:

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{t}_a\mathbf{p}_a^{\mathrm{T}} + \mathbf{E} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}, \tag{2.3}$$

dove **E**(*I*×*V*) è la matrice dei residui, **T**(*I*×*A*) la matrice degli *score*, **P**(*V*×*A*) la matrice dei *loading* e $A \leq \min(I,V)$, viene detto numero di variabili latenti, le quali descrivono la parte rilevante della variabilità dei dati (Facco, 2005).
In dettaglio, gli *score* sono combinazioni dei dati originari secondo:

$$\mathbf{t}_i = \mathbf{X}\mathbf{p}_i \,. \tag{2.4}$$

La matrice degli *score* **T**, che ha per righe i vettori $\mathbf{t}_i$, rappresenta le coordinate dei dati sullo spazio individuato dalle variabili latenti. Gli *score* contengono le informazioni su come i campioni si relazionano tra loro.
La matrice **P** dei *loading* ha per righe i vettori $\mathbf{p}_i$, gli autovettori della matrice di covarianza $\mathrm{cov}(\mathbf{X})$, e contiene le informazioni su come le variabili si relazionano tra loro; i suoi elementi sono i coseni direttori di ciascuna variabile latente. Poiché gli *score* sono tra loro ortogonali e i *loading* ortonormali, le variabili latenti sono tra loro non correlate.
Le coppie $\mathbf{t}_i$ e $\mathbf{p}_i$ possono essere disposte in ordine decrescente dei rispettivi autovalori, i quali sono misure della varianza spiegata dalla *a*-esima variabile latente. Tale varianza può essere intesa come una quantità di informazioni del *set* originario di dati nello spazio definito dalle variabili latenti, se esiste grande correlazione tra le variabili originarie. Il dato viene dunque rappresentato da un numero di variabili inferiore a quello originario (usualmente, $A << \min(I,V)$), senza perdere informazioni rilevanti e sistematiche, qualora *A* sia scelto opportunamente. I residui raccolti in **E** corrispondono alle informazioni non rappresentate dal modello.
La relazione esterna sulla matrice delle variabili di qualità di prodotto **Y**(*I*×*M*), dove *M* rappresenta il numero di variabili finali di qualità per ogni batch ed è del tutto analoga alla precedente:

$$\mathbf{Y} = \sum_{a=1}^{A} \mathbf{u}_a \mathbf{q}_a^{\mathrm{T}} + \mathbf{F} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \,, \tag{2.5}$$

dove **F**(*I*×*M*) è la matrice dei residui, **U**(*I*×*A*) la matrice degli *score*, **Q**(*M*×*A*) la matrice dei *loading* riferiti alla matrice **Y**. Il modello impone di minimizzare $\|\mathbf{E}\|$ e $\|\mathbf{F}\|$. La relazione interna lega gli *score* $\mathbf{t}_a$ della matrice **X** con quelli $\mathbf{u}_a$ di **Y** e può essere di tipo lineare:

$$\mathbf{u}_a = b_a \mathbf{t}_a \,, \tag{2.6}$$

in cui i coefficienti di regressione sono:

$$b_a = \frac{\mathbf{u}'_a \mathbf{t}_a}{\mathbf{t}'_a \mathbf{t}_a}. \tag{2.7}$$

Uno dei più comuni metodi utilizzati per calcolare i parametri di un modello PLS è l'algoritmo noto con il nome di NIPALS (*Nonlinear Iterative Partial Least Squares*; Geladi e Kowalski, 1986). I vettori degli *score* $\mathbf{t}_a$ sono calcolati per ogni dimensione $a$ = 1, 2,…, $A$ del modello PLS, in modo che la combinazione lineare delle variabili in **X**, attraverso degli opportuni pesi (detti *weights*), data da $\mathbf{t}_a = \mathbf{X}\mathbf{w}_a$, e la combinazione lineare delle variabili in **Y**, data dalla $\mathbf{u}_a = \mathbf{Y}\mathbf{q}_a$, massimizzino la covarianza tra **X** e **Y**, spiegata da ciascuna dimensione $a$ della PLS. I pesi $\mathbf{w}_a$ e $\mathbf{q}_a$ sono introdotti per mantenere l'ortogonalità degli *score*. L'analisi correla gli *score* della matrice **X** con gli *score* della matrice **Y** attraverso l'Equazione (2.7), essendo $b_a$ un elemento del vettore dei coefficienti di regressione della relazione interna $\mathbf{b}_a$. L'algoritmo quindi calcola *score*, *loading*, pesi e coefficienti di regressione secondo una procedura sequenziale.
Esistono tecniche statistiche che servono a trovare i vettori delle direzioni di massima variabilità nei dati della matrice **X**. Il metodo PLS attua una rotazione di questi vettori al fine di rappresentare meglio **Y** e predire meglio le variabili di qualità del prodotto. Tali vettori ruotati vengono chiamati variabili latenti.

### *2.2.2 Calibrazione e convalida*

In primo luogo si hanno a disposizione dei dati di processo. Su questo *set* viene costruito il modello di calibrazione che lega fra loro le variabili dei vari campioni. Successivamente vengono considerati nuovi dati, al di fuori di quelli del modello, con i quali viene fatta la convalida del modello. Quando si ha un nuovo vettore-campione $\mathbf{x}_{new}$ lo si proietta all'interno del modello, predicendo lo *score* secondo:

$$\hat{\mathbf{t}}_{new} = \frac{\mathbf{x}_{new}\mathbf{W}}{\mathbf{P}^{\mathrm{T}}\mathbf{W}}, \tag{2.8}$$

in cui **W** è la matrice dei pesi. Ne deriva un vettore dei residui $\mathbf{e}_{new}$:

$$\mathbf{e}_{new} = \mathbf{x}_{new} - \hat{\mathbf{x}}_{new}, \tag{2.9}$$

dove:

$$\hat{\mathbf{x}}_{new} = \hat{\mathbf{t}}_{new}\mathbf{P}^{\mathrm{T}}. \tag{2.10}$$

Ottenuto il vettore-proiezione $\hat{\mathbf{x}}_{new}$, è possibile ottenere la predizione $\hat{\mathbf{y}}_i$, che è il valore della variabile qualità stimato a partire dal nuovo dato $\hat{\mathbf{x}}_{new}$ secondo una regressione lineare.

### *2.2.3 Trattamento preliminare dei dati*

Al fine di estrarre le caratteristiche di correlazione e non semplicemente di covarianza, la matrice dei dati di processo **X** e la matrice della qualità finale **Y** devono essere pre-trattate. Le operazioni di seguito vengono descritte per **X**, ma valgono anche per **Y**. Eseguendo un *autoscaling*, la matrice di covarianza delle misure corrisponde alla matrice di correlazione. L'*autoscaling* consiste in un centramento al valor medio (*mean centering*) e una riduzione a varianza unitaria (*scaling*). Il *mean centering* consiste nel sottrarre la media per ogni variabile (Kourti, 2003):

$$\bar{\mathbf{x}}_v = \frac{\sum_{i=1}^{I} x_{i,v}}{I}, \tag{2.11}$$

in cui $x_{i,v}$ è l'elemento della matrice $\mathbf{X}[I\times(V\cdot K)]$ situato nella riga $i$ e nella colonna $vk$.
Lo *scaling* compensa le differenze di unità di misura diverse tra variabili, in modo da dare a tutte lo stesso peso. Si effettua dividendo tutte le misure di una variabile per la deviazione standard della variabile stessa, in modo che la varianza per tutte le variabili risulti unitaria:

$$\mathrm{var}(\mathbf{x}_v) = \frac{\sum_{i=1}^{I} \left(x_{i,v} - \bar{\mathbf{x}}_v\right)^2}{I} \tag{2.12}$$

e

$$\sigma = \sqrt{\mathrm{var}(\mathbf{x}_v)}\,. \tag{2.13}$$

Tutte le simulazioni PLS effettuate in questa Tesi utilizzano l'*autoscaling* come trattamento preliminare sui dati.

### *2.2.4 Statistiche di controllo*

È necessario definire delle statistiche che quantifichino la capacità di rappresentare i dati da parte del modello PLS, nello spazio all'interno e all'esterno del modello. In questo modo, per i dati disponibili e per eventuali nuovi dati che vengono proiettati sul modello, è possibile definire la loro normalità (in termini di condizioni operative) sulla base dei valori delle statistiche, rispetto ad un limite di controllo definito nella fase di calibrazione del modello. Di

seguito si fa riferimento alle sole statistiche inerenti alla matrice **X,** ma le considerazioni possono essere estese anche alla matrice **Y**.

### 2.2.4.1 Statistica SPE

Lo spazio esterno al modello è caratterizzato dalla statistica $\mathrm{SPE}_i$, errore quadratico medio (*squared prediction error*). SPE serve a rappresentare la mancanza di accuratezza statistica nel regredire i dati. Esso è la somma dei quadrati di ciascun campione (riga) di **E**, ovvero per l'*i*-esimo campione:

$$\mathrm{SPE}_i = \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{x}_i\left(\mathbf{I}-\mathbf{P}\mathbf{P}^{\mathrm{T}}\right)\mathbf{x}_i^{\mathrm{T}} = \sum_{v=1}^{V} e_{i,v}{}^2 = \sum_{v=1}^{V}\left(x_{i,v}-\hat{x}_{i,v}\right)^2, \tag{2.14}$$

dove $\mathbf{e}_i$ è un vettore riga della matrice dei residui e **I** la matrice identità. La statistica SPE indica quanto bene ogni campione viene rappresentato dal modello e, in termini geometrici, il valore $\sqrt{\mathrm{SPE}_i}$ rappresenta la distanza euclidea dell'*i*-esimo punto dall'iperpiano di dimensioni ridotte costituito dalle variabili latenti.
Per questa statistica il limite $\mathrm{SPE}_{\alpha,\mathrm{Lim}}$ è definito da Jackson e Mudholkar (Jackson, 1991):

$$\mathrm{SPE}_{\alpha,\mathrm{Lim}} = \theta_1\left(\frac{z_\alpha\sqrt{2\theta_2 h_0^2}}{\theta_1}+1+\frac{\theta_2 h_0\left(h_0-1\right)}{\theta_1^2}\right)^{\frac{1}{h_0}}, \tag{2.15}$$

in cui:

$$\theta_n = \sum_{r=A+1}^{R}\lambda_r^n \text{ per } n=1,2,3 \tag{2.16}$$

e

$$h_0 = 1-\frac{2\theta_1\theta_3}{3\theta_2^2}, \tag{2.17}$$

e infine $z_\alpha$ la deviazione normale standard per la percentuale di confidenza (1-α).

### 2.2.4.2 Statistica $T^2$

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

$$T_i^2 = \mathbf{t}_i \mathbf{\Lambda}^{-1} \mathbf{t}_i^{\mathrm{T}} , \tag{2.18}$$

dove $\mathbf{\Lambda}^{-1}$ è l'inversa della matrice diagonale degli autovalori $\lambda_i$ (Wise e Gallagher, 1996). In questo caso viene quindi investigato lo spazio all'interno del modello, in particolare lo spazio degli *score*. Secondo Jackson (1991) i limiti per il controllo nello spazio degli *score* sono definiti da un'ellissoide di fiducia che ha come semiassi:

$$s_a = \sqrt{\lambda_a T_{A,I,\alpha}^2} , \ \forall\, a = 1, 2, ..., A , \tag{2.19}$$

dove $T_{A,I,\alpha}^2$ è:

$$T_{A,I,\alpha}^2 = \frac{A(I-1)}{I-A} F_{A,(I-A),\alpha} = T_{\mathrm{Lim}}^2 , \tag{2.20}$$

nella quale compare la distribuzione *F*, il cui valore dipende dal numero di variabili latenti *A*, dal numero di campioni *I* e dal limite di confidenza (1-α). $T_{\mathrm{Lim}}^2$ è il limite di controllo per la statistica $T^2$.

## *2.2.5 Selezione del numero di variabili latenti*

Quando si esegue l'operazione di approssimazione definita nella (2.5), il residuo **F** deve essere minimizzato per aumentare la rappresentatività del modello. In particolare, secondo Geladi e Kowalski (1986), deve essere minimizzata $\|\mathbf{F}_a\|$ all'interno della relazione mista:

$$\mathbf{F}_a = \mathbf{F}_{a-1} - b_a \mathbf{t}_a \mathbf{q}_a^{\mathrm{T}} . \tag{2.21}$$

La scelta del numero di variabili latenti con cui costruire il modello è quindi molto importante.

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

$$RMSECV_m = \sqrt{\frac{PRESS_m}{I}} , \tag{2.22}$$

in cui *PRESS* (*prediction residual sum of squares*) si calcola come:

$$PRESS_m = \sum_{i=1}^{I} \left( y_{i,m} - \hat{y}_{i,m} \right)^2 . \tag{2.23}$$

Abitualmente l'aggiunta di variabili latenti al modello fa decrescere il valore dell'errore nel *set* di calibrazione. Quando però il numero di variabili latenti è eccessivo si descrive una varianza poco rilevante del *set* di calibrazione o addirittura del rumore, che rappresenta la parte non sistematica della variabilità. Questo fa sì che l'errore sul *set* di convalida cresca. La scelta del numero di variabili latenti ottimale per costruire il modello pertanto deve essere fatta ponendo attenzione ad entrambi questi aspetti: la sensitività dell'errore in funzione del numero di variabili latenti sia in calibrazione che in convalida.
Le analisi PLS effettuate in seguito utilizzano questa metodologia per la scelta del numero di variabili latenti da trattenere nel modello.

### *2.2.6 Selezione delle variabili: indice VIP*

Nello sviluppo di un modello di regressione lineare risulta di fondamentale importanza, ai fini della qualità della stima, la selezione delle variabili predittive utilizzate per la costruzione dello stimatore. Infatti è probabile che non tutte le variabili a disposizione per la costruzione di un modello di regressione siano effettivamente utili al miglioramento della stima. Una variabile predittiva può non possedere una dipendenza di tipo lineare con la variabile dipendente, e la costruzione di un modello, che utilizza tale variabile, "forza" la relazione funzionale esistente tra le due verso una relazione lineare, con conseguente scadimento delle prestazioni nella stima. Allo stesso modo l'utilizzo di variabili con basso rapporto segnale-rumore provocherebbe il fenomeno dell'*overfitting* con regressione del solo rumore dovuto al processo di misurazione. Sono stati sviluppati, perciò, dalla comunità scientifica diversi metodi di selezione per superare i problemi illustrati. Il metodo VIP (*variable importance in the projection*), sviluppato da Chong e Jun (2005), utilizzato in questa Tesi, effettua la selezione delle variabili solo dopo aver costruito il modello PLS contenente tutte le variabili predittive a disposizione. La selezione viene effettuata calcolando l'indice VIP per la *v*-esima variabile predittiva, che è calcolabile dalla formula:

$$VIP_v = \sqrt{\frac{V \sum_{a=1}^{A} \left( b_a{}^2 \mathbf{t}^{\mathrm{T}}{}_a \cdot \mathbf{t}_a \left( \frac{w_{a,v}}{\left\| \mathbf{w}_a \right\|} \right)^2 \right)}{\sum_{a=1}^{A} \left( b_a{}^2 \mathbf{t}^{\mathrm{T}}{}_a \cdot \mathbf{t}_a \right)}} , \tag{2.24}$$

in cui $\mathbf{w}_a$ rappresenta il peso associato alla *a*-esima variabile latente e $w_{a,v}$ il suo *v*-esimo elemento. In particolare, il metodo seleziona la variabile da ammettere nel nuovo *set* di predittori ridotto solo se l'indice risulta superiore a 1. In caso contrario la variabile viene giudicata inessenziale o non correlata con la variabile indipendente e viene perciò esclusa dal *set* originario.

### *2.2.7 Predizione della qualità finale in processi batch*

Un modo per predire la qualità in un processo batch mediante il metodo PLS è la stima del parametro di qualità che un prodotto ha al termine del batch. Nei processi batch le variabili di processo vengono misurate in linea e controllate attraverso un sistema di regolazione, ottenendone alla fine il profilo temporale. Il tempo assume quindi un ruolo fondamentale in questi processi e i dati di processo vengono organizzati in una matrice tridimensionale $\underline{\mathbf{X}}$ ($I \times V \times K$) dove $I$ è il numero di batch, $V$ le variabili di processo e $K$ gli istanti temporali in cui avviene il campionamento delle variabili. $\underline{\mathbf{X}}$ contiene le misurazioni della traiettoria delle variabili di processo rilevate all'interno della durata del batch. Però per applicare il metodo PLS è necessaria una matrice bidimensionale. Proprio a tal scopo si può effettuare allora lo srotolamento (*unfolding*) della matrice tridimensionale dei dati di processo al fine di trasformarla in una matrice bidimensionale ampliata, riuscendo così a considerare anche la dinamica del batch (Nomikos e MacGregor, 1994). La procedura per realizzare l'*unfolding* viene descritta al §2.2.7.1.

#### 2.2.7.1 *Unfolding*

Le possibilità per l'*unfolding* sono due:

- *unfolding* nel senso delle variabili (*variable-wise unfolding*): ogni sezione orizzontale ($V \times K$) viene disposta sotto a quella precedente e si ottiene una matrice $\mathbf{X}[(K \cdot I) \times V)]$ che corrisponde a trattare i dati di ogni variabile (nelle colonne) in tutti i batch e in tutti gli istanti temporali. Il metodo è rappresentato in Figura 2.1.

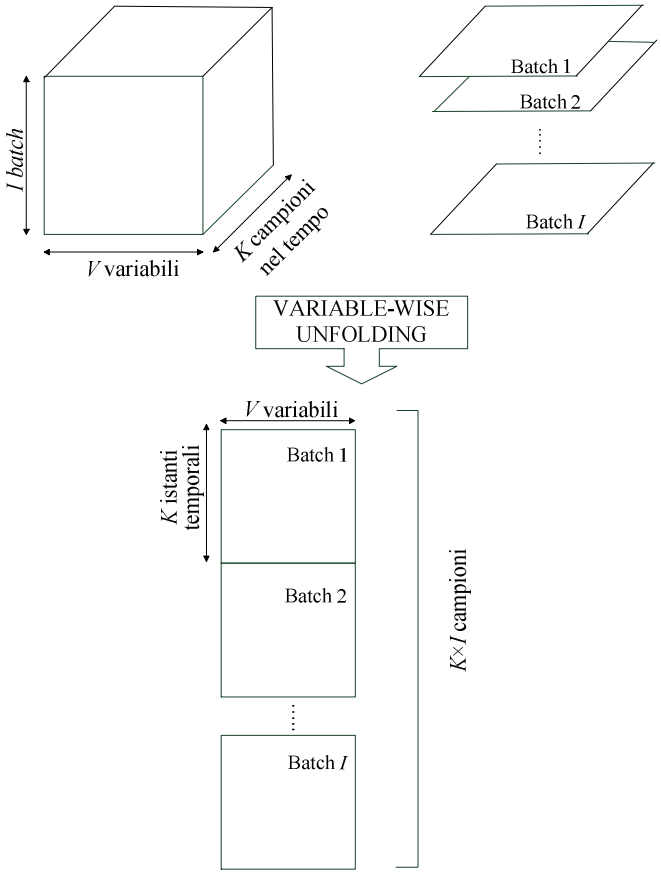



Figura 2.1.vsd

**Figura 2.1.** *Rappresentazione dell'unfolding nel senso delle variabili per la matrice* $\underline{\mathbf{X}}$.

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

- *unfolding* nel senso dei batch (*batch-wise unfolding*): si dispongono le $K$ sezioni verticali ($I \times V$) affiancate le une alle altre e si ottiene una matrice $\mathbf{X}[(I \times (V \cdot K)]$ in cui ciascuna riga contiene i dati di tutte le variabili di un batch per tutti gli istanti temporali, come rappresentato in Figura 2.2.

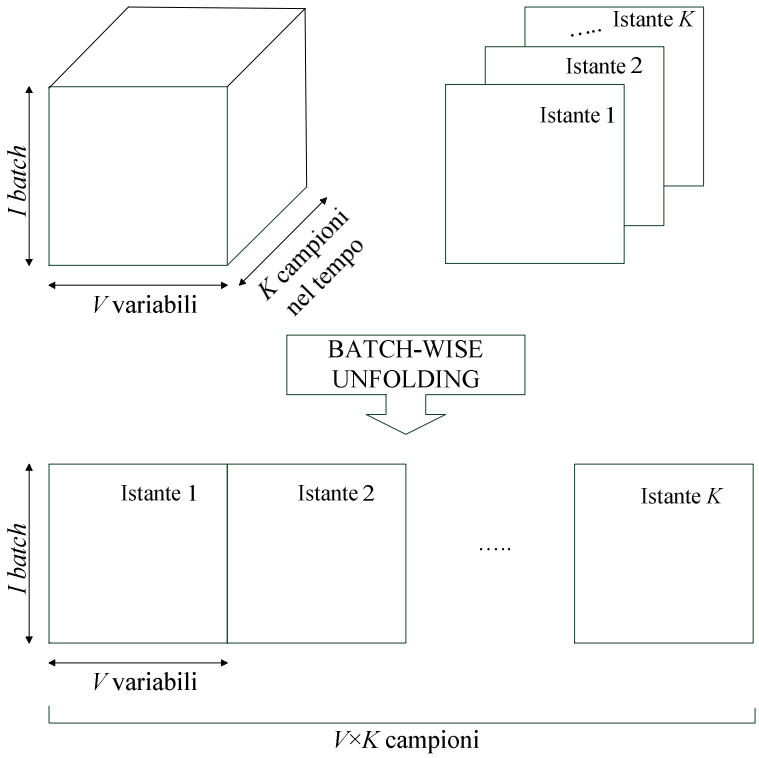



Figura 2.2.vsd

**Figura 2.2.** *Rappresentazione dell'unfolding nel senso dei batch per la matrice* $\underline{\mathbf{X}}$.

In questo caso, applicando PLS si considera la variazione nel tempo delle traiettorie delle variabili in tutti i batch rispetto alla traiettoria media della variabile nel batch stesso, e quindi si considera la dinamica.

In questa Tesi si fa riferimento al metodo *batch-wise unfolding*.

#### 2.2.7.2 Predizione del parametro di qualità in tempo reale

Un altro modo per predire la qualità col metodo PLS è la stima in tempo reale. Nella predizione in linea della qualità finale di un nuovo batch $\mathbf{x}_{new}$ si deve poter aver la possibilità di effettuare la predizione ad ogni istante di campionamento disponibile. L'applicazione del metodo statistico multivariato PLS si riferisce al vettore $\mathbf{x}_{new}$, il quale deve contenere i dati dell'intero batch, cioè per tutte le $VK$ variabili. Durante il processo, all'istante $k$ il vettore $\mathbf{x}_{new}$ contiene solo le informazioni disponibili fino a $k$; da ($k$+1) a $K$, $\mathbf{x}_{new}$ non è completo, poiché mancano le osservazioni future. Nomikos e MacGregor (1994) hanno proposto diversi metodi di riempimento di $\mathbf{x}_{new}$. Una possibilità di riempimento è di assumere che i dati mancanti abbiano future deviazioni dal valore medio uguali a quelle dell'ultimo istante di campionamento. Quest'ultimo metodo è quello utilizzato nella Tesi. La procedura da seguire per fare la predizione in linea è la seguente:

1. viene costruito il modello di calibrazione con un *set* di dati di processo e dati di qualità del prodotto (*set* di calibrazione);

2. all'istante *k* arriva $\mathbf{x}_{new}$ che deve essere legato a $\mathbf{y}_i$ all'istante *k* per costituire le due matrici di convalida;
3. viene effettuato l'*autoscaling* come pretrattamento sui dati e il riempimento della matrice delle variabili di processo;
4. viene calcolata dal modello la proiezione $\hat{\mathbf{x}}_{new}$ secondo l'Equazione (2.10) e successivamente viene effettuata la predizione della qualità per ottenere $\hat{\mathbf{y}}_i$ secondo l'Equazione (2.11);
5. viene calcolato l'errore di predizione assoluto:

$$\mathbf{errore}_i = \left|\mathbf{y}_i - \hat{\mathbf{y}}_i\right|. \tag{2.25}$$

# Capitolo 3

# Stima del titolo virale finale

Il Capitolo mostra i risultati che sono stati ottenuti dalle analisi PLS effettuate sui dati di processo. Dopo un'analisi preliminare sulle variabili iniziali, vengono presentati due metodi per eseguire la stima del titolo virale finale: una modellazione di soli batch in specifica e una modellazione locale che considera batch in e fuori specifica, entrambe sviluppate sulle variabili di fermentazione.

## 3.1 Sviluppo della metodologia di lavoro

La produzione del reovirus avviene mediante fermentazioni batch, in reattori da 300 L o da 600 L. Generalmente, il vantaggio che può essere attribuito alle produzioni di tipo discontinuo è l'elevata flessibilità, in quanto possono essere effettuate produzioni stagionali in campagne, ottenendo una grande varietà di prodotti nelle stesse apparecchiature. D'altra parte diventa complesso il monitoraggio di processo e il controllo della qualità del prodotto, e spesso ci si affida a ricette predefinite sviluppate seguendo l'esperienza, eseguite sempre allo stesso modo.

Nel processo in esame si riscontrano alcune problematiche comuni ai processi batch. Il problema principale è legato alla variabilità della qualità del prodotto finale. Infatti, nonostante in ciascuna produzione si conducano le medesime azioni e si cerchino di garantire le stesse condizioni operative, la qualità del prodotto finale varia, comportando una certa percentuale di fuori specifica. Viene quindi sviluppata una metodologia di lavoro, che evidenzia diverse possibilità di intervento per migliorare la gestione del processo. Essa viene illustrata in Figura 3.1.

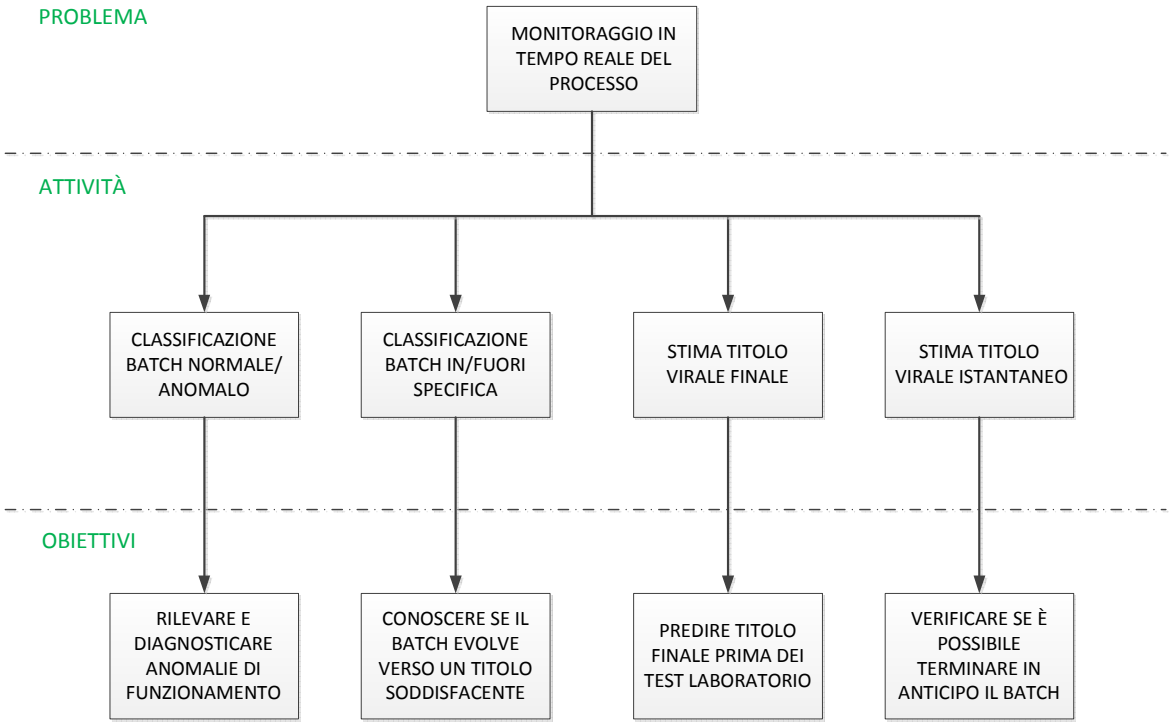



Figura 3.1.vsd

**Figura 3.1.** *Logica di sviluppo delle attività per il monitoraggio in tempo reale del processo per la produzione di reovirus.*

In Figura 3.1 si strutturano quattro diverse attività che si possono sviluppare per migliorare la conduzione del processo di fermentazione:

- classificazione dei batch come normali o anomali; in questo caso si conduce il monitoraggio in tempo reale delle variabili di processo misurate durante la fermentazione, per rilevare eventuali anomalie di processo ed eventualmente diagnosticarne la causa;
- classificazione dei batch come in o fuori specifica; ciò corrisponde al monitoraggio in tempo reale dello stato di conduzione del batch, per conoscere se il batch analizzato evolve verso la specifica di produzione;
- stima del titolo finale del reovirus; in questo caso si predice, in tempo reale e a batch concluso, il titolo di virus presente alla fine del batch, per poter avere una stima del titolo prima degli esiti dei test di laboratorio;
- stima del titolo istantaneo del reovirus; corrisponde alla predizione del titolo virale in ogni istante, per verificare se è possibile terminare in anticipo il batch.

In questa Tesi viene sviluppata l'attività 3 di stima del titolo del reovirus, grandezza che rappresenta la qualità del prodotto alla fine della fermentazione. La stima del titolo viene fatta mediante lo sviluppo di modelli, applicabili durante la fermentazione o a batch concluso, per predire, con sufficiente accuratezza, il titolo di ciascun batch. Per la stima si può tollerare un errore assoluto di ± 0.36 $TCID_{50}$/mL, stabilito dalla precisione con cui viene valutato il titolo nei test di laboratorio. Si dovrà in particolare verificare se il batch sia in specifica o fuori specifica, in base al fatto che il titolo sia rispettivamente maggiore o minore ad un certo

valore di soglia. Attualmente il titolo virale di una produzione è noto 15 giorni dopo la conclusione del batch. Si vedrà come, con i modelli sviluppati, si sia in grado di conoscere il titolo al più tardi al termine delle 61 h di fermentazione.
Inizialmente si sviluppa un modello per la predizione del titolo dei batch che evolvono in condizioni di specifica. Successivamente, vengono presi in considerazione anche i batch che evolvono verso condizioni di fuori specifica.

## 3.2 Relazione tra il titolo finale e le variabili iniziali per batch in specifica

All'interno del processo di produzione di reovirus sono coinvolte variabili iniziali e variabili di fermentazione. Le variabili iniziali sono tutte quelle variabili di processo che si ottengono prima del processo di fermentazione, appartenenti alle matrici $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$, $\mathbf{X}_4$, $\mathbf{X}_5$ presentate nelle Tabelle 1.1, 1.2, 1.3, 1.4, 1.5 al Capitolo 1, mentre le variabili di fermentazione sono quelle proprie dello stadio di fermentazione. In primo luogo vengono considerate le variabili iniziali per individuare eventuali correlazioni col titolo, pur sapendo che gran parte dell'informazione sarà contenuta nelle variabili di fermentazione, perché è in essa che avviene la moltiplicazione del virus e quindi "si costruisce" la qualità finale. Ciò non pregiudica che possano esserci informazioni utili portate dalle variabili iniziali.
A questo scopo, viene sviluppato un modello PLS costruito sulle sole matrici $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$, $\mathbf{X}_4$, $\mathbf{X}_5$. Vengono studiati i *loading* e i pesi $\mathbf{W}$ del modello PLS, costruito sui dati delle variabili iniziali e del titolo finale disponibili per ogni batch. Le variabili coinvolte sono differenti se si considerano batch del fermentatore da 300 L o batch del fermentatore da 600 L. In Tabella 3.1 vengono presentate le variabili utilizzate nell'analisi, differenziando fra quelle legate al fermentatore da 300 L e quelle legate al fermentatore da 600 L. Per l'analisi iniziale si considerano solo i batch in specifica, per i quali sono disponibili i dati delle variabili iniziali. Per il fermentatore da 300 L si utilizzano dati di 12 batch per 26 variabili, raccolti nella matrice di dati di processo $\mathbf{X}_{300in}(12\times26)$ e nella matrice di qualità $\mathbf{Y}_{300in}(12\times1)$. Per il fermentatore da 600 L si utilizzano dati di 16 batch per 26 variabili e si costruiscono la matrice di dati di processo $\mathbf{X}_{600in}(16\times26)$ e la matrice di qualità $\mathbf{Y}_{600in}(16\times1)$.

**Tabella 3.1.** *Variabili iniziali utilizzate nell'analisi esplorativa. Sono indicate con ✓ le variabili disponibili nei fermentatori da 300 L e 600 L.*

| N° Variabile | Descrizione | 300 L | 600 L |
|---|---|---|---|
| 1 | N° uova consegnate | ✓ | ✓ |
| 2 | N° uova scartate | ✓ | ✓ |
| 3 | N° uova rotte | ✓ | ✓ |
| 4 | N° uova morte | ✓ | ✓ |
| 5 | N° uova usate | ✓ | ✓ |
| 6 | N° cellule / 1 mL terreno | ✓ | ✓ |
| 7 | N° cellule / embrione | ✓ | ✓ |
| 8 | N° cellule / mL terreno fermentatore | ✓ | ✓ |
| 9 | Volume matrice virus infettante (mL) | ✓ | ✓ |
| 10 | Titolo virus ($TCID_{50}$) / mL matrice | ✓ | ✓ |
| 11 | MOI | ✓ | ✓ |
| 12 | Volume fermentatore (L) | ✓ | |
| 13 | Fornitore 1 | ✓ | ✓ |
| 14 | Fornitore 2 | ✓ | ✓ |
| 15 | Fornitore 3 | | ✓ |
| 16 | Gruppo di incubazione 1 | ✓ | ✓ |
| 17 | Gruppo di incubazione 3 | ✓ | ✓ |
| 18 | Età pollo (settimane) | ✓ | ✓ |
| 19 | Durata incubazione (h) | ✓ | ✓ |
| 20 | Disinfettante 1 | ✓ | ✓ |
| 21 | Disinfettante 2 | ✓ | ✓ |
| 22 | Tempo da speratura a carico camion (h) | ✓ | ✓ |
| 23 | Durata trasporto (h) | ✓ | ✓ |
| 24 | N° uova incubate | ✓ | ✓ |
| 25 | N° uova consegnate in totale | ✓ | ✓ |
| 26 | N° uova non fertili | ✓ | ✓ |

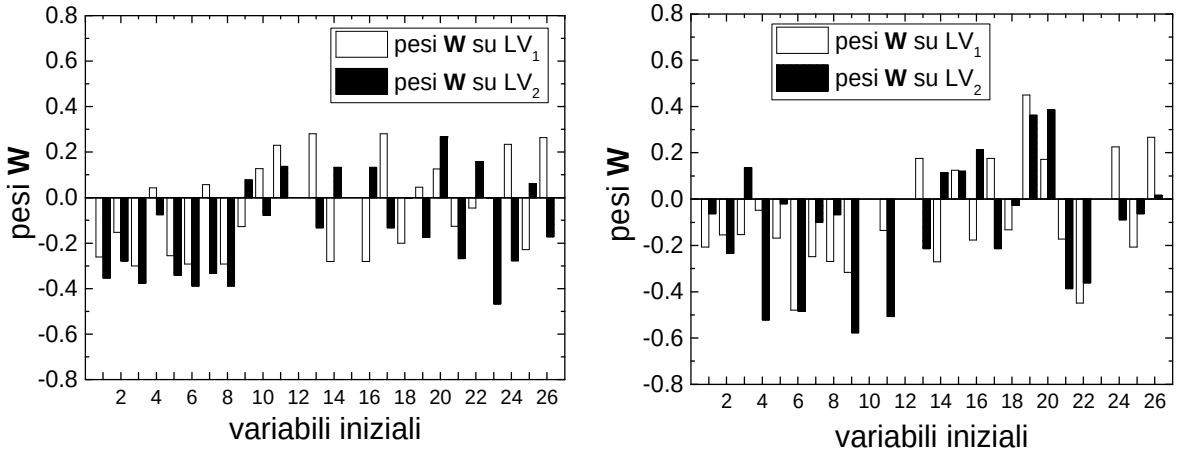
Vengono sviluppati due modelli, entrambi a due variabili latenti, per trattare separatamente i dati appartenenti ai fermentatori da 300 L e da 600 L, i cui risultati in termini di varianza spiegata da $LV_1$ e $LV_2$ sulle matrici di dati di processo e sulle matrici di qualità sono riportati in Tabella 3.2.

**Tabella 3.2.** *Varianza spiegata da $LV_1$ e $LV_2$ sulle matrici dei dati di processo $\mathbf{X}_{300in}$ e $\mathbf{X}_{600in}$, e sulle matrici di qualità $\mathbf{Y}_{300in}$ e $\mathbf{Y}_{600in}$.*

| | $\mathbf{LV_1}$ | $\mathbf{LV_2}$ |
|---|---|---|
| $\mathbf{X}_{300in}$ | 36% | 15% |
| $\mathbf{Y}_{300in}$ | 51% | 17% |
| $\mathbf{X}_{600in}$ | 25% | 13% |
| $\mathbf{Y}_{600in}$ | 31% | 25% |

Dalla Tabella 3.2 si osserva che la varianza spiegata sulla matrice di qualità, cumulata sulle 2 variabili latenti, nel caso del modello costruito con batch del fermentatore da 600 L, non è elevata, attorno al 56%; nel caso del modello che usa i dati dei batch del fermentatore da 300 L, la varianza spiegata cumulativamente sulle 2 variabili latenti è pari a 68%. Quest'ultima percentuale è abbastanza grande, ma si ritiene sia semplicemente causata dall'effetto globale di tutte le variabili. Dal modello PLS si ottengono i *loading* e i pesi **W** su $LV_1$ e $LV_2$, che

danno informazioni analoghe. Pertanto, in Figura 3.2 si riportano i diagrammi dei pesi **W** per i modelli sui due fermentatori. I pesi **W** vengono calcolati per ogni variabile iniziale, nelle 2 variabili latenti.




(a) (b)

**Figura 3.2.** *Diagramma dei pesi* **W** *su $LV_1$ e $LV_2$ (a) per il modello per il fermentatore da 300 L e (b) per il modello per il fermentatore da 600 L.*

La Figura 3.2a e la Figura 3.2b mostrano che i pesi **W** delle variabili iniziali assumono valori molto bassi; infatti, su entrambe le variabili latenti, si rimane su valori di un ordine di grandezza inferiori rispetto ai valori che assumono i pesi **W** ottenuti da un modello PLS costruito sulle variabili di fermentazione (approfondito al §3.3). I risultati ottenuti in termini di qualità di regressione in calibrazione dai modelli per i due reattori sono riportati in Tabella 3.3.

**Tabella 3.3.** *RMSEC e $R^2$ dei modelli PLS per il fermentatore da 300 L e per il fermentatore da 600 L.*

| | **RMSEC** | $R^2$ |
|---|---|---|
| **300 L** | 0.54 | 0.68 |
| **600 L** | 0.64 | 0.56 |

Dalla Tabella 3.3 si nota che l'errore quadratico medio di regressione nella predizione in calibrazione (RMSEC, *root-mean squared error of calibration*) è elevato per entrambi i modelli. Il coefficiente di correlazione multipla ($R^2$), che indica la percentuale di varianza totale della qualità spiegata dall'equazione di regressione, e che quindi è indice della capacità di predizione del modello in calibrazione, è invece basso per entrambi. Sulla base di quanto ottenuto dall'analisi dei pesi **W** e in merito alle prestazioni della regressione, si può quindi affermare che le variabili iniziali sono scarsamente correlate al titolo finale. Sebbene esse non siano indicative per il titolo finale, dalla loro analisi si possono ottenere alcune informazioni.

La Figura 3.2a, considerata unitamente a ciò che emerge da ulteriori analisi statistiche non riportate, mostra che le uova provenienti dal fornitore 1 [13] sono affette da una percentuale di uova non fertili [26] maggiore rispetto alle uova provenienti dal fornitore 2 [14]. La stessa correlazione, nonostante sia meno evidente, si ritrova per il fermentatore da 600 L.
Si decide a questo punto di non considerare le variabili iniziali, in quanto non sufficientemente rappresentative della variabile di qualità, e di sviluppare una modellazione per stimare il titolo virale finale dei batch in specifica che utilizza le variabili della fermentazione.

## 3.3 Predizione del titolo finale dalle variabili di fermentazione per batch in specifica

Dall'analisi sulle variabili iniziali si è visto che non vi sono variabili particolarmente significative per la stima del titolo virale finale. Si passa quindi allo studio delle variabili della fermentazione, sviluppando dei modelli PLS per la predizione del titolo virale finale dei soli batch in specifica, a partire dai dati relativi alle 24 variabili di fermentazione appartenenti alla matrice $\underline{\mathbf{X}}_6$ (Tabella 1.6 al Capitolo 1). I dati di queste variabili sono temporali, in quanto le variabili vengono misurate ogni 30 s.
Lo studio si applica ai soli batch in specifica; pertanto si rende necessario classificare preventivamente i batch in tempo reale come in o fuori specifica. Ad ogni istante *t* di acquisizione delle misure delle variabili di fermentazione si valuta, con dei modelli appositamente sviluppati, se il batch sta evolvendo verso delle condizioni di specifica o di fuori specifica. Se il batch viene classificato in specifica allora si potrà procedere alla stima del titolo virale finale. Tale stima può essere condotta in tempo reale, oppure solamente una volta concluso il batch. Qualora il batch venga classificato come fuori specifica, ciò che interessa maggiormente è la diagnostica, studiando le variabili responsabili del fuori specifica. La logica con cui effettuare la classificazione e la stima del titolo dei batch in specifica è schematizzata in Figura 3.3.

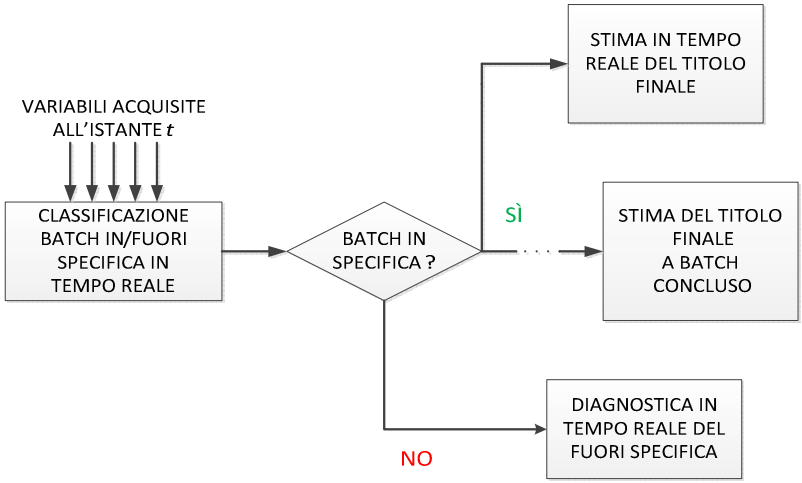



Figura 3.3.vds

**Figura 3.3.** *Logica con cui attuare la stima del titolo finale dei batch in specifica.*

In questa logica si inserisce il lavoro della Tesi, focalizzato sullo sviluppo di modelli PLS dinamici per la stima del titolo finale. Per la stima si sviluppano 2 modelli PLS dinamici, ovvero 2 sensori virtuali, per trattare singolarmente i batch che vengono condotti nel reattore da 300 L e in quello da 600 L. Ciò deriva dal fatto che, essendo prese in esame molte variabili legate al sistema di controllo (variabili controllate, *set point* e *controller output*), processi che avvengono in volumi differenti sono contraddistinti da dinamiche molto diverse e vanno quindi trattati separatamente. Inoltre, è presente nei due fermentatori un diverso sistema di controllo per la temperatura del fermentatore, che introduce un'ulteriore variabilità, difficilmente modellabile. Nei paragrafi seguenti sono presentati nel dettaglio i modelli per i due reattori.

### *3.3.1 Predizione del titolo finale nel fermentatore da 300 L*

Per stimare il titolo finale di batch condotti nel fermentatore da 300 L, si costruisce un modello prendendo in considerazione i dati dei batch conclusisi in specifica. Essi sono 30 batch, ma per creare un *set* di calibrazione che sia il più robusto e omogeneo possibile, vengono eliminati i batch particolarmente diversi dagli altri. Questi batch devono avere almeno una delle seguenti caratteristiche:

- batch con titolo finale eccessivamente alto rispetto alla media;
- batch che presentano anomalie su una o più variabili di fermentazione.

Sulla base di questo criterio, dall'insieme dei 30 batch se ne eliminano 3. Per la costruzione del modello PLS si utilizzano le misure delle 24 variabili di fermentazione ottenute con 1 campionamento ogni 10 min ($\Delta_{camp}$=10 min), considerando in totale 366 istanti temporali, in modo da conservare gran parte dell'informazione temporale, ma anche eliminare molti dati ridondanti. Utilizzando i dati dei 27 batch selezionati, si costruisce la matrice di dati di

processo $\mathbf{X}_{300IS}[27\times(24\cdot 366)]$, trattata con *batch-wise unfolding*. La relativa matrice di qualità è $\mathbf{Y}_{300IS}(27\times 1)$, e contiene il titolo finale del reovirus per ciascun batch selezionato.

### 3.3.1.1 Selezione delle variabili di fermentazione

Inizialmente viene eseguita un'analisi preliminare su tutte le variabili della fermentazione, per individuare le correlazioni tra di esse e con il titolo finale, in modo da selezionare le variabili maggiormente indicative. Per questo viene costruito un modello PLS a due variabili latenti, su tutte le variabili della fermentazione, e cioè a partire dalle matrici $\mathbf{X}_{300IS}$ e $\mathbf{Y}_{300IS}$ precedentemente autoscalate. In Tabella 3.4 è riportata la varianza spiegata da $LV_1$ e $LV_2$ su $\mathbf{X}_{300IS}$ e $\mathbf{Y}_{300IS}$.

**Tabella 3.4.** *Varianza spiegata da $LV_1$ e $LV_2$ sulla matrice dei dati di processo $\mathbf{X}_{300IS}$ e sulla matrice di qualità $\mathbf{Y}_{300IS}$.*

| | $\mathbf{LV_1}$ | $\mathbf{LV_2}$ |
|---|---|---|
| $\mathbf{X}_{300IS}$ | 14% | 3% |
| $\mathbf{Y}_{300IS}$ | 50% | 48% |

Dal modello PLS, si ottengono i *loading* e i pesi **W**, per ciascuna variabile e in ogni istante temporale; essi vengono mediati rispetto al tempo, ottenendo un valore per ciascuna variabile di fermentazione, in ogni variabile latente del modello. In Figura 3.4 è riportato il diagramma dei pesi **W**, pressoché analogo al diagramma dei *loading*.

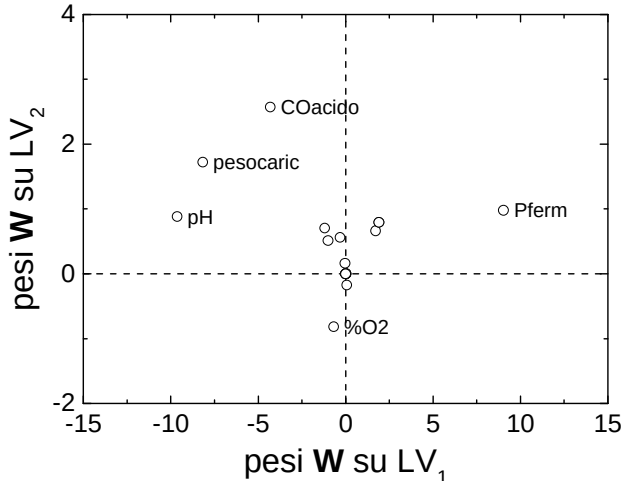



Figura 3.4.opj

**Figura 3.4.** *Diagramma dei pesi* **W***, su $LV_1$ e su $LV_2$, ottenuto dal modello PLS che utilizza le matrici $\mathbf{X}_{300IS}$ e $\mathbf{Y}_{300IS}$.*

Dalla Figura 3.4 si possono ricavare le principali correlazioni tra le variabili di processo:

- le variabili pH, *output* controllore della portata di acido (COacido) e peso della carica del fermentatore (pesocaric) sono correlate su entrambe le variabili latenti; quindi un ipotetico

aumento del pH durante la fermentazione sarebbe legato ad una maggiore richiesta di acido e ad un aumento del peso della carica nel fermentatore;

- il pH, principalmente, ma anche il peso della carica del fermentatore (pesocaric) e l'*output* controllore della portata d'acido (COacido), sono anticorrelati alla pressione del fermentatore (Pferm) sulla prima variabile latente; questo significa che se all'interno del fermentatore si dovesse verificare un aumento di pH, questo sarebbe legato ad una diminuzione della pressione del fermentatore;
- l'*output* controllore della portata di acido (COacido) è anticorrelato all'ossigeno disciolto (%O2) sulla seconda variabile latente; se quindi all'interno del fermentatore dovesse aumentare la richiesta di acido, questo sarebbe legato ad una diminuzione della percentuale di ossigeno disciolto.

Si nota che alcune variabili (velocità dell'agitatore e relativo *set point*, *set point* del pH, *set point* dell'$O_2$ disciolto, *set point* e CO della pressione del fermentatore, *set point* e CO della temperatura della camicia, *output* della valvola di controllo dei gas esausti e i 2 CO relativi alla temperatura di sterilizzazione) hanno peso circa pari a 0 e quindi la loro variabilità non è correlata alla variabilità del titolo finale.

Se si confronta la Figura 3.4 con la Figura 3.2a, si nota che i pesi **W**, nel caso si considerino le variabili della fermentazione piuttosto che quelle iniziali, si attestano su valori ampiamente maggiori. Questo dimostra come le variabili di fermentazione siano maggiormente correlate al titolo virale finale rispetto alle variabili iniziali. I risultati ottenuti confermano che è nello stadio di fermentazione che "viene costruito" il titolo finale del reovirus; pertanto ha senso continuare a sviluppare la modellazione utilizzando le variabili coinvolte in quello stadio.

Per sviluppare il sensore PLS è opportuno selezionare le variabili di fermentazione più significative per la predizione del titolo. Il concetto di predittività è inteso come la variabilità delle variabili di fermentazione correlata alla variabilità del titolo finale. Si fa quindi riferimento all'analisi condotta mediante l'indice VIP (*variable importance in the projection*). Poiché le variabili dei CO presentano un andamento discontinuo nel tempo, per poter riuscire ad estrarre informazioni utili da queste variabili, si considera per l'analisi VIP il loro valore integrale nel tempo. Dal modello PLS costruito su $\mathbf{X}_{300IS}$ (con i CO integrati nel tempo) e $\mathbf{Y}_{300IS}$ si ricava l'indice VIP per ciascuna variabile e in ogni istante di campionamento, come riportato in Figura 3.5. Si noti che l'indice VIP, per ogni variabile, è riportato nei 366 istanti di campionamento.

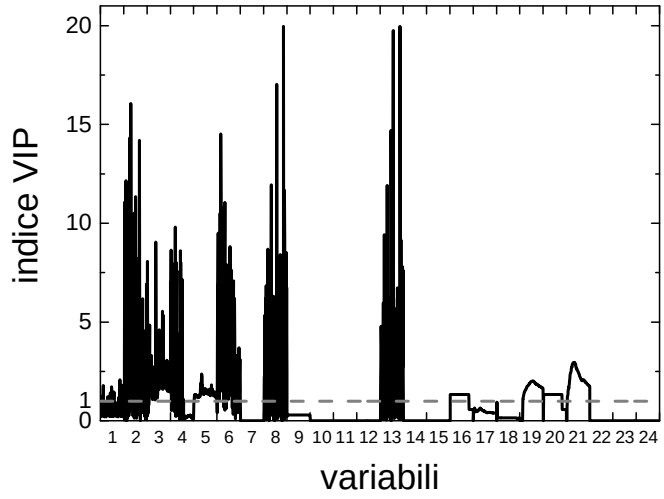



Figura 3.5.opj

**Figura 3.5.** *Indice VIP per ogni variabile di fermentazione, nei 366 istanti di campionamento.*

Dalla Figura 3.5 si può capire che le variabili più importanti per la predizione del titolo sono:

- temperatura del fermentatore [1];
- temperatura della camicia di raffreddamento [2];
- pH [3];
- percentuale di $O_2$ disciolto [4];
- peso della carica del fermentatore [5];
- pressione del fermentatore [6];
- portata d'aria [8];
- *set point* della portata d'aria [13];
- *output* valvola controllo gas esausti [16];
- *output* controllore della portata di acido [19];
- *output* controllore della pressione [20];
- *output* controllore dell'aria [21].

Poiché la Figura 3.5 rappresenta l'evoluzione temporale dell'indice VIP per ogni variabile, si può vedere come il pH risulti essere maggiormente importante nelle ore centrali della fermentazione, mentre la concentrazione di $O_2$ disciolto è una variabile predittiva solo nella prima metà del batch. La pressione, con relativo CO, e la temperatura della camicia sono maggiormente importanti ad inizio batch. Le altre variabili di processo non evidenziano un profilo ben delineato nel tempo e influenzano il titolo circa nello stesso modo per tutto il batch. Per quanto riguarda le variabili di tipo CO, si nota come l'*output* controllore della portata d'acido e l'*output* controllore della portata d'aria siano maggiormente importanti nelle ore centrali del batch. Guardando i valori dell'indice VIP, si può affermare che la portata d'aria e il relativo *set point*, la temperatura della camicia di raffreddamento e la pressione risultano fondamentali per la stima. Fra le variabili di fermentazione individuate con peso

sostanzialmente nullo (Figura 3.4), si conferma che le variabili con pesi **W** molto bassi possiedono anche indice VIP prossimo a 0; esse non sono pertanto importanti nemmeno per quanto riguarda la predizione del titolo. In conclusione, le 12 variabili selezionate con l'analisi dell'indice VIP sono quelle che vengono considerate per lo sviluppo del "Modello$_{300}$".

### 3.3.1.2 Selezione del numero di variabili latenti

Il modello costruito sulle variabili di processo selezionate con l'indice VIP (Modello$_{300}$) viene usato per stimare il titolo. La matrice di dati di processo usata dal Modello$_{300}$ contiene dati di 27 batch e 12 variabili nel tempo, ed è costruita secondo il *batch-wise unfolding*. Il modello PLS viene testato secondo un approccio di tipo *leave one out* (Wold, 1978; Montgomery, 2005). Ciò vuol dire che dei 27 batch considerati, uno viene ciclicamente inserito in convalida; si realizzano così 27 test, nei quali un batch di volta in volta è il batch da convalidare, e tutti gli altri 26 costituiscono il *set* di calibrazione. Anche in questo caso si utilizza un numero di istanti temporali equivalenti a 1 campionamento ogni 10 min ($\Delta_{camp}$=10 min) per considerare i dati delle variabili. La matrice dei dati di processo usata per costruire il Modello$_{300}$ è $\mathbf{X}_{300mod}[27\times(12\cdot366)]$ e la relativa matrice di qualità è $\mathbf{Y}_{300IS}(27\times1)$. Per ciascun batch in convalida viene calcolato l'errore assoluto sulla predizione del titolo virale che compie il modello PLS:

$$\text{errore}_i\left(\text{TCID}_{50}/\text{mL}\right)=\left|y_i-\hat{y}_i\right|, \tag{3.1}$$

dove $y_i$ rappresenta il titolo virale finale per il generico batch di convalida $i$ e $\hat{y}_i$ è il titolo virale finale predetto dal modello PLS. Gli errori vengono poi mediati per ottenere l'errore medio di predizione del titolo sul *set* di convalida, espresso come:

$$\text{errore medio}\left(\text{TCID}_{50}/\text{mL}\right)=\frac{\sum_{i=1}^{I}\left|y_i-\hat{y}_i\right|}{I}, \tag{3.2}$$

dove $I$ è il numero totale dei batch considerati dal modello, e in questo caso è $I_{300IS}$=27. Fondamentale nella modellazione è la scelta del numero di variabili latenti da trattenere nel modello. La Figura 3.6 rappresenta uno studio di sensitività in cui si riporta l'errore medio in funzione del numero di variabili latenti, nel caso in cui il modello sia costruito con tutte le variabili della fermentazione o con le sole variabili derivanti dallo studio dell'indice VIP. Si confronta quindi il Modello$_{300}$ con un modello PLS costruito sulle matrici $\mathbf{X}_{300IS}$ e $\mathbf{Y}_{300IS}$.

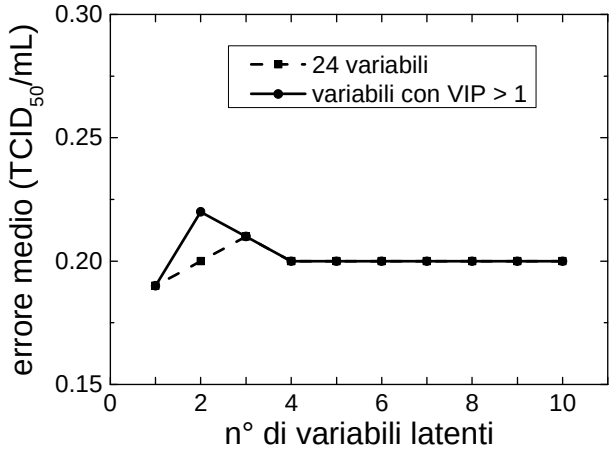



Figura 3.6.opj

**Figura 3.6.** *Sensitività dell'errore medio in funzione del numero di variabili latenti per il Modello$_{300}$ e per un modello PLS che utilizza* $\mathbf{X}_{300IS}$ *e* $\mathbf{Y}_{300IS}$.

Dalla Figura 3.6 si nota come l'errore medio di stima si attesti su 0.20 $TCID_{50}$/mL e quindi sia ampiamente inferiore a 0.36 $TCID_{50}$/mL, indipendentemente dal numero di variabili latenti con cui si costruisce il modello. Considerando le prime 10 variabili latenti, l'errore si dimostra pressoché invariante. Il minimo valore dell'errore si ha per una sola variabile latente scelta. Però, poiché il modello per una variabile latente non è sufficientemente affidabile, si sceglie di utilizzare 2 variabili latenti per lo sviluppo del Modello$_{300}$. Secondo la Figura 3.6, l'errore non varia eccessivamente nemmeno secondo le variabili di processo utilizzate nella costruzione del modello. Confrontando il modello costruito su tutte le variabili di fermentazione con quello costruito sulle sole variabili di fermentazione desunte dall'analisi con l'indice VIP, si decide di utilizzare tutte le 24 variabili di fermentazione e quindi il Modello$_{300}$ sarà costruito utilizzando $\mathbf{X}_{300IS}$ e $\mathbf{Y}_{300IS}$.

### 3.3.1.3 Analisi delle segnalazioni di non rappresentatività

È necessario verificare la validità dei risultati ottenuti dalla predizione PLS in convalida, e questo viene fatto analizzando le statistiche SPE e $T^2$ inerenti alla matrice $\mathbf{X}_{300IS}$. Nello specifico, se $T^2_i > T^2_{\text{Lim}}$ (Equazioni 2.14 e 2.15) o se $\text{SPE}_i > \text{SPE}_{\alpha,\text{Lim}}$ (Equazioni 2.18 e 2.20), il batch non è ben rappresentato dal modello. Le segnalazioni di non rappresentatività sono dunque il numero di batch, in percentuale, per i quali si verifica almeno una di queste condizioni. Il limite di fiducia (1-α) viene posto pari a 99%. I limiti di fiducia per le statistiche SPE e $T^2$ sono costruiti a partire dall'assunzione che gli errori di predizione sul *set* di calibrazione siano disposti secondo una distribuzione di tipo Gaussiano. Facendo uso di un t–test, è possibile affermare che la distribuzione degli errori di calibrazione del Modello$_{300}$ può considerarsi approssimativamente normale. Visto il risultato ottenuto, ha allora senso operare uno studio sulle segnalazioni di non rappresentatività. Utilizzando il Modello$_{300}$, le segnalazioni di non rappresentatività sono pari a 14.8%. Questa percentuale è dovuta

principalmente alla statistica SPE, ed è causata dal fatto che certe variabili di fermentazione presentano, per alcuni batch, dei picchi nel loro andamento temporale; questo fa sì che il batch non sia adeguatamente rappresentato dal modello.
Si prova allora da un punto di vista empirico ad osservare l'effetto della scelta di diversi istanti di campionamento sull'errore medio e sulle segnalazioni di non rappresentatività. In Tabella 3.5 è riportato l'errore medio di predizione e le segnalazioni di non rappresentatività per il $Modello_{300}$, a seconda dell'intervallo di campionamento delle variabili. Il numero di variabili latenti scelte rimane pari a 2, secondo l'analisi precedente.

**Tabella 3.5.** *Errore medio e segnalazioni di non rappresentatività per il $Modello_{300}$ in funzione di 4 diversi intervalli di campionamento delle variabili.*

| | $\Delta_{camp}$=10 min | $\Delta_{camp}$=30 min | $\Delta_{camp}$=60 min | $\Delta_{camp}$=120 min |
|---|---|---|---|---|
| **errore medio** | 0.20 | 0.21 | 0.21 | 0.20 |
| **segnal. non rappr.** | 14.8% | 11.1% | 18.5% | 7.4% |

Dalla Tabella 3.5 si può affermare che l'errore di predizione sostanzialmente non varia con $\Delta_{camp,}$ se si rimane entro $\Delta_{camp}$=120 min. Le segnalazioni di non rappresentatività, invece, diminuiscono con l'aumentare di $\Delta_{camp}$. Ciò è dovuto al fatto che sono presenti dei picchi nel profilo temporale di alcune variabili di fermentazione. Generalmente questo non ha effetto sulle prestazioni della predizione, infatti l'errore è pressoché costante.
Si sceglie il $Modello_{300}$, costruito usando le 24 variabili della fermentazione, 2 variabili latenti e $\Delta_{camp}$=120 min come modello ottimale.

### 3.3.1.4 Predizione del titolo finale con il modello ottimizzato

Con i dati selezionati secondo quanto emerso negli studi precedenti, si costruisce il $Modello_{300}$ ottimizzato con le matrici $\mathbf{X}_{300ott}[27\times(24\cdot30)]$ e $\mathbf{Y}_{300IS}(27\times1)$. In Tabella 3.6 è riportata la varianza spiegata in media dal modello, per ogni variabile latente, su entrambe le matrici.

**Tabella 3.6.** *Valori, medi fra le varie prove, di varianza spiegata dalle diverse variabili latenti per la matrice dei dati di processo $\mathbf{X}_{300ott}$ e per quella di qualità $\mathbf{Y}_{300IS}$, usando il $Modello_{300}$ ottimizzato.*

| | $LV_1$ | $LV_2$ | TOTALE |
|---|---|---|---|
| $\mathbf{X}_{300ott}$ | 15% | 5% | 20% |
| $\mathbf{Y}_{300IS}$ | 49% | 39% | 88% |

Dalla Tabella 3.6 si può notare che, cumulativamente, le 2 variabili latenti riescono a spiegare un'alta percentuale della variabilità dei dati della matrice di qualità $\mathbf{Y}_{300IS}$. Per $\mathbf{X}_{300ott}$ si nota che, cumulativamente, esse riescono a spiegare solo il 20% della variabilità dei dati. La

percentuale è bassa perché l'informazione contenuta in $\mathbf{X}_{300ott}$ correlata a $\mathbf{Y}_{300IS}$ è poca. L'errore di predizione del titolo in calibrazione si attesta su 0.06 $TCID_{50}$/mL e ciò dimostra che il modello di calibrazione è robusto. Sempre considerando i risultati in calibrazione, in Tabella 3.7 è riportato il valore di RMSEC e $R^2$ del $Modello_{300}$ ottimizzato.

**Tabella 3.7.** *RMSEC e $R^2$ del modello ottimizzato per il fermentatore da 300 L.*

| RMSEC | $R^2$ |
|---|---|
| 0.14 | 0.98 |

Dalla Tabella 3.7 si nota che il RMSEC è molto basso, mentre l'indice $R^2$ è molto elevato; questo risultato è sinonimo della bontà di regressione del modello in calibrazione. Per giudicare propriamente le prestazioni del modello in convalida, si effettua una caratterizzazione dell'errore di predizione in convalida del titolo virale finale. Si analizzano 3 aspetti importanti, che devono essere minimizzati:

- errore medio, come riportato nella (3.2);
- percentuale di batch i cui titoli vengono predetti al di sotto del limite di specifica (si ricorda che tutti i batch considerati sono in specifica e come tali dovrebbero essere predetti);
- percentuale di batch i cui titoli vengono predetti con un errore superiore a 0.36 $TCID_{50}$/mL, cioè superiore all'errore di precisione dei test di laboratorio.

I risultati sono riportati in Tabella 3.8.

**Tabella 3.8.** *Caratterizzazione dell'errore in convalida per il $Modello_{300}$ ottimizzato: errore medio, batch predetti fuori specifica e titoli predetti con errore > 0.36 $TCID_{50}$/mL.*

| Errore medio ($TCID_{50}$/mL) | Batch predetti fuori specifica | Titoli predetti con errore > 0.36 |
|---|---|---|
| 0.20 | 0% | 11.1% |

La Tabella 3.8 mostra che l'errore medio, che si attesta su 0.20 $TCID_{50}$/mL, è ampiamente sotto l'errore limite e nessun batch viene erroneamente predetto fuori specifica. Infine, solo una bassa percentuale di titoli viene predetta con un errore superiore al limite di precisione. Pertanto, con il modello sviluppato la predizione del titolo virale alla conclusione del batch, nel fermentatore da 300 L, è possibile.

### 3.3.1.5 Confronto fra modelli PLS lineari e non lineari con il modello ottimizzato

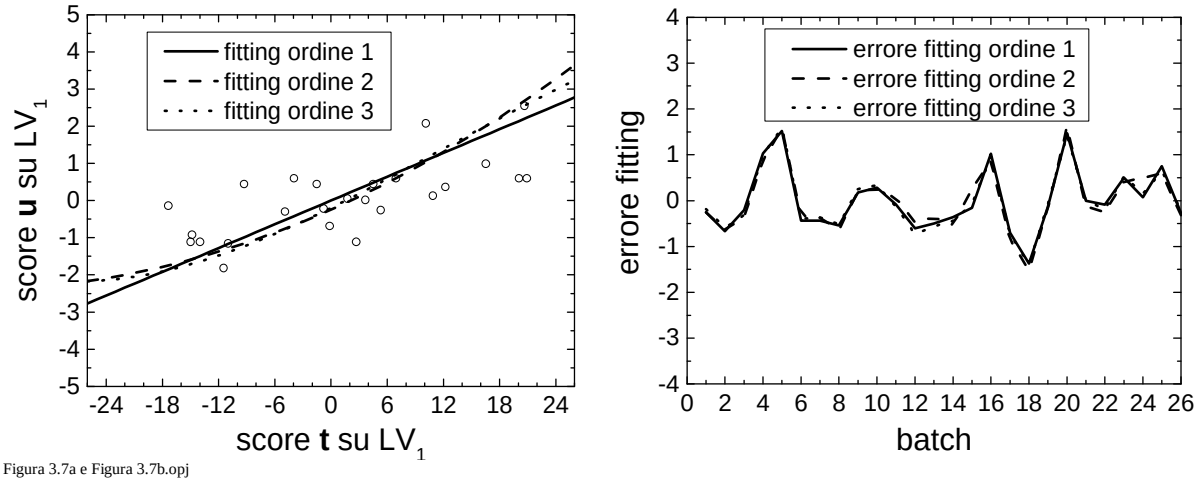
Le analisi fin qui condotte sono state realizzate con un modello PLS che ha una relazione interna di tipo lineare fra *score* **u** e *score* **t**. Si vogliono confrontare le prestazioni del $Modello_{300}$ ottimizzato con modelli sviluppati su dati del tutto analoghi, che però possiedono

una relazione interna di ordine 2 e 3. In Tabella 3.9 si riporta la caratterizzazione dell'errore di predizione in convalida per i 3 modelli.

**Tabella 3.9.** *Caratterizzazione dell'errore in convalida per il Modello$_{300}$ ottimizzato: errore medio, batch predetti fuori specifica e titoli predetti con errore > 0.36 TCID$_{50}$/mL, nel caso di PLS lineare, PLS quadratico e PLS cubico.*

| Tipo di modello | Errore medio (TCID$_{50}$/mL) | Batch predetti fuori specifica | Titoli predetti con errore > 0.36 |
|---|---|---|---|
| **PLS lineare** | 0.20 | 0% | 11.1% |
| **PLS quadratico** | 0.19 | 0% | 11.1% |
| **PLS cubico** | 0.19 | 0% | 14.8% |

Dalla Tabella 3.9, si osserva che un modello PLS non lineare non porta degli evidenti vantaggi. A conferma di ciò viene proposto, in Figura 3.7a, il diagramma degli *score* ottenuto nella convalida del batch n° 19 di $\mathbf{X}_{300ott}$. La Figura 3.7a riporta gli *score* **u** contro gli *score* **t** per la prima variabile latente, che è quella che spiega la maggior parte della varianza contenuta in entrambe le matrici trattate, e inoltre sono presentate le relative curve di *fitting* ottenute dalla regressione con polinomi di grado 1, 2 e 3. In Figura 3.7b è proposto l'errore assoluto di *fitting* che viene compiuto dal metodo di regressione con i tre tipi di polinomio, in relazione a quanto riportato in Figura 3.7a.




(a) (b)

**Figura 3.7.** *Rappresentazione delle caratteristiche della PLS lineare e non lineare: (a) score* **u** *e score* **t** *su LV$_1$ e relative curve di fitting applicando il Modello$_{300}$ ottimizzato al batch n° 19 di* $\mathbf{X}_{300ott}$ *e (b) rappresentazione del relativo errore di fitting.*

La disposizione degli *score* in Figura 3.7a denota chiaramente una forma sufficientemente ben approssimabile con una retta, e infatti le 3 curve di *fitting* non si scostano molto l'una dall'altra.

L'errore di *fitting*, secondo quanto osservabile in Figura 3.7b, sembra non essere sostanzialmente diverso nei 3 casi, e quindi sulla base dei risultati ottenuti si conferma l'utilizzo di un modello PLS lineare.

### 3.3.1.6 Predizione del titolo finale in tempo reale con il modello ottimizzato

Il metodo PLS può essere utilizzato non solo al fine di saper predire accuratamente il titolo virale finale *a batch concluso*, ma anche saper predirlo *in tempo reale*. A questo scopo, viene eseguita la predizione del titolo in tempo reale, secondo l'ottica proposta in Figura 3.3. Sviluppando un modello di questo tipo è possibile conoscere il titolo finale ancora prima della conclusione del batch. Per l'analisi in linea si utilizza il $\text{Modello}_{300}$ ottimizzato, che usa le matrici $\mathbf{X}_{300\text{ott}}$ e $\mathbf{Y}_{300\text{IS}}$. La stima in tempo reale della qualità viene concepita nella logica descritta al §2.2.7.2. Viene quindi costruito il modello di calibrazione con i dati di 26 batch per ogni prova, e un batch alla volta viene convalidato. Per quel batch, ad ogni istante, viene stimato il titolo virale finale, ottenendo un errore di predizione per ogni istante temporale considerato. L'errore di stima al generico istante *t* fa riferimento a quanto riportato nella (2.25), e quindi viene posto come segue:

$$\text{errore}\,(t)\ \left(\text{TCID}_{50}/\text{mL}\right) = \left|\text{y}_i - \hat{\text{y}}_i(t)\right|, \tag{3.3}$$

dove $\hat{\text{y}}_i(t)$ è la stima del titolo finale per il generico batch di convalida *i* all'istante *t*. In Figura 3.8 si riporta l'andamento dell'errore di stima nel tempo di fermentazione per il caso relativo al batch n° 2 di $\mathbf{X}_{300\text{ott}}[27\times(24\cdot30)]$.

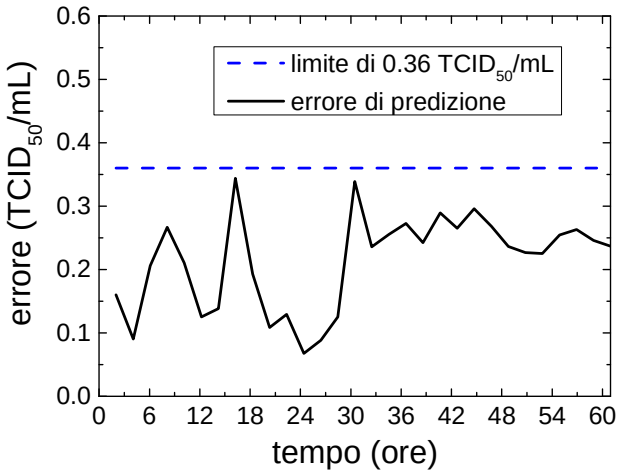



Figura 3.8.opj

**Figura 3.8.** *Andamento nel tempo dell'errore di predizione del titolo finale per il* $\text{Modello}_{300}$*, che utilizza le matrici* $\mathbf{X}_{300\text{ott}}$ *e* $\mathbf{Y}_{300\text{IS}}$ *in riferimento alla convalida del batch n° 2 di* $\mathbf{X}_{300\text{ott}}$.

L'errore assume un profilo fortemente oscillante (ossia incerto) fino alla 30ª ora di fermentazione. Da quel momento, le oscillazioni tendono a smorzarsi e la predizione si mantiene affetta da un errore praticamente costante fino al termine del batch. È quindi possibile, a partire dalla 30ª ora di fermentazione circa, stimare il titolo virale finale, con un errore che si mantiene sempre al di sotto del limite di 0.36 $TCID_{50}$/mL. Questo risultato mostra che la parte della fermentazione maggiormente importante, nella quale viene "costruita" la qualità finale del prodotto, è la prima metà del batch. Da quel momento in poi all'interno del batch non si sviluppano meccanismi che aggiungono informazioni utili al fine di predire il titolo finale. Le considerazioni che si possono trarre dalla Figura 3.8 sono valide anche se si analizzano altri batch di convalida.

## *3.3.2 Predizione del titolo finale nel fermentatore da 600 L*

Ottenuto il modello ottimale per stimare il titolo nel fermentatore da 300 L, si ripropone un'analisi simile anche per il reattore da 600 L, sviluppando un modello di predizione a partire dai dati dei 27 batch in specifica a disposizione. Di questi se ne eliminano 5, secondo il criterio descritto al §3.3.1. Si costruisce quindi la matrice dei dati di processo contenente i dati dei 22 batch selezionati, per 24 variabili campionate con $\Delta_{camp}$=10 min. La matrice è denominata $\mathbf{X}_{600IS}[22\times(24\cdot366)]$, trattata con *batch-wise unfolding*. La relativa matrice di qualità è $\mathbf{Y}_{600IS}(22\times1)$ e contiene il titolo finale del reovirus per ciascun batch.

### 3.3.2.1 Selezione delle variabili di fermentazione

L'analisi condotta per il fermentatore da 300 L al §3.3.1.1 viene svolta in modo analogo sul fermentatore da 600 L. Anche in questo caso si fa uno studio preliminare sui dati dei batch in specifica, raccolti nelle matrici $\mathbf{X}_{600IS}$ e $\mathbf{Y}_{600IS}$, con lo scopo di rintracciare correlazioni fra le variabili di fermentazione, e col titolo. In Tabella 3.10 è riportata la varianza spiegata da $LV_1$ e $LV_2$ sulle matrici $\mathbf{X}_{600IS}$ e $\mathbf{Y}_{600IS}$ con cui è costruito il modello PLS a due variabili latenti.

**Tabella 3.10.** *Varianza spiegata da $LV_1$ e $LV_2$ sulla matrice dei dati di processo $\mathbf{X}_{600IS}$ e sulla matrice di qualità $\mathbf{Y}_{600IS}$.*

| | $\mathbf{LV_1}$ | $\mathbf{LV_2}$ |
|---|---|---|
| $\mathbf{X}_{600IS}$ | 8% | 15% |
| $\mathbf{Y}_{600IS}$ | 79% | 13% |

Le due variabili latenti spiegano cumulativamente il 92% dell'informazione contenuta in $\mathbf{Y}_{600IS}$, quindi una percentuale molto alta. Dal modello si ottengono, ancora una volta, i *loading* e i pesi **W** per ogni istante di campionamento considerato nel modello. Essi riportano approssimativamente le stesse informazioni, pertanto nell'analisi si considerano i pesi **W**,

rappresentati in Figura 3.9. Il diagramma riporta i valori medi nel tempo dei pesi **W** calcolati per ogni variabile di processo, in ciascuna delle 2 variabili latenti.

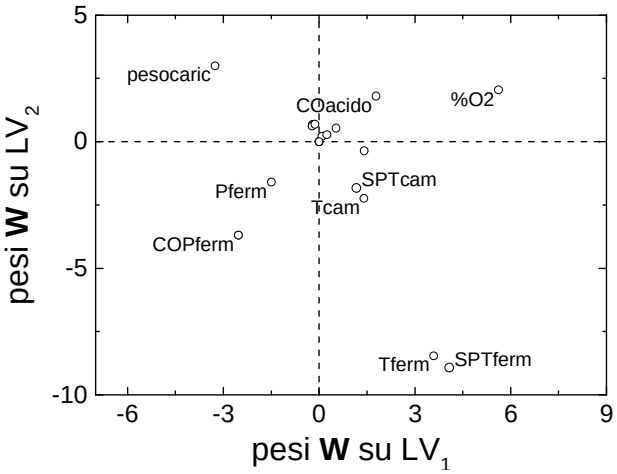



Figura 3.9.opj

**Figura 3.9.** *Rappresentazione del diagramma dei pesi* **W**, *su $LV_1$ e $LV_2$, ottenuto dal modello che utilizza* $\mathbf{X}_{600IS}$ *e* $\mathbf{Y}_{600IS}$.

È possibile osservare alcune correlazioni tra le variabili della fermentazione:

- l'ossigeno disciolto (%O2) e l'*output* del controllore della portata dell'acido (COacido) sono correlati tra loro, mentre sono anticorrelati alla pressione (Pferm) e all'*output* controllore della pressione (COPferm) su entrambe le variabili latenti; questo significa che nel momento in cui all'interno del fermentatore dovesse aumentare la pressione, questo sarebbe legato ad una diminuzione della percentuale di ossigeno disciolto;
- l'ossigeno disciolto (%O2) è anticorrelato alla temperatura del fermentatore (Tferm) e al relativo *set point* (SPTferm) sulla seconda variabile latente; la correlazione evidenzia come se la temperatura diminuisce, la percentuale di ossigeno disciolto aumenta;
- il peso della carica del fermentatore (pesocaric) è anticorrelato alla temperatura della camicia (Tcam) e a quella del fermentatore (Tferm) su entrambe le variabili latenti; la correlazione evidenzia che un aumento del peso della carica immessa nel fermentatore è legato ad una diminuzione della temperatura del fermentatore;
- il peso della carica del fermentatore (pesocaric) è anticorrelato all'*output* controllore della pressione (COPferm) sulla seconda variabile latente; la correlazione mostra che un aumento del peso della carica immessa è legato ad una diminuzione del CO della pressione;
- il peso della carica del fermentatore (pesocaric) è anticorrelato all'ossigeno disciolto (%O2) e al CO della portata d'acido (COacido) sulla prima variabile latente; se il peso della carica immessa aumenta, il CO della portata d'acido diminuisce;
- la temperatura del fermentatore (Tferm), correlata con la temperatura della camicia di raffreddamento (Tcam), è anticorrelata alla pressione (Pferm) e all'*output* controllore della

pressione (COPferm) sulla prima variabile latente; se la temperatura del fermentatore aumenta, la pressione del fermentatore diminuisce.

Dalla Figura 3.9 si nota che sono presenti alcune variabili di fermentazione che risultano poco significative. Esse sono: velocità dell'agitatore e relativo *set point*, *set point* del pH, *set point* dell'$O_2$ disciolto, *set point* della pressione del fermentatore e i 2 CO relativi alla temperatura di sterilizzazione.

Come già si era visto nel fermentatore da 300 L, anche in questo caso si può dire che le variabili di fermentazione sono le variabili rappresentative della variabile di qualità. Fra esse ce ne sono alcune di maggiormente predittive del titolo virale finale, individuabili mediante l'analisi dell'indice VIP. Per l'analisi, le variabili di tipo *controller output* vengono considerate con il loro valore integrale nel tempo, in quanto presentano un profilo temporale caratterizzato da oscillazioni, che renderebbero difficile l'interpretazione dei risultati. In Figura 3.10 si riporta l'indice VIP calcolato per le variabili dello stadio di fermentazione in funzione del tempo di fermentazione.

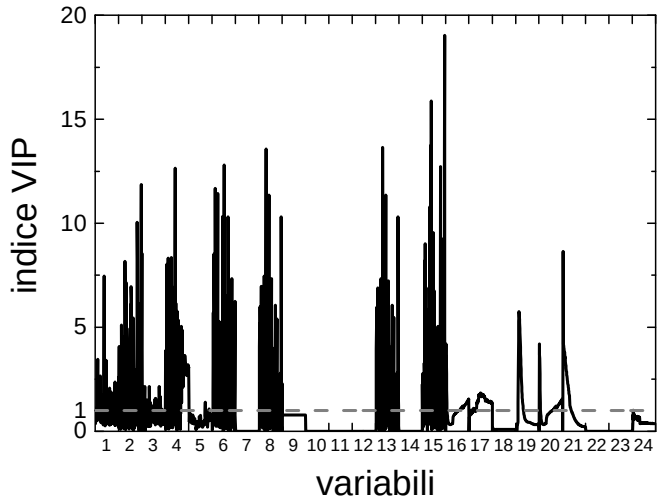



Figura 3.10.opj

**Figura 3.10.** *Indice VIP per le 24 variabili di fermentazione, nei 366 istanti di campionamento.*

Le variabili che possiedono indice VIP > 1 sono:

- temperatura del fermentatore [1];
- temperatura della camicia di raffreddamento [2];
- pH [3];
- percentuale di $O_2$ disciolto [4];
- pressione del fermentatore [6];
- portata d'aria [8];
- *set point* della portata d'aria [13];
- *set point* della temperatura della camicia di raffreddamento [15];

- apertura della valvola di scarico dei gas esausti [16];
- *output* controllore della temperatura del reattore [17];
- *output* controllore della portata di acido [19];
- *output* controllore della pressione [20];
- *output* controllore dell'aria [21].

Poiché per ogni variabile si vede lo sviluppo dell'indice nel tempo, è possibile vedere l'effetto temporale delle variabili sul titolo. La Figura 3.10 mostra che la percentuale di $O_2$ disciolto, pur rimanendo sempre importante, possiede un andamento dell'indice VIP che decresce nel tempo. Questo significa che questa variabile possiede una capacità predittiva che diminuisce nel corso della fermentazione. Le altre variabili di processo predittive non presentano invece un profilo definito e rimangono significative per tutto il batch. La temperatura della camicia, la percentuale di $O_2$ disciolto, la pressione, la portata d'aria, il *set point* dell'aria, il *set point* della temperatura della camicia, il CO dell'aria sono le variabili contraddistinte dai valori più alti dell'indice VIP. Esse sono quindi le variabili che possiedono la capacità predittiva maggiore. Per quanto riguarda l'importanza dei CO al fine di predire il titolo, l'*output* della valvola dei gas esausti, equivalente al CO della pressione, è importante ad inizio e fine batch, mentre l'*output* controllore della temperatura del fermentatore è importante soprattutto da metà batch. Risultano inoltre importanti gli *output* controllori della portata d'acido e dell'aria; l'*output* controllore dell'aria, in particolare, è importante all'inizio del batch. Per il $\text{Modello}_{600}$ si usano quindi le 13 variabili selezionate.

### 3.3.2.2 Selezione del numero di variabili latenti

Viene eseguito uno studio di sensitività per valutare come varia l'errore di predizione medio compiuto dal modello PLS in funzione del numero di variabili latenti selezionate. Il $\text{Modello}_{600}$ viene costruito con i dati di processo appartenenti ai 22 batch in specifica selezionati, 13 variabili di fermentazione considerate con $\Delta_{camp}$=10 min, in quanto si ritiene ragionevole ridurre gli istanti di campionamento. I dati selezionati sono raccolti nella matrice $\mathbf{X}_{600mod}[22\times(13\cdot366)]$, costruita secondo *batch-wise unfolding*. La matrice di qualità è invece $\mathbf{Y}_{600IS}(22\times1)$. Il modello PLS viene valutato analizzando l'errore di stima del titolo secondo un approccio *leave one out* allo stesso modo di quanto eseguito per il modello con i batch del fermentatore da 300 L al §3.2.1.2. L'errore, si ricorda, è un errore medio di predizione del titolo sul *set* di convalida, espresso come nella (3.2). Nel caso specifico il numero totale $I$ di batch considerati è $I_{600IS}$, pari a 22.

Al fine di stabilire il numero di variabili latenti da trattenere nel modello, in Figura 3.11 si riporta lo studio di sensitività dell'errore medio in funzione del numero di variabili latenti. Si confronta il $\text{Modello}_{600}$ con un modello PLS che usa $\mathbf{X}_{600IS}$ e $\mathbf{Y}_{600IS}$, ovvero tutti i dati delle variabili di fermentazione.

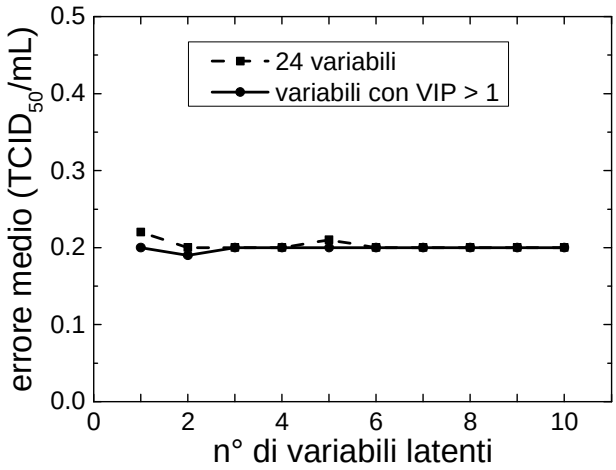



Figura 3.11.opj

**Figura 3.11.** *Studio di sensitività dell'errore medio in funzione del numero di variabili latenti per il Modello$_{600}$ e per un modello PLS su* $\mathbf{X}_{600IS}$ *e* $\mathbf{Y}_{600IS}$.

Considerando le prime 10 variabili latenti, si nota che la differenza nei risultati è poco percepibile. L'errore medio si mostra sostanzialmente invariante rispetto al numero di variabili latenti e si mantiene su valori di circa 0.20 $TCID_{50}$/mL, ampiamente al di sotto del limite di 0.36 $TCID_{50}$/mL. Si decide di utilizzare le variabili ricavate dall'analisi dell'indice VIP, e un numero di variabili latenti pari a 2, dove l'errore medio presenta un leggero valore di minimo.

### 3.3.2.3 Analisi delle segnalazioni di non rappresentatività

I risultati ottenuti dal modello PLS in merito alla predizione del titolo di reovirus, mostrati in Figura 3.11, evidenziano un errore medio di stima che si mantiene circa pari a 0.20 $TCID_{50}$/mL. Per un'analisi completa è necessario interrogarsi sull'effettiva validità dei risultati. Si va quindi ad analizzare le segnalazioni di non rappresentatività, definite come al §3.2.1.3. Gli errori ottenuti in calibrazione, nella maggior parte delle prove eseguite col Modello$_{600}$, si distribuiscono secondo una curva che può ben approssimare una Gaussiana. Ha quindi senso fare uno studio in dettaglio riguardo alle segnalazioni di non rappresentatività.

Utilizzando il Modello$_{600}$, costruito con 2 variabili latenti, le 13 variabili maggiormente predittive e $\Delta_{camp}$=10 min, emerge (Tabella 3.11) che il 18.2% dei batch di $\mathbf{X}_{600mod}$ provoca l'insorgere di segnalazioni di non rappresentatività. Questa percentuale, attribuibile tutta alla statistica SPE, non è eccessivamente alta. Il modello ha quindi una buona capacità di rappresentare i dati di processo. Essendo però il limite di confidenza al 99% si dovrebbe cercare di abbassarla.

La percentuale di rilevazioni in cui la statistica SPE supera il suo limite è causata dal fatto che, per alcuni batch, certe variabili mostrano dei picchi nel loro profilo temporale; questo fa sì che il modello incontri delle difficoltà nella rappresentazione del batch. Si osserva da un punto di vista empirico, l'effetto di diversi $\Delta_{camp}$ selezionati sull'errore medio di stima e sulle

segnalazioni di non rappresentatività. Si confrontano quindi le prestazioni di modelli costruiti con matrici a diversi intervalli di campionamento sulle variabili. In Tabella 3.11 è riportato l'errore medio e le segnalazioni di non rappresentatività in funzione del $\Delta_{camp}$ con cui si selezionano gli istanti di campionamento delle variabili nel modello.

**Tabella 3.11.** *Errore medio e segnalazioni di non rappresentatività per il Modello$_{600}$ in funzione del $\Delta_{camp}$.*

| | **$\Delta_{camp}$=10 min** | **$\Delta_{camp}$=30 min** | **$\Delta_{camp}$=60 min** | **$\Delta_{camp}$=120 min** |
|---|---|---|---|---|
| **errore medio** | 0.20 | 0.20 | 0.20 | 0.27 |
| **segnal. non rappr.** | 18.2% | 22.7% | 27.3% | 31.8% |

Dalla Tabella 3.11 si vede che l'errore medio di predizione sostanzialmente non varia con il $\Delta_{camp}$ e si mantiene su 0.20 $TCID_{50}$/mL, tranne per intervalli di campionamento superiori a 60 min, per i quali l'errore comincia a crescere a causa della perdita di eccessiva informazione contenuta nei dati. All'aumentare del $\Delta_{camp}$ la percentuale delle segnalazioni di non rappresentatività cresce perché vengono perse informazioni importanti per la rappresentazione del batch. La situazione di compromesso ottimale rimane quella che considera $\Delta_{camp}$=10 min. Il Modello$_{600,}$ costruito usando le variabili ottenute dall'analisi dell'indice VIP, 2 variabili latenti e 1 campionamento ogni 10 min, è il modello ottimale sviluppato. Esso viene definito Modello$_{600}$ ottimizzato, utilizzato nelle successive analisi.

### 3.3.2.4 Predizione del titolo finale con il modello ottimizzato

Viene testato il Modello$_{600}$ ottimizzato valutando le sue prestazioni in termini di stima del titolo virale finale. In Tabella 3.12 è riportata la varianza media spiegata per ogni variabile latente, usando il Modello$_{600}$ ottimizzato, per i dati di $\mathbf{X}_{600mod}$ e $\mathbf{Y}_{600IS}$.

**Tabella 3.12.** *Valori medi, fra le varie prove, di varianza spiegata dalle diverse variabili latenti per $\mathbf{X}_{600mod}$ e $\mathbf{Y}_{600IS}$ usando il Modello$_{600}$ ottimizzato.*

| | **$LV_1$** | **$LV_2$** | **TOTALE** |
|---|---|---|---|
| $\mathbf{X}_{600mod}$ | 8% | 15% | 23% |
| $\mathbf{Y}_{600IS}$ | 82% | 11% | 93% |

Dalla Tabella 3.12 si può notare che, cumulativamente, le 2 variabili latenti riescono a spiegare una percentuale molto alta della variabilità dei dati della matrice di qualità. Su $\mathbf{X}_{600mod}$, però, le 2 variabili latenti riescono a spiegare solo il 23% della variabilità dei dati. La percentuale è bassa, perché è poca l'informazione contenuta in $\mathbf{X}_{600mod}$ correlata a $\mathbf{Y}_{600IS}$.
L'errore di predizione del titolo in calibrazione si attesta in media su 0.05 $TCID_{50}$/mL e ciò dimostra che il modello di calibrazione è solido. Sempre considerando i risultati ottenuti in calibrazione, in Tabella 3.13 è riportato il RMSEC e il $R^2$ del Modello$_{600}$ ottimizzato.

**Tabella 3.13.** *RMSEC e $R^2$ del modello ottimizzato per il fermentatore da 600 L.*

| RMSEC | $R^2$ |
|---|---|
| 0.26 | 0.93 |

Il RMSEC risulta piuttosto basso, mentre l'indice $R^2$ risulta elevato; questo risultato fornisce una conferma in merito alla bontà della regressione attuata dal modello in calibrazione. Testando il modello in convalida, i risultati ottenuti per il modello ottimale in termini di errore medio, batch erroneamente predetti fuori specifica e titoli predetti con un errore > 0.36 $TCID_{50}$/mL, sono riportati in Tabella 3.14.

**Tabella 3.14.** *Caratterizzazione dell'errore in convalida per il $Modello_{600}$ ottimizzato; errore medio, batch predetti fuori specifica e titoli predetti con un errore > 0.36 $TCID_{50}$/mL.*

| Errore medio ($TCID_{50}$/mL) | Batch predetti fuori specifica | Titoli predetti con errore > 0.36 |
|---|---|---|
| 0.20 | 0% | 9.1% |

I risultati sono soddisfacenti, similmente a quelli ottenuti per il $Modello_{300}$ mostrati in Tabella 3.4. L'errore medio è ampiamente sotto l'errore limite di precisione, attestandosi su 0.2 $TCID_{50}$/mL. Nessun titolo viene predetto in modo da classificare il batch come fuori specifica e solo una percentuale modesta di titoli (inferiore a 10%) viene predetta con un errore superiore all'errore limite di precisione.

### 3.3.2.5 Confronto fra modelli PLS lineari e non lineari con il modello ottimizzato

La relazione interna tra *score* **u** e *score* **t** del modello PLS è del primo ordine. In questo paragrafo si confrontano le prestazioni del modello PLS ottimale, costruito usando una relazione interna lineare, quadratica e cubica. In Tabella 3.15 è riportato l'errore medio di predizione per i 3 modelli.

**Tabella 3.15.** *Caratterizzazione dell'errore in convalida per il $Modello_{600}$ ottimizzato: errore medio, batch predetti fuori specifica e titoli predetti con un errore > 0.36 $TCID_{50}$/mL, nel caso di PLS lineare, PLS quadratico e PLS cubico.*

| Tipo di modello | Errore medio ($TCID_{50}$/mL) | Batch predetti fuori specifica | Titoli predetti con errore > 0.36 |
|---|---|---|---|
| **PLS lineare** | 0.20 | 0% | 9.1% |
| **PLS quadratica** | 0.20 | 0% | 4.5% |
| **PLS cubica** | 0.20 | 0% | 4.5% |

Dalla Tabella 3.15 si osserva che fra i vari casi non ci sono differenze rilevanti tali da giustificare l'utilizzo di un modello PLS non lineare. Questo risultato era emerso anche

nell'analisi al §3.2.1.5 nella Tabella 3.5 per il $Modello_{300}$ ottimizzato. Non appare quindi sufficientemente vantaggioso adottare una PLS non lineare.

#### 3.3.2.6 Predizione del titolo finale in tempo reale con il modello ottimizzato

Un modello PLS è in grado di predire la qualità non solo a batch concluso, ma anche in tempo reale, se opportunamente costruito, analogamente a quanto realizzato al §3.3.1.6. Si utilizzano i dati con cui si è costruito il $Modello_{600}$ ottimizzato e si sviluppa un modello PLS per la predizione in linea. In Figura 3.12 si riporta il profilo dell'errore di predizione, inteso come dall'Equazione (3.3), in funzione del tempo di fermentazione, per il batch n° 11 di $\mathbf{X}_{600mod}$.

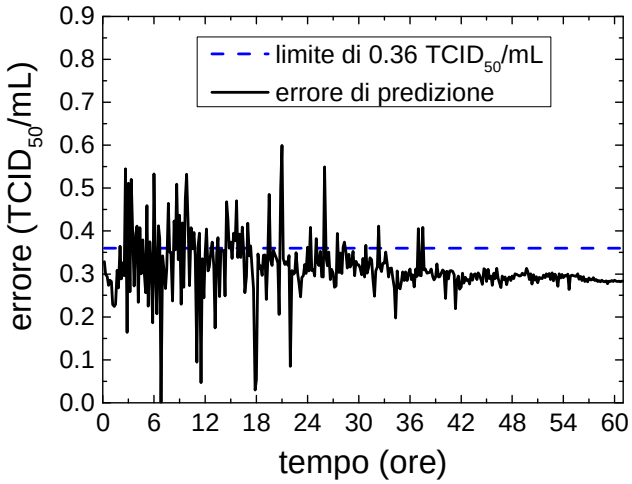



Figura 3.12.opj

**Figura 3.12.** *Errore di predizione nel tempo di fermentazione per il $Modello_{600}$, che utilizza le matrici $\mathbf{X}_{600mod}$ e $\mathbf{Y}_{600IS}$, in riferimento al batch n° 11 di $\mathbf{X}_{600mod}$.*

Dalla Figura 3.12 si nota che la predizione è altamente incerta tra la 3ª e la 35ª ora di fermentazione, in quanto l'errore è soggetto a forti oscillazioni. Questa caratteristica è riscontrabile anche nei profili dell'errore di predizione per gli altri batch di $\mathbf{X}_{600mod}$ con cui si testa il modello. Ciò significa che l'informazione raccolta nelle prime ore di fermentazione non è sufficiente per determinare la qualità finale. A partire dalla 30ª ora circa, l'errore comincia ad essere soggetto ad oscillazioni sempre più smorzate. Dalla 40ª ora in poi, l'errore si attesta intorno ad un unico valore, che non supera più il valore limite di 0.36 $TCID_{50}$/mL. L'esito ottenuto permette di ipotizzare che è nella prima metà della fermentazione che si costruisce il titolo virale finale.

### *3.3.3 Conclusioni sulla predizione del titolo finale con batch in specifica*

L'obiettivo di questa prima parte del lavoro era di sviluppare dei modelli per predire con sufficiente accuratezza il titolo finale di batch in specifica. I modelli sono stati sviluppati per entrambi i fermentatori, analizzando le informazioni derivanti da tutte le variabili di processo.

I risultati ottenuti al §3.2 hanno mostrato che non conviene utilizzare le variabili iniziali per sviluppare i modelli PLS. È stata però trovata una correlazione eseguendo l'analisi preliminare sulle variabili iniziali di processo: le uova provenienti dal fornitore 1 sono affette da una percentuale di uova non fertili maggiore rispetto alle uova provenienti dal fornitore 2. Poiché le variabili iniziali non sono risultate sufficientemente correlate al titolo, si sono considerate per le successive analisi solo le variabili di fermentazione, che sono risultate invece essere le variabili maggiormente rappresentative del titolo. Considerando inizialmente quanto ottenuto dalla stima del titolo a batch concluso, i risultati della predizione del titolo sono molto positivi:

- la stima è accurata, con un errore medio che si attesta ampiamente al di sotto di 0.36 $TCID_{50}$/mL;
- tutti i batch vengono predetti in specifica;
- solo per pochi batch il titolo viene predetto con un errore superiore a 0.36 $TCID_{50}$/mL.

In Tabella 3.16 vengono riassunti i risultati migliori ottenuti in termini di caratterizzazione dell'errore e segnalazioni di non rappresentatività, per i modelli ottimali realizzati nei 2 fermentatori.

**Tabella 3.16.** *Risultati migliori ottenuti per il $Modello_{300}$ ottimizzato e per il $Modello_{600}$ ottimizzato: caratterizzazione dell'errore medio e segnalazioni di non rappresentatività.*

| | **Errore medio ($TCID_{50}$/mL)** | **Batch predetti fuori specifica** | **Titoli predetti con errore > 0.36** | **Segnalazioni di non rappresentatività (SPE)** |
|---|---|---|---|---|
| **$Modello_{300}$ ottimizzato** | 0.20 | 0% | 11.1% | 7.4% |
| **$Modello_{600}$ ottimizzato** | 0.20 | 0% | 9.1% | 18.2% |

Analizzando la predizione del titolo virale finale in tempo reale, si può ipotizzare che entro le prime 30÷35 ore dall'inizio del batch venga costruita la qualità finale del reovirus. Da lì in avanti la stima del titolo finale è possibile con una precisione caratterizzata da un errore di predizione inferiore al valore limite di 0.36 $TCID_{50}$/mL. Questo risultato ottenuto dall'analisi PLS è significativamente importante. Esso infatti trova riscontro nei lavori compiuti da Grande e Benavente (2000) in merito alla biochimica e alla biologia molecolare dei processi che usano reovirus aviari. Gli autori hanno analizzato in dettaglio i processi biochimici della replicazione del reovirus in monostrato, quindi in una condizione diversa dal processo in sospensione che avviene nel fermentatore. In particolare, sono state studiate sperimentalmente le cinetiche di crescita del virus all'interno della fase di replicazione intracellulare, e i risultati ottenuti hanno mostrato che il tempo ottimale che permette di ottenere il massimo valore del titolo virale varia tra le 21 e le 28 h. Essi hanno definito la durata ottimale della fase di fermentazione in cui il virus è a contatto con il monostrato cellulare pari a 24 h. È dunque

all'interno delle prime 24 ore circa che si può ipotizzare avvengano i processi chimici e biologici dello stadio di fermentazione che servono a caratterizzare il titolo finale del reovirus. Il processo considerato negli studi di Grande e Benavente è diverso da quello che è l'oggetto di studio di questa Tesi, però si ritiene possibile che le 24 ore definite dagli studi in merito alla durata della fase di incubazione possano coincidere con le 30÷40 ore che sono emerse dall'analisi PLS sulla stima del titolo virale finale in tempo reale. Per questo motivo si è suggerito all'azienda di programmare una campagna sperimentale per studiare come varia il titolo durante la fermentazione con l'obiettivo di validare i risultati ottenuti ed eventualmente riuscire a ridurre le ore di fermentazione garantendo la specifica di qualità.

## 3.4 Predizione con batch in e fuori specifica

Nel processo di produzione di reovirus non vi sono solo fermentazioni che evolvono in prodotto finale in specifica, bensì anche fuori specifica. È importante essere in grado di predire accuratamente il titolo dei batch in specifica, ma anche di quelli fuori specifica, al fine di poter predire il titolo di tutti i batch, potenzialmente senza ricorrere alla classificazione. In questo Paragrafo vengono presentati i modelli costruiti per la stima, a batch concluso e in tempo reale, del titolo virale finale nel caso di batch in specifica (IS) e fuori specifica (FS). Vengono presi in considerazione i dati temporali delle variabili di fermentazione, appartenenti alla matrice $\underline{\mathbf{X}}_6$ presentata in Tabella 1.6 al Capitolo 1, campionati con $\Delta_{camp}$=10 min. La modellazione PLS dinamica viene realizzata mediante la costruzione di un modello locale. In Figura 3.13 si propone la modalità con cui viene creato il *set* di calibrazione del modello locale.

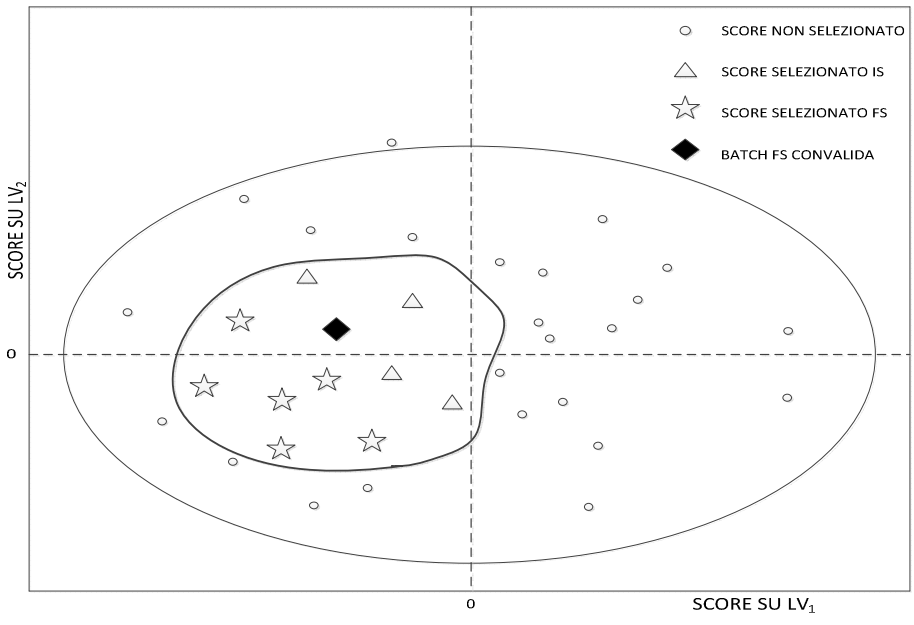



Figura 3.13.vsd

**Figura 3.13.** *Criterio di selezione dei batch più "vicini" al batch di convalida.*

Seguendo la Figura 3.13, il modello locale per la stima viene costruito con i seguenti passi:

- si costruisce il modello PLS dai dati del *set* di calibrazione, costituiti da batch in specifica e batch fuori specifica;
- il batch da convalidare viene proiettato sul diagramma degli *score* **t** ottenuto dal modello;
- vengono isolati sul diagramma degli *score* **t** i 10 batch storici che risultano più "vicini" al batch da convalidare. Questa vicinanza è intesa in termini di distanza euclidea nel piano *N*-dimensionale degli *score*, in cui *N* è pari al rango della matrice dei dati di processo calibrata nel modello;
- si sviluppa il modello locale utilizzando in calibrazione i dati appartenenti ai 10 batch selezionati e si convalida nuovamente lo stesso batch.

La scelta di adottare un modello locale deriva dal fatto che, considerando nello studio anche i batch fuori specifica, che introducono titoli bassi rispetto alla media, l'intervallo di valori assunti dal titolo aumenta inevitabilmente. I batch fuori specifica aggiunti sono in numero minore rispetto a quelli in specifica. Il modello locale utilizza un *set* di calibrazione più specifico perché costituito dai dati dei soli batch del modello maggiormente simili al batch da convalidare. Si riesce così a predire il titolo dei batch con maggiore accuratezza.

In questo studio viene creato un modello locale per il reattore da 600 L, descritto in dettaglio nei paragrafi seguenti. La predizione del titolo viene eseguita sia a batch concluso che in linea.

### *3.4.1 Predizione del titolo finale dei batch del reattore da 600 L*

Viene realizzato un modello locale PLS al fine di predire, a batch concluso e in tempo reale, il titolo finale di tutti i batch appartenenti al reattore da 600 L. I batch considerati sono tutti quelli disponibili, 38 in totale, di cui 27 sono in specifica e 11 sono fuori specifica. Per sviluppare un sensore virtuale robusto è necessario eliminare i dati dei batch particolarmente diversi dagli altri. Per quanto riguarda i batch in specifica si utilizzano quindi le matrici $\mathbf{X}_{600IS}[22\times(24\cdot366)]$ e $\mathbf{Y}_{600IS}(22\times1)$, già selezionate nell'analisi di stima con i batch in specifica. Analoga analisi va fatta per i batch fuori specifica, eliminando un solo batch. Con i dati dei 10 batch fuori specifica selezionati, contenenti le 24 variabili di fermentazione campionate con $\Delta_{camp}$=10 min, si costruiscono le matrici $\mathbf{X}_{600FS}[10\times(24\cdot366)]$ e $\mathbf{Y}_{600FS}(10\times1)$. Per la modellazione PLS vengono concatenate le matrici contenenti i dati dei batch in specifica e fuori specifica, ottenendo la matrice di dati di processo $\mathbf{X}_{600}[32\times(24\cdot366)]$, trattata con *batch-wise unfolding*, e la matrice di qualità $\mathbf{Y}_{600}(32\times1)$.

$$\mathbf{X}_{600}=\begin{bmatrix}\mathbf{X}_{600IS}\\ \mathbf{X}_{600FS}\end{bmatrix},\mathbf{Y}_{600}=\begin{bmatrix}\mathbf{Y}_{600IS}\\ \mathbf{Y}_{600FS}\end{bmatrix}. \tag{3.4}$$

Nella costruzione del modello PLS, le matrici vengono autoscalate su media e varianza globali.

### 3.4.1.1 Selezione delle variabili di fermentazione per un modello sui batch fuori specifica

Dal momento che sono stati inseriti i batch fuori specifica, è necessario effettuare una nuova analisi sulle variabili di fermentazione, perché i batch fuori specifica contengono informazioni che possono cambiare le correlazioni tra le variabili e con il titolo finale. Introducendo i batch fuori specifica, la modellazione viene sviluppata sulle matrici $\mathbf{X}_{600}$ e $\mathbf{Y}_{600}$. Si costruisce con esse un modello PLS a 2 variabili latenti. In Tabella 3.17 è riportata la varianza spiegata da $LV_1$ e $LV_2$ su $\mathbf{X}_{600}$ e su $\mathbf{Y}_{600}$.

**Tabella 3.17.** *Varianza spiegata da $LV_1$ e $LV_2$ sulla matrice dei dati di processo $\mathbf{X}_{600}$ e sulla matrice di qualità $\mathbf{Y}_{600}$.*

| | $\mathbf{LV_1}$ | $\mathbf{LV_2}$ |
|---|---|---|
| $\mathbf{X}_{600}$ | 14% | 7% |
| $\mathbf{Y}_{600}$ | 46% | 39% |

La varianza spiegata, rispetto al caso che considera i soli batch in specifica studiato al §3.3.2.1, è più bassa. Questo significa che l'introduzione dei batch fuori specifica nell'analisi, comporta una maggior variabilità che è difficile da modellare.

Vengono ottenuti dall'analisi i valori dei *loading* e dei pesi **W**, in funzione del tempo di fermentazione. In Figura 3.14 viene riportato il diagramma dei pesi **W** per ogni variabile, che si ottiene mediando i pesi **W** rispetto al numero di istanti temporali.

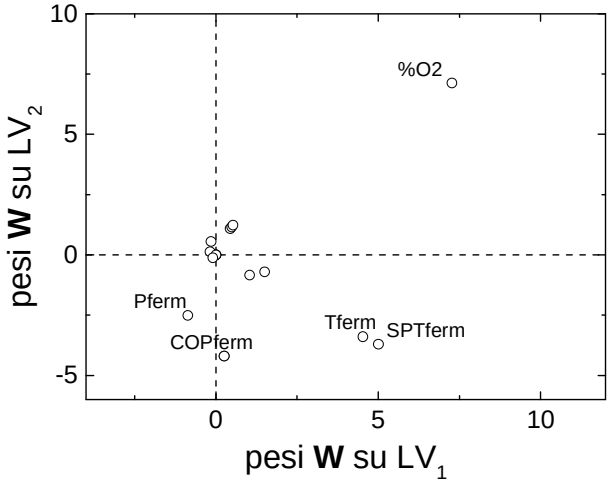



Figura 3.14.opj

**Figura 3.14.** *Diagramma dei pesi* **W** *su $LV_1$ e $LV_2$ ottenuto dal modello che usa le matrici $\mathbf{X}_{600}$ e $\mathbf{Y}_{600}$.*

Le correlazioni osservabili in Figura 3.14, confermate anche guardando il diagramma dei *loading*, sono:

- il CO della pressione (COPferm), equivalente all'apertura della valvola di scarico dei gas esausti (non rappresentato perché sovrapposto), è correlato alla temperatura del fermentatore (Tferm) su entrambe le variabili latenti, e al *set point* della temperatura del fermentatore (SPTferm); se quindi all'interno del fermentatore avvenisse un aumento di temperatura, questo sarebbe collegato ad un aumento dell'apertura della valvola di scarico dei gas esausti;
- l'$O_2$ disciolto (%O2) è anticorrelato alla pressione del fermentatore (Pferm) su entrambe le variabili latenti; si può quindi affermare che un aumento della pressione all'interno del fermentatore è legato ad una diminuzione della percentuale di ossigeno disciolto;
- la pressione del fermentatore (Pferm) è anticorrelata alla temperatura del fermentatore (Tferm) sulla prima variabile latente; se all'interno del fermentatore la temperatura dovesse aumentare, vi sarebbe una diminuzione della pressione;
- la temperatura del fermentatore (Tferm), il *set point* della temperatura del fermentatore (SPTferm) e il CO della pressione (COPferm), sono anticorrelati all'$O_2$ disciolto (%O2) sulla seconda variabile latente; dalla correlazione si evince che se la temperatura del fermentatore dovesse aumentare, questo sarebbe collegato ad una diminuzione della percentuale di ossigeno disciolto.

Le variabili che risultano poco significative dall'analisi sono: velocità dell'agitatore e relativo *set point*, *set point* del pH, *set point* dell'$O_2$ disciolto, *set point* della pressione del fermentatore e i 2 CO relativi alla temperatura di sterilizzazione.

Confrontando la Figura 3.14 con la Figura 3.9, si può affermare che, se si considerano batch in specifica e batch fuori specifica, i valori dei pesi sono inferiori in questo caso, ovvero l'informazione contenuta nelle variabili correlate al titolo è minore rispetto al caso che analizza solo i batch in specifica. Si può individuare anche un cambio di correlazione tra la temperatura del fermentatore e l'*output* controllore della pressione; in questo caso, temperatura del fermentatore e CO di pressione sono correlati.

Osservando la Figura 3.14, si nota che le variabili di fermentazione presentano comunque un peso rilevante sulle 2 variabili latenti e quindi si può affermare che esse sono correlate al titolo. Viene quindi studiato l'indice VIP per capire quali siano le variabili che possiedono un'elevata capacità predittiva del titolo. Viene costruito un modello PLS sulle 24 variabili di fermentazione, utilizzando le matrici $\mathbf{X}_{600}$ e $\mathbf{Y}_{600,}$ conservando 2 variabili latenti. I *controller output*, diversamente dalle altre variabili di fermentazione, vengono considerati con il loro valore integrale nel tempo. In Figura 3.15 è proposto l'indice VIP calcolato in funzione delle 61 ore di fermentazione, per ogni variabile di fermentazione, con un modello PLS che utilizza $\mathbf{X}_{600}$ e $\mathbf{Y}_{600}$.

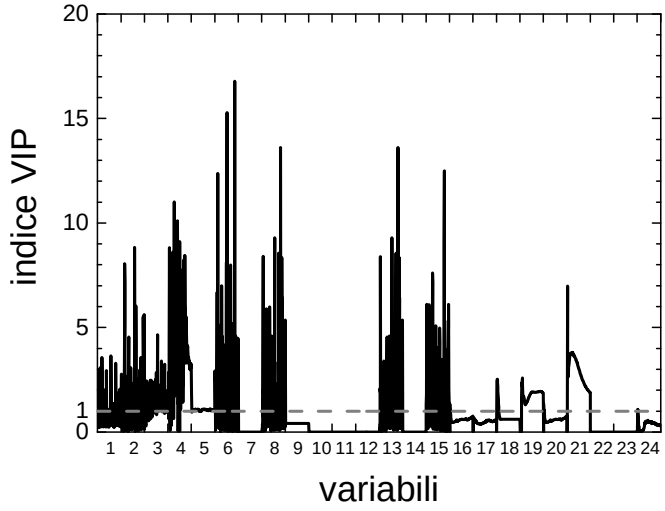



Figura 3.15.opj

**Figura 3.15.** *Indice VIP per ogni variabile di fermentazione, nei 366 istanti di campionamento.*

Le variabili maggiormente predittive sono:

- temperatura del fermentatore [1];
- temperatura della camicia di raffreddamento [2];
- pH [3];
- percentuale di $O_2$ disciolto [4];
- pressione del fermentatore [6];
- portata d'aria [8];
- *set point* della portata d'aria [13];
- *set point* della temperatura della camicia [15];
- *output* controllore della portata di acido [19];
- *output* controllore dell'aria [21].

In particolare, si osserva come il pH sia importante per tutta la durata del batch. La percentuale di $O_2$ disciolto è anch'essa una variabile predittiva molto importante, anche se l'indice VIP decresce nel tempo. Analogamente l'*output* controllore dell'aria risulta predittivo soprattutto all'inizio del batch. Pertanto, le variabili selezionate tramite l'analisi VIP sono 10 e con esse viene sviluppata la modellazione.

### 3.4.1.2 Scelta del numero di variabili latenti per un modello sui batch fuori specifica

Per determinare quale sia il numero ottimale di variabili latenti da scegliere per sviluppare la modellazione, si valuta l'errore medio di predizione del titolo in funzione del numero di variabili latenti, analogamente a quanto fatto al §3.3.2.2. Vengono sviluppati dei modelli PLS sui dati dei batch in e fuori specifica del reattore da 600 L. La matrice dei dati di processo del modello PLS è $\mathbf{X}_{600glob}[32\times(10\cdot366)]$ e la relativa matrice di qualità è $\mathbf{Y}_{600}(32\times1)$.

Un passo fondamentale dello studio è determinare il numero di variabili latenti da usare nello sviluppo del modello locale per la stima del titolo. Dovendo costruire 2 modelli, uno con tutti i dati e il successivo modello locale, è necessario determinare il numero ottimale di variabili latenti per entrambi. In Figura 3.16 si riporta quindi lo studio di sensitività in cui si valuta l'errore medio di predizione del titolo in funzione del numero di variabili latenti, usando il modello PLS che utilizza le matrici $\mathbf{X}_{600glob}$ e $\mathbf{Y}_{600}$.

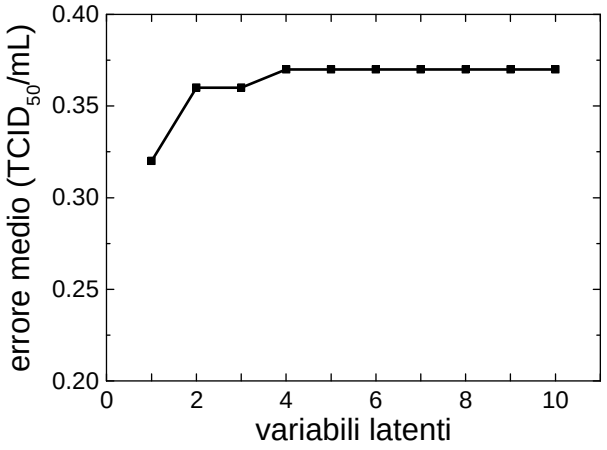



Figura 3.16.opj

**Figura 3.16.** *Errore medio in funzione del numero di variabili latenti per il modello PLS che utilizza* $\mathbf{X}_{600glob}$ *e* $\mathbf{Y}_{600}$.

L'errore medio si mantiene su valori elevati, intorno al valore limite di 0.36 $TCID_{50}$/mL. Il minimo si trova in prossimità di 1 variabile latente, ma il dato non è affidabile. La Figura 3.16 suggerisce di scegliere 2 o 3 variabili latenti. Si sceglie pertanto di confrontare queste 2 ipotesi dal punto di vista della varianza spiegata mediamente su $\mathbf{Y}_{600}$. In Tabella 3.18 si riporta la varianza spiegata in media da ciascuna variabile latente su $\mathbf{Y}_{600}$, e la relativa varianza cumulativa.

**Tabella 3.18.** *Valori medi di varianza spiegata dalle diverse variabili latenti su* $\mathbf{Y}_{600}$ *e relativa varianza cumulativa per il modello PLS che utilizza* $\mathbf{X}_{600glob}$ *e* $\mathbf{Y}_{600}$.

| | $\mathbf{LV_1}$ | $\mathbf{LV_2}$ | $\mathbf{LV_3}$ | **TOTALE** |
|---|---|---|---|---|
| **2 LV** | 45% | 44% | / | 89% |
| **3 LV** | 44% | 46% | 8% | 98% |

Si scelgono 3 variabili latenti per il modello PLS che usa $\mathbf{X}_{600glob}$ e $\mathbf{Y}_{600}$, perché la varianza cumulata spiegata è elevata pur rimanendo al di sotto di 99%, evitando così di descrivere del rumore associato ai dati.

Scelto il numero ottimale di variabili latenti per il modello costruito su tutti i dati disponibili, si esegue un'analisi del tutto analoga per il $Modello_{600}$ locale. In Figura 3.17 si presenta lo studio di sensitività che riporta l'errore medio di predizione in funzione del numero di

variabili latenti scelte, in riferimento al Modello$_{600}$ locale, costruito sui dati del fermentatore da 600 L, selezionati da $\mathbf{X}_{600glob}$ e $\mathbf{Y}_{600}$.

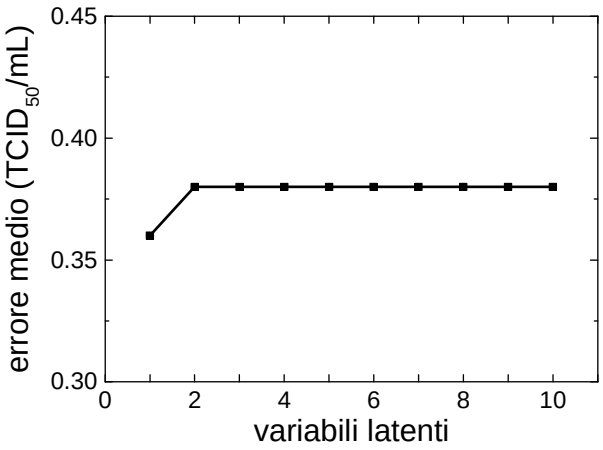



Figura 3.17.opj

**Figura 3.17.** *Errore medio in funzione del numero di variabili latenti per il Modello$_{600}$ locale costruito sui dati dei batch selezionati da* $\mathbf{X}_{600glob}$ *e* $\mathbf{Y}_{600}$.

Si nota che l'errore medio si mantiene ancora intorno a valori superiori all'indice di precisione di 0.36 $TCID_{50}$/mL. Rispetto a quanto si può osservare dalla Figura 3.16, l'errore medio in questo caso è più elevato, però l'ottimizzazione di modello effettuata nelle successive analisi permetterà di ottenere risultati di stima migliori con il modello locale. Il dato per 1 variabile latente selezionata non è affidabile, mentre si ritiene ragionevole scegliere 2 o 3 variabili latenti. A supporto dell'analisi si sceglie di studiare la varianza media spiegata sulla matrice di qualità da ciascuna variabile latente e la varianza cumulativa associata, confrontando il Modello$_{600}$ locale a 2 e 3 variabili latenti, in Tabella 3.19.

**Tabella 3.19.** *Valori medi di varianza spiegata dalle diverse variabili latenti sulla matrice di qualità e relativa varianza cumulativa per il Modello$_{600}$ locale.*

| | **$LV_1$** | **$LV_2$** | **$LV_3$** | **TOTALE** |
|---|---|---|---|---|
| **2 LV** | 75% | 23% | / | 98% |
| **3 LV** | 74% | 25% | 1% | > 99% |

In base a quanto emerso in Tabella 3.19, la varianza spiegata cumulativamente nel caso si scelgano 3 variabili latenti è superiore al 99%; questo significa che il modello descrive anche il rumore associato ai dati. Si sceglie pertanto di utilizzare 2 variabili latenti per il Modello$_{600}$ locale.

### 3.4.1.3 Prestazioni del modello locale

Viene sviluppato il Modello$_{600}$ locale, a 2 variabili latenti, per valutare le sue prestazioni in termini di caratterizzazione dell'errore medio di predizione, definito secondo la (3.2). Lo

studio relativo all'errore di predizione medio ottenuto in convalida viene condotto prendendo in considerazione 3 aspetti:

- il valore dell'errore di predizione, mediato sull'intero *set* di convalida, comprendente batch in specifica e batch fuori specifica;
- i titoli predetti in modo sbagliato, cioè la percentuale di batch fuori specifica il cui titolo viene erroneamente predetto in specifica, e la percentuale di batch in specifica il cui titolo viene erroneamente predetto fuori specifica;
- la percentuale di batch il cui titolo viene predetto con un errore superiore al limite pari a 0.36 $TCID_{50}$/mL.

I risultati forniti dal modello vanno convalidati, e a tal proposito si effettua uno studio che considera le statistiche SPE e $T^2$, in convalida, al fine di rilevare la percentuale di segnalazioni di non rappresentatività. Il limite di fiducia $(1 - \alpha)$ viene posto pari al 99%.

Inizialmente si utilizzano nel modello le variabili ricavate dall'analisi dell'indice VIP. Successivamente si selezionano empiricamente diverse combinazioni delle 5 variabili ritenute più importanti all'interno dello stadio di fermentazione, e maggiormente correlate al titolo finale, cioè pH, percentuale di $O_2$ disciolto, pressione del fermentatore, portata d'aria e *set point* della temperatura della camicia di raffreddamento. Questa operazione viene effettuata con l'obiettivo di creare un modello specifico, con una matrice di dati rappresentativa dei dati di processo, al fine di predire con accuratezza il titolo e avere possibilmente una bassa percentuale relativa alle segnalazioni di non rappresentatività. Nella Tabella 3.20 si riportano i risultati ottenuti dallo studio di caratterizzazione dell'errore di predizione e dall'analisi delle segnalazioni di non rappresentatività per il $Modello_{600}$ locale, costruito utilizzando le variabili con indice VIP superiore a 1, o con diverse combinazioni delle 5 variabili giudicate più significative nella fermentazione.

**Tabella 3.20.** *Caratterizzazione dell'errore di predizione e segnalazioni di non rappresentatività del $Modello_{600}$ locale utilizzando le variabili con indice VIP > 1 o diverse combinazioni delle 5 variabili giudicate più significative.*

| | **Errore medio ($TCID_{50}$/mL)** | **Batch predetti in modo sbagliato** | **Titoli predetti con errore > 0.36** | **Segnalazioni di non rappresentatività (SPE)** |
|---|---|---|---|---|
| **variabili VIP > 1** | 0.38 | 53.1% | 43.8% | 15.6% |
| **pH,$O_2$,P,Aria,$SPT_{cam}$** | 0.35 | 40.6% | 43.8% | 15.6% |
| **pH,$O_2$,P,Aria** | 0.31 | 37.5% | 40.6% | 28.1% |
| **$O_2$,P,Aria,$SPT_{cam}$** | 0.36 | 37.5% | 40.6% | 21.9% |
| **pH,$O_2$,Aria** | 0.35 | 31.3% | 37.5% | 25% |
| **pH,$O_2$,P** | 0.35 | 37.5% | 43.8% | 31.3% |
| **$O_2$,P,Aria** | 0.34 | 25% | 40.6% | 43.8% |

L'errore medio di predizione non si discosta molto dal valore dell'errore limite in ogni caso; la causa è da imputare alla stima del titolo per i batch fuori specifica, che risulta piuttosto

difficile. Appare chiaro che è necessario trovare una soluzione di compromesso fra i 4 aspetti analizzati per scegliere il modello PLS ottimale. Il $Modello_{600}$ locale ottimizzato è quello costruito con le variabili pH, percentuale di $O_2$ disciolto e portata d'aria insufflata. Per questo modello, l'errore medio di predizione si attesta, anche se leggermente, al di sotto del limite di 0.36 $TCID_{50}$/mL e una percentuale dei titoli del 30% viene predetta in modo errato. Si sottolinea che il risultato ottenuto non è negativo, in quanto la percentuale di batch fuori specifica il cui titolo viene predetto in specifica è solo del 30%. Meno del 40% dei titoli viene predetta con un errore superiore al limite e poco più del 20% dei batch è affetto da segnalazione di non rappresentatività. I risultati sono i migliori ottenibili con i dati che si hanno a disposizione.

### 3.4.1.4 Predizione del titolo finale in tempo reale con il modello locale

Come già visto per i batch in specifica al §3.3.2.6, la stima della qualità finale con il metodo PLS può essere eseguita anche in tempo reale. In questo modo si è in grado di conoscere con sufficiente precisione il valore del titolo finale del reovirus prima della fine del batch, durante la fermentazione stessa. Applicando il modello locale in linea per il batch in specifica n° 10 di $\mathbf{X}_{600glob}$ e per il batch fuori specifica n° 26 di $\mathbf{X}_{600glob}$, si calcola l'errore (3.3) in funzione del tempo di fermentazione, che viene riportato in Figura 3.18.

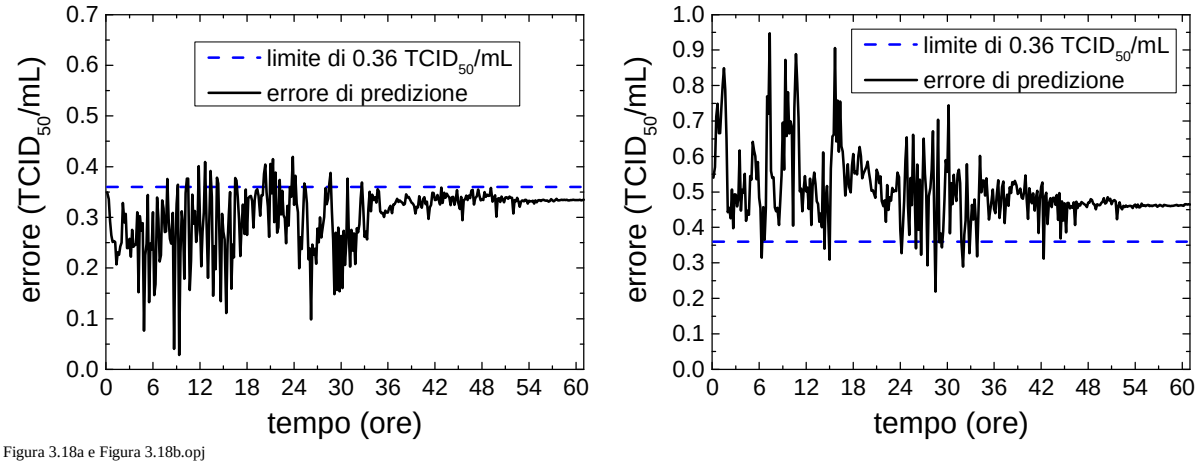



Figura 3.18a e Figura 3.18b.opj

(a) (b)

**Figura 3.18.** *Andamento nel tempo dell'errore di predizione del titolo finale per il $Modello_{600}$ locale, in riferimento (a) al batch in specifica n° 10 di $\mathbf{X}_{600glob}$ e (b) al batch fuori specifica n° 26 di $\mathbf{X}_{600glob}$.*

Volendo confrontare la Figura 3.18a con la Figura 3.18b, si nota innanzitutto che, per il batch fuori specifica, l'errore si mantiene su valori superiori per tutta la durata del batch; in particolare, l'errore è sempre superiore a 0.36 $TCID_{50}$/mL, anche alla fine del batch. Questa caratteristica nell'andamento dell'errore si osserva anche per gli altri batch fuori specifica considerati. Si ricorda comunque che il risultato è accettabile, in quanto, nonostante l'errore

sia superiore all'indice di precisione delle misure sperimentali, esso non è tale da portare a predire il batch come in specifica, salvo un 30% dei casi.
In Figura 3.18a si vede che l'errore, fino a circa la 35ª ora, è fortemente caratterizzato da ampie oscillazioni. Dalla 35ª ora circa in poi, l'errore tende a stabilizzarsi. Anche guardando la Figura 3.18b, l'errore è caratterizzato da forti oscillazioni, che si manifestano fino alla 40ª ora circa; da quel momento, l'errore tende ad assestarsi.
Si può quindi ipotizzare che, anche nel caso si considerino i batch fuori specifica, sono importanti le prime 30÷40 ore di fermentazione, in cui avvengono i processi necessari affinché venga "costruita" la qualità del titolo finale del reovirus. Da quel momento in avanti il titolo finale potrebbe essere pressoché già determinato; la stima diventa quindi possibile, con un'accuratezza migliore nel caso di batch in specifica.

### *3.4.2 Predizione del titolo finale dei batch del reattore da 300 L*

Lo sviluppo di un modello per la stima del titolo virale finale, considerando i dati dei batch in e fuori specifica, può essere fatto anche per i dati appartenenti al fermentatore da 300 L. La modellazione viene eseguita in modo del tutto analogo a quanto riportato al §3.4.1. I batch che vengono considerati sono tutti quelli disponibili, 37 in totale. Di essi 30 sono in specifica e 7 sono fuori specifica. Per sviluppare un sensore virtuale è necessario costruire un solido *set* di calibrazione e quindi eliminare i dati dei batch particolarmente diversi dagli altri. Per quanto riguarda i batch in specifica si utilizzano le matrici $\mathbf{X}_{300IS}[27\times(24\cdot366)]$ e $\mathbf{Y}_{300IS}(27\times1)$, già utilizzate nell'analisi di stima con i batch in specifica. Un'analoga analisi va fatta per i batch fuori specifica e si eliminano 3 batch. Con i dati dei 4 batch fuori specifica selezionati, contenenti le 24 variabili di fermentazione campionate con $\Delta_{camp}$=10 min, si costruiscono le matrici $\mathbf{X}_{300FS}[4\times(24\cdot366)]$ e $\mathbf{Y}_{300FS}(4\times1)$. Per la modellazione PLS vengono concatenate le matrici contenenti i dati dei batch in specifica e fuori specifica, ottenendo la matrice di dati di processo $\mathbf{X}_{300}[31\times(24\cdot366)]$, trattata con *batch-wise unfolding*, e la matrice di qualità $\mathbf{Y}_{300}(31\times1)$.

$$\mathbf{X}_{300}=\begin{bmatrix}\mathbf{X}_{300IS}\\ \mathbf{X}_{300FS}\end{bmatrix},\ \mathbf{Y}_{300}=\begin{bmatrix}\mathbf{Y}_{300IS}\\ \mathbf{Y}_{300FS}\end{bmatrix}. \tag{3.5}$$

Nella costruzione del modello PLS, si ricorda, le matrici vengono autoscalate su media e varianza globali.
In primo luogo, si esegue un'analisi preliminare realizzando un modello PLS al fine di trovare le correlazioni fra le variabili, e fra le variabili e il titolo. La modellazione viene effettuata selezionando 2 variabili latenti. In Tabella 3.21 è riportata la varianza spiegata da $LV_1$ e $LV_2$ su $\mathbf{Y}_{300}$.

**Tabella 3.21.** *Varianza spiegata da $LV_1$ e $LV_2$ sulla matrice di qualità* $\mathbf{Y}_{300}$.

| $\mathbf{LV_1}$ | $\mathbf{LV_2}$ |
|---|---|
| 55% | 38% |

Si calcolano i *loading* e i pesi **W**, ottenendo un valore per ogni istante di tempo. Il calcolo si esegue per ogni variabile di fermentazione, in ognuna delle 2 variabili latenti. Utilizzando i pesi **W**, con il loro valor medio rispetto al tempo, si realizza il diagramma dei pesi **W**, mostrato in Figura 3.19, ottenuto col modello PLS costruito su $\mathbf{X}_{300}$ e $\mathbf{Y}_{300}$.

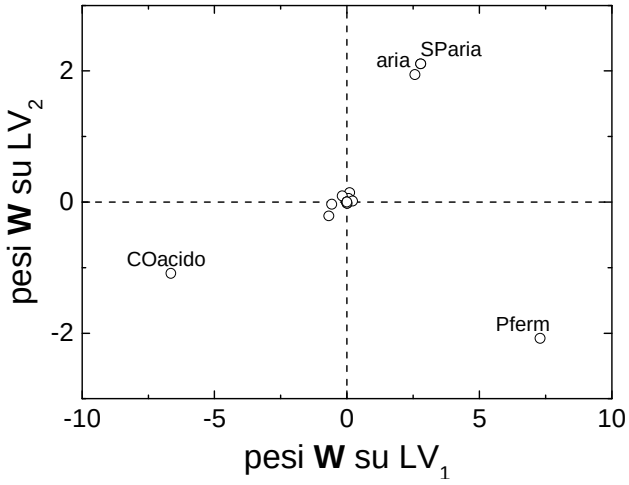



Figura 3.19.opj

**Figura 3.19.** *Diagramma dei pesi* **W** *su $LV_1$ e $LV_2$ ottenuto dal modello che utilizza* $\mathbf{X}_{300}$ *e* $\mathbf{Y}_{300}$.

Dal diagramma dei pesi **W** (Figura 3.19), anche se le stesse considerazioni si possono trarre anche dal diagramma dei *loading*, si notano le correlazioni:

- la portata d'aria (aria) e il relativo *set point* (SParia), sono anticorrelate all'*output* della portata di acido (COacido) su entrambe le variabili latenti; un possibile aumento della portata d'aria insufflata all'interno del fermentatore è quindi legato ad una diminuzione del CO della portata d'acido;
- la pressione del fermentatore (Pferm) è anticorrelata alla portata d'aria (aria) sulla seconda variabile latente; un aumento della pressione all'interno del fermentatore è legato ad una diminuzione della portata d'aria immessa;
- il CO della portata d'acido (COacido) è anticorrelato alla pressione del fermentatore (Pferm) sulla prima variabile latente; la correlazione permette di osservare che se il CO della portata d'acido aumenta, la pressione del reattore diminuisce.

Confrontando ciò che si osserva in Figura 3.19 con la Figura 3.4, in cui un'analoga analisi veniva fatta considerando solo i batch in specifica, si nota che le correlazioni principali si mantengono. Le relazioni di processo trovate fra le variabili nei due casi sono quindi le medesime. Poiché i valori dei pesi delle variabili di fermentazione su $LV_1$ e $LV_2$ sono

abbastanza elevati, si può dire che c'è una certa correlazione al titolo. Per quantificare tale correlazione si fa riferimento all'analisi dell'indice VIP. Pertanto, in Figura 3.20, si riporta l'indice VIP calcolato in funzione del tempo di fermentazione, per ogni variabile di fermentazione, col modello a 2 variabili latenti che utilizza le matrici $\mathbf{X}_{300}$ e $\mathbf{Y}_{300}$. Si ricorda che per quanto riguarda le variabili CO, viene considerato il loro valore integrale nel tempo.

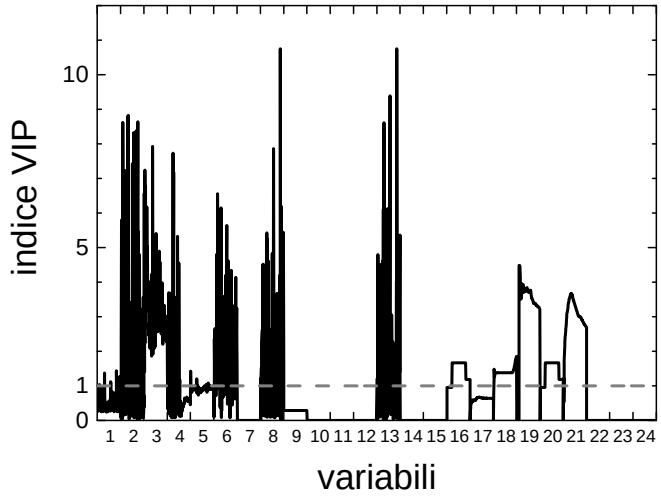



Figura 3.20.opj

**Figura 3.20.** *Indice VIP per ogni variabile di fermentazione, nei 366 istanti di campionamento.*

La Figura 3.20 evidenzia le variabili con VIP superiore a 1:

- temperatura della camicia di raffreddamento [2];
- pH [3];
- percentuale di $O_2$ disciolto [4];
- pressione del fermentatore [6];
- portata d'aria [8];
- *set point* della portata d'aria [13];
- *output* valvola controllo gas esausti [16];
- *output* controllore della portata di base [18];
- *output* controllore della portata di acido [19];
- *output* controllore della pressione [20];
- *output* controllore dell'aria [21].

Vengono selezionate queste 11 variabili per sviluppare la modellazione successiva.

Utilizzando per la modellazione PLS i dati appartenenti ai batch in e fuori specifica del fermentatore da 300 L, conviene non affidarsi ad un modello di tipo locale per avere dei risultati soddisfacenti. Questa considerazione viene motivata dal fatto che i batch fuori specifica disponibili per il fermentatore da 300 L sono pochi, quindi non risulta ottimizzata la selezione dei batch per il modello locale, che non migliora la rappresentazione del batch. Se si

prendono in considerazione invece i batch appartenenti al fermentatore da 600 L, essendo il numero di batch fuori specifica superiore, il modello locale permette di ottenere risultati migliori.

Analogamente a quanto realizzato al §3.4.1.3, si costruiscono le matrici $\mathbf{X}_{300glob}[31\times(11\cdot366)]$ e $\mathbf{Y}_{300}(31\times1)$. Con esse viene sviluppato il modello PLS a 3 variabili latenti i cui risultati ottenuti, in merito a caratterizzazione dell'errore di predizione e analisi delle segnalazioni di non rappresentatività, vengono riportati in Tabella 3.21. Vengono presentati diversi casi, a seconda che si utilizzino tutte le variabili ricavate dall'analisi dell'indice VIP, o varie combinazioni delle 5 variabili ritenute più importanti e rappresentative del titolo. Le variabili maggiormente significative che sono state individuate sono: temperatura della camicia di raffreddamento, pH, percentuale di $O_2$ disciolto, portata d'aria insufflata, *controller output* della portata d'acido.

**Tabella 3.21.** *Caratterizzazione dell'errore di predizione e segnalazioni di non rappresentatività per il modello PLS costruito con* $\mathbf{X}_{300glob}$ *e* $\mathbf{Y}_{300}$*, utilizzando le variabili ricavate dall'analisi dell'indice VIP o varie combinazioni delle 5 variabili di fermentazione ritenute più significative.*

| | **Errore medio ($TCID_{50}$/mL)** | **Batch fuori specifica predetti in specifica** | **Titoli predetti con errore > 0.36** | **Segnalazioni di non rappresentatività (SPE)** |
|---|---|---|---|---|
| **variabili VIP > 1** | 0.24 | 100% | 25.8% | 29% |
| **pH,$O_2$,$T_{cam}$,Aria,$CO_{acido}$** | 0.23 | 100% | 16.1% | 32.3% |
| **pH,$O_2$,$T_{cam}$,Aria** | 0.23 | 100% | 22.6% | 19.4% |
| **pH,$O_2$,$T_{cam}$** | 0.25 | 75% | 22.6% | 19.4% |
| **pH,$O_2$,Aria** | 0.22 | 50% | 19.4% | 25.8% |

Osservando la Tabella 3.21, si capisce che è necessario trovare una soluzione di compromesso fra i 4 aspetti analizzati per scegliere il modello PLS ottimale. Il modello PLS costruito utilizzando come variabili pH, percentuale di $O_2$ disciolto e portata d'aria insufflata, risulta essere quello con prestazioni migliori. L'errore medio di predizione si attesta ben al di sotto del limite di 0.36 $TCID_{50}$/mL, una percentuale minore del 20% di titoli viene predetta con un errore superiore al limite, e solo poco più del 20% dei batch è affetto da segnalazione di non rappresentatività. Infine, aspetto più determinante, metà dei titoli dei batch fuori specifica vengono predetti in modo corretto. Si noti che le 3 variabili individuate sono le stesse variabili selezionate per il caso del reattore da 600 L, quindi si conferma come esse siano le variabili maggiormente rappresentative del titolo, indipendentemente dal volume del fermentatore.

Se viene realizzata la stima del titolo virale finale in tempo reale, analogamente a quanto realizzato al §3.4.1.4, il modello PLS costruito permette di predire il titolo finale in tempo reale con sufficiente accuratezza solo a partire dalla 30$^a$ ora. Ancora una volta, negli istanti iniziali le oscillazioni dell'errore sono molto marcate, ovvero la stima è possibile, ma non è

sufficientemente precisa. Dalla 30ª ora, circa, le oscillazioni si smorzano e la stima è più attendibile. Come per il caso del fermentatore da 600 L, avendo introdotto nel modello dei dati di batch fuori specifica, l'errore si attesta su valori abbastanza elevati per i batch fuori specifica.

### *3.4.3 Conclusioni sulla predizione del titolo finale dei batch in e fuori specifica*

Sono stati realizzati dei modelli PLS costruiti su dati di batch in specifica e fuori specifica appartenenti al fermentatore da 300 L e a quello da 600 L, con lo scopo di stimare il titolo finale del reovirus. In primo luogo vengono riassunti i risultati ottenuti dai modelli PLS ottimali, costruiti per i 2 reattori, in merito alla stima del titolo dei batch in e fuori specifica, a batch concluso. Essi sono riportati in Tabella 3.22.

**Tabella 3.22.** *Caratterizzazione dell'errore di predizione e segnalazioni di non rappresentatività ottenuti dalla modellazione PLS di dati di batch in e fuori specifica, distintamente per i 2 reattori.*

| | **Errore medio ($TCID_{50}$/mL)** | **Batch predetti in modo sbagliato** | **Titoli predetti con errore > 0.36** | **Segnalazioni di non rappresentatività (SPE)** |
|---|---|---|---|---|
| **300 L** | 0.22 | 9.7% | 19.4% | 25.8% |
| **600 L** | 0.35 | 31.3% | 37.5% | 25% |

Si può dire che i risultati ottenuti sono soddisfacenti, anche se l'introduzione dei batch fuori specifica nel *set* di calibrazione ha peggiorato le prestazioni del modello nella stima, specialmente per quanto riguarda il fermentatore da 600 L. Conviene pertanto affidarsi alla metodologia di classificazione perché la predizione del titolo è più precisa. Ciò è dovuto al fatto che si hanno a disposizione pochi dati per rappresentare adeguatamente il comportamento dei batch fuori specifica nel modello PLS.

In secondo luogo, si considera la stima del titolo eseguita in tempo reale, per i 2 modelli sviluppati. I risultati ottenuti mostrano che, a partire dalla $30^a \div 40^a$ ora, è possibile stimare il titolo virale finale, sia di batch in specifica, sia di batch fuori specifica. Questo risultato è ancora in accordo con gli studi condotti su scala di laboratorio da Grande e Benavente (2000), come avveniva per il caso riguardante i soli batch in specifica. Quindi, anche introducendo i batch fuori specifica all'interno dello studio, le possibili analogie nella conduzione del processo su monostrato o in sospensione non cambiano, e appare giustificato suggerire all'azienda di programmare una campagna sperimentale al fine di studiare come varia il titolo durante la fermentazione per validare i risultati ottenuti.

# Conclusioni

Nella Tesi è stato affrontato il problema della stima della qualità di prodotto in un processo industriale per la produzione di reovirus impiegati nella formulazione di vaccini aviari. In particolare, in questo processo una frazione dei batch prodotti negli anni 2013 e 2014 evolve verso fuori specifica. Inoltre, la misura della qualità (il titolo virale finale del reovirus) è nota da prove di laboratorio solo 15 giorni dopo la conclusione del batch. L'obiettivo della Tesi è stato quello di sviluppare dei modelli per predire il titolo virale finale in modo accurato e in tempi minori rispetto a quelli necessari per i test di laboratorio. I modelli sono stati sviluppati per predire accuratamente il titolo sia di batch in specifica, sia di batch fuori specifica. È stata proposta una metodologia generale per il monitoraggio del processo, una parte della quale è appunto la predizione della qualità del prodotto.

Per lo sviluppo di un modello robusto di stima, in primo luogo è stata effettuata un'analisi preliminare sulle variabili coinvolte in tutti gli stadi del processo. Dallo studio è risultato che:

- le variabili iniziali, come numero di uova scartate, volume di matrice di virus infettante e durata dell'incubazione, sono scarsamente correlate al titolo finale, e possiedono una limitata capacità di predizione se confrontate con le variabili della fermentazione;
- vi è un fornitore di uova le quali, prima dell'invio all'impianto, vengono incubate in modo più efficiente, perché il numero di uova fertili è superiore rispetto ad altri fornitori; si è pertanto suggerito all'azienda di tener controllata la qualità dell'uovo (presenza di embrioni morti o fragili, stato di vita dell'embrione);
- lo stadio in cui viene "costruita" in modo pressoché completo la qualità del prodotto è la sola fermentazione.

È stata sviluppata una modellazione sulle variabili di fermentazione, tramite il metodo statistico PLS, per la predizione del titolo virale finale nei batch in specifica, concepita a seguito di una metodologia (non discussa in questa Tesi) che si è dimostrata in grado di classificare in tempo reale i batch come in specifica o fuori specifica. È stato quindi sviluppato un sensore virtuale per ciascuno dei 2 fermentatori disponibili (uno da 300 L e uno da 600 L).

I risultati della predizione del titolo finale a batch concluso sono stati soddisfacenti. Infatti:

- la predizione del titolo finale è accurata, con un errore medio che si attesta su 0.20 $TCID_{50}$/mL, ampiamente al di sotto dell'errore con cui viene valutato il titolo nei test di laboratorio (0.36 $TCID_{50}$/mL);
- solo per pochi batch (< 15%) il titolo viene predetto con un errore superiore a 0.36 $TCID_{50}$/mL.

Uno studio analogo è stato condotto considerando assieme i dati di batch in specifica e batch fuori specifica. In questo caso, i risultati ottenuti per la stima a batch concluso sono stati soddisfacenti, seppur l'errore medio di predizione sia risultato pari a 0.22 $TCID_{50}$/mL per il reattore da 300 L, e 0.35 $TCID_{50}$/mL per il reattore da 600 L. In questa analisi sono state individuate 3 variabili (pH, percentuale di $O_2$ disciolto e portata d'aria insufflata) che permettono di ottenere i risultati migliori nella stima del titolo finale a batch concluso.
La stima del titolo finale è stata condotta anche in tempo reale, sviluppando modelli per il reattore da 300 L e per il reattore da 600 L. Si sono ottenute predizioni attendibili del valore di titolo virale finale già a partire dalla $30^a \div 40^a$ ora dall'inizio del batch (cioè da circa metà batch), sia per batch in specifica che per batch fuori specifica. Si è pertanto ritenuto che entro le prime 30÷40 h di lavorazione venga determinata la qualità finale del prodotto. Si è riscontrato che questo risultato è coerente con lo studio svolto da Grande e Benavente (2000), dove è stata ritenuta pari a 24 h la durata ottimale della fase di fermentazione, ancorché su monostrato cellulare. Si è ritenuta possibile un'analogia con le 30÷40 h che sono emerse dall'analisi PLS in tempo reale. Per questo motivo si è suggerito all'azienda di programmare una campagna sperimentale per studiare come varia il titolo durante la fermentazione, con l'obiettivo di convalidare i risultati ottenuti ed eventualmente riuscire a ridurre il tempo di fermentazione, garantendo ugualmente la specifica di qualità del prodotto finale.

# Appendice

# Figure e codici contenuti nella Tesi

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

## A.1 Figure del Capitolo 1

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

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

| **Figura** | ***File*** |
|---|---|
| Figura 1.1 | `Figura 1.1.vsd` |
| Figura 1.2a | `Figura 1.2a.opj` |
| Figura 1.2b | `Figura 1.2b.opj` |
| Figura 1.3 | `Figura 1.3.opj` |
| Figura 1.4 | `Figura 1.4.vsd` |
| Figura 1.5 | `Figura 1.5.vsd` |
| Figura 1.6 | `Figura 1.6.vsd` |

## A.2 Figure del Capitolo 2

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

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

| **Figura** | ***File*** |
|---|---|
| Figura 2.1 | `Figura 2.1.vsd` |
| Figura 2.2 | `Figura 2.2.vsd` |

## A.3 Figure del Capitolo 3

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

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

| **Figura** | ***File*** |
|---|---|
| Figura 3.1 | `Figura 3.1.vsd` |
| Figura 3.2a | `Figura 3.2a.opj` |
| Figura 3.2b | `Figura 3.2b.opj` |
| Figura 3.3 | `Figura 3.3.vsd` |
| Figura 3.4 | `Figura 3.4.opj` |
| Figura 3.5 | `Figura 3.5.opj` |
| Figura 3.6 | `Figura 3.6.opj` |
| Figura 3.7a | `Figura 3.7a.opj` |
| Figura 3.7b | `Figura 3.7b.opj` |
| Figura 3.8 | `Figura 3.8.opj` |
| Figura 3.9 | `Figura 3.9.opj` |
| Figura 3.10 | `Figura 3.10.opj` |
| Figura 3.11 | `Figura 3.11.opj` |
| Figura 3.12 | `Figura 3.12.opj` |
| Figura 3.13 | `Figura 3.13.vsd` |
| Figura 3.14 | `Figura 3.14.opj` |
| Figura 3.15 | `Figura 3.15.opj` |
| Figura 3.16 | `Figura 3.16.opj` |
| Figura 3.17 | `Figura 3.17.opj` |
| Figura 3.18a | `Figura 3.18a.opj` |
| Figura 3.18b | `Figura 3.18b.opj` |
| Figura 3.19 | `Figura 3.19.opj` |
| Figura 3.20 | `Figura 3.20.opj` |

## A.4 Codici di calcolo

In Tabella A.4 sono riportati i codici di calcolo e i *file* da cui sono presi i relativi dati di *input*.

**Tabella A.4.** *Codici di calcolo per il Capitolo 3.*

| **Codici di calcolo** | **Dati di *input*** | **Descrizione** |
|---|---|---|
| `pesiPLSdatiniz300.m` | `X300datinizspec.mat`<br>`Y300fermspec.mat` | Codice per l'analisi preliminare sulle variabili iniziali di processo per il fermentatore da 300 L |
| `pesiPLSdatiniz600.m` | `X600datinizspec.mat`<br>`Y600fermspec.mat` | Codice per l'analisi preliminare sulle variabili iniziali di processo per il fermentatore da 600 L |
| `pesiPLSferm300.m` | `X300fermspec.mat`<br>`Y300fermspec.mat`<br>`X300fermspecvwu.mat` | Codici per la stima a batch concluso e in tempo reale applicata ai batch in specifica appartenenti al fermentatore da 300 L |
| `vipindexPLSferm300.m` | | |
| `stimaPLSferm300.m` | | |
| `nonlinearePLSferm300.m` | | |
| `stimaonlinePLSferm300.m` | | |
| `pesiPLSferm600.m` | `X600fermspec.mat`<br>`Y600fermspec.mat`<br>`X600fermspecvwu.mat` | Codici per la stima a batch concluso e in tempo reale applicata ai batch in specifica appartenenti al fermentatore da 600 L |
| `vipindexPLSferm600.m` | | |
| `stimaPLSferm600.m` | | |
| `nonlinearePLSferm600.m` | | |
| `stimaonlinePLSferm600.m` | | |
| `pesiPLSferm300loc.m` | `X300fermspec.mat`<br>`Y300fermspec.mat`<br>`X300fermspecvwu.mat`<br>`X300fermnonspec.mat`<br>`Y300fermnonspec.mat`<br>`X300fermnonspecvwu.mat` | Codici per la stima a batch concluso e in tempo reale applicata ai batch in e fuori specifica appartenenti al fermentatore da 300 L |
| `vipindexPLSferm300loc.m` | | |
| `stimaPLSferm300loc.m` | | |
| `stimaonlinePLSferm300loc.m` | | |
| `pesiPLSferm600loc.m` | `X300fermspec.mat`<br>`Y300fermspec.mat`<br>`X300fermspecvwu.mat`<br>`X300fermnonspec.mat`<br>`Y300fermnonspec.mat`<br>`X300fermnonspecvwu.mat` | Codici per la stima a batch concluso e in tempo reale applicata ai batch in e fuori specifica appartenenti al fermentatore da 600 L |
| `vipindexPLSferm600loc.m` | | |
| `stimaPLSferm600loc.m` | | |
| `stimaonlinePLSferm600loc.m` | | |

# Nomenclatura

$a$ = indicatore generico per il numero di variabili latenti (-)

$A$ = numero di variabili latenti (-)

$b_i$ = generico coefficiente di regressione (-)

$CO_{PI}$ = *controller output* del regolatore PI (-)

$\mathbf{e}_i$ = vettore riga della matrice dei residui **E** (-)

$e_{i,v}$ = elemento della matrice **E** (-)

$\mathbf{e}_{new}$ = vettore riga contenente i residui del campione $\mathbf{x}_{new}$ (-)

$errore_i$ = errore assoluto di predizione della qualità per l'*i*-esimo batch ($TCID_{50}$/mL)

$\mathbf{errore}_i$ = vettore dell'errore di predizione assoluto della qualità in linea (-)

**E** = matrice degli errori nei metodi statistici multivariati per la matrice **X** (-)

$F_{A,(k-A),\alpha}$ = distribuzione statistica F (-)

**F** = matrice degli errori nei metodi statistici multivariati per la matrice **Y** (-)

$h_0$ = coefficiente numerico della formula di Jackson-Mudholkar (-)

$i$ = indicatore generico di un'osservazione (-)

$I$ = numero totale di osservazioni (campioni o batch) (-)

$I_{300}$ = numero di batch in e fuori specifica da 300 L considerati (-)

$I_{600}$ = numero di batch in e fuori specifica da 600 L considerati (-)

$I_{300IS}$ = numero di batch in specifica da 300 L considerati (-)

$I_{600IS}$ = numero di batch in specifica da 600 L considerati (-)

**I** = matrice identità (-)

$k$ = generico istante temporale del processo (-)

$k_C$ = guadagno del regolatore PI (-)

$K$ = istanti temporali di campionamento totali (-)

$M$ = numero di variabili di qualità del prodotto (-)

$\mathbf{M}_r$ = generica matrice delle variabili di processo di rango r (-)

$N$ = rango della matrice di processo usata per costruire il modello locale (-)

$\mathbf{p}_i$ = generico vettore colonna della matrice dei *loading* **P** (-)

$\mathbf{p}_r$ = *loading* della generica matrice $\mathbf{M}_r$ (-)

**P** = matrice dei *loading* (-)

$PRESS_m$ = errore di predizione sulla somma dei quadrati dei residui (-)

**Q** = matrice dei *loading* generica per la matrice **Y** (-)

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

$R$ = rango di una generica matrice (-)

$R^2$ = coefficiente di correlazione multipla (-)

$s_a$ = generico semiasse dell'ellissoide di confidenza nel diagramma degli *score* (-)
$\mathrm{SPE}_i$ = errore di predizione al quadrato per il generico campione *i* (-)
$\mathrm{SPE}_{\alpha,Lim}$ = limite dell'errore di predizione al quadrato (-)
$t$ = generico istante temporale della fermentazione (-)
$t_{\mathrm{finale}}$ = istante temporale finale della fermentazione (-)
$\mathbf{t}_i$ = generico vettore colonna della matrice degli *score* **T** (-)
$\mathbf{t}_r$ = *score* della generica matrice $\mathbf{M}_r$ (-)
$\hat{\mathbf{t}}_{new}$ = predizione del vettore degli *score* per un nuovo campione $\mathbf{x}_{new}$ (-)
$\mathbf{T}$ = matrice degli *score* sulle variabili di processo (-)
$T^2$ = statistica di Hotelling (-)
$T^2_{A,k,\alpha}$ = limite di confidenza per il diagramma degli *score* e $T^2$ (-)
$T_i^2$ = generica distanza dall'origine del diagramma degli *score* nel loro piano (-)
$T^2_{\mathrm{Lim}}$ = limite della statistica $T^2$ di Hotelling
$\mathbf{u}_i$ = generico vettore colonna della matrice degli score **U** (-)
$\mathbf{U}$ = matrice degli *score* per **Y** (-)
$v$ = indicatore generico per le variabili (-)
$V$ = numero totale delle variabili di processo misurate (-)
$\mathrm{VIP}_v$ = indice VIP per la *v*-esima variabile di processo (-)
$\mathbf{w}_i$ = generico vettore colonna della matrice dei pesi **W** (-)
$w_{i,v}$ = elemento della matrice **W** (-)
$\mathbf{W}$ = matrice dei pesi **W** (-)
$\mathbf{x}_i$ = vettore riga di **X** (-)
$x_{i,v}$ = elemento della matrice **X** (-)
$\hat{x}_{i,v}$ = stima del vettore $x_{i,v}$ (-)
$\mathbf{x}_{new}$ = generico vettore di nuovi dati (-)
$\hat{\mathbf{x}}_{new}$ = predizione del vettore $\mathbf{x}_{new}$ (-)
$\mathbf{x}_v$ = vettore colonna della matrice **X** (-)
$\bar{\mathbf{x}}_v$ = vettore dei valori medi per ogni colonna della matrice **X** (-)
$\mathbf{X}$ = matrice bidimensionale delle variabili di processo misurate (-)
$\underline{\mathbf{X}}$ = matrice tridimensionale delle variabili di processo misurate (-)
$\mathbf{X}_1$ = matrice delle variabili relative ai trattamenti delle uova pre-impianto (-)
$\mathbf{X}_2$ = matrice delle variabili relative a pretrattamento e raccolta delle uova (-)
$\mathbf{X}_3$ = matrice delle variabili di processo misurate durante lo stoccaggio (-)
$\mathbf{X}_4$ = matrice delle variabili di processo coinvolte durante l'iniezione del virus (-)
$\mathbf{X}_5$ = matrice delle variabili di processo misurate a inizio e fine fermentazione (-)
$\underline{\mathbf{X}}_6$ = matrice delle variabili della fermentazione misurate in linea (-)
$\mathbf{X}_{300}$ = matrice con variabili di fermentazione e batch in e fuori specifica da 300 L (-)
$\mathbf{X}_{600}$ = matrice con variabili di fermentazione e batch in e fuori specifica da 600 L (-)

$\mathbf{X}_{300in}$ = matrice di dati delle variabili iniziali per il reattore da 300 L (-)
$\mathbf{X}_{600in}$ = matrice di dati delle variabili iniziali per il reattore da 600 L (-)
$\mathbf{X}_{300IS}$ = matrice di dati con variabili di fermentazione e batch in specifica da 300 L (-)
$\mathbf{X}_{600IS}$ = matrice di dati con variabili di fermentazione e batch in specifica da 600 L (-)
$\mathbf{X}_{300FS}$ = matrice con variabili di fermentazione e batch fuori specifica da 300 L (-)
$\mathbf{X}_{600FS}$ = matrice con variabili di fermentazione e batch fuori specifica da 600 L (-)
$\mathbf{X}_{300glob}$ = matrice con variabili dell'indice VIP e batch in e fuori specifica da 300 L (-)
$\mathbf{X}_{600glob}$ = matrice con variabili dell'indice VIP e batch in e fuori specifica da 600 L (-)
$\mathbf{X}_{300mod}$ = matrice di dati di processo utilizzata dal $Modello_{300}$ (-)
$\mathbf{X}_{600mod}$ = matrice di dati di processo utilizzata dal $Modello_{600}$(-)
$\mathbf{X}_{300ott}$ = matrice di dati di processo utilizzata dal $Modello_{300}$ ottimizzato (-)
$y_{i,m}$ = elemento della matrice **Y** (-)
$\hat{y}_{i,m}$ = stima del vettore $y_{i,m}$ (-)
$\mathbf{y}_i$ = generico vettore di dati di qualità (-)
$\hat{\mathbf{y}}_i$ = stima del generico vettore di dati di qualità (-)
**Y** = matrice della variabile di qualità (-)
$\mathbf{Y}_{300}$ = matrice di qualità di batch in e fuori specifica da 300 L (-)
$\mathbf{Y}_{600}$ = matrice di qualità di batch in e fuori specifica da 600 L (-)
$\mathbf{Y}_{300in}$ = matrice di qualità per il reattore da 300 L nell'analisi con le variabili iniziali (-)
$\mathbf{Y}_{600in}$ = matrice di qualità per il reattore da 600 L nell'analisi con le variabili iniziali (-)
$\mathbf{Y}_{300IS}$ = matrice di qualità con batch in specifica da 300 L (-)
$\mathbf{Y}_{600IS}$ = matrice di qualità con batch in specifica da 600 L (-)
$\mathbf{Y}_{300FS}$ = matrice di qualità con batch fuori specifica da 300 L (-)
$\mathbf{Y}_{600FS}$ = matrice di qualità con batch fuori specifica da 600 L (-)
$z_\alpha$ = deviazione normale standard (-)

Apici

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

Lettere greche

$\alpha$ = limite di fiducia (-)
$\mathbf{\Lambda}$ = matrice diagonale degli autovalori (-)
$\boldsymbol{\lambda}$ = vettore delle varianze degli *score* delle variabili latenti (-)
$\lambda_a$ = autovalore della matrice $\mathbf{\Lambda}$ associato alla *a*-esima componente principale (-)
$\Delta_{camp}$ = intervallo di campionamento per le variabili della fermentazione (min)

$\varepsilon$ = errore normalizzato entrante al regolatore (-)
$\theta_i$ = coefficienti della formula di Jackson-Mudholkar (-)
$\sigma$ = varianza (-)
$\sigma^2$ = deviazione standard (-)
$\tau_I$ = tempo dell'azione integrale (s)

Acronimi

CO = *controller output*
FS = fuori specifica
IS = in specifica
LV = variabili latenti
MOI = *multiplicity of infection*
NIPALS = *nonlinear iterative partial least squares*
NOC = normali condizioni operative
PBS = *phosphate buffered saline*
P&I = *piping and instrumentation diagram*
PID = proporzionale integrale differenziale
PLS = metodo della proiezione su strutture latenti
RMSEC = *root-mean square error of calibration*
RMSECV = *root-mean square error of cross validation*
RPM = *revolutions per minute*
SPE = errore di predizione al quadrato
VIP = *variable importance in the projection*

# Riferimenti bibliografici


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

Facco, P. (2005). Monitoraggio di un processo semicontinuo di polimerizzazione con metodi statistici multivariati. *Tesi di Laurea in Ingegneria chimica*, DIPIC, Università di Padova.

Geladi, P. e B. Kowalski (1986). Partial least-squares regression: a tutorial. *Anal. Chim. Acta*, **185**, 1.

Grande, A. e J. Benavente (2000). Optimal conditions for the growth, purification and storage of the avian reovirus S1133. *J. Viro. Met*., **85**, 43-54.

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

Kourti, T. (2003). Multivariate dynamic data modelling for analysis and statistical process control of batch processes, start-ups and grade transitions. *J. Chemometrics*, **17**, 93-109.

Mandenius, C. F. e R. Gustavsson (2014). Mini-review: soft sensors as means for PAT in the manufacture of bio-therapeutics. *J. Chem. Technol. Biotechnol.*, 1-13.

Montgomery, D.C. (2005). *Introduction to statistical quality control* ($5^{th}$ ed.). John Wiley & Sons, Inc. (U.S.A.).

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

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

Robertson, M.D. e G.E. Wilcox (1986). Avian reovirus. *Vet. Bull*., **56**, 155-174.

Tomba, E., M. De Martin, P. Facco, J. Robertson, S. Zomer, F. Bezzo e M. Barolo (2013). General procedure to aid the development of continuous pharmaceutical processes using multivariate statistical modeling – An industrial case study. *Int. J. Pharm.*, **444**, 25-39.

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, S. (1978). Cross validatory estimation of the number of components in factor and principal component models. *Technometrics*, **20**, 397-405.

# Ringraziamenti

Ed eccomi infine giunto al momento dei ringraziamenti, doveroso. Esser arrivato fin qui, finalmente al termine di questo lungo e faticoso percorso ingegneristico intrapreso 5 anni fa, culminato con la realizzazione di questa Tesi, è sicuramente uno dei più grandi traguardi che avrei mai potuto immaginare. E se ci sono arrivato il merito va prima di tutto ai miei genitori, Ivonne e Gilberto, che ancor più di me forse hanno sempre creduto in questa scelta, fin dall'inizio, e che in questi anni mi sono stati costantemente vicino, facendo grandi sacrifici e cercando in ogni momento di far sì che potessi studiare e andare avanti nel mio percorso alle migliori condizioni possibili. Ringrazio poi mio fratello, Manuel, per l'aiuto da sempre concessomi nei momenti critici, la grande disponibilità e l'immensa pazienza. E ringrazio anche Spenck, sì proprio il cagnolino, il cui contributo in questi anni non è certo stato trascurabile, anzi. Ringrazio i miei secolari amici, della quotidianità di Vicenza, Riccardo, Ermanno e Luke, per gli splendidi momenti e i fine settimana passati insieme ormai da 10 anni. Un sentimentale grazie poi a Silvia, che si è dimostrata una persona molto importante nella ma vita. E non posso assolutamente non ringraziare i miei amici aspiranti ingegneri di Padova, Mattia Z., Mattia V., Nicola B., Nicola C., Francesco, Marta e Francesca, fondamentali compagni di avventure/sventure, con cui ho condiviso tutto ciò che è successo in questi anni di Università, dai momenti più felici a quelli più tristi.

E passiamo ora alle persone che hanno reso possibile fattivamente questo complesso lavoro di Tesi. Essere entrato, seppur temporaneamente, a far parte del CAPE-Lab è stata sicuramente un'esperienza che non dimenticherò mai. Mettersi in gioco in un ambiente a contatto con persone di assoluta professionalità non è stato facile e mi ha sicuramente trasmesso molto, sotto svariati punti di vista. Ringrazio quindi Martina, per l'estrema disponibilità concessami, il tempo dedicatomi, il prezioso aiuto fornitomi e per avermi seguito e incoraggiato praticamente ogni giorno in questi 7 mesi. Ringrazio Pierantonio, per la risolutezza e la simpatia, per i consigli illuminanti, per avermi dato la possibilità di apprendere da lui la nuova disciplina a cui mi sono dedicato in questi 7 mesi. Ringrazio il Prof. Bezzo per avermi convinto, perché senza la sua proposta non avrei mai intrapreso questo lavoro di Tesi. E ringrazio il mio relatore, il Prof. Barolo, per avermi costantemente seguito e indicato la via, nonostante i suoi mille impegni professionali, per la sua precisione, la sua onestà, il suo rigore, la sua saggezza. Se questa Tesi è stato un successo, il merito va certamente anche all'azienda Merial, con cui è stato concepita. Voglio quindi portare un sentito grazie alla mia correlatrice aziendale, Dott.ssa Donatella, e a Michele, per esser stati sempre gentili, aperti, disponibili e presenti nel momento del bisogno, nonostante le tante incombenze aziendali. Ringrazio infine il direttore, l'Ing. D'Onofrio, per aver reso possibile questa collaborazione.