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

DATA ANALYTICS FOR POWDER FEEDING MODELLING ON CONTINUOUS SECONDARY PHARMACEUTICAL MANUFACTURING PROCESSES

Relatore: ***Prof. Massimiliano Barolo***
Correlatori: ***Ing. Pierantonio Facco***
Dr. Simeone Zomer

Laureando: ANTONIO BENEDETTI

ANNO ACCADEMICO 2017-2018

Ai miei nonni Ruggero, Antonia, Sergio, Rosina,
per avermi trasmesso il significato del sacrificio
e per avermi sostenuto in ogni istante.

Abstract

The primary manufacturing of active pharmaceutical ingredients for oral solid dosage medicines is characterized by a high variability of the final powder form, that is likely to affect the downstream manufacturing operations. The effects caused by the raw materials variability in the secondary continuous processes are often ignored in the early stage of the drug and development design, since the approach to predict powder behaviour in the equipment is still knowledge-based and semi-empirical. The lack of first principle understanding of the powder processability and flowability along the process is an obstacle that is complicated to be overcame. Moreover, the univariate approach to the general understanding of the powder phenomena does not produce a comprehensive solution to the problem. This causes several problematics in the start-up operations of new continuous processes that are likely to cause delays in the manufacturing campaigns.

In this Thesis, a data-driven procedure of investigating raw materials variability in an industrial database is presented, together with a multivariate statistical modelling approach for the first unit operation of continuous tableting lines, i.e. the loss-in-weight feeder. The main objective of the Thesis is to explore the capabilities of using statistical pattern recognition techniques to identify and model hidden patterns of similarities in a powder materials dataset and analyse the general problem of how materials variability can affect powder feeding modelling. Firstly, a general procedure to reorganise a materials dataset, explore the general structure, recognize patterns and build up a classification system for new incoming materials is developed. The application on a case study dataset of raw materials powders showed excellent results in terms of patterns identification and classification, maximizing the amount of information that can be extracted from a restricted number of materials descriptors. Secondly, some few different scenarios of how to use a data-driven approach for the prediction of a targeted quality variable for predicting feeding performance are introduced and a case study example is presented. The lack of first principle understanding of powder flowability in feeding equipment is addressed from a multivariate statistical approach, combining data from the equipment setup, materials properties data and process data from the feeder sensors in order to explain the correlation between these variables and the feed factor profile. The results of this analysis seem promising in terms of the speed up of the continuous tabletting manufacturing development.

The Thesis is the result of a 6-month Erasmus internship at GlaxoSmithKline R&D center of Ware (UK) as part of a collaboration research project between the CAPE-Lab group at the University of Padova and GlaxoSmithKline R&D.

Riassunto

L'industria chimica tradizionale, come ad esempio l'industria petrolchimica, l'industria chimica delle *commodities* e quella fine, è continua per sua natura. Altre industrie, come ad esempio l'industria chimica alimentare, cambió velocemente da una realtà manifatturiera discontinua ad una più moderna realtà continua per incrementare l'efficienza e allo stesso tempo soddisfare un mercato caratterizzato da grandi volumi di produzione. I vantaggi di operare un processo in maniera continua sono vari e non difficili da intuire, tra i quali i più menzionati sono una maggiore flessibilità e facilità nelle fasi di *scale up/down*, una diminuzione dei tempi di produzione, una massimizzazione della produzione stessa con una riduzione degli scarti e dell'energia consumata e, infine, una velocizzazione dell'intera catena produttiva e di distribuzione [1]. Nonostante questo, l'industria farmaceutica continua a produrre gran parte dei suoi prodotti secondo un approccio tradizionale discontinuo. Questo ritardo tecnologico è principalmente dovuto alla complessa regolamentazione, prevista enti regolamentatori esterni, che spesso scoraggia le compagnie farmaceutiche ad investire verso nuovi sistemi più avanzati e redditizi [2] [3]. A maggior ragione, l'alto margine di profittabilità tradizionalmente ottenuto anche senza investire nell'ottimizzazione dei processi, ha spinto le aziende a focalizzare gli investimenti verso altre aree di ricerca. Tuttavia, l'incremento dei costi e dei rischi associati alla ricerca e sviluppo di nuove molecole, insieme all'introduzione di farmaci generici altamente competitivi sul mercato, ha portato ad un calo del ritorno economico ed una conseguente necessità di revisionare e migliorare i processi già esistenti.

Grazie anche ad una serie di iniziative promosse dalle agenzie regolatorie come la Food and Drug Administration (FDA) e la European Medicines Agency (EMA), da alcuni decenni le aziende farmaceutiche, in parallelo al grande sforzo profuso dalla ricerca accademica, stanno gradualmente esplorando la possibilità di convertire alcuni impianti manifatturieri alla produzione in continuo [5][6]. Questo sviluppo tecnologico è guidato da un postulato fondamentale, chiamato comunemente quality-by-design (QbD) che significa propriamente qualità attraverso la progettazione, esplicitando chiaramente come la progettazione di processo e di prodotto contega al suo interno il risultato stesso della qualità e delle specifiche ricercate. Questo significa che per conoscere e controllare la qualità finale del prodotto si deve raggiungere una conoscenza comprensiva della relazione tra proprietà delle materie prime, parametri di processo e specifiche di qualità. In questa nuova concezione di sviluppo di prodotto, la

modellazione matematica acquisisce un ruolo chiave nell'integrare l'esperienza nel settore pregressa e la progettazione di prodotto [7].

In questo caso rientra anche la produzione di compresse (*oral solid dose*, comunemente denominate OSD), che ancora ricopre la maggior parte delle forme farmaceutiche in commercio. Tutte le linee continue per la produzione di compresse cominciano con un'alimentazione costante delle polveri in ingresso, che costituiscono il principio attivo o l'eccipiente prevista dalla formulazione del farmaco finale, attraverso un'apparecchiatura standard chiamata alimentatore o dosatrice a perdita-in-peso (*loss-in-weight feeder*).

Questa apparecchiatura è particolarmente progettata per assicurare una certa stabilità e accuratezza in un ampio intervallo di materiali polvirulenti differenti. Tuttavia, le prestazioni di uno specifico alimentatore che opera in condizioni normali, dipende fortemente dalle proprietà delle materie prime alimentate, con particolare attenzione alle proprietà che caratterizzano la scorrevolezza della polvere. Per esempio, materiali dalla facile scorrevolezza hanno mostrato una migliore in termini di variabilità della portata alimentata, ma possono comunque nascondere qualche problematica che comporta un'oscillazione della portata se la polvere è soggetta ad una pressione improvvisa (fenomeno chiamato *flushing*). Contrariamente, materiali difficilmente scorrevoli (anche noti come coesivi), hanno mostrato diverse problematiche, che comportano una instabilità nell'alimentazione dei materiali richiesti dalla formulazione, anche in condizioni operative normali. Questo è dovuto alla loro capacità di aderire alla superficie delle pareti della tramoggia, della coclea (vite) e della tramoggia di carico o formare aggregati (effetto *bridge*) [8].

È dunque importante classificare un nuovo materiale prima di iniziare a processarlo nell'apparecchiatura sulla base delle proprietà di scorrimento. Allo stato attuale, una corretta classificazione deve essere attribuita da un ingegnere o uno scienziato dei materiali e, solo successivamente, porta alla stima delle prestazioni all'interno dell'apparecchiatura grazie alla conoscenza di un esperto del processo. Dunque, lo stato dell'arte in questo campo si basa fortemente su un approccio empirico tradizionale che si affida alla conoscenza accumulata nel passato. Tuttavia, mancano completamente strumenti che possano predirre sistematicamente e quantitativamente le prestazioni di un nuovo materiale sconosciuto all'interno della linea continua. Inoltre, le proprietà dei materiali legate alla scorrevolezza sono ancora poco oggettivamente e complessivamente definite a causa dell natura intrinsica multi-dimensionale del problema, in cui diversi parametri e tecniche di caratterizzazione non sono sufficienti a descrivere completamente la dinamica associata alla processabilità nelle varie apparecchiature di processo [9]. A maggior ragione, la complessità di descrivere la movimentazione dinamica delle polveri è confermata da una storica mancanza di

conoscenza dei principi primi che stanno alla base dei fenomeni fisici associati alla stessa [10].

Lo scopo di questa Tesi è esplorare le capacità di un supporto alla progettazione di processo e di prodotto basato sull'analisi dei dati che favorisca lo sviluppo e la selezione di nuovi materiali per le linee secondarie continue per la produzione di compresse.

La parte principale del progetto ruota attorno allo sviluppo di una procedura generale per supportare l'analisi e l'espansione di un *database* di materiali farmaceutici polvirulenti usando un approccio basato sull'analisi dei dati per riconoscere ed eventualmente modellare schemi (*patterns*) ricorrenti nascosti nei dati stessi. Con l'ausilio di un caso studio basato su un *database* industriale, viene proposta una metodologia sistematica per riorganizzare i dati, esplorare il sistema, identificare schemi non noti e modellare un sistema di classificazione per definire classi di scorrimento delle polveri. Il livello di conoscenza che può essere ottenuto da questa analisi strutturata è propenso ad apportare un alto valore aggiunto nello sviluppo di linee secondarie continue. La procedura è inoltre facilmente estendibile ad altre aree di interesse industriale sia nella realtà farmaceutica che in altre realtà manifatturiere.

Le metodologie applicate e suggerite vanno dall'applicazione di tecniche multivariate comuni come l'analisi delle componenti principali [11] per l'analisi esplorativa e la visualizzazione dei dati a tecniche più complesse come l'utilizzo di algoritmi di *clustering* [39] per l'identificazione di *patterns* nascosti nei dati o tecniche di classificazione che spaziano dall'analisi multivariata lineare [49] a tecniche di modellazione non-lineare delle classi basate su applicazioni machine learning [43][44]. L'applicazione al caso studio ha mostrato ottimi risultati per quanto riguarda l'identificazione di classi di scorrevolezza, identificando quattro possibili classi, partendo da un set limitato di proprietà disponibili. La modellazione delle classi per la costruzione di un modello per la classificazione ha invece dimostrato come sistemi di classificazione non-lineare risultino più appropriati per descrivere la conformazione dei limiti fra classi.

Nella seconda e ultima parte della Tesi, diversi approcci multivariati per la modellazione di un alimentatore a perdita-in-peso vengono proposti. Lo scopo di questi modelli è fornire degli strumenti flessibili e facilmente interpretabili per la predizione di eventuali variabili di qualità. Le applicazioni vanno dalla semplice caratterizzazione o simulazione di un profilo per investigare la processabilità di alcuni materiali nell'apparecchiatura ad applicazioni più complesse come la progettazione di sistemi di monitoraggio in linea e l'identificazione di possibili guasti o deviazioni dalle condizioni operative normali.

In questo caso si è cercato di costruire dei modelli statici [12] e dinamici [34] di questa unità di processo tramite la regressione di una variabile importante per la qualità (*feed*

factor) a partire dai parametri di ingresso, quali setup dell'apparecchiatura, proprietà dei materiali e variabili dinamiche registrate dall'apparecchiatura durante la fase operativa. Oltre ai vari spunti modellistici presentati, un esempio applicativo su un caso studio ripreso da alcuni dati sperimentali permette di apprezzare come dei modelli dinamici, basati sui dati raccolti in passato, possano produrre delle stime del profilo di *feed factor* che risultano essere molto prossime ai valori reali. La convalida esterna del modello mostra infatti dei risultati più che soddisfacenti per lo scopo per cui è stato concepito ed apre la porta all'utilizzo di questo approccio per scopi modellistici più sofisticati.

Table of Contents

Introduction

Many traditional chemical industries, such as petrochemical, fine and bulk chemicals, are continuous by nature, others (like the food industry) rapidly changed from batch to continuous manufacturing to improve their efficiency and adapt to a large volume market. The advantages of a continuous way of operating are various, e.g. an increasing flexibility and facility in scaling up, a decreasing manufacturing time, a maximization of the production with a minimization of wastes and energy consumption and a consequent speeding up of the supply chain [1]. Despite that, the pharmaceutical industry is still manufacturing a wide majority of products using a batch environment. This technology delay is mainly due to a lack in the regulatory sphere for continuous pharmaceutical processes that discouraged the company to invest in the past pushing forward to more advanced and efficient production routes [2] [3]. Nonetheless, the remarkable characteristic of the pharmaceutical industry has been its incredibly high reported profitability even with a relatively scarce optimization of the production processes. This fact prevented the companies to focus the investments in this area of research for several decades. However, the increasing costs and risks in the drug discovery and development field and the new competition for the introduction of low-price generic substitutes, caused a downturn of the return of investments and a consequent need of revising the processes to maximize the profit margin [4].

Thus, thanks to several initiatives launched by regulatory agencies such as the Food and Drug Administration (FDA) and the European Medicines Agency (EMA), in the last decades the pharmaceutical industries, in parallel with a wide research effort of the academic organizations, gradually explored the possibilities of continuous manufacturing [5] [6]. This technology development has been driven by the fundamental postulate, commonly named quality-by-design (QbD), which clearly states that the product quality must be “embedded” inside the product itself since its design process. This statement means that, in order to know and control the final quality of a product, a comprehensive knowledge of the relationships between material properties, process parameters and quality specifications must be achieved. In this new concept of product development, mathematical modelling acquires a leading role to integrate experience-based knowledge and product design [7].

This is also the case of tablet manufacturing, that still cover most of the used drug product forms in the market. All continuous tablet manufacturing processes begin with a constant feeding of the powder materials required by the formulation into the system using a standard equipment, called loss-in-weight feeder.

Loss-in-weight feeders have been designed to ensure a certain feeding stability and accuracy over a wide range of different powder materials. However, the performance of a particular feeder, during normal operating conditions, highly depend on the raw material properties, specifically, material flow properties. For example, easy flowable materials have shown better performances in terms of feed rate variability but they can be responsible of "flushing" phenomena, when the powder is subjected to sudden pressure drops, that may lead to relevant oscillation in the flow rates. Contrarily, cohesive powders have shown several problems that may lead to a relevant instability even during normal operations, because of their capability to adhere to the hopper walls or bridge across the same or, eventually, to stick to the screw surface [8].

Therefore, it is important to classify a new material based on its flow properties before starting to process it in the equipment.

Nowadays, a correct classification of a specific powder needs to be performed by material scientists and, ultimately, it would lead to a qualitative estimation of the feeding performance carried out by process experts. Consequently, the state of art in this field heavily relies on the empirical knowledge accumulated in the past and the lack of a systematic and quantitative tool, to predict the feeder performance, might cause several unexpected problems when new materials are firstly introduced in the continuous line.

Moreover, the flow-related behaviour of powders is still poorly understood because of its multi-dimensional intrinsic characteristic, in which several parameters and characterization techniques are not enough to fully describe the powder dynamics in the specific manufacturing context and they often present a high non-linearity descriptive space [9]. A univariate method of investigation and classification based on a single parameter or single characterization test might not be sufficient to fully represent the real behaviour of the material. Nonetheless, the complexity of the description of powder systems is confirmed by the historical shortage of first principle understanding of the physical phenomena associated to the same [10].

The purpose of this thesis project is to explore the capability of data analytics to support the early stage of input material selection and the consequent product and process design of secondary continuous manufacturing processes. The core of the project is dedicated to the development of general procedure to aid the analysis and the development of a pharmaceutical raw materials database using a data-driven approach to recognize and model hidden patterns in the data. A systematic methodology to reorganize the data, explore the system, identify unknown patterns and design a supervised recognition system is proposed and illustrated going through an example of an industrial application.

The degree of knowledge about the system that can be obtained by this investigation is likely to provide a support in the development of secondary continuous line, especially as concerns the feeding unit operation and the possibility of using a data-driven approach to understand and predict the flowability performance of new powder materials in the equipment.

To this end, in the last part of the project a possible modelling application is presented about how to combine input material information and other input process data (e.g. equipment setup) in order to develop a multivariate statistical modelling approach to predict the feeder performance.

Chapter 1

Chemometrics modelling and data analysis

In this Chapter the mathematical description of the chemometrics methods used in this Master Thesis research project is given. The two main aspects described are the pattern recognition techniques in reference to clustering and classification and the linear multivariate regression techniques in the context of the prediction of the quality performances of a response variable. The wide majority of the data analysis carried out in this research project is based on multivariate statistical techniques; thus, a relevant section of this chapter is dedicated to latent variable modelling either for pattern determination and exploratory analysis, process optimization and quality prediction. However, a specific section is dedicated to a relatively novel technique in non-linear classification, originally introduced in the machine learning community, and called support vector machines (SVM). This classification technique is then compared to some more classical statistical pattern recognition models, such as partial least-squares discriminant analysis (PLS-DA) and *k* nearest neighbours (*k*-NN). Although all the chemometrics techniques discussed have broader applications, the discussion here refers solely to the context of the related applications presented in the next chapters.

1.1 Multivariate statistical analysis techniques

Nowadays industrial processes have thousands of sensors that record and store a massive amount of data and even a single unit operation might have dozens of them. Moreover, input materials are always characterized by a significant number of material properties to try to establish the link between input and output quality. The same laboratory instruments used for the characterization measure tens of thousands variables. The results are significantly large datasets that may result impossible to be fully explored using a univariate approach with a consequent wastefulness of the available resources. It is in this context that multivariate statistical analysis (MSA) becomes a powerful tool for the *off-line* or *in-line* data analysis.

MSA techniques, also known as latent variables (LVs) techniques, allow the construction of models that represent the system taken into consideration through the

definition of a new restricted set of variables, latent variables precisely, generated from the original data available, but able to capture in an optimal way the variability structure and the correlation of the data.
The most known and used methods are the principal components analysis (PCA) [11], to explore the correlation between variables in a single dataset and the projection on latent structures or partial least squares regression (PLS), to develop regression models and determine the correlation between two blocks of data: the so called predictor (**X**) variables and predicted (**Y**) variables [12] [13] [14]. The areas of application of MSA in the pharmaceutical industry are numerous, e.g. general understanding of raw material attributes and batch differentiations [15][16], increasing the knowledge of the process [17] and aiding the optimization of the process parameters [18], real time prediction of quality attributes [19], and calibration models for process analytical technology (PAT) control systems [20] [21] [22].

1.1.1 Principal components analysis

The concept of PCA has its first appearance in 1901 by Karl Pearson, although the first to present in its formal general procedure as we know it today was Harold Hotelling in 1933. However, the delineation of the basis for the systematic application of the method to reasonably sized problem had to wait for the advancement of the computers and it is attributed to J. Edward Jackson in 1981 [23].
The basic idea behind PCA is based on the fact that, in any dataset, the key informations are embedded in just some dominating variables and in the way the change respect to one another, known as co-variance [24]. The other sources of variability instead, do not add any relevant information to the analysis and this is easy to be understood in the chemical systems where noise in the process or instrument measurements and redundant variables are always present. Thus, it is necessary a method to compress the essential information in such a way that the general trend of the data might be easily explained and displayed, whereas the noise is excluded in some sort of signal averaging. PCA is exactly the method that is able to extract the information of covariance and correlation between the original variables, identifying the linear combinations that better describe the variability in the data [25].
Mathematically, considering $\mathbf{X}$ $[N \times V]$ a matrix of N rows and V columns, in which the rows are the samples or elements (e.g. powder materials, experiments, samples) and

the columns are the variables of the system (e.g. material properties, process variables), PCA decomposes X as the sum of *r* vectors **t** and **p**:

$$\boldsymbol{X} = \boldsymbol{t_1 p_1^T} + \boldsymbol{t_2 p_2^T} + \ldots + \boldsymbol{t_r p_r^T}\,, \tag{1.1}$$

where *r* is the rank of the matrix and must less or equal than the smaller dimension of **X**, i.e., $r \leq \min\{N, V\}$.

The vectors **t** and **p** collect the t_i and p_i elements and are ordered progressively by the amount of variance captured. The vectors **t** are known as score vectors and they describe how the samples are related each other. The vectors **p** are known as loading vectors and they contain the information of how the variables relate each other. Usually the resulting PCA model considers just *z* components of the respective vectors that describes the wide majority of the variance of the data, so that the remaining small variance factors are consolidated into a residual matrix **E**:

$$\mathbf{X} = \mathbf{t_1 p_1^T} + \mathbf{t_2 p_2^T} + \ldots + \mathbf{t_k p_k^T} + \mathbf{E}\,, \tag{1.2}$$

where **E** is also known as the matrix of the errors, which is generated by the reconstruction of the original **X** [26].

The value of *z* can be determined by different approaches but for this research project the cumulative percent variance and the average eigenvalue methods have been used according to the description given by Valle et al. [27].

The score and loading vectors are calculated starting from the eigenvector decomposition of the covariance or correlation matrix of the original **X** defined as:

$$\mathrm{cov}(\mathbf{X}) = \frac{\mathbf{X^T X}}{\mathrm{N} - 1}\,, \tag{1.3}$$

in which the relative eigenvectors λ, constituted of the λ_i eigenvalues, are in fact the loading vectors **p** so that:

$$\mathrm{cov}(\mathbf{X})\mathbf{p} = \lambda\mathbf{p}\,. \tag{1.4}$$

The scores vectors **t** are obtained from combination of the original dataset following:

$$\mathbf{t} = \mathbf{Xp}\,. \tag{1.5}$$

As a matter of fact, the scores **t** form an orthogonal set and are collected in the respective matrix **T**, while the loading **p** are orthonormal and are collected in the respective matrix **P**. These important properties guarantee that scores and loadings are orthogonal each other and so the resulting principal components are not correlated. In this way, the original problem can be re-formulated as in Figure 1.1:

$$\mathbf{X} = \mathbf{TP^T} + \mathbf{E}, \qquad \mathbf{X_{appr}} = \mathbf{TP^T}, \qquad \mathbf{E} = \mathbf{X} - \mathbf{X_{appr}}\,. \tag{1.6}$$

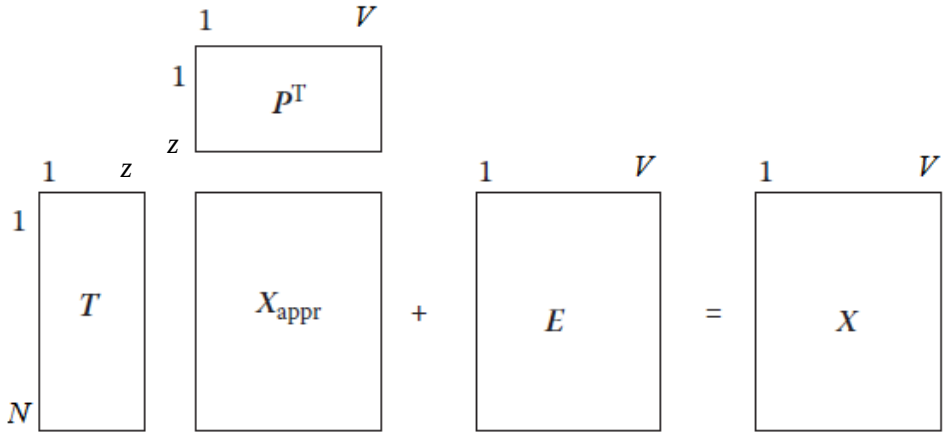


Figure 1.1 *Approximate reconstruction of an original dataset X from PCA scores T and loadings matrix P using z latent variables. E is the so-called error or residual matrix.*

Geometrically, PCA identified in the principal components the directions of maximum variability of the data. Because of their orthogonality, it is possible to build up a new fictitious space with a reduced number of latent variables in which the samples are projected. Generally, two or three latent variables already explained a relevant proportion of the variance in the data, so that it is easy to represent the data structure into a two or three-dimensional scatter of the new reduced space. From this point of view, the loadings explain the rotation of the new coordinate system respect to the original one and the scores are the coordinates of the samples within the new space [28]. An intuitive example of how it is possible to reduce three original dimensions in a new system of two latent variables is given in Figure 1.2, but the concept is easily translated when the original space is multidimensional.

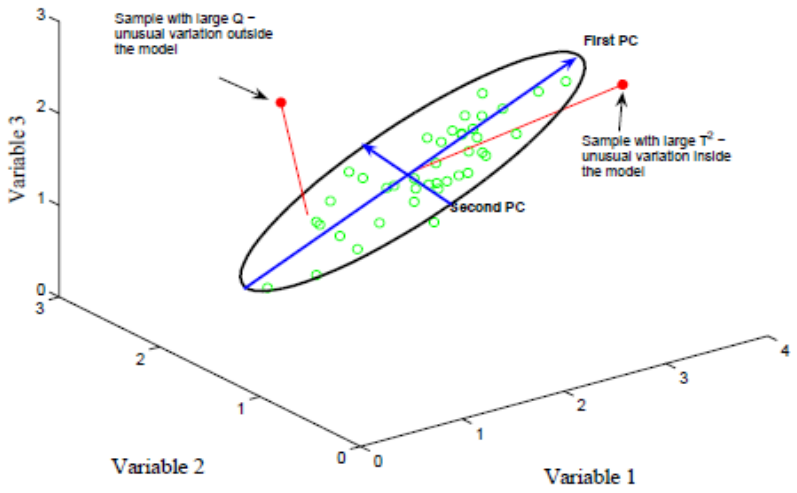


Figure 1.1 *Geometrical representation of PCA reduction from a three-dimensional space to a two-dimensional space according to the variance information explained by the data [26].*

1.1.2 Data preprocessing

Data preprocessing is a fundamental step before starting any kind of MSA, especially in chemical and process data where the unit dimensions might be really different among the process variables or material properties characterization. Thus, data need to be pretreated and the most common operations are mean-centering and variance scaling. Mean-centering subtracts to each column of the original matrix the relative average in such a way to translates the axes coordinate system to the centroid of the data. Variance scaling instead has the aim of levelling off the effect of the variance captured when different unit dimensions have considerable impact on the system variation and this is achieved dividing each column for the standard deviation of the same. The application of these techniques is a standard requirement for manufacturing or process data and it is common refer as autoscaling. Figure 1.3 shows the geometrical meaning and impact of each of these techniques on the final PCA result.

However, the pretreatment is always selected based on the nature of the original data and autoscaling is not necessarily the best choice.

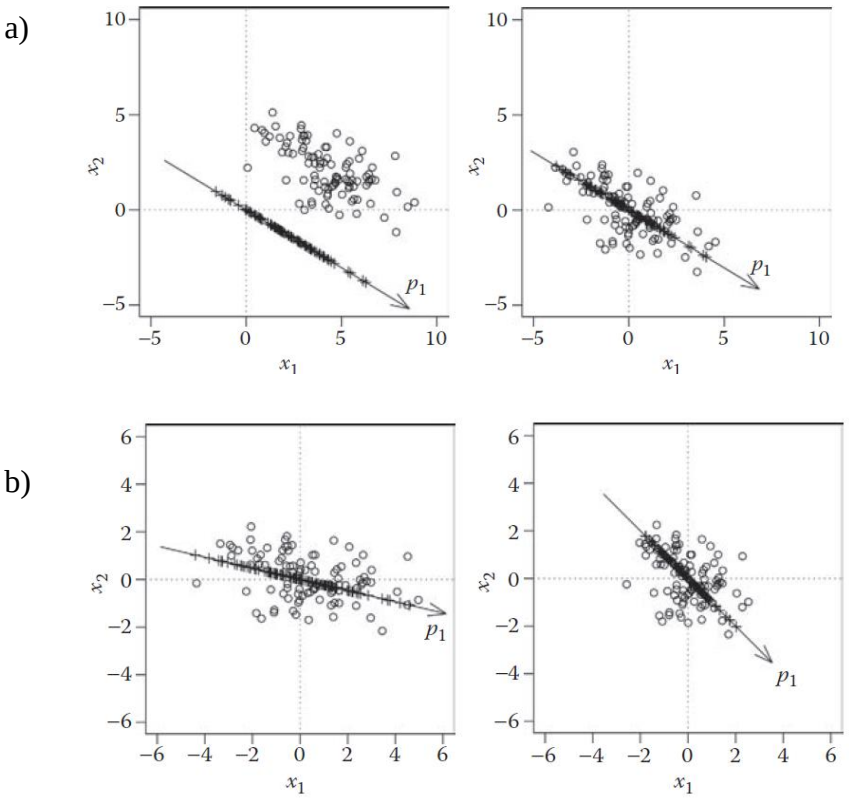


Figure1.2 *(a) The two plots at the top show the effect of mean-centering preprocessing on PCA scores. (b) The two plots at the bottom show the effect of scaling preprocessing on PCA scores already mean-centered. From the top-left corner to the bottom-right corner the overall representation of autoscaling is highlighted [29].*

1.1.3 Partial least-squares

Partial Least-Squares, also known as Projection on Latent Structures was introduced in the late seventies by the statistician Herman Wold for the numerical analysis of chains of matrices in econometrics [30] [14]. The method was then extended to the chemometrics field by his son Svante Wold and many others [31] [14] [32] in the eighties.

PLS is defined as a multivariate linear regression method to correlate and find the relationship between two blocks of data **X** and **Y**, with the final purpose of generating a predictive model to estimate the relative variables in **Y**. As suggested by the name, the projection technique concept is similar to an extension of the PCA to a regression problem between two matrices, but with an important difference in the application of the algorithm. As a matter of fact, if in the PCA the single block is decomposed to sequentially extract the principal components maximizing the amount of variance captured by each of them, in the PLS the principal components selection needs to take into account the directions that explain the larger amount of variance in the predictor block **X** [N x V_X] that is highly and linearly related with variance in the interesting properties in the predicted block **Y** [N x V_Y].

Geometrically, PLS is a projection of the predictors and predicted elements into a new common space in which the correlation between **X** and **Y** is maximized and in which the dimensionality is defined by the new appropriate set of latent variables (*lv).*

From the mathematical perspective, the problem can be reformulated as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E}_{\mathbf{X}}\,, \tag{1.7}$$

$$\mathbf{Y} = \mathbf{J}\mathbf{Q}^{\mathbf{T}} + \mathbf{E}_{\mathbf{Y}}\,, \tag{1.8}$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}\,, \tag{1.9}$$

where **T** [N x lv] and **J** [N x lv] are the score matrix of **X** and **Y**, $\mathbf{P}^{\mathbf{T}}$ [N x lv] and $\mathbf{Q}^{\mathbf{T}}$ [N x lv] are the loading, $\mathbf{E}_{\mathbf{X}}$ [N x V_X] and $\mathbf{E}_{\mathbf{Y}}$ [N x V_Y] are the residuals matrix to reconstruct the original **X** and **Y**. **W** [N x lv] is an additional block of vectors, known as weights, that are required to maintain orthogonal scores.

As a result, PLS is a powerful linear regression method that allows the handling of a large number of initial variables because of its capability of reducing the original space preserving the maximum linear correlation between input and output with a relatively small number of PLS components used for regression.

1.1.4 Multiple responses PLS

One of the several advantages of using PLS instead of other linear regression methods is the ability to model and analyze several *Y* responses together, which is particularly

useful when the *Y's* are correlated and it is likely to have an overall representation of the quality variables. This concept is well explained by S. Wold et al. [12] and it is something to keep in mind when a problem required the modelling of a set of responses. If the multivariate **Y** block is measuring different variables that are fairly independent each other, a single PLS model would result to have many components and thus it is difficult to be analyzed. In this case building up separate models for each *Y* with fewer latent variables is a smart solution to simplify the interpretation. If this is not the case and the *Y*'s are correlated there are no reasons to prefer separate single models. To understand the degree of correlation in the response block, it is suggested to perform a PCA on the **Y** matrix, in such a way that the practical rank of **Y** is obtained and if it is small compared to the number of *Y* variables, it means that the variables are highly correlated and a single PLS with multiple responses is warranted. Contrarily, if the practical rank is not close to one and it required the use of several latent variables, so that the resulting score plot clusters the *Y's* in different groups, separate PLS models for each group should be developed [12].

Modelling a profile response, instead of singular separated variables, gives the advantage of removing background noise uniformly based on the fact that the correlation between input variables and output variables is modelled in a single model. An example of a situation where a multiple responses PLS model is required to predict the quality profile of a target variable is given in the case study presented in Chapter 4.

1.1.5 Multi-way PLS

Multi-way Partial Least-Squares, also referred as MwPLS, is the natural extension of a PLS model to handle three-dimensional matrices of data. Manufacturing data developed along three dimensions are very common in all the type of industries that are dealing with batch processes, where a prescribed amount of materials is processed for a finite time. Thus, the two-way data matrix consisting of $i=1,2,...,I$ batch runs and $j=1,2,...,J$ process measurements can be extended over a $k=1,2,...,K$ time data points, resulting in a three-dimensional matrix $\underline{\mathbf{X}}$ $[I \times J \times K]$, in which the batches are organized along the vertical side, the process variables along the horizontal side and their time progression along the last dimension. Analogously, one or more quality response variables, $m=1,2,...,M$, can be disposed in a three-way response matrix $\underline{\mathbf{Y}}$ $[I \times M \times K]$ as shown in Figure 1.4.

Although several direct modelling approaches to three dimensional matrices have been proposed, the PLS model is a bilinear model and for this reason it requires an unfolding procedure to reorganize the data in two dimensions and finally perform a normal PLS. The choice on the direction of unfolding is determined by the source of variability that

is more likely to be modelled and, particularly, for batch data analysis the most common unfolding techniques are variable-wise unfolding and batch-wise unfolding [33].

Variable-wise unfolding rearranges the data in a such a way that all the batches are disposed one below another and each sampling point becomes a sample (row) of the resulting two-dimensional **X** or **Y** matrix as:

$$\underline{\boldsymbol{X}}\,(I \times J \times K) \Rightarrow \boldsymbol{X}\,(KI \times J)\,, \tag{1.10}$$

On the other hand, batch-wise unfolding reshapes the data to obtain a final two-dimensional matrix X or Y where the batches are placed side by side in a such a way that a complete batch is a sample as:

$$\underline{\boldsymbol{X}}\,(\mathrm{I} \times \mathrm{J} \times \mathrm{K}) \Rightarrow \boldsymbol{X}\,(\mathrm{I} \times \mathrm{JK})\,. \tag{1.11}$$

In the first case, only the variances and instantaneous cross-covariances of the variables are taken into account and, as a consequence, the dynamic of the batches is not modelled.

In the second case, due to the fact that the batches are one aside the other, variances and instantaneous cross-covariances are now captured by the model at every sampling time.

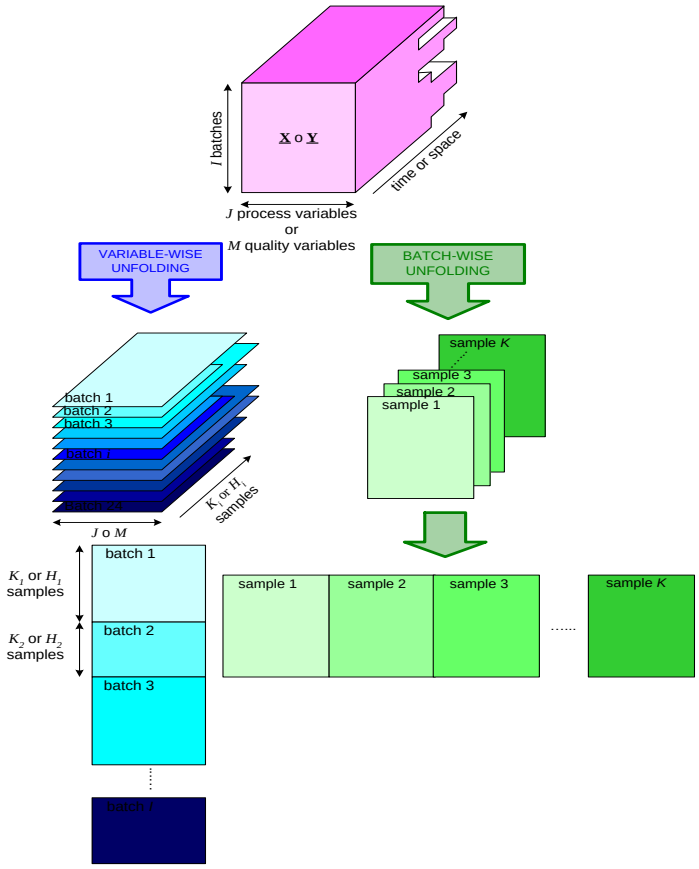


Figure 1.3 *A three-dimensional data matrix and two possible ways of unfolding it.*

The implication of this unfolding technique is that the dynamics of the variation around the average trajectory at each point time are explained, thus when a PLS model is applied between **X** and **Y** a prediction of the average trajectory of a quality response variable for future batches based on a historical database can be obtained [34].
The capabilities of this latter approach are spread across a wide range of applications, e.g. they include the possibility of on-line soft sensing to predict end-quality point during the batch evolution, modelling dynamic system for process analysis or control system design [19][35]. However, the major drawback of this procedure is that the batch runs need to all have the same length in order to be unfolded. This problem can be solved applying several different strategies of batch alignment or synchronization, depending on the origin of the processed data [36].

1.2 Statistical pattern recognition

The availability of an enormous quantity of data and the need to extract from them as many information as possible is one of the challenge of the next future in order to make an effective and profitable use of them.
Nowadays the problems of automatic identification, characterization, classification and grouping of patterns in dataset where nothing or little is known about, are common and frequent in all the scientific disciplines such as engineering, chemistry, biology, drug discovery and many more. However, giving a general but accurate definition of what is a pattern is not trivial. Watanabe specifies a pattern "as opposite of a chaos; it is an entirely vaguely defined, that could be given a name" [37]. It appears clear that in the recognition of a certain pattern the decision-making part plays a fundamental role and it involves the use of an arbitrary intelligence to attribute a certain qualification to the patterns identified.
Depending on the amount of initial information available for a given dataset, the exercise of pattern recognition may lead to one of the two following tasks:

- unsupervised pattern recognition (e.g. clustering), in which multiple patterns needs to be discovered and grouped based on the similarity of the feature's descriptors;
- supervised pattern recognition (e.g. discriminant analysis), also known as classification, in which new objects need to be assigned to predefined known classes.

The statistical approach to these situations can be defined as a more general problem where each element is represented in terms of *d* features or measurements, and it can be viewed as a point projected into a *d*-dimensional space. For unsupervised pattern

recognition, the goal is to identify compact and disjoint regions of elements in the *d*-dimensional space and select the set of features that better described each of them. For supervised pattern recognition, the set of features that describes a certain pattern needs to be learnt from the training set, so that the objective is to establish the decision boundaries in the feature space that better separate elements belonging to different classes. In both the cases the effectiveness of the feature set recognition is given by how well patterns from different classes can be divided [38].
In practice, applications may happen that both these tasks need to be performed one after the other, even if this is not the most common situation. However, an example of this scenario is given in the case study presented in Chapter 3.

1.2.1 Unsupervised pattern recognition

Unsupervised pattern recognition techniques, usually referred as "clustering techniques", are very common in the data mining context where the information available for a database of samples is very limited. The final goal of these techniques is to systematically organize into groups the elements of the matrix in such a way that the distances within elements of each singular group is minimized, whereas the distances between the different groups is maximized. The main concept of clustering can be summarized as the formation of groups of records with similar features. The problem is to find the most suitable criteria, that has to be translated into an algorithm, to quantitatively define the concept of similarity as a mathematic statement for measuring proximity between records and grouping them into meaningful clusters.
The importance of the capability of identifying and labelling records appears obvious, but, as sharply pointed out by Gelbard et al. [38], there are several reasons that explain why this approach is still seen with some perplexities in the industrial applications. Firstly, the variety of techniques developed in this field raise the issue of standardization because different algorithms produce different outcomes and there is any standard procedure to compare them. Secondly, a consistent effort and responsibility has to be dedicated to the interpretation of the various clusters results and it requires the support of some expertise of the system analyzed. As a matter of fact, the clustering process might be unpredictable and not always the results match the expectations. Finally, the quality of the clustering algorithms is not clearly measurable and a comparison between several techniques needs to be performed in order to select the most suitable method for the specific problem.

However, cluster analysis is a very powerful tool to rapidly organize large amounts of complex data if its use is smartly applied considering the boundary conditions of the original system.

1.2.1.1 Hierarchical Clustering Analysis

Hierarchical cluster analysis (HCA) is an agglomerative method of clustering in which the data are organized in a nested sequence of groups according to a linkage rule. The general concept behind agglomerative methods is that it begins with each object forming its own cluster to then progress by combining the initials clustering into larger ones based on a distance selection criterion between the elements and the clusters.
Even in this case, there are several distance measures between samples that can be used and a wide choice of linkage algorithms to guide the clustering formation.
Initially, a distance measure needs to be defined and the choice can affect the decision of applying it to the original elements' space dimension or to the reduced space if a preliminary PCA is performed. The two most popular distances used in the chemometrics applications are:

- Euclidean distance $$b_{x,y} = \sqrt{\sum_{d=1}^{D}(\boldsymbol{x}_d - \boldsymbol{y}_d)^2}\,, \qquad (1.12)$$
- Mahalanobis distance $$b_{x,y} = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T \cdot \boldsymbol{Z^{-1}} \cdot (\boldsymbol{x}-\boldsymbol{y})}\,, \qquad (1.13)$$

where **x** and **y** are two vectors in a *d*-dimensional space that represent two generic records *x* and *y* with *d*=1,...,*D* measurements, and **Z** is the variance-covariance matrix whose elements are the covariance between each pair of variables. The Euclidean distance is the simple geometrical concept of distance between two objects in a 2-dimensional space that can be easily generalized to a *d*-dimensional space, where *d* is used to define a generic dimension that is equal to the D measurements in this specific case. In this case, each variable is considered equally important in determining the final distance between observations. It can be used even in the case of data compression, e.g. PCA reduction, and in this case the distance in the new reduced space is calculated between the new set of coordinates given by the scores.
On the other hand, the Mahalanobis distance can directly take into account and compress the effect of correlation between variables by decreasing the weight associated to such variables because it involves the calculation of the inverse of the covariance matrix. It is usually used to avoid the principal component reduction and the consequent principal component selection on the original dataset, but the computation

of the variance-covariance matrix might cause problems if the original dataset present a high degree of multicollinearity [40].
Once the distances between elements have been calculated, the new distance between clusters need to be computed and the clusters need to be linked using what it is normally referred as “cluster method” and its “linkage rule”. Once again, there are several possible alternatives to obtain the inter-cluster distances.
A short description of the most used methods is given below and Table 1 summarizes the methods’ differences according to the Eigenvector® documentation (http://wiki.eigenvector.com/index.php?title=Main_Page) :

- **nearest neighbour**: The distance between any two clusters is defined as the minimum of all possible pair-wise distances of objects between the two clusters; the two clusters with the minimum distance are grouped together. This method tends to perform well with data that form elongated "chain-type" clusters.
- **furthest neighbour**: The distance between any two clusters is defined as the maximum of all possible pair-wise distances of objects between the two clusters; the two clusters with the minimum distance are grouped together. This method has better results with data that form "round", distinct clusters.
- **pair-group average**: The distance between any two clusters is defined as the average distance of all possible pair-wise distances between objects in the two clusters; the two clusters with the minimum distance are grouped together. This method shows similar performances with both "chain-type" and "round" clusters.
- **centroid**: The distance between any two clusters is defined as the difference in the multivariate means (centroids) of each cluster; the two clusters with the minimum distance are grouped together.
- **median**: The distance between any two clusters is defined as the difference in the weighted multivariate means (centroids) of each cluster, where the means are weighted by the number of objects in each cluster; the two clusters with the minimum distance are joined together. Because of the weighted distance, this method is supposed to perform better than the Centroid method if the number of elements is expected to vary substantially between clusters.
- **Ward's method**: This method does not require calculation of the cluster centres; it links the two existing clusters such that the resulting pooled within-cluster variance (with respect to each cluster's centroid) is minimized.

Table 1.1 *Cluster methods description based on inter-custering distances calculation and linkage rule.*

Method	Distance Between Existing Clusters	Linkage Rule
Nearest Neighbour	Minimum of pair-wise distances between any two objects in each cluster	Join 2 nearest clusters
Furthest Neighbour	Maximum of pair-wise distances between any two objects in each cluster	Join 2 nearest clusters
Pair-Group Average	Average distance between all pair of objects in each cluster	Join 2 nearest clusters
Centroid	Distance between the means (centroids) of each cluster	Join 2 nearest clusters
Median	Distance between the weighted means (centroids) of each cluster	Join 2 nearest clusters
Ward's Method	N/A	Join clusters such that the resulting within-clusters variance (with respect to the centroids) is minimized

One of the greatest features of HCA is the possibility of displaying the overall results of the clustering in the form of a dendrogram, that is the mathematic formalization of a tree representation to illustrate the relationship between observations, according to the different distance levels.

However, the easy reading level of this graphic representation and the variety of different grouping methods might lead to an inconsistent pattern identification and therefore it is advisable to test all the possible method and compare the different results. This task usually includes a quality evaluation of the clustering outcomes guided by some possible statistical indicators. A dendrogram is mathematically defined as the graphical representation of a cophenetic matrix (matrix of similarities between clusters). Thus, a comparison between the cophenetic matrix and the original distances matrix of the unmodeled data can be calculated by means of a cophenetic correlation coefficient (c_h), defined as:

$$c_h = \frac{\sum_{x<y}(b_{xy} - \bar{b})(t_{xy} - \bar{t})}{\sqrt{[\sum_{x<y}(b_{xy} - \bar{b})^2][\sum_{i<j}(t_{xy} - \bar{t})^2]}} \qquad (1.14)$$

where b_{xy} is the Euclidean distance between the x-th and the y-th observations, t_{xy} is the cophenetic distance, that is height of the node at which two elements are first grouped, and $\bar{b}$ and $\bar{t}$ are respectively the average of b_{xy} and t_{xy}. A value of the cophenetic correlation coefficient close to 1 means that the distances in the cophenetic matrix are consistently preserved with respect to the original one [41]. However, the estimation of this coefficient to compare different algorithms is not necessarily an absolute

indicator of which one of the methods is the “best” one. Generally, a good method should preserve the original information about the similarity between samples and this means that the cophenetic coefficient should be high, but this does not imply that the method with the higher coefficient is the most well-performing in terms of correct clustering. This concept will be better exemplified in the case study of Chapter 3.

1.2.2 Supervised pattern recognition

Supervised pattern recognition, commonly called “classification”, is a standard task for most of the scientific disciplines because it responds to the demand of classifying new elements according to a list of known groups. The classification of a new molecule starting from its functional groups or the identification of a genetic disease by identifying the sequence of the protein-code are just some possible examples of real situation where large dataset of classified samples can be used to create a training set for a future automatic classification. Mathematically, the formalization of the so called training set is a mathematical model between some measurements or variables that defines *G* classes previously determined. A practical example, that anticipates the case study of Chapter 3, is about whether it is possible or not to identify the classes of flowability of new powder materials based on their material characterization properties. Usually a limited existing dataset where the classes are established is available but it is not always possible to test a relevant amount of a new powder to understand their flowability. Much easier is to characterize few relevant indicators and try to identify the pattern similarity compared to the existing classes in the training set.
Several classification methods, also called “classifiers”, have been proposed in the last fifty years of research, but the statistical approach to the problem is the one that has been most strongly investigated and used in practice. Because of the increasing complexity required by the novel applications, more sophisticated techniques like artificial neural networks and methods imported from statistical learning theory are taking hold. However, in spite of the numerous methods proposed in the literature, it is not possible to state that a method is superior to the others and it is able to solve *any* kind of problem of supervised pattern recognition. Furthermore, a more complex method is not a synonym of a major suitability for complicate systems and vice versa. As for unsupervised pattern recognition, the choice of the method is highly related to

the nature of the original dataset to be modelled and a comparison between many algorithms is the only way to determine the one to be used.

The performance of a classification method can be evaluated using a simple statistical parameter like the percentage of samples correctly classified (*% CC*).

Generally, the procedure to determine the quality of predictions of a model is called "validation" and it can be done following two subsequent steps:

- internal cross-validation of the training set;
- external validation on an independent test set.

The first consists in an iterative process where the original matrix of data is split in subsets, each of them constituted of one or more samples. A model is then trained using a subset as training set to train the supervised learning algorithm and the remaining data are used for testing the model performance. This procedure is then repeated several times randomly partitioning the original data for the next iteration and calculating the cross-validation error associated. The average cross-validation error is finally used as a statistical indicator to prevent overfitting during training. The strategy of the data partition can be randomly selected or it can be standardized using some common techniques like *k*-fold, leave-one-out and many others. Usually, this model assessment technique is not suggested for a stand-alone use unless the number of original samples is limited.

The second next step of external validation on a "blind test" is always recommended and it consists on considering a series of samples to be unknown class membership since the beginning and now the model, built up using just the training set, is applied to the external set of data to test the percentage of correct classification. In this case a common and robust rule of thumb to determine the test size is 80% of the data should be used for training and 20% for external validation. However, once again this is largely dependent on the dimension of the original dataset and its class partition [42] [38].

As already mentioned, existing classifiers algorithms are numerous and they follow a wide variety of classification approaches that generate different features and related suitable applications. Nevertheless, in the statistical approach classifiers can be distinguished as local or global, if only a small part or all samples are taken into account for class assignment; class-modelling, when a specific and delimited space for each class is identified and object projected outside are not classified; parametric or non-parametric, if the form of the analytical distribution is known or not, distance-based if the classification is based on a distance criteria between elements; linear or non-linear, on the basis of how class boundaries are derived; probabilistic if they assume a certain probability distributions of the samples [43].

1.2.2.1 Support vector machines

Support Vector Machines (SVM) are a non-linear data modelling technique developed by Vapnik in 1995, in the context of machine learning classification and, few years later, it has been extended to solve non-linear regression problems [44] [45]. In the following years, this new approach for solving classification problems has gained a very high popularity not only in the machine learning framework, but also in the chemometrics community. The reasons behind this unique success are the interesting novel features and promising performance of the method compared to other traditional approaches, that ensure a flexibility in solving problems with a high degree of complexity.
Among all the attractive features of SVM, the key characteristic is that class definitions are obtained "drawing" the boundaries between classes using the samples that are laying in the proximity as "supports". The mathematical formulation revolves around the structural risk minimisation (SRM) concept:

$$R_e \leq R_{emp} + \sqrt{\frac{d_{VC}\left(log\left(\frac{2N}{d_{VC}}\right)+1\right) - log\left(\frac{\eta}{4}\right)}{N}} \qquad (1.15)$$

where N is the number of samples in the training set, d_{VC} is the Vapnik-Chervonenkis dimension that roughly relates to the complexity of the boundaries, R_e is the expected risk that relates to the true error, R_{emp} is the empirical risk that relates to misclassification error as observed during model building, and lastly, η is 1 – the probability that the upper bound defined for R_e holds. In particular, R_{emp} relates to the error measured on the training sample itself, while R_e relates to the "true" underlying error [45]. Consequently, the best classifier is the one that minimises the upper bound on R_e and not the one that minimises the error on the training data as in the traditional empirical risk minimisation (ERM), employed by conventional neural networks.
Furthermore, SRM formalises the important conception that the complexity of the classifiers should always be linked to the size of the training set available (the number of samples N), in such a way that the complexity has to be controlled in order to avoid underfitting or overfitting during modelling. This key feature of SVMs allows to relax and generalise the class definition problem, which is the main goal in statistical learning [47].
Despite the complex formulation of the SVM algorithm, its mathematical derivation can be summarized in three parts:

- basic definition for linearly separable classes;

- extension to the non-linearly separable case thanks to the use of kernels functions;
- generalised solution and implementation of trade-off parameters to control complexity.

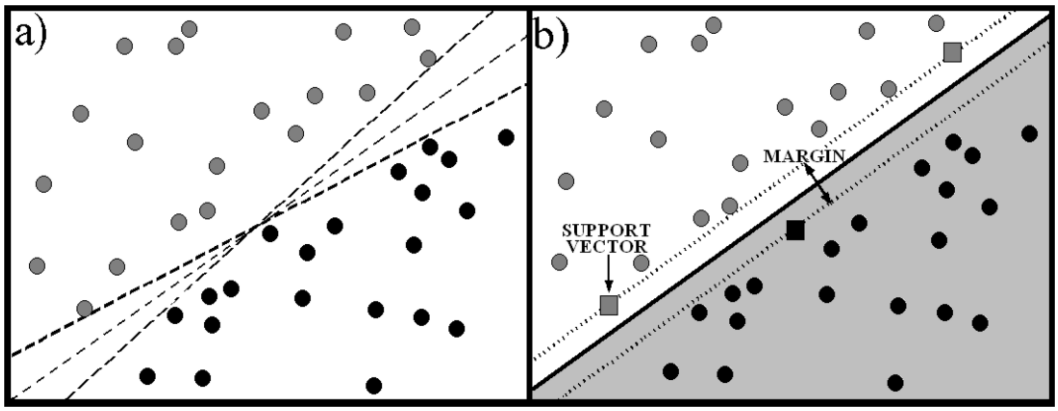


Figure 1.4 *(a) Two classes example of the numerous possible separating hyperplanes and (b) identification of the optimal hyperplane for a two class linearly separable case using the closest samples (square marks) as support vectors to maximise the margin (black dotted lines) [46].*

In the first case, a binary classification problem for linearly separable classes can be considered as in the example of Figure 1.5. The classification function to determine the boundary between the two classes can be constructed as a hyperplane, i.e. a line if two-dimensional, a plane if three-dimensional and, generalizing, a n-dimensional hyperplane if the space is n-dimensional.

It is not difficult to understand that an infinite number of hyperplane that split the two classes respecting the class membership can be found. Thus, the problem is to determine which is the one that separates them in an optimal manner and that corresponds to the situation where the distance between closest samples and the hyperplane is maximal. Mathematically, the optimal hyperplane must be equally spaced from the two classes and the space between the classes takes the name of margin. Hence, the problem is reduced to an optimisation problem of minimization of the margin expressed by a Lagrange function through a quadratic convex programming problem. As previously mentioned, to construct this margin only the observations that lies on the margin will be used and they are actually named support vectors since they solely determine the solution. As a result, all the other samples could potentially be removed from the training set and the boundaries definition would remain the same.

Extending the case to a non-linear separable problem, the boundaries identification task becomes more complicated. An intuitive example is given in Figure 1.6 where two

classes are present but it is not possible to find a suitable hyperplane to split them as required.

In fact, this complex situation is not so uncommon as the descriptors of the system are linearly not correlated. However, SVM are able to handle this situation, in a very smart and efficient way, adding an extra step to the optimisation problem previously explained. In order to do that, the data needs to be "projected" into a very high-dimensional space (compared to the original data space) by the mean of a feature function $\varphi(x)$. Then, the algorithm finds the optimal hyperplane through the margin minimization always using the closest samples as supports. The feature function $\varphi(x)$ is selected from a restricted family of functions called kernels and this dimensional stratagem is known as kernel trick. Figure 1.6 schematizes the concept through the simplest example of a 2-dimensional space projected to a 3-dimensional space, but as it has just been outlined the kernel trick is applied to a much higher dimensional space increasing the capabilities of modelling very complex distributions. Several different kernel functions can be chosen in order to model the boundaries shape but the most popular choice is the so called Radial Basis Function (RBF) because of its singular tuneable parameter (the radial width γ).

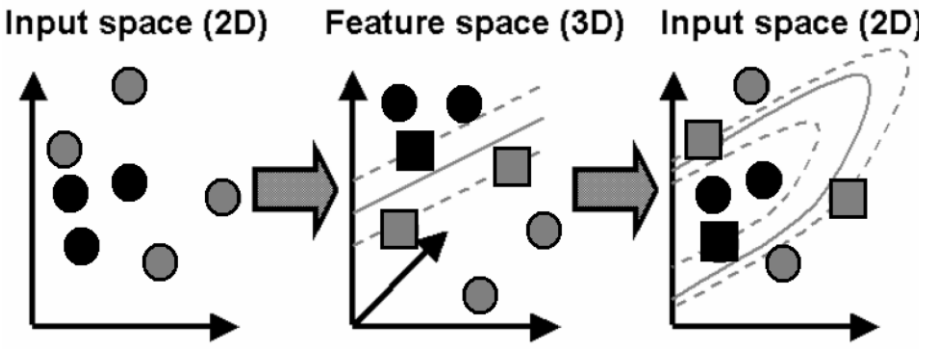


Figure 1.5 *Boundary construction for a non separable case after projection into a higher dimensional space, where the optimal hyperplane is defined. The squares indicate the support vectors [46].*

The last step of the algorithm procedure is the introduction of a trade-off parameter to control the complexity of the system since the kernel trick might lead to very intricate boundaries with the risk of overfitting. As a matter of fact, the radial width of the kernel γ is not enough to shape the boundaries taking into account simultaneously the margin maximisation and training error minimisation accordingly to the SRM concept. Thus, an additional parameter *C*, called penalty error or cost, is introduced to balance which of the two aspect should be emphasised. A lower penalty error will accentuate the margin maximisation, whereas a higher penalty error will increase the bias towards the training error minimisation. This new parameter is particularly important when the

system is characterized by highly non-linear boundaries and the risk of degrading the performance of predicting new samples membership is likely to happen.

An intuitive representation of the effect of both γ and *C* on the boundaries modelling construction is given in figure 1.7 and figure 1.8, keeping firstly *C* and then γ constant respectively to isolate the singular effect of the parameters.

Although the method was initially developed for two classes separation problem, several different procedures have been proposed to treat multiclass problems since they are common to be encountered in practical applications. However, the general approach to the classifiers definition does not change [48].

For a more rigorous mathematical description of the SVM algorithm, refer to [48] [46] [47] [44].

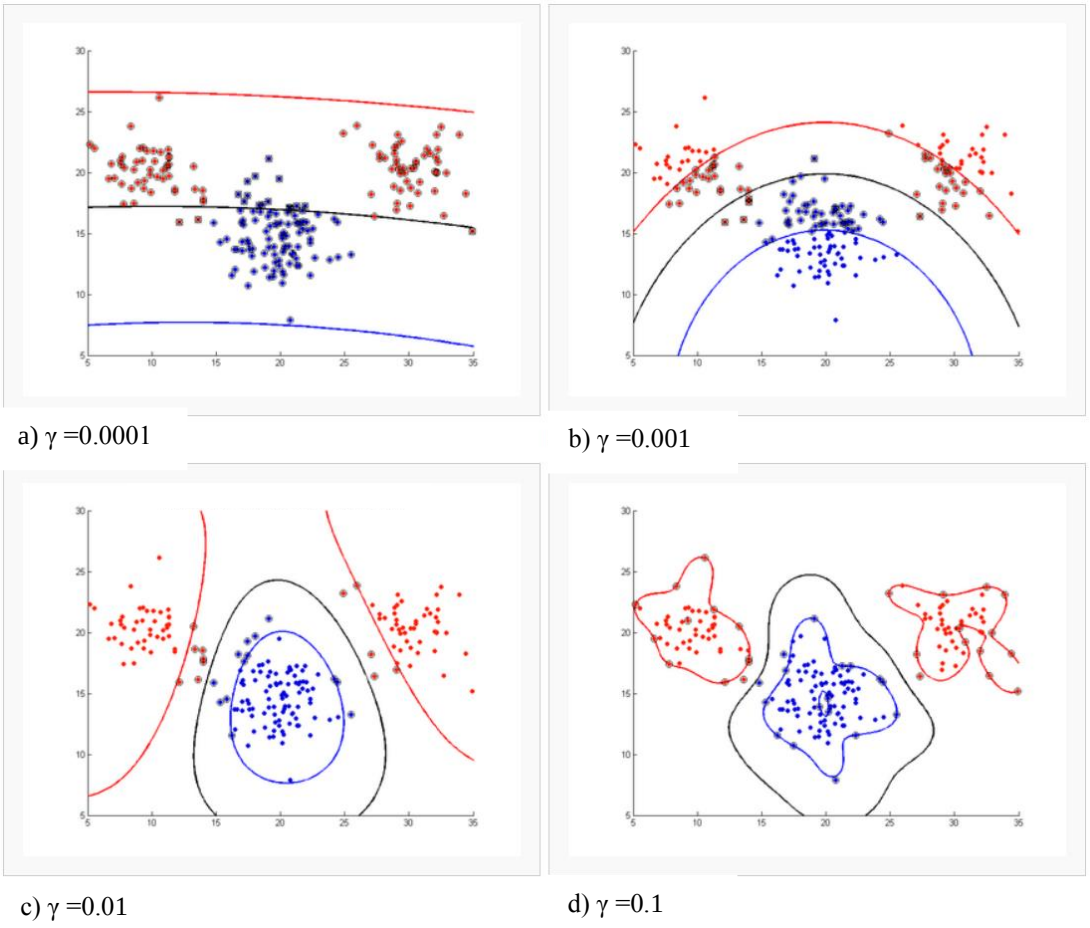
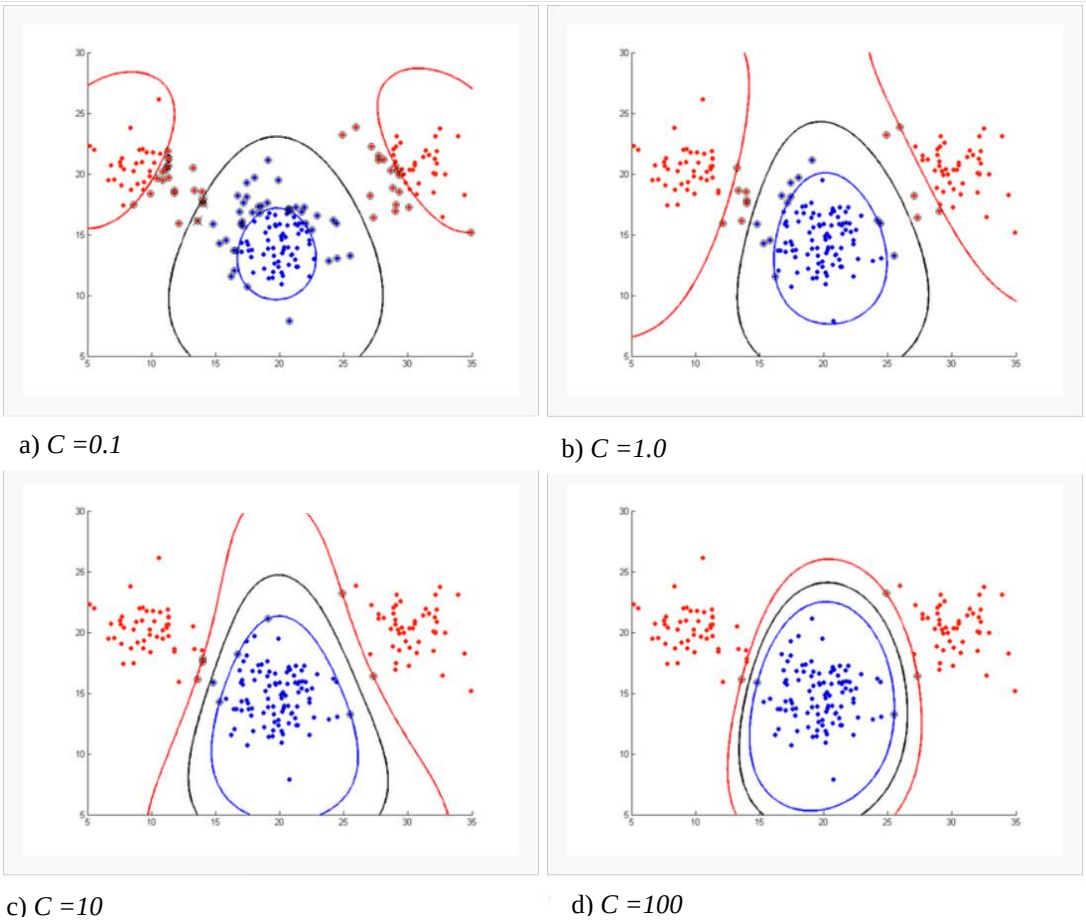


Figure 1.6 *Intuitive two classes example of margin distortion effect varying γ parameter with cost constant C=1.0 and (a) γ =0.0001, (b) γ =0.001, (c) γ =0.01, (d) γ =0.1. (http://wiki.eigenvector.com/index.php?title=Svmda)*



a) *C =0.1* b) *C =1.0*

c) *C =10* d) *C =100*

Figure 1.7 *Intuitive two classes example of margin distortion effect varying C parameter with γ constant γ=0.01 and (a) C=0. 1, (b) C =1, (c) C=10, (d) C=100. (http://wiki.eigenvector.com/index.php?title=Svmda)*

1.2.2.2 Partial least-squares discriminant analysis

A partial least-squares discriminant analysis (PLS-DA) is nothing more than an extension of PLS for problems of classification and discrimination [49]. A multivariate regression model is constructed maximizing the covariance between the matrix of the samples features **X** and a class membership matrix **Y**. For multi-class problems, the class membership matrix **Y** is obtained by binary column encoding of the assigned class, where 1 indicates that a sample belongs to that class and 0 does not belong to that class [50].

The resulting model is able to project onto a lower dimensional space the relationship between data that explains their class affinity and this is done by a linear multivariate regression with the advantage of the capability of handling numerous co-linear input descriptors. Moreover, a PLS discriminant model, contrarily to conventional regression, does not need to model a response to fit exactly the data, so that the inner connection

between the respective assigned class is more likely to be identified and learnt by the original training set.

1.2.2.3 *K* nearest neighbours

A very common reference method to test the performance among classifiers is the *k*-NN, or *k*-nearest neighbours. This is mainly due to its formulation simplicity and its non-linear approach to supervised pattern recognition of complex dataset. Compared to the methods above, *k*-NN requires a much simpler algorithm and a very reduced computational effort. As a matter of fact, it is a distance-based method where a sample is classified according to the majority class of its *k*-nearest neighbours in the original data space [43]. As a result, the algorithm just need to calculate and analyse the distances between all the possible pairs of samples and identify which class shows the prevailing appearance among the selected *k* closest neighbours.

Despite its simplicity, this method has several advantages because it does not require the formulation of numerous underlying assumptions, e.g. probability density function or normality of noise distributions and so on, and it has also the great feature of handling multiclass problem in a non-linear environment. Moreover, the original space can be reduced by principal component technique like PCA to take into account co-linear variables and noise reduction.

On the other hand, this method has several limitations that need to be considered before applying it to any kind of dataset. Firstly, it is very sensitive to the choice of the distance measure and data pre-treatment. Secondly, the numbers of samples in each class of the training set should be approximately equal to not compromise the consistency of the majority vote paradigm. In addition, unless otherwise implemented, each variable assumes the same importance and the spread or variance in a class is not taken into account. Finally, ambiguous or outlying samples can hardly influence the resultant classification and the method becomes not sensitive to complex class boundary definitions [42].

Chapter 2

Continuous direct compaction tablet manufacturing and powder feeding

This Chapter gives an overview of continuous manufacturing of solid oral dosage to contextualize the framework of the data analysis applications that are described into details in the next chapters. A particular focus is given to the next generation of continuous direct compression lines for tablet manufacturing, defining extensively the powder feeding equipment and its modes of operation. Finally, a generic description of the modelling flowchart approach, used in this Master Thesis project, concludes the Chapter.

2.1 Continuous manufacturing of solid oral dosage pharmaceuticals

2.1.1 General overview

Oral solid dosage (OSD) forms, such as tablets and capsules, are still the most common products form commercialized in the pharmaceutical market after almost two centuries since their introduction. This is not just the result of an established marketing confidence of this kind of products, but tablets and capsules still represent about half of all new medicines licensed by FDA [51] [52].

The pharmaceutical manufacturing process of a solid drug is the result of two sequential phases:

- upstream, or so called primary manufacturing, in which the active pharmaceutical ingredient (API) is produced by several alternatives of synthesis methods;
- downstream, or so called secondary manufacturing, in which the API is combined with other pharmacologically inert solid substances, known as excipients, to complete the final dosage form.

In the context of secondary tablet manufacturing, different methods can be used based on the inherent material properties of all the compounds to meet the final goal of ensuring the best quality product. Historically, the most common routes of manufacturing are implemented in a sequence of batch modes techniques, even if, as

already mentioned, over the last twenty years the pharmaceutical industry is pushing forward the development of continuous lines. Among all the solutions, three major routes can be identified: direct compression, dry granulation and wet granulation (Figure 2.1). The choice of the most adapt way of manufacturing is related to the material properties of the powders required by the specific drug formulation.

This project aims to aid the development of a continuous direct compression line, therefore, this alternative only will be presented in the next section, particularly focusing on powder feeding equipment.

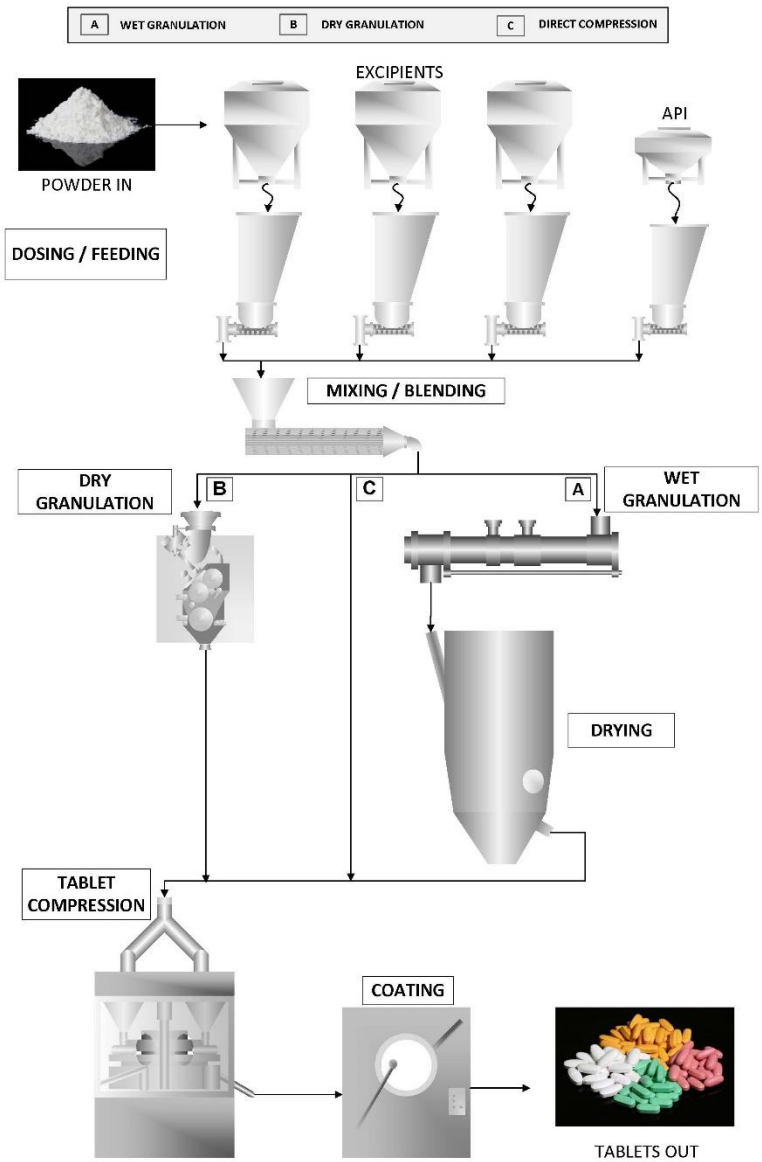


Figure 2.1 *Continuous OSD manufacturing process flow (https://www.tabletscapsules.com/enews_tc/2016/editorial/images/tc_ask_10_10_16full.jpg).*

2.1.2 Continuous direct compression process

Direct compression is the simpler route to implement a continuous compaction manufacturing plant. Therefore, it has generated interested in developing it in a commercial scale. Contrary to dry and wet granulation, it does not require a size enlargement of particles during processing because of the suitable granular flow and dissolution properties of the API. Thus, the core technologies of this process become the initial continuous powder feeding and mixing phase, and the final tablet compression. A schematic block flow diagram of the direct compression process is illustrated in Figure 2.2.

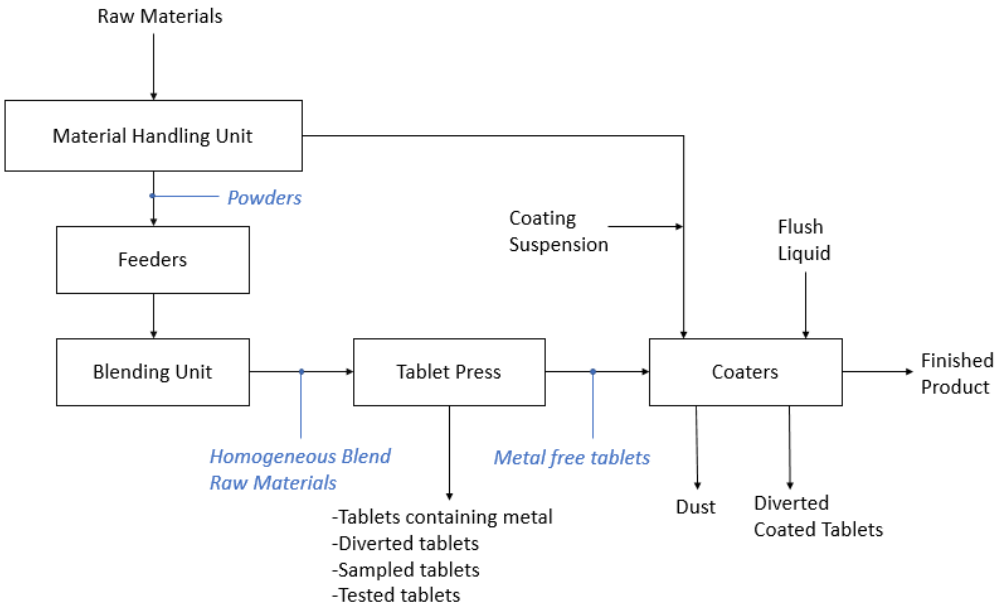


Figure 2.2 *Direct compression block flow diagram.*

Multiple feeders at the beginning of the line provide the API and the excipients according to the formulation. Feeders feed in the same point the continuous blender and it is very important to constantly reach the set point feed rate to avoid content uniformity issues in the final tablet. The mass flow rate set point depends on the recipe of the specific product. Problems arisen in the feeding phase of the process can compromise the final quality specifications, i.e. the content of API in the tablets. Eventually, the blending unit is designed to damp small disturbances in the mass flowrate fed to the unit and ensure a homogeneous concentration within certain acceptability limits. Once the powder is homogeneously blended, the material is supplied downstream to a rotary tablet press and the blend uniformity is measured by an integrated near-infrared (NIR)

spectroscopy probe just before entering the rotary press to eventually reject the product or adjust the process via an advanced feedback control system. The tablets are obtained by a simple compaction of powder and during the compression weight, thickness and hardness specifications are monitored to eventually reject them when they are out of the limits. At this point, tablets within specification targets are sent to a metal check device and they are finally ready to be coated.

The tablet coater can be manually loaded in the line or automatically loaded by a vacuum transport system. The transfer from the press to the coater permits an auto-dedusting effect and dust is sucked via a dedicated vacuum line. The tablet coater consists of multiple rotating wheels that can operate in parallel with different aliquots of tablets. The wheels rotate at high speed allowing a centrifugal redistribution of the tablets against the perforated wall of the coater. When the tablets are well distributed, a central nozzle sprays the coating fluid and some air nozzles force the tablets away from the wall and dry the tablets' surface at the same time.

Finally, the product is discharged and it is ready to be packaged and commercialized.

2.2 The loss-in-weight feeder

The loss-in-weight feeder, as shown in Figure 2.3, simply consists in:

- a hopper;
- a single or twin screws;
- a load cell;
- a feedback control system.

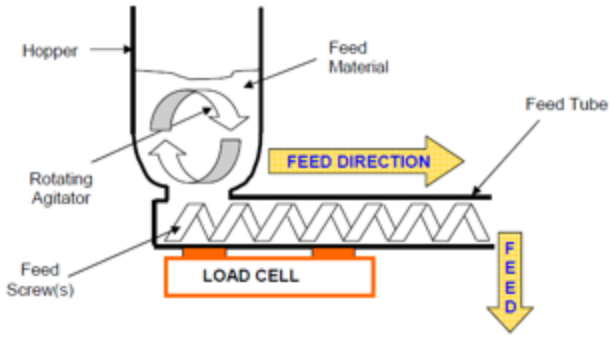


Figure 2.3 *Sketch of a loss-in-weight feeder.*

The loss-in-weight feeding technology is based on achieving the target mass flow rate by re-weighting constantly the total system (feeder, hopper and material contained in it) to determine weight loss per unit time, which is equal to the mass flowrate fed into the process. The feedback control system adjusts the controlled variable (i.e. mass

flowrate) by acting on the manipulated variable (i.e. the motor speed). A transmission provides a controlled speed and torque conversions from the motor to the screw(s) through a gearbox system. The relation between motor speed and screw speed is linear, thus from a conceptual point of view speaking about motor speed or screw speed does not make any difference.

This control approach ensures a high accuracy, but has some drawbacks, especially at high feed rates and long-term campaigns. To avoid the use of large-volume hoppers, a periodic refill of the hopper is required. It is during the refill phase that significant deviation in the feed rate may occur due to intrinsic inaccuracies in the control. The problem of the refill can be avoided if large and expensive systems are designed, so that refill phases are minimized. Nevertheless, this causes several problems in terms of cost for structural changes to the equipment and to the plant itself and the problem persists at high feed rate or long term campaigns [54].

During the refill, the weight-based gravimetric control must be temporarily bypassed by a volumetric control because the introduction of new material perturbs and compromises the load cell mode of operation. In this situation, the load cell is technically "blind" and the control strategy must change to face the different phenomena that produce some variations in the process conditions.

Figure 2.4 helps to understand the different cyclic operating phases of a feeder plotting the weight versus time. It is important to anticipate that cycles can have a different evolution in time and amount of material refilled because the mass flowrate is not perfectly constant in the industrial operation of the plant.

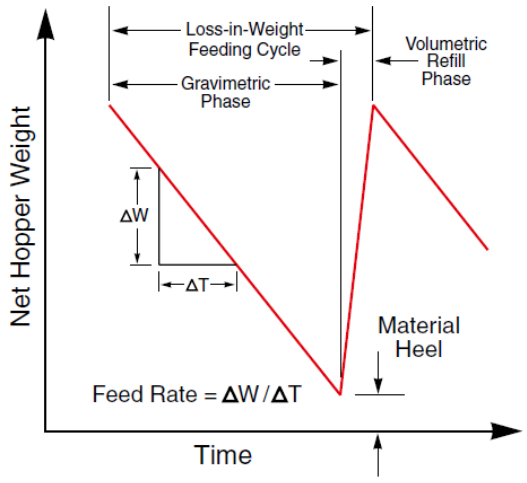


Figure 2.4 *Cyclic operating phases of a loss-in-weight feeder system looking at the net hopper weight profile* [54].

Moreover, the refill starts rather before the hopper does empty, and that occurs for two main reasons. Firstly, because it is necessary to assure a constant supply of powder to the process to reach the target required without discontinuities. Secondly, the incoming aerated powder from the refill may increase the pressure and may cause uncontrolled flooding in the output rate, if the level of powder in the hopper is too low.

Nevertheless, it is reasonable to expect an increase of the density of the heel of material that remains in the hopper during the refill due to the fact that the new material entering the hopper has the effect of compacting the powder already there. This change in the density profile has been confirmed by several tests and manufacturing trial, and causes a progressive overfeeding of material to the downstream process that can vary in a range between +1% for relatively constant density material and +10-15% for powders whose density can vary substantially [53]. The geometry and the design of the feeder and the refill system contribute significantly to the entity of the overfeeding variation, but the control strategy still plays the major role.

Feeder manufacturers proposed several control strategies and equipment design solutions to solve the problem experimenting various and different routes and approaches to face the challenges presented during the refill of the hopper.

Traditionally, the control loop was maintained opened throughout the refill phase and the screw speed kept constant to the last value assigned by the controller just prior to entering the volumetric phase. This approach drives the system out of control during the entire refilling phase and even for some time after refill completion because the weighing system and the feedback control needs some seconds to accumulate some measured data and stabilize the gravimetric control.

The new generation of loss-in-weight feeders aims to maintain the gravimetric accuracy even during the refill, implementing a complementary control system to gradually adjust the screw speed to counterbalance the effects of what is happening in that situation. This result can be achieved by storing in the controller's memory an array of new variables called *feed factor* that it is defined as the ratio between the mass flowrate flowing out from the output of the screw and the screw velocity as shown in eq. 2.1:

$$ff = \frac{\dot{m}}{v_s} = \left[\frac{g/_s}{rev/_s}\right] = \left[\frac{g}{rev}\right] \tag{2.1}$$

This new variable is somehow representative of the density of powder since it measures the grams of powder released in a revolution of the screw, thus higher feed factor values means higher density of the powder since the powder contained in a single revolution increases.

This new array of values is continuously evaluated during the gravimetric exercise of the feeder operation. Problems arise when the load cell is blind, so no information about the mass flowrate can be directly retrieved. In this case, the estimation of the situation inside the feeder through the feed factor evaluation should maintain the system under control. However, estimating the real situation inside the feeder is not easy since many phenomena concur at the same time. As a matter of fact, when the gravimetric operation of the equipment has begun, the density of the material inside the hopper should slightly decrease in time because less material in the higher portions of the hopper is adding pressure. The weight constantly decreases to maintain constant the feed rate of powder (Figure 2.5a). Therefore, the motor speed is relatively constant in a first gravimetric phase and successively increasing when the level decreases (Figure 2.5b) and less dense powder must be supplied. In this situation, the feed factor profile gradually decreases with the weight of powder in the hopper until the density profile is gradually changing, but depending on the inherent material properties of the powder, it will drop down, more or less depending on the flow properties, dramatically when the powder reaches a certain level in the hopper. Figure 2.5d shows the time profile of the feed factor for a typical cohesive material. In this case, the feed factor decreases fast and in a constant manner until the refill happens. At that point, the compression of the powder at the heel of the hopper during the refill would lead to a rise in the density, as already mentioned. Then, the velocity of the motor speed should be adjusted, and particularly quickly decreased (Figure 2.5b), to minimize the mass flow error associated (Figure 2.5c).
Observing Figure 2.5 from an overall perspective, the profiles of the variables involved in the dynamic operation of the equipment can be understood and the effect of the speed array correction on motor speed and mass flowrate can be appreciated. However, industrial experience shows that providing the controller with the right feed factor profile during refilling is not so trivial, especially when operating with different materials in terms of characterization properties, i.e., PSD, density, flowability and so on. Moreover, to minimize the variability in the screw speed profile associated to the inversion of tendency caused by the change in the density along the hopper and subsequent to the refill, planning a proper refill strategy that varies depending on the material is fundamental.

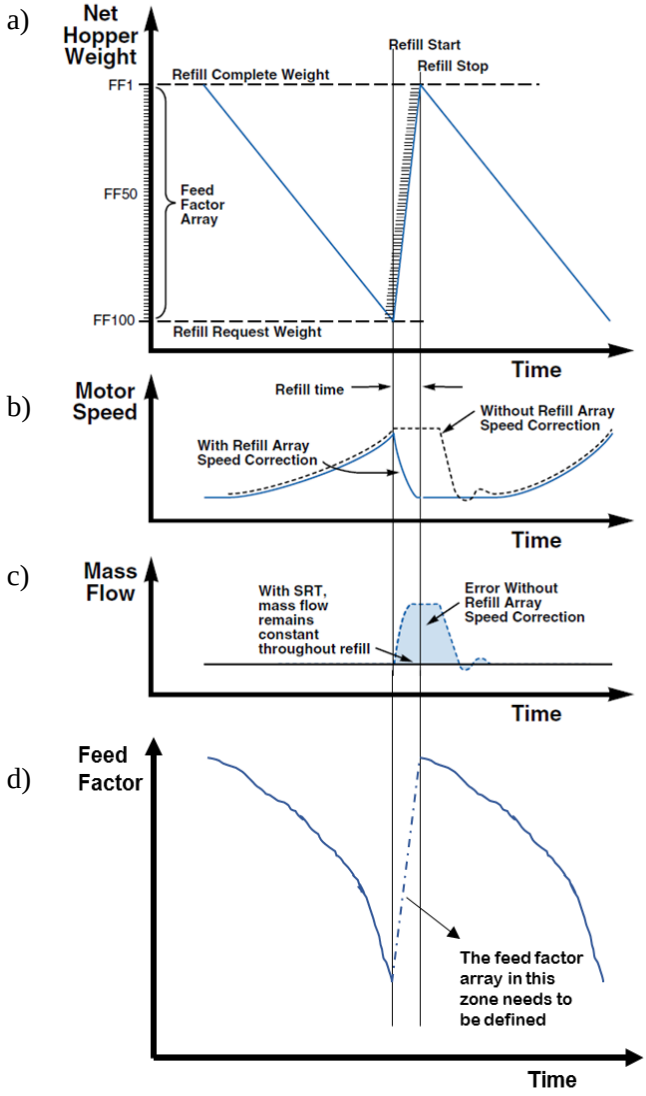


Figure 2.5 *Dynamic profiles of the variables involved during the feeder's operation: a) weight; b) motor speed; c) mass flow; d) feed factor. (Image partially adapted from [54])*

2.3 Data analysis and modelling flowchart

As already demonstrated by several publications, a comprehensive knowledge of the dynamic phenomena involved in the normal continuous operating conditions of the equipment is the result of a strong correlation between the design and the setup of the equipment and the material properties of the powder flowing through it [55][56][57]. Thus, to predict the variation of the density in the gravimetric phase, the possibility of developing a "mechanistic" first-principles model has been evaluated, but because of the complexity interactions of several parameters in the system, this approach seems to not produce valuable results in a short-term period, especially when the design of the

equipment is not similar to conventional powder storage and conveying units [58] [59] [1].

Any possible application of discrete element method (DEM) simulations of the system have been discarded since the beginning, because of their prohibitive evaluation time associated to multiple units required in a single continuous manufacturing line [60] [61] [62] [63].

Furthermore, a model based approach developed from experimental and manufacturing campaigns datasets would suggest a more systematic way to develop a predictive, easy-to-update and flexible tool. It is in this context that a data-driven modelling strategy to, firstly, cluster and classify new materials and, secondly, predict the feed factor profile during the gravimetric phase, has raised interest.

In the next chapter, a structured approach for data-mining of an essential powder API materials dataset is proposed. The aim of this case study is to outline the pattern similarities in terms of powder flowability in order to evaluate the processability of new materials for secondary continuous tabletting line. The general modelling flowsheet starts from an unsupervised pattern recognition of the data to define possible classes of surrogate materials. Then, some possible supervised pattern recognition models are tested to classify new incoming materials.

In the last chapter, a possible strategy for multivariate statistical modelling of a loss-in-weight feeder is delineated. The equipment setup data and material properties characterization of the processed powder are combined in the input of the model to predict the feed factor profile during a gravimetric run. This approach may help studying the effect of raw materials variability into the equipment, with potential positive implications during process developments studies of new APIs.

Chapter 3

Powder materials clustering and classification

In this Chapter, a general procedure to aid the analysis of pharmaceutical raw materials database using a data-driven approach is illustrated and supported by an example of application on an industrial dataset of powder materials candidates for secondary continuous manufacturing processes.

3.1 Introduction

Active pharmaceutical ingredients (API) synthesis and finish routes during primary manufacturing lead to a final form of powder materials that is difficult to be fully predicted, especially when one moves from an early to a late stage of process development. The reasons for material variability are various and intricate to be identified because they lie in every detailed choice made at each stage of the process design life cycle. This leads to unexpected differences in the API material properties that are not always easy to be understood and corrected by a reverse-engineering approach. The effect of raw material variability in the downstream processes is even harder to be forecast and many manufacturing problems appear in the transition from conceptual to detailed design of secondary continuous manufacturing processes, compromising the final product quality and deliverability.

One of the major limitations in this scenario is the empirical or semi-empirical approach to evaluate the material processability in the equipment of the line. A significant amount of a new raw material needs to be available in order to test its behaviour across the line or alternatively to be fully characterized by laboratory analysis. The bottleneck in this approach is that in the early stage of process development only a very small amount of drug product is available and it might not be enough even to perform a full set of characterization tests. Even in the case that a sufficient amount of powder is available for a limited number of tests, the problem is to decide which kind of test should be preferred to investigate how particle properties might affect process performance in the equipment, and how to assess the impact of multiple characterization measurements upon downstream processability [64].

Thus, a systematic and reliable procedure to understand the relationship between raw material properties and process performance is required to establish the design space and to enhance the robustness of the process control strategy, meeting the requirement of the Quality by Design paradigm [7]. In this connection, Oka et al. [65] theorise that, in the context of secondary continuous manufacturing, collecting material information "paves the road for the development of an adaptive, self-learning material database, which aids in future process development".

However, there is no trace in the literature of the formalization of a general procedure to develop and investigate raw material database in order to extract the maximum amount of information and support the product and process development.

This is an area of great interest for many companies in the pharmaceutical sector, particularly when they are referred to the API field rather than excipients. However, the question arises about how it is possible to develop a structured approach to concretely build up a comprehensive material database starting from limited information. Which is the correct approach to explore, mine and model an industrial powder materials dataset in order to support the input materials selection of new active ingredient candidates in the early phase of process design? Which are the inherent common features of raw material datasets, and how should they be taken into account in the data analysis and modelling phase?

The purpose of this Chapter is to address these questions through the analysis of a primitive dataset of powder candidates for a continuous direct compaction process where limited characterization tests are available for each of them. A standard systematic procedure to investigate and model these common structured experimental datasets is delineated as in Figure 3.1 and the generalization of the approach to a more complete dataset that includes more properties and materials is also discussed.

3.2 Materials and methods

The proposed methodology consists of four consecutive steps, also called stages:

1. dataset organization;
2. exploratory analysis;
3. unsupervised pattern recognition;
4. supervised pattern recognition.

In each step, a series of activities are suggested in order to maximize the amount of information that can be extracted from the dataset and create specific models for cluster identification and class delineation. The level of subjectivity in the decision-making process for each recommended action is strongly related to the dataset structure, the degree of complexity that is likely to be introduced by the modeller, the experience of

the people involved in the analysis, and the requirements of the final users (e.g. engineers, formulators or project stakeholders). Nevertheless, the proposed procedure is thought to be a general and flexible approach that can be adopted with none or minor changes to any pharmaceutical raw material database in order to help the development of the manufacturing process.

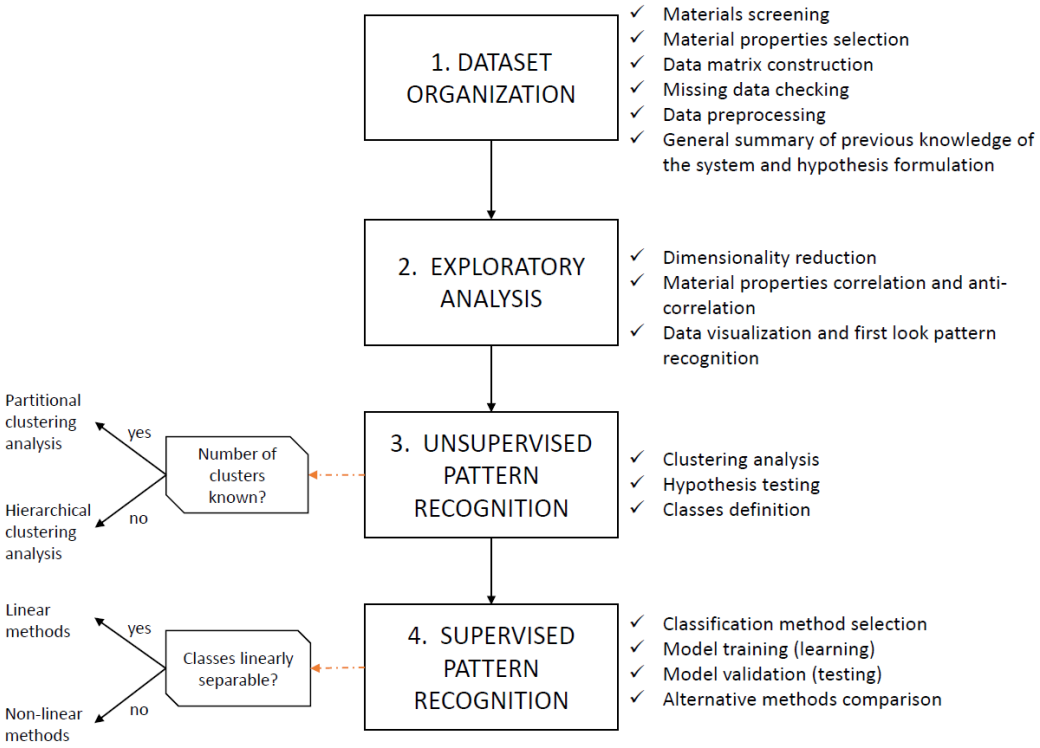


Figure 3.1 *Schematic of the general procedure to aid the analysis of a pharmaceutical raw materials database using a data-driven approach in agreement with the QbD framework.*

In the first step, all the data for the raw materials of the area of interest need to be collected and systematically organized to create a significant dataset of materials with various features and properties. Then, the materials must be screened and the relevant material properties must be selected depending on the final aim of the analysis. Materials that have multiple missing entries should be rejected or, eventually, the missing measurements must be collected, to avoid the recognition of artificial patterns due to automatic imputation of extrapolated data. However, in case of a single missing property for a relative small proportion of samples in the dataset, several strategies to deal with missing entries can be adopted, but their use is advised only if strictly necessary [68]. The size of the resulting dataset should be approximately of 50 materials to ensure the robustness of the application of statistical methods of investigation. Actually, there are no strict rules about the size required by the dataset for applying

pattern recognition techniques, but it is strongly recommended that the number of variables has to be correlated to the number of samples, such that the greater the number of variables included, the larger should the dataset size be. In unsupervised pattern recognition, a rule of thumb suggested by Formann (1984) and commonly used is that the minimal sample size should be 2^k, where k is the number of descriptors [66]. Thus, if a latent variable reduction is used before performing a cluster analysis, the expected number of principal components for approximately 10-15 descriptors is around 3-5 and, consequently, a safe approximation on the minimum number $2^5=32$ can be exceeded to 50. Prior to any analysis, the data need to be preprocessed to level the variables importance and centre the columns. Lastly, a general recap of the previous knowledge of the materials investigated needs to be done with a special focus on the materials that are well known to be troublesome or straightforward. This latter point will help the hypothesis formulation of the expected results of the investigation or, eventually, it will highlight some gaps in the status of the materials characterization.

The second step involves an exploratory analysis of the dataset created in the first step. Dealing with materials experimental database is problematic in terms of investigating the possible correlations between properties and representing the samples in a multi-dimensional space, when a classic univariate approach is used. Then, latent variables techniques are much more than useful, since they allow the creation of simple multivariate charts to emphasize the correlation and anti-correlation between properties and to allow the visualization of the samples in a tangible reduced space. For these reasons, a preliminary PCA is suggested at this stage, followed by a careful analysis of scores and loading plots to understand the general structure of the data and, eventually, recognize some possible evident, or less evident, patterns.

The third step of the procedure aims to structurally identify clusters of elements in the data through the application of unsupervised pattern recognition techniques of research. At this stage, the modelling aspect can be highly customizable by the user since dozens of clustering techniques are available. Despite that, a complex approach, for a relatively small and simple database of pharmaceutical raw materials, does not offer any guarantees in terms of final outcomes. For this reason, the choice can be restricted to two different standard categories of algorithms: i) partitional clustering techniques, if the number of clusters is known a priori; ii) agglomerative clustering techniques, if the number of clusters is unknown. Several different possible clusters might result from the clustering identification and they can all have a statistical meaning in terms of patterns similarities. Nonetheless, only one or few of this cluster models are physically meaningful and an objective correct evaluation needs to be done by a cross comparison of some statistical indices of the methods performance and an *a posteriori* clustering analysis of the clusters with the support of an expert of the system. Testing the

hypothesis formulated in the first step is also helpful in order to assess the outcomes of clustering analysis and define the class partition for each material considered in the study. Potentially, this step of the procedure might be not necessary if in the dataset exists a categorical variable that leads to the definition of groups of materials with similar features.

The fourth step is the design of a suitable classifier using supervised pattern recognition techniques. The goal of this step is to train a classification model using the labelled training samples obtained from the previous step in order to classify new materials that are going to be added to the database in the future. Even in this case, a wide variety of different approaches can be used to choose the best classifier, but to be confident about the choice, a comparison between alternative methods is always suggested. In particular, if the boundaries between classes are not clearly linearly separable, a comparison between the performance of linear and non-linear classifiers is required.

3.3 Step 1: dataset organization

The first step of the procedure is a general dataset reorganization and preprocessing.

Forty-seven powder materials were used in this study, of which are APIs forty-four and only three are excipients. Prior to the analysis, materials were pre-screened from a slightly larger dataset in order to have a complete characterization of at least these eight measurements: particle size distribution (d10, d50, d90), specific surface area (SSA), bulk density (BD), tapped density (TBD), CARR index (CARR) and Hausner ratio (HR).

The choice of using these specific characterization descriptors was not driven by sophisticate criteria. The simple fact that these tests are the only available for the largest number of APIs explains the reason by itself. As a matter of fact, these are basic tests that are usually carried out for all the drugs in the early stage of development since they do not require a high amount of powder and they are usually sufficient to describe the general crystal properties of the powder and hypothesize the expected bulk behaviour.

There are several evidences in the literature and pharmaceutical knowledge that powder packing efficiency and powder flow are a function of particle shape, particle size and surface properties [16] [67]. Thus, a more comprehensive database should include more properties related to particle shape and surface forces in such a way that more information about powder processability might be extracted even without using a lot of resources to investigate flowability and packing efficiency. However, there are several obstacles that have prevented this optimal scenario up to date. For example, particle size data are easy to be obtained and they are frequently collected even during the manufacturing process, but there is a lack of suitable and objective techniques for

particle shape analysis characterization of bulk powders. The most accepted and used (e.g. scanning electron microscopy or optical microscope) are often time-consuming and limited to relative small numbers of sampled particles [67]. Surface properties, as for example surface energy and electrostatic charging, are known to play a crucial role in the description of the electrostatic effects that can reduce the flowability in continuous manufacturing since it can be conditioned by the formation of electrostatic bridges, but the tests usually require a laborious procedure and a relevant quantity of powder [64]. The powder flow itself can be measured by dozens of indices, but in addition to the trials just mentioned, it is well recognised that there is not a *single* test that can fully describe the flow properties of a powder since the powder itself it is exposed to various levels of shear, normal stress and charging along the process. The problem is therefore to determine which of them are more pertinent in order to describe their behaviour in the equipment minimizing the available resources.

Nevertheless, the importance of extending the approach to a more thorough dataset is recognized and the collection of few important properties regarding particle shape and surface forces is suggested even if it is not going to influence the general investigation methodology presented and the evaluation of the possible benefits produced are beyond the aim of this work.

Six of the selected materials (M35, M36, M37, M38, M39, M40) were missing only the SSA test, but they were considered anyway to study the possible effect of a single missing variable for a relatively small part of the database. Several different and sophisticated methods can be used to deal with missing entries in a dataset, but for simplicity a mean of the column for that variable was used in order to avoid failures in extrapolation given by the standard best guess replacement used by the PLS Toolbox in which the missing data are replaced using the projection and loadings of the model constructed from the known data. Nevertheless, it is important to recognize that the replacement by a single value reduces artificially the variance of the imputed variable and alters the correlation between variables. However, because it is applied to about 1 % of the data with random distribution, this approach is still considered satisfactory [68].

The main hypothesis that comes out from a first analysis of the properties available for this primitive dataset is that a combination of particle size (PSD), surface (SSA), packing (BD and TBD) and compressibility indices (HR, CARR) should already bring relevant information about powder flow and processability. The API materials collected in this database should span the entire range of flowability from cohesive to easy flow powders, and this is an important fact that should reflect in the data analysis outcomes. It is common practice in powder technology to split the classes of flowability in four:

very cohesive and easy flow (also called free flowing) are the outer classes, poor flow and medium-poor flow are the respective intermediate classes.

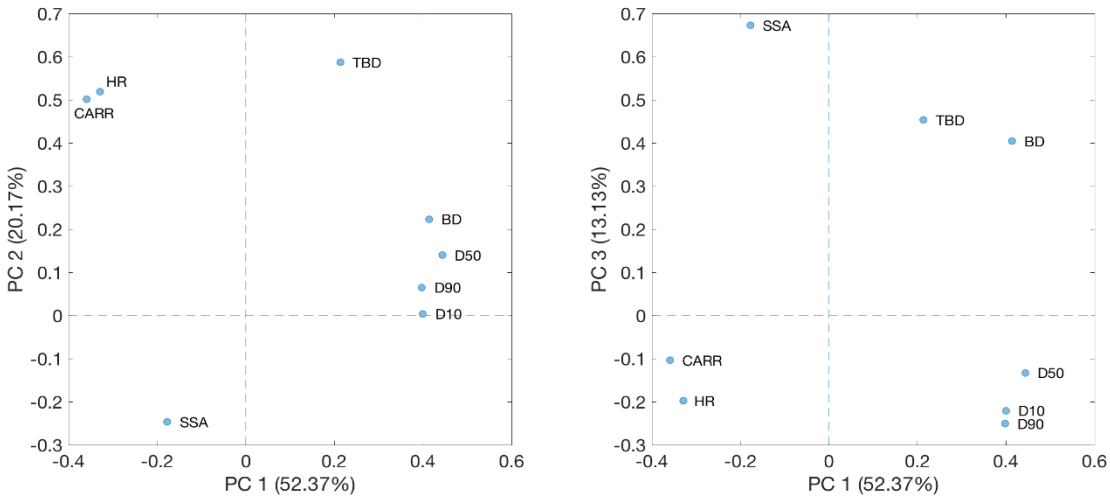
Table 3.1 *General features of the material dataset analysed in the case study.*

n° of total samples	47
n° of API samples	44
n° of excipients samples	3
n° of measurements for each sample	8
% missing data	~1 %

A synthesis of the general features of the material dataset is reported in Table 3.1.
Before proceeding to the second step of data exploration, an autoscale (i.e., mean centering and scaling to unit variance) preprocessing is required in order to give to all the variables the same importance and to centre the data off-set.

3.3 Step 2: exploratory data analysis

The second step of the proposed procedure involves an exploratory analysis of the data to visualise the samples in a dimensionally reduced space and understand the general correlation and anti-correlations between properties using a multivariate statistical approach, e.g. a simple PCA.
Three principal components (PCs) were selected since they describe 85.7 % of the cumulative variance of the data with PC1, PC2 and PC3 accounting for 52.3 %, 20.17 % and 13.1 % of the variance respectively.



(a) (b)

Figure 3.2 *PC1 vs PC2 (a) and PC1 vs PC3 (b) loading plots of the PCA exploratory model.*

Loadings scatter plots representing PC1/PC2 and PC1/PC3 are reported in Figure 3.2. The loading plots indicate which variable/s are explaining most of the variance along a certain principal component and the positive or negative correlation between the properties. In this case, the main factors of variance along PC1 are the PSD descriptors and the bulk density, that are highly and positively correlated. The Hausner ratio and CARR index are anti-correlated to the latter along PC1, but together with tapped bulk density, they are the main factors of variability along PC2. Specific surface area (SSA) shows a minor contribution to the variance of the first two principal components, but it dominates PC3.

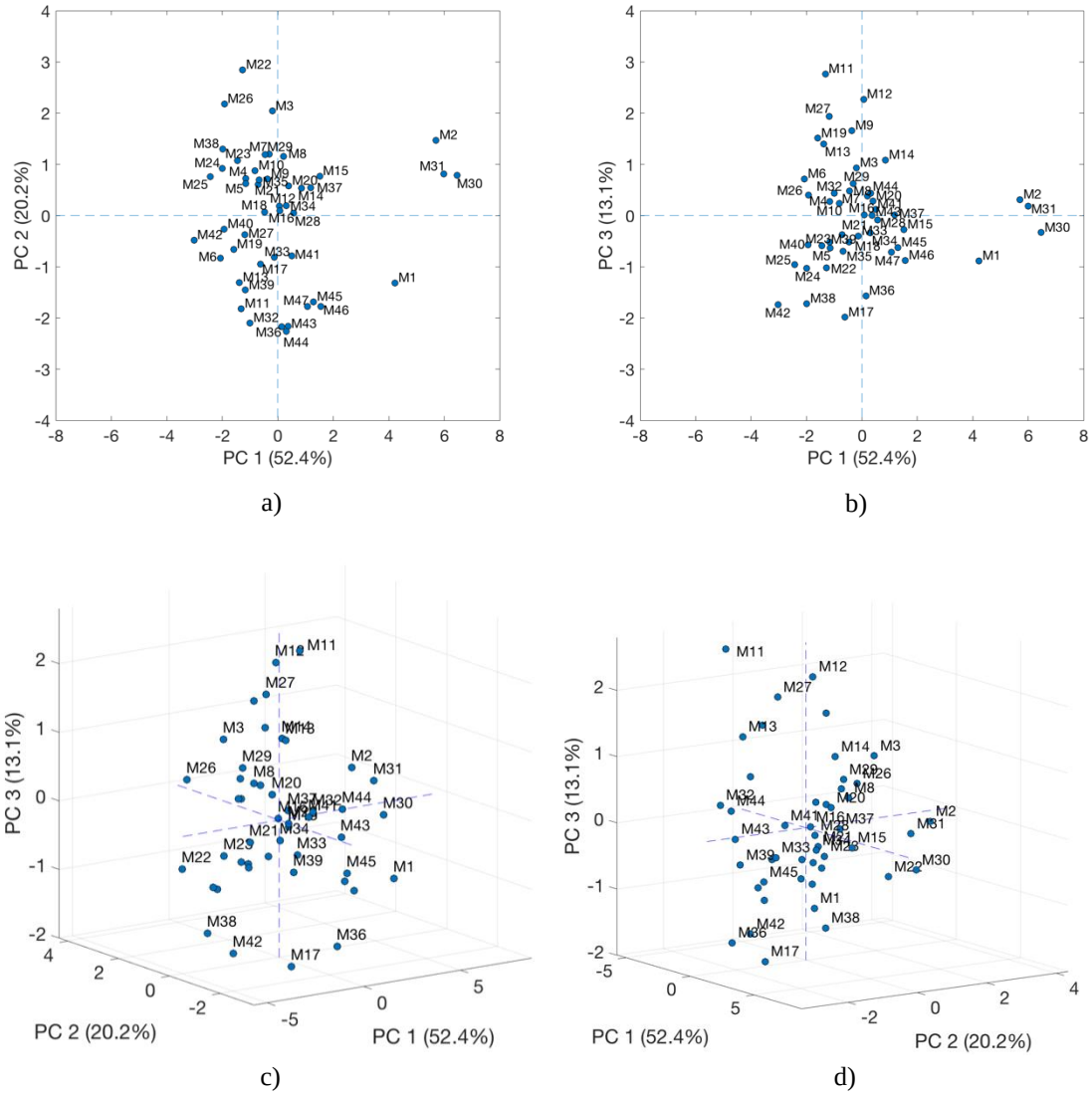


Figure 8.3 *PC1 vs PC2 (a) and PC1 vs PC3 (b) scores plots of the PCA exploratory model. The graphs (c) and (d) are a three-dimensional representation of the scores from two different angular perspectives.*

Scores scatter plots representing PC1/PC2 and PC1/PC3 can be used to identify similarities between materials and visually recognize if the samples cluster with some pattern or they distribute undistinguishably in the reduced space. However, since three PCs are used, an alternative representation is given by a three-dimensional scatter plot as shown in Figure 3.3. In this case, it can be immediately noticed that samples M1, M2, M30, M31 are clustering quite far from all the other materials, whereas at least two different clusters can be observed along the second principal component axis. In fact, a more accurate examination shows that four different clusters can be approximately identified if PC1/PC2 scores are analysed: i) a first cluster on the right side; ii) a second cluster in the top-left quarter; iii) a third cluster including most of the elements that are in the top-right quarter and that are close to the axis origin; iv) a fourth cluster including most of the samples in the bottom-left and bottom-right quarter. A cross-reading analysis of the three-dimensional scores plot of figure confirmed this first bi-dimensional interpretation, but a systematic unsupervised pattern recognition procedure needs to be applied in order to establish pattern similarities among the data in a rigorous way.

3.4 Step 3: unsupervised pattern recognition

The third step of the proposed procedure aims to quantitatively identify and define unknown patterns in the data using an unsupervised pattern recognition approach.
The support of an expert in the field of the specific materials in question is required to analyse the output results.
As discussed in Chapter 1, several different clustering methods can be found in the literature, but for an industrial dataset of materials like this one, a hierarchical clustering analysis seems more reasonable than partitional clustering because the number of clusters in the system is usually unknown and the possible estimation from the exploratory analysis is not fully trustworthy.
The current clustering methodology involves three level of decision: i) the space dimension of investigation, ii) the distance criteria definition, and iii) the clusters formation's algorithm selection.
Firstly, dimensionality reduction of the original *d*-dimensional space needs to be evaluated since the hierarchical clustering methods can be applied both to the original projection space of the samples and to the reduced multivariate space (like in the PCA reduction obtained in the previous step). Thus, the choice here is done based on the output considerations from the exploratory analysis, particularly which kind of

advantages are caused by a principal component reduction prior to a clustering identification for the specific data structure of the original database. In this case, the main advantage is noise elimination because 14.3 % of the variance in the data can be attributed to noise and it is well known that clustering algorithm are sensitive to noisy data. Moreover, the collinearity of some of the material properties was solidly confirmed by the physical interpretation of the loading plots and these two facts enhance the level of confidence in using a multivariate reduction prior to a clustering analysis.

These two reasons are quite common for any kind of materials database, but it is highly recommended to carefully analyse the data to determine some possible deviations from this scenario in order to exclude special artefacts in the data or a superficial level of comprehension of the PCA results.

Secondly, the distance function between elements in the new space of reference given by the scores coordinates needs to be defined. As previously stated, the choice might involve several geometrical distance criteria (see Chapter 2), but in the specific case a Euclidean distance is the simplest and the most suitable choice since the space was already compressed to take into account the effect of correlation between variables.

Finally, the selection of the cluster method and its linkage rule is the crucial step to search for pattern similarities in the data. In hierarchical agglomerative clustering, there is not a method that is objectively superior to the others because the clusters' distribution is highly dependent on the inherent structure of the dataset and on the level of base knowledge of the system. The suggestion proposed in this study is a systematic comparison of the six most common HCA cluster methods through the simultaneous analysis of two indicators: i) the evaluation of the preservation of "original" information using the cophenetic correlation coefficient, and ii) a knowledge-based analysis of dendrogram outcomes with the help of a system expert (e.g. materials scientists).

To this end, a simple script to test all the methods was developed and for each of them a dendrogram, as summarized in Figure 3.4 and a cophenetic correlation coefficient, as reported in Table 3.2 were computed considering the scores coordinates as the "original" reference, because the methods were applied to the reduced space.

The cophenetic correlation coefficients show that all the cluster methods seem to maintain intact the similarities between samples in a similar manner, because all the coefficients are around 0.8, with the only exception of the Ward cluster method that is around 0.7, but this discrepancy is not significant since the value is still close to a theoretical maximum value of 1.

Table 3.2 *Cophenetic correlation coefficient calculated for each of the six cluster methods.*

Ward	k-NN	Furthest neighbour	Pair-group average	Median	Centroid
0.73	0.86	0.81	0.82	0.84	0.82

On the other hand, a very different scenario can be interpreted by the dendrograms representation of Figure 3.4. The median (Figure 3.4e), the centroid (Figure 3.4f) and the k-NN (Figure 3.4b) methods produce meaningless clustering identification, since they are able to distinguish only the materials that are far from the centre of the PC1/PC2 scores plot. These results can be considered unsatisfactory since they do not add any physical value to the definition of hidden patterns in the data. The average-paired distance method (Figure 3.4d) is able to identify four clusters and this can match well with the initial hypothesis, but a more accurate analysis shows that one of the cluster consist of the single material M11, that is a micronized material with a very high specific surface area that explains the reason why it is clustering in an outer position along the PC3. However, this API is not supposed to be so different from the closest samples in terms of processability and it is expected to be paired with other close cohesive material with similar patterns in the other directions. On the contrary, the Ward (Figure 3.4a) and the furthest neighbour (Figure 3.4c) methods reveal some interesting and similar partitions of the dataset since they both identify four different clusters according to the consideration that came out in the exploratory analysis. The only difference in the final clusters is in the samples M6, M11, M13, M19 and M27 that are mixed up in the adjacent classes C2 and C3. In this case, the furthest neighbour clustering algorithm is preferred because of the resulting higher cophenetic correlation coefficient together with a more meaningful explanation by system experts, that confirmed the superior affinity of the mismatched samples with the class C3 rather than C2. A summary of the differences between clusters is reported in Table 3.3 and Table 3.4.

The last important action, which must be taken before proceeding with the next step of the procedure, is a detailed interpretation analysis of the selected clustering output and the hypothesis testing of the preliminary consideration on the system. In Figure 3.5, an updated two-dimensional and three-dimensional representation of the data in the reduced space is reported, together with the new clusters division obtained by the furthest neighbour linkage method.

It is now possible to state that four groups of materials are forming four different and evident clusters according to the space distribution assumed in first instance in step 2, but with a certain irregular profile of the related boundaries between classes.

fact that the boundaries are overlapping each other in some regions of the space is an exact demonstration of a non-linearity in the pattern structure. However, this non-linear boundaries' shape is not so explicitly marked and trusted since the number of samples is still very low, but it suggests that linear methods of classification might have low quality performance and it is well worthy to test both types of classifiers in the next step of the procedure.

Table 3.3 *Summary analysis of the clusters identified by the hierarchical clustering with furthest neighbour linkage method.*

Class name	n° of materials	Materials
C1	4	M1,M2,M30,M31
C2	14	M11,M13,M17,M19,M27,M32,M36,M39,M43, M44,M45,M46,M47,M6
C3	15	M12,M14,M15,M16,M20,M28,M29,M3, M33,M34,M37,M41,M7,M8,M9
C4	14	M10,M18,M21,M22,M23,M24,M25,M26, M35,M38,M4,M40,M42,M5

Table 3.4 *Summary analysis of the clusters identified by the hierarchical clustering with Ward linkage method.*

Class name	n° of materials	Materials
C1	4	M1,M2,M30,M31
C2	9	M17,M32,M36,M39,M43,M44,M45,M46,M47
C3	20	M12,M14,M15,M16,M20,M28,M29, M3,M33,M34,M37,M41,M7,M8,M9, M11,M13,M19,M27,M6
C4	14	M10,M18,M21,M22,M23,M24,M25,M26, M35,M38,M4,M40,M42,M5

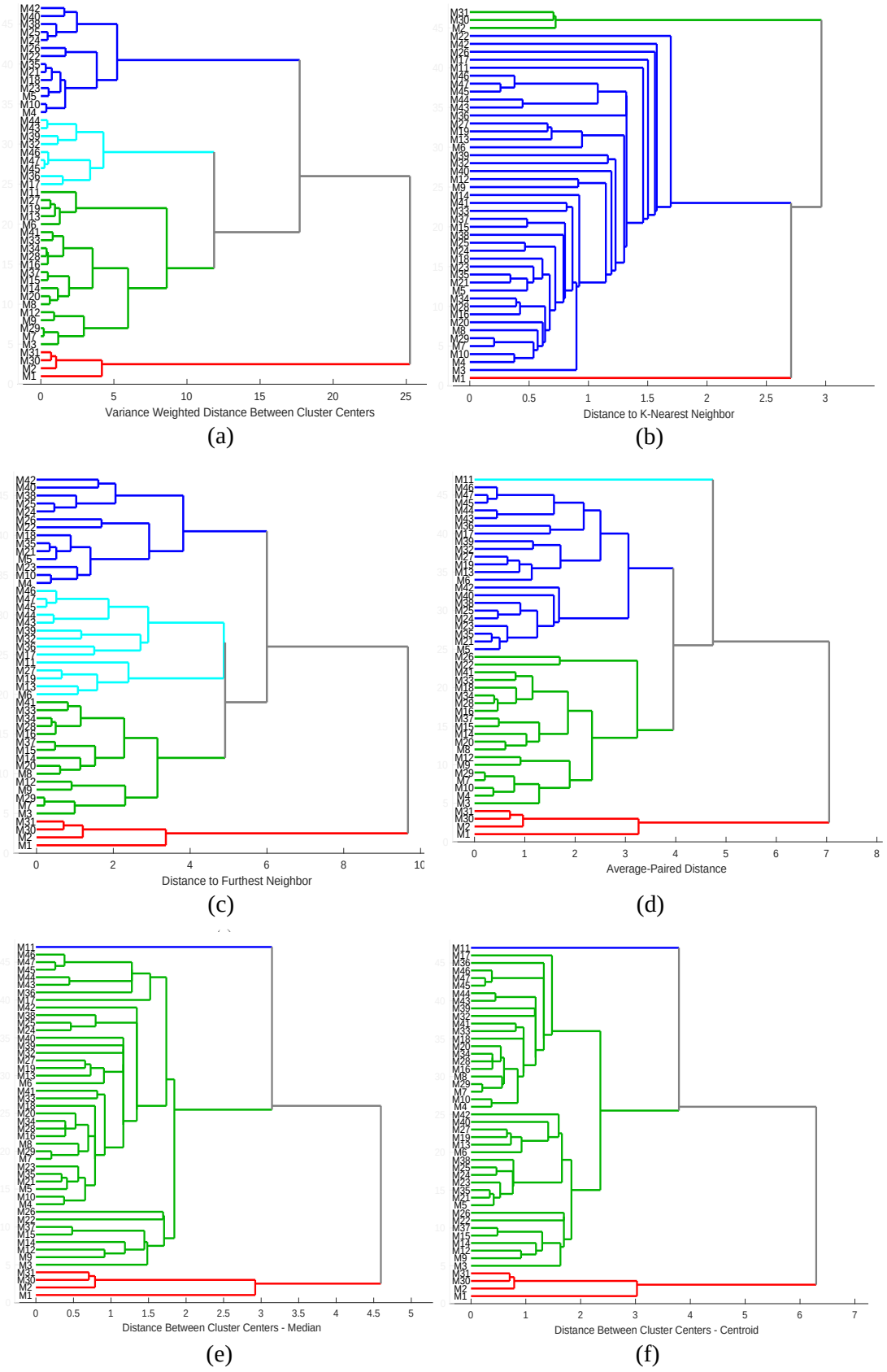


Figure 3.4 *Dendrograms of the six different clustering methods: Ward (a), k-NN (b), furthest neighbour (c), pair-group average (d), median (e) and centroid (f).*

A more technical analysis of the clusters' characteristics has been done with the help of the materials scientists involved in the data collection and drug design and formulation, with the final goal of understanding the principles that lead to pattern formation and how this information can be used in the early stage of process development. From the very beginning, it appeared quite clear that the four groups are reflecting the four metrics of flowability that were expected. In particular, the four materials of class C1 are well known to be the easy flow materials that never caused any processability problems in the past. Some of the materials in the furthest class C4 are known to be very cohesive because their bulk behaviour was intensively studied after that some problems, (e.g. stickiness; formation of electrostatic bridges), appeared when they were

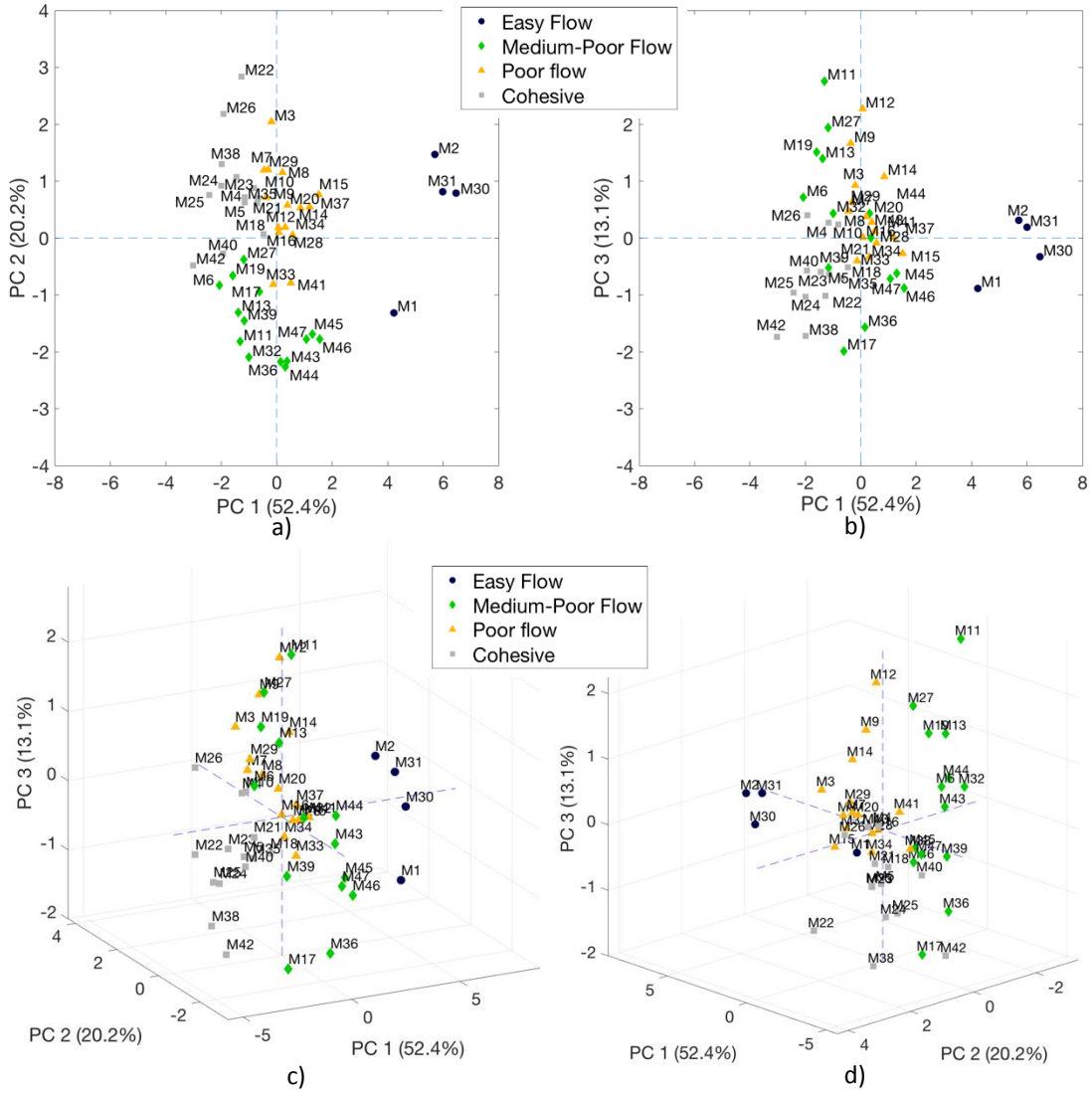


Figure 3.5 *PC1 vs PC2 (a) and PC1 vs PC3 (b) scores plots with the four different clusters differentiation as identified by the furthest neighbour method. (c) and (d) are highlighting the clusters in the three-dimensional representation of the scores from two different angular perspectives*

used in some of the downstream equipment. Similar considerations can be done for some of the powders in class C3, but the phenomena appeared with less intensity and frequency than in the previous case. A slightly better situation can be found for the materials included in class C2, since their flow and packing behaviour is not optimal as for the samples in class C1, but it is still fairly good and not considered to be problematic for downstream processability.

Despite this qualitative experience-based clusters analysis, a more rigorous study has been done to highlight how these considerations are likely to be reflected in some of the descriptors of the system. In fact, now that the materials have been grouped together, it is possible to analyse some features of the system and draw some conclusions on the patterns similarities. In order to do that, a summary of the class mean values of the most important features for the identified clusters are reported in Table 3.5.

Table 3.5 *Summary analysis of the features of the clusters identified by the HCA with furthest neighbour linkage method.*

Class name	n° of materials	$CARR_{mean}$	HR_{mean}	Class label
C1	4	23.4	1.3	"Easy flow"
C2	14	35.5	1.6	"Medium-poor flow"
C3	15	43.6	1.8	"Poor flow"
C4	14	55.7	2.3	"Cohesive"

CARR index and Hausner Ratio are traditionally used to assess powder flow because they measure the change in density when the powder is subjected to force by tapping. Thus, it makes perfectly sense that at least the class mean value of these properties increases when the class of flowability moves from easy flow to cohesive. This is well shown in Table 3.5, but it not necessarily true that a material with a slightly higher CARR index, or Hausner ratio, is more cohesive than another one. This would mean that powder flow can be evaluated just from a univariate perspective of one of these two variables, but this is exactly the opposite of what this kind of approach is going to demonstrate.

A good example to show this concept is considering two materials with similar CARR index, but clustering in different classes, e.g. M30 and M44. M44 has a slightly lower CARR index than M30, but the latter is clustered with no doubts as easy flow or free flowing whereas the first is considered medium-poor flow. Why does it sound contradictory and it is likely to lead to a wrong conclusion? The correct reason can be found if the other variables of these materials are observed together with the trend of the respective classes. For example, the PSD values are greatly higher for M30 and for the relevant class than for the other materials. Of course, this kind of analysis is

extremely simplified if a multivariate approach is used like in this case, but the trend of the average of CARR index and HR is still a good indicator and it seems to perfectly match the qualitative analysis results conducted by the experts.
Some further considerations can be done if the cluster observation is conducted by looking at some other properties that are not present in the initial dataset because they are available just for a limited number of samples, e.g. flow function coefficient (*ffc*) from shear cell data that are available just for 15 samples as reported in Table 3.6.

Table 3.6 *Flow function coefficients (ffc) from shear cell data on 15 materials and respective classes. The samples with the symbol * are all samples with missing entry in the SSA value.*

Material	Class	*ffc*
M2	C1	7.7
M39*	C2	2
M43	C2	5.9
M44	C2	6.1
M45	C2	5.6
M46	C2	6.1
M47	C2	5.6
M12	C3	3.4
M14	C3	4.8
M28	C3	3.8
M10	C4	2.3
M18	C4	2.5
M40*	C4	1.7
M42	C4	3.1
M5	C4	3.2

It can be observed that, even in this case, the *ffc* value of the materials are following an incremental distribution among classes, as for CARR and HR, when moving from "easy flow" to "cohesive" with the only exception of material M39 that is classified as "medium-poor flow" but it is in fact a cohesive material with a very low *ffc*. This discrepancy is probably due to the fact that material M39 was one of the materials with a missing value of SSA and as a consequence the replacement of the variable with the mean of the respective column is not appropriate for this API. Furthermore, it is important to clarify that the classes, that have been identified by the unsupervised pattern recognition, are not the same that are generally used to classify powder flowability accordingly to *ffc* only and the Janike classification, but the ranges have been slightly moved towards higher value [55]. However, the shear cell data collected

are not enough to define some possible correlations between the original variables included and the *ffc*.

In conclusion, the hierarchical cluster analysis recognized four different patterns in the data that are easy to be visualized in the reduced three-dimensional space and that seem to have a physical meaning that is coherent with the four grades of powder flowability assumed in the first step. This four groups identification is accepted as a starting point to train a classification method that is likely to classify new samples based on this pattern similarities.

3.5 Step 4: supervised pattern recognition

The fourth and last step of the proposed procedure involves the development of a classification model to perform a supervised pattern recognition and define standard classes of similarities for the identification of future "unknown" samples.

The design of a suitable classifier is one of the most difficult tasks because the process of selection and design must be done according to the inherent features of the system under investigation. The proposed procedure of analysis allows us to challenge this last step with the confidence that sufficient knowledge about the dataset and its main characteristics has been gained from the previous steps. Consequently, the modelling approach is simplified and the decision-making process will become more agile even for a non-expert user of classification models.

The crucial question mark at this point is to use either a linear or a non-linear model of classification, and which one, among the numerous possible choices, is the best performing. An iterative approach to test multiple types of classifiers at the same time seems a solid strategy to face the problem.

In this case, several evidences about the non-linearity of the patterns' structure should address the modelling options towards non-linear models of classification, but, as previously stated, a comparison between linear and non-linear models might be a good way of confirming this conclusion. Support vector machines (SVM) and *k*-nearest neighbours (*k*-NN) were selected as possible equivalent non-linear methods, whereas PLS-DA was picked as a linear alternative method. Hence, an iterative strategy of 100 iterations is developed and at each step the data are randomly split in an 80% - 20% proportion to create a training or calibration set and a testing or validation set respectively. The models are then built, learning from the training set the patterns of the classes labelled in the learning phase, and testing the efficacy in prediction using the external validation set of samples in the consequent validation phase. To evaluate the performance of all methods in both the phases, a very simple index as the percentage of

samples correctly classified *(% CC*) is defined and the final average across the iterations (*% CC*$_{avg}$) is calculated together with the standard deviation σ.

The SVM model is built on the original preprocessed data, without LVs reduction, since one of the most attractive feature of this classifier is the capability of handling numerous variables, even in a non-linear space and with a certain collinearity between descriptors. A non-linear radial basis function (RBF) kernel is used and the control on complexity is obtained following the PLS Toolbox® package of optimization of model parameters, where cost and gamma are optimized through a "Random Subsets" cross-validation procedure with 5 data splits and 1 iteration, over a grid of appropriate parameters.

The *k*-NN model is built on the original preprocessed data for the same reason of the SVM and the number of nearest neighbours selected is 5, since the smallest group of samples that form a class (C1) is made up of four components. This choice is expected to ensure a more robust class definition for all the training partitions with at least three materials from C1 and this should be true for the wide majority of the sampled iterations.

Even for the PLS-DA model, the algorithm is applied to the original preprocessed data because this linear classification technique is based on the multivariate partial-least squares discrimination and the consequent latent variable reduction to obtain the class regression. The number of principal components selected is 3 and it has been determined by observing the consistency of the variance explained by some examples of PLS-DA models constructed using random partitions of the data and the information about collinearity between variables obtained by the PCA performed in the exploratory analysis (step 2).

All the algorithms have been implemented to give probability estimates of the class membership accordingly to the respective different calculation for each of the classifiers. These probability calculation methods are exhaustively explained in http://wiki.eigenvector.com/index.php?title=Sample_Classification_Predictions.

Therefore, two different result scenarios can be obtained: i) a first scenario where the correct classification for a given sample is based on choosing the class that has the highest probability, regardless of the magnitude of that probability; ii) a second scenario where the correct classification is based on the rule that each sample belongs to a class if the probability is greater than a specified threshold probability (e.g. 50% for this case).

In this latter scenario, a sample is misclassified even in the case that two classes have the same probability (technically referred as unassigned sample) or that two different class attributions result from the classification exercise. As a matter of fact, when a threshold on the probability is defined, the sensitivity of the different classification

methods to the ambiguous samples is highlighted and some weaknesses of some algorithms are likely to arise.

The results of the scenario 1 are reported in Table 3.7 and it is possible to observe that the performance of SVMs in calibration are superior than *k*-NN and PLS-DA because of the direct boundary construction approach of the Vapnik's method that is modelling the margin edges to respect the given classification of the training set. Rather similar are instead the validation results of the three models, where PLS-DA is the one showing the lower average *% CC* and the higher standard deviation, but the dissimilarities are not so marked as expected.

Table 3.7 *Results of the iterative procedure for testing the classifiers in the scenario 1 (Most Probable). The average of the correct classified samples percentage across the iterations for each method is reported together with the standard deviation in both the calibration and validation phases.*

	Calibration			***Validation***		
	SVM	***k*-NN**	**PLS-DA**	**SVM**	***k*-NN**	**PLS-DA**
$\%\ CC_{avg}$	99	89	92	91	92	88
σ	1	4	3	9	9	10

On the other hand, the resulting situation of the scenario 2, as reported in Table 3.8, is entirely different and the dissimilarities between linear and non-linear methods are evident. As a matter of fact, the performance of SVM and *k*-NN both in calibration and validation are very similar to the same methods' performance in scenario 1, but the average *% CC* value of the PLS-DA has dropped of 15-20% compared to its performance in the first scenario, and even to the performance of the other methods in the same scenario. The worst results emerge in validation where the *% CC* is around 0.70 and the standard deviation assumes the highest value of 0.15.

Table 3.8 *Results of the iterative procedure for testing the classifiers in the scenario 2 (Class Pred Strict). The average of the correct classified samples percentage across the iterations for each method is reported together with the standard deviation in both the calibration and validation phases.*

	Calibration			***Validation***		
	SVM	***k*-NN**	**PLS-DA**	**SVM**	***k*-NN**	**PLS-DA**
$\%\ CC_{avg}$	98	84	77	89	86	70
σ	2	5	5	9	12	15

These results are actually confirming that the non-linearity of the patterns' shape is likely to be better modelled by non-linear methods, such as SVM and *k*-NN, rather than

linear models, like PLS-DA, where the degree of uncertainty on the ambiguous samples is very high.
Furthermore, it can be concluded that a simple non-linear model like *k*-NN, that is computationally inexpensive and very easy to be implemented and interpreted, is not providing much worse results than SVM. However, it is important to highlight, once again, that the results of comparison are true for this case study only and the performance of the classifiers cannot be generalized to other datasets.

3.6 Conclusions

In this chapter, a general data-driven procedure to investigate a raw materials database of pharmaceutical ingredients have been presented and supported by the application on an industrial case study. The proposed procedure aimed at maximizing the amount of information that can be extracted from a database of material properties through the formalization of a series of consecutive steps and actions that need to be taken from the researchers and analysts when they are approaching similar problems from the very beginning, where none or little information about the data is known. The procedure is finalized to the identification and modelling of hidden patterns in the data through unsupervised and supervised pattern recognition techniques. The final goal is to build a classification model that is able to facilitate the comprehension of new materials that are going to be added to the system in such a way that the following phases of product and process design can be boosted and sped up.
The data analysis introduced in this work, is based on a flexible statistical approach that can be highly customised by the user depending on his/her background and final scope. The techniques used in each step of the analysis can space from the multivariate statistical approach to the machine learning framework, but some important requirements need to be met in order to be confident with the modelling decisions. The support of one or more experts of the materials studied, e.g. materials scientists or formulators, is crucial in this first approach of exploration and model design, since the results need to have a sensible physical meaning and a reasonable applicability.
The developed methodology has been successfully applied to an industrial case study of API candidates for secondary continuous manufacturing processes.
The data on material properties measurements have been collected and reorganized in the first step with the aim of finding possible patterns that were indicative of flowability metrics of the powder, without having data of flow tests measurements. A set of eight commonly available powder descriptors for particle size, surface properties and packing behaviour have been selected to obviate the lack of first principle mechanistic understanding of the of the system, but supported by some scientific evidences of the

correlations between these properties and the powder bulk behaviour. Some general assumptions of the possible classification model outputs have been formulated according to the basic knowledge of the materials experts. The data have been pretreated and prepared to the next steps of data analysis.

In the second step, an exploratory analysis using a multivariate latent variables approach showed good results in highlighting the relationship between the input variables and simplifying the interpretability of the data structure in a three-dimensional reduced space of investigation.

The linear relationships between some descriptors, e.g. PSD and bulk density or HR, tap density and CARR index, were confirmed, together with the natural formation of clusters of data samples in the latent space that seem to respect the classes of flowability hypothesized in the previous step. The general non-linear framework of the patterns structure was emphasized to address the modelling strategy of the next step in the right direction.

The third step of the procedure has identified four different patterns in the data using the hierarchical clustering analysis, an unsupervised pattern recognition technique particularly used in absence of certainty about the number of clusters. All the possible declinations of the agglomerative clustering algorithm have been tested in order to select the clusters analysis that is more meaningful in physical terms.

The last step has shown why it is important to evaluate several different models of classification and, particularly, how the choice should always involve the comparison between linear and non-linear methods. The non-linearity of patterns framework was confirmed by the superior performance in both calibration and validation of non-linear classifiers such as SVM and *k*-NN.

The outlined procedure can be used in the earlier stage of product and process development of secondary continuous line in order to aid the input materials selection, where the API variability of the input powders is very high and the systematic reorganization of the background knowledge is required to maximize the profit outcomes in using it. In this context, the identification of possible materials surrogates for design study of new products or the continuous improvement of existing ones is a sensible area of interest. It is also very helpful to understand some possible correlations among properties and build some possible model of regression between variables or to highlight the lack of experimental knowledge about some aspects of the powder characterization process.

Furthermore, the outputs of this procedure can be used as input for a following stage of data-driven modelling of the unit operations of the continuous lines, facilitating the development of models to predict the performance of the equipment and the related final targeted quality of the products.

Chapter 4

A multivariate statistical approach for powder feeding modelling

In this Chapter, a multivariate statistical approach for powder feeding modelling is presented together with some examples of how it is possible to integrate the previous knowledge gained on the system with the raw materials database investigation illustrated in the previous Chapter.

4.1 Introduction

In Chapter 2, the importance of building a data-driven model for a loss-in-weight feeding unit in order to increase the equipment and process understanding and to enhance the control strategy has been highlighted.

In particular, before developing a modelling strategy for the volumetric phase of refill, maximizing the amount of information and knowledge about the gravimetric phase is crucial. In this phase, the feeder is supposed to be controlled by its feedback control system and a constant mass flowrate target is deemed to be achieved. In fact, during normal gravimetric operating conditions the feeding system is far away from the theoretical stability assumed. This happens because every material has specific features that are then reflected in the "flowability" concept and this affects the control system performance. As mentioned in Chapter 2, the variation of the density profile across the equipment is different from material to material and, in particular, easily flowable materials are likely to experience less variation in the density profile than cohesive powders.

This means that the feed factor of easily flowable materials should be constant and more stable for the largest part of the gravimetric exercise than what happen with poorly flowable materials, that are supposed to have a very rapid drop of the profile and an associated high internal variability.

This variability in the feed factor is compensated for by acting on the speed of the conveying screws to stabilize the mass flowrate of the powder at the desired value, and only a small variation in the output variable is observed. This fluctuation is considered "normal" and under control when it is within the range of $\pm$ 5% of the set-point, because

it can be dumped by the subsequent blending unit operation without compromising the final tablet quality. When the feed factor profile is highly instable or drops too rapidly, the control system is likely to lose the control of the mass flowrate outside the confidence intervals. No matter if the materials are easily or poorly flow, at a certain point of the gravimetric run, the feed factor drops too quickly to be efficiently counterbalanced by the control system. Luckily, the dropping point is quite close to the end of the gravimetric run for materials with a good flowability, and a bit worse is the situation for cohesive materials. In both cases, there is the need of predicting the feed factor profile and variability in order to identify the dropping point and plan a refill phase that anticipates that occurrence, ensuring a more stable transition between the different phases.

Therefore, it is important to develop a model able to predict the feed factor for a single material or that can possibly predict the feeder performance of different materials in the line, in such a way as to plan a consistent refill strategy.

This Chapter provides some insights about the data available and the modelling possibilities that can be applied to meet the requirements that have been illustrated.

4.2 Feeder data

In this section, a recap of the feeder data that are commonly available in the relevant section of a commercial continuous direct compaction line are summarized.

Table 4.1 *Available categorical variables concerning the process condition and the setup of a commercial loss-in-weight feeder.*

Variable symbol	Variable name
s	Screw type
gb	Gear box ratio
$\dot{m}_{sp}$	Mass flowrate set-point

Generally, a loss-in-weight feeder is not characterized by a significant number of parameters regarding the process conditions and the setup of the equipment. The most common ones are summarized in Table 4.1 and consist in screw type, gear box and mass flowrate set-point. Few different screw types and gear box settings can be changed in the initial configuration of the set of multiple feeders, and the choice is strictly related to the engineering consideration on the main features of the type of material that is going to be processed. On the other hand, the mass flowrate set-point is the only parameter that is given as input to the equipment, and its value is determined by the formulation of the drug that is processed in the line. Typical values range from few kg/h (for tablets with low API concentrations) up to 20 kg/h for excipients.

Similarly, only few sensors are installed in these pieces of equipment and the most common variables recorded (or calculated in real time) are reported in Table 4.2. In fact, the mass flowrate and the feed factor are variables that are not recorded by a physical sensor, but are calculated by the controller logic as the derivative of the decreasing weight over the time, and the ratio of the mass flowrate over the screw speed, respectively.

Table 4.2 *Available continuous variables recorded by a commercial loss-in-weight feeder.*

Variable symbol	Variable name
$\dot{m}$	Mass flowrate
v_s	Screw speed
w	Weight
ff	Feed Factor

Other data, which are not collected by the feeder but can be relevant to delineate a feeder model, are powder properties presented in Table 4.3. Most of the powder properties reported in the table are the same of the ones used in the dataset discussed in the previous Chapter. The data collected by researchers and materials scientists can also be used as input for a feeder model in order to investigate or predict the variability in performance given by different powders. Some other useful material properties, that can especially affect the feeder operation, are suggested here as a valuable integration to better investigate which set of properties is giving the largest contribution.

In conclusion, to meet the requirements illustrated in the previous section, it is suggested to collect the data by the so called "gravimetric experiments", that consist in experiments where the hopper is filled up with the maximum amount of powder and it is then discharge in the gravimetric mode until empty. Then, the data recorded for each experiment can be stored as "batch data" and treated as such.

Table 4.3 *Examples of powder materials properties that are usually collected in the drug product design phase and that are available for modelling purposes.*

Variable symbol	Variable name
d_{10}	Particle size d_{10}
d_{50}	Particle size d_{50}
d_{90}	Particle size d_{90}
ffc	Flow function coefficient
WFA	Wall friction angle
BD	Bulk density
TBD	Tapped density
HR	Hausner ratio
CARR	CARR index
SSA	Specific surface area

4.3 A "static" modelling approach for the prediction of the feed factor profile

As previously mentioned, when dealing with loss-in-weight feeders, the time evolution of the feed factor becomes particularly important. For simplicity, the hopper of the equipment is conceptually divided in 10 theoretical zones of the same weight, also referred to as "bins". For each bin an average value of the feed factor (*ff*) across that bin is calculated and this information can be used to delineate an approximate feed factor curve of 10 data points. This information is traditionally used to pass a "calibration" array of feed factor values before starting the manufacturing operations in the line. This is expected to help the control system to find an optimal initial screw speed of the feeder in such a way that the system can reach the set-point quickly and, eventually, to signal a possible wrong configuration of the feeder setup that may require a very high or low number of motor speed when the feeder is close to the end of the gravimetric phase. The calibration array is also somehow supporting the control system, but the way it interfaces to it is generally not disclosed by the equipment manufacturing companies.
Nevertheless, having a model that predicts an array of feed factor values based on the historical data would be extremely useful. This kind of model could also be used to investigate if the equipment setup can affect the feeder performance and, eventually, establish the best setup for a specific material.
The proposed methodology consists in a PLS multivariate linear regression method to define the relation between a block $\mathbf{X_1}$, which is a dataset where the rows are the experiments and the columns are the variables that explain the feeder setup (e.g. the ones reported in Table 4.1), and the block $\mathbf{Y_1}$, which is the dataset that contains the respective arrays of the average feed factors.

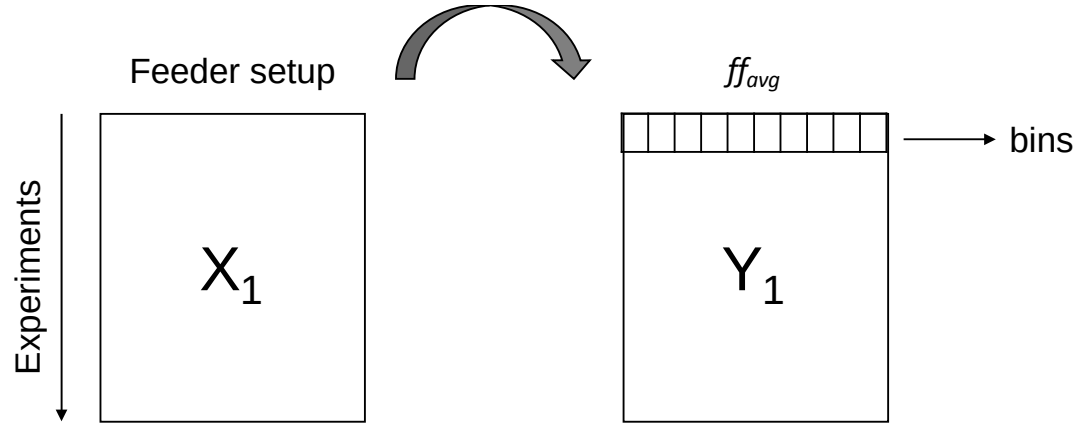


Figure 4.9 *Schematic use of PLS to estimate the feed factor array from the feeder setup information on data from a single material.*

Figure 4.1 represents a scheme of this possible modelling strategy and it can be noticed that, in fact, this model is a multiple response PLS model that is likely to give a better overall prediction of the feed factor profile than multiple single model for each bin because of the internal correlation of the multiple *Y's*. Moreover, the noise in the measurements of the average *ff* values will be removed uniformly along the profile, since the number of principal components is selected based on the correlation between the entire profile and not just on the single bin value.

Matrix $\mathbf{Y_1}$ is supposed to have a strongly linear correlation between the *Y's* profiles for different experiments. However, if this does not occur it would mean that there are some other sources of variability that have not been included in the matrix of the predictors $\mathbf{X_1}$ and that they need to be considered. As a matter of fact, the variability can be given by some experimental conditions (e.g. humidity of the air, temperature, and so on), and analysis of the loadings of the model can aid the identification of the source of variation in the design space.

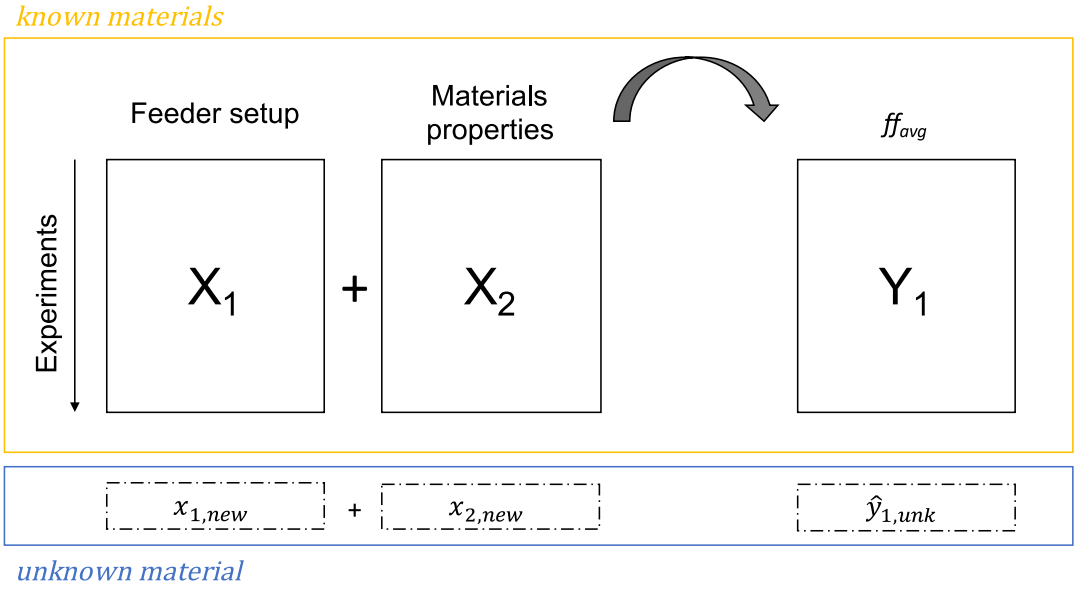


Figure 4.10 *Schematic use of a PLS model for the prediction of the feed factor array. The model is the result of the input combination of a matrix X_1 of feeder setup information and a matrix X_2 of materials properties data. The model can be used for the prediction of a new material feed factor array if the feeder setup and materials properties information are provided. Eventually, the model can be used for exploratory purposes.*

This model can easily be extended to include not only the data about a single material at a time, but considering multiple materials at the same time. The modification regards the input matrix of predictors $\mathbf{X_1}$ only, and basically consists in combining a second block of data $\mathbf{X_2}$ with the raw material properties of the powders, e.g. the ones suggested in Table 4.3 as shown in Figure 4.2. In this case, the capabilities of a multivariate approach become really evident since the PLS model is able to handle as many input

variables as available, and this fact will greatly simplify the a posteriori analysis of the results.

The number of materials that are included in the model input is a choice of the modeller and depends on the final purpose of the analysis. For example, if the aim of the model is an exploratory analysis of the effect of the different materials features on the response variable, then a global model that includes as many different materials as available is probably a good option. However, this model is likely to have very poor performances in prediction since the feed factor shape is affected by the processability of the materials. In order to overcome this problem, the selection of the materials can be done following a "similarity" approach, in the sense that only materials that have similar features are used as inputs for the prediction of a new material x_{new} with unknown feed factor response $\hat{y}_{1,unk}$. The similarity rule between materials can be based on the classification model developed in the previous Chapter, thus selecting only the materials that come from the same class of "flowability", or only a subset of the *k* nearest neighbours within the same class to increase the specificity of the model itself.

4.4 A "dynamic" modelling approach for the prediction of the feed factor profile

The previous PLS modelling approach was named "static" because in fact the information about the dynamic evolution of the experiments is lost in the modelling simplification of using a reduced array of ten average feed factor values. Even if the predicted multiple responses are modelled considering the correlation of *Y's* and not just as independent values, there is no trace of the natural dynamic relation between the values of *Y's* at different time instants.

Furthermore, the way the **Y** data are pretreated to obtain an average value of feed factors for each of the ten bins, is an oversimplification of what really happens across the feeder during a gravimetric experiment. As a matter of fact, the average of target variable profile cannot give any precise information about some possible anomalous phenomena that happened in some specific zones of the feeder for some particular materials, e.g. the recurrence of a feed factor dropping point in a limited zone or of a high internal variability of the feed factor that is "masked" by the average.

Hence, a new multivariate modelling approach should be adopted to make good use of the data available for the experiments in such a way as to consider the entire progression of each batch.

The natural extension of PLS to deal with three-dimensional multivariate data is the multi-way PLS model with batch-wise unfolding of the dataset in order to include the

batch dynamics. A schematic procedure of how to rearrange the blocks of data to develop a MwPLS model for the above problem is shown in Figure 4.3.

A single model for a specific material can be developed as for the first scenario of the previous section, but here a block of dynamics variables $\mathbf{X_3}$, as the ones presented in Table 4.2, needs to be added to the equipment setup information in order to model a dynamic response $\mathbf{Y_2}$. Then, $\mathbf{Y_2}$ becomes a matrix that contains the entire profile along the time of the feed factor and not just some intermediate average points.

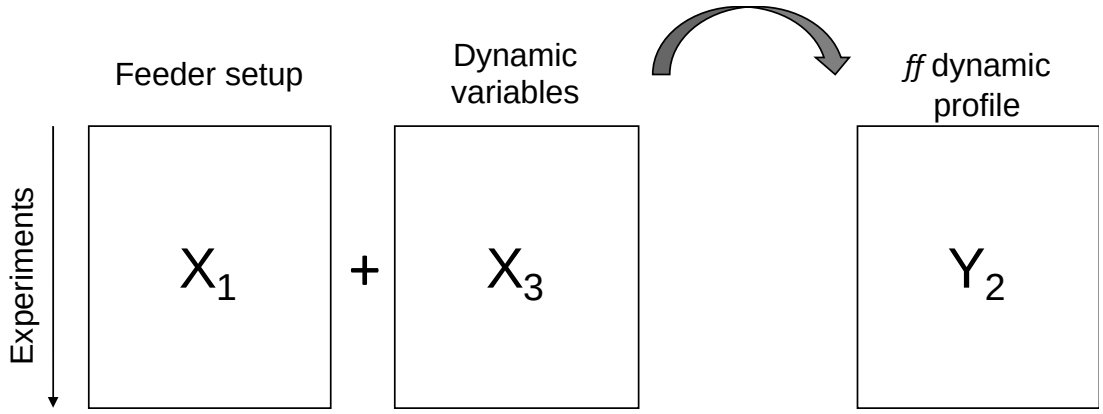


Figure 4.11 *Schematic procedure of a MwPLS model for the prediction of the entire profile of the feed factor.*

This approach allows for the prediction of the dynamic variation around the average trajectory of the targeted variable based on the historical database of gravimetric experiments for a single material. It has also the advantage of extending the application to other technical purposes, such as for example the online monitoring of the feeder equipment but, in order to achieve this target in commercial manufacturing operations, several improvements need to be done.

Obviously, the fact that dynamic experimental data are required to develop this predictive model is a limitation for the practical use of the model itself, especially if the interest is focused on the prediction of new materials behaviour in the equipment. As in the previous case, to overcome this problem a possible approach is to improve the model including data from multiple materials and a block of data for the materials properties in order to include in the model some information about the design space that is likely to cause the variation in the feeder performance across the range of different materials. The representation of this extension of the model is given in Figure 4.4.

The benefits of having a model like this one are very similar to the benefits generated by the "static" multiple materials model of Figure 4.2, but with some advantages given by the fact that the introduction of the dynamic block of variables can explain some possible deviations from the expected behaviour of a new material that is typical of the group of similar materials used as predictors.

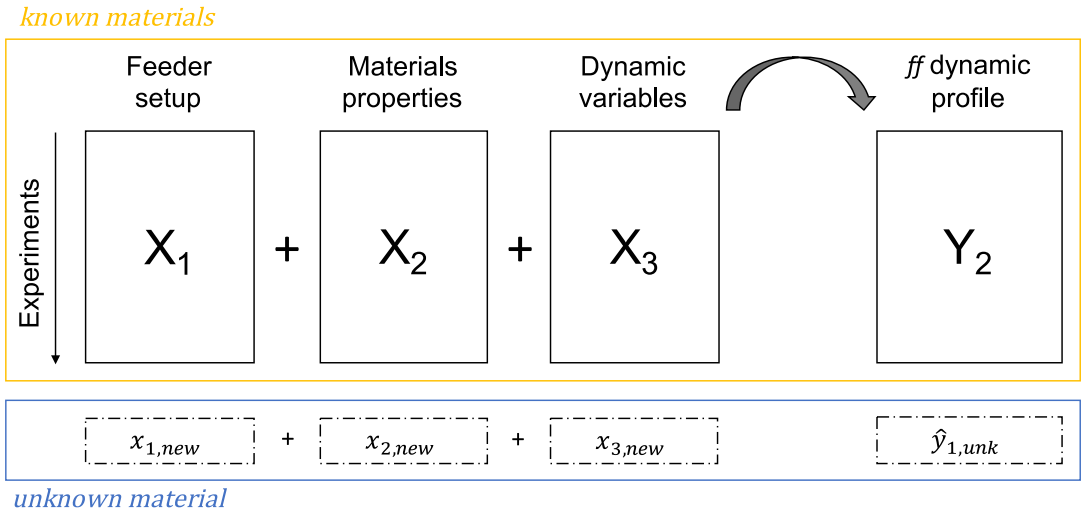


Figure 4.12 *Schematic procedure of a MwPLS model extended to multiple materials. Purposes and uses can be considered similar to the model proposed and schematized in Figure 4.2, but with some advantages.*

Nevertheless, it has the limitation that producing a new set of dynamic variables $x_{3,new}$ for the unknown material might require some adjustments in the input variable selection to include in the model and the specific regression of some of them.
All the figures here are just a simplify scheme of the models' structure but a proper rearrangement of the datasets needs to be done according to Nomikos and MacGregor [34].

4.5 Case study: a dynamic model for predicting feeder performances on a single material

In the first part of the section 4.4, a dynamic modelling approach to predict the feed factor profile of a single material has been proposed (Figure 4.3). In this section, a case study on a pharmaceutical manufacturing application is illustrated. The model has been developed with the aim of providing a simple data-driven tool for feed factor prediction and that can be used by operators, materials scientists and other researchers involved in drug design and development. In particular, the has been designed to reduce the gravimetric feeder trials that are usually performed before a manufacturing campaign.

However, with minor changes the area of application can be extended to more specific purposes, e.g. on-line faults detection.

4.5.1 Available data

The available data are from four gravimetric experiments conducted prior to a testing campaign on the feeding part of a new continuos tabletting line, and they refer to a loss-in-weight feeder with a volume capacity of approximately $2.5x10^{-3}$ m^3. The API material tested has the powder characterization properties reported in Table 4.4 and corresponds to material M18 in the dataset of the case study of Chapter 3.

Table 4.4 *Material properties of the API powder tested in the loss-in-weight feeder.*

d_{10}	12.9	[-]
d_{50}	31.1	[-]
d_{90}	63	[-]
ffc	2.5	[-]
WFA	31.7	[°]
BD	0.29	[g/cm^3]
TBD	0.54	[g/cm^3]
HR	1.86	[-]
CARR	46	[-]
SSA	0.39	[g/m^2]

The material was previously classified (see results Chapter 3) as a "cohesive" powder. Therefore, the expected feed factor profile should rapidly decrease during the gravimetric run and a noisy irregular shape of the profile of both the feed factor and the mass flowrate is expected.

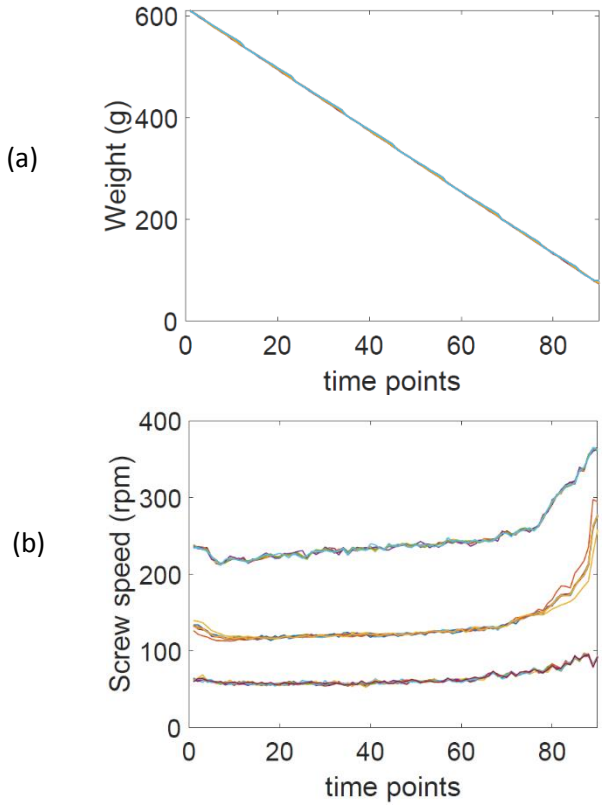
The original experiments were conducted at three different mass flowrate setpoints: 5 kg/h, 10 kg/h and 20 kg/h. In fact, two of the four original experiments used the same mass flowrate setpoint of 10 kg/h. Obviously, the batches have different lengths because of the different flowrates and a batch alignment is required in order to develop a MwPLS model using a batch-wise unfolding. The presence of a monotonic decreasing variable, such as the weight, suggested the indicator variables technique for resampling the trajectories with respect to this variable as proposed by Nomikos and MacGregor [70]. Since these are the only data available, thirty additional realizations of the same batches, i.e. ten for each mass flowrate setpoint, have been simulated by adding white noise to the original signals according to the standard deviation of each variable. The profiles of the two experiments at the same mass flowrate target have been averaged to produce the respective new ten batches. A total number of thirty-four batches can be used to build the model. A summary of the available and realized batches is given in Table 4.5.

Table 4.5 *Number of available and realized batches at 5 kg/h, 10kg/h and 20 kg/h*

Mass flowrate setpoint (kg/h)	n° of available batches	n° of realizations batches
5	1	10
10	2	10
20	1	10

The batches have been ordered and numbered progressively starting from the original ones. The variables data collected by the feeder are the one reported in Table 4.2, i.e. weight, screw speed, mass flowrate and feed factor. To understand the general dynamic evolution of each batch in better terms, the time profiles for all the batches are shown in Figure 4.5.

Lastly, all the experiments have been done using the same screw type and the same gear box ratio. Hence, no considerations or comparisons can be done on the effect of different screw design.



(c)

(d)

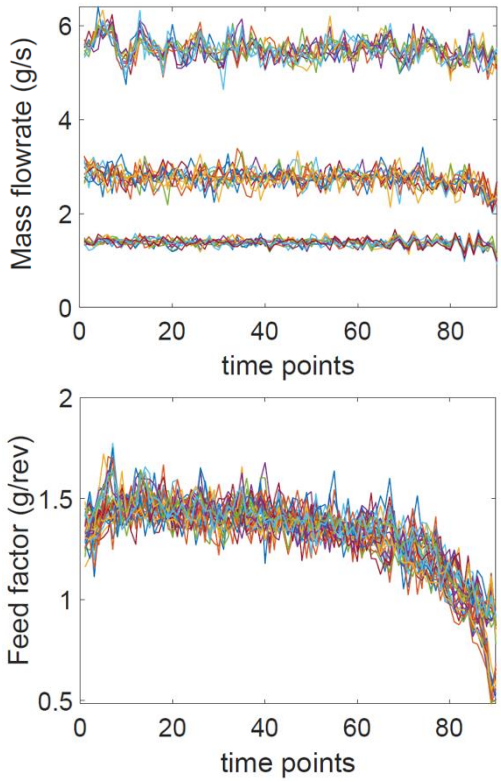


Figure 4.5 *Time profiles of the variables for the loss-in-weight for both the original and the realized batches: (a) weight; (b) screw speed; (c) mass flowrate; (d) feed factor.*

4.5.2 Modelling strategy

In this case study, the only equipment setup parameter, which can be used as input for the MwPLS model, is the mass flowrate set-point. Since only feeder data from one material are available, the material characterization properties will not be included as inputs. In this context, the choice is mostly related to the selection of the dynamic variables to include as predictors to model the feed factor profile. The model will be used by the users to predict the feed factor at a specific mass flowrate, and this means that including all the dynamic variables in the model will create a problem in the definition of unknown related profiles such as the screw speed and the mass flowrate. However, if the weight is used in the model as the only known dynamic variable of the system, the users can simply specify the weight as a linear array of 90 elements with the first value equal to the maximum weight at the beginning of the experiment, and the last value equal to the minimum weight reached at the end of the experiment. In this way, the model will "map" the relationship between mass flowrate setpoint and weight profile based on the training data that are used to calibrate the model.

For that purpose, four batches (batch #3, batch #12, batch #19, batch #27) at different flowrates were randomly selected to be used as external validation dataset and the

remaining thirty batches are used to calibrate the model. Finally, the data were rearranged according to Nomikos and MacGregor [34].

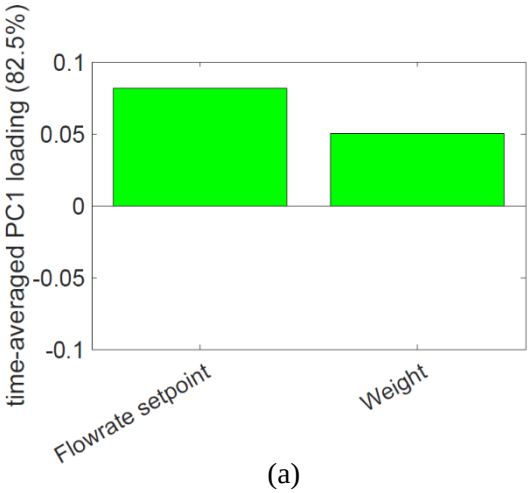
4.5.3 Results

A MwPLS model was built on the calibration dataset and two PC were selected to explain 96.3 % of the variability in the **X** block, and 39.7% of the variability in the **Y** block (see Table 4.6).

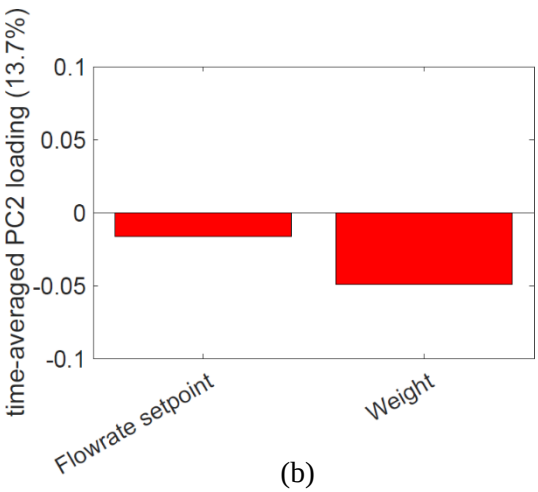
Table 4.6 *MwPLS model on the 30 batches calibration dataset: model diagnostic.*

	X Block		Y Block	
Component	**Variance (%)**	**Cumulative variance (%)**	**Variance (%)**	**Cumulative variance (%)**
1	82.5	82.5	17.7	17.7
2	13.7	96.2	21.9	39.7

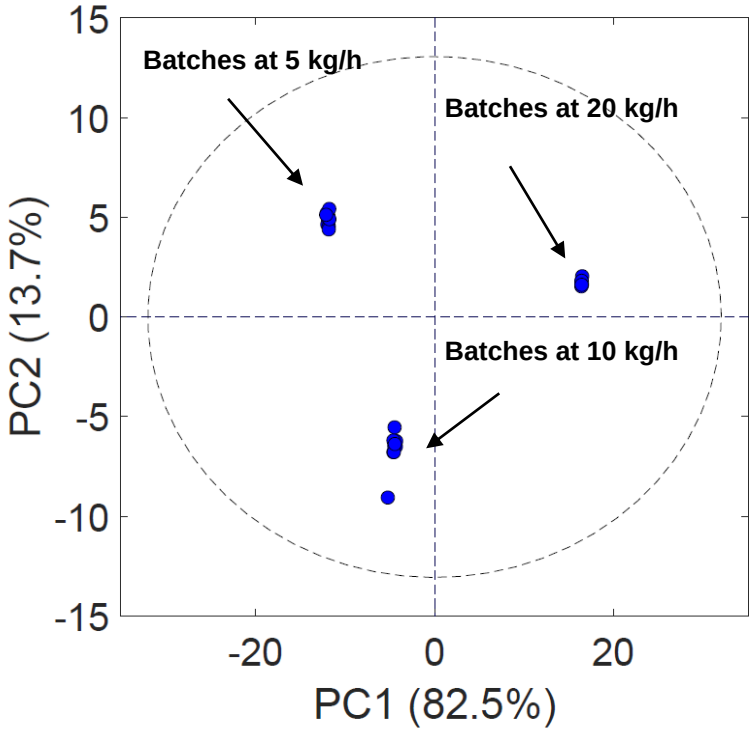
The time averaged loading plots (Figure 4.6a and 4.6b) show that the mass flowrate explains most of the variance in the first principal component, while the weight explains most of the variance in the second principal component. These results match the initial expectations since they are the only input parameters selected.



(a)



(b)



(c)

Figure 4.6 *MwPLS model on 30 batches calibration dataset: (a) PC1 time-averaged loading plot; (b) PC2 time-averaged loading plot; (c) PC1 vs PC2 score plot.*

The score plot (Figure 4.6c) shows clearly that the batches are clustering according to the mass flowrate setpoint and this indicates that the response variable profile, i.e. the feed factor, varies with reference to the mass flowrate target. Once again, this result confirms the existent knowledge of the system and the supposed discrimination power of the model based on the mass flowrate.

The estimation performance for the feed factor profile shows a very good curve fitting in the model calibration with an average R^2 = 0.91. If these performances will be confirmed also in the external validation, the capability of the model can be considered perfectly appropriate for the scope for which has been conceived.

Therefore, the four validation batches were projected onto the existing model and the results on the score plot can be observed in Figure 4.7.

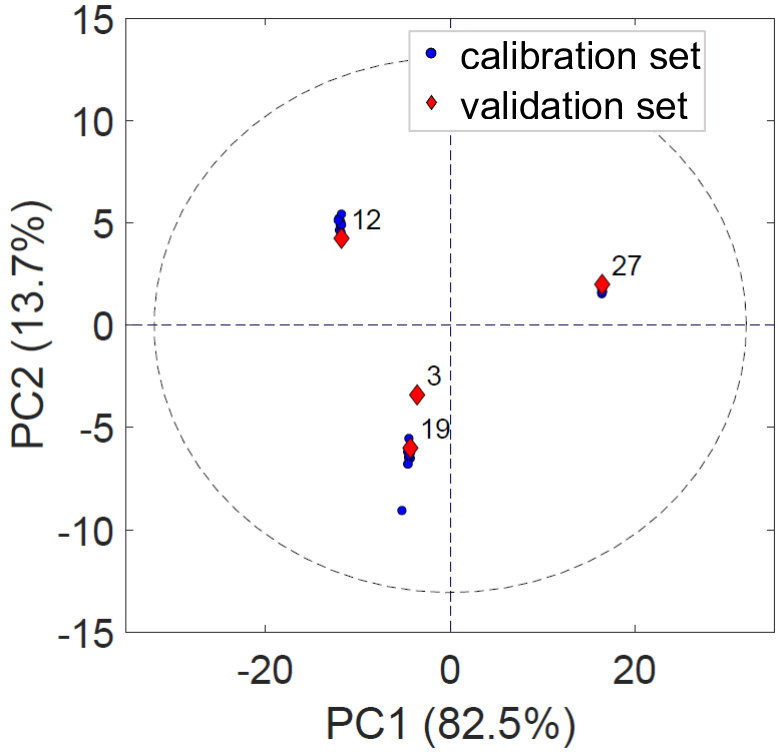


Figure 4.7 *Score projection of the four external validation batches on the MwPLS model.*

The score's projection shows how the validation batches cluster close to the proper mass flowrate especially along PC1. Batch #3 shows a slight drift along the PC2. In fact, this batch was excluded on purpose from the calibration dataset since it is one of the original batches available at 10 kg/h. This means that, according to the way the new batches have been realized, no "noisy" profiles are matching this one (the white noise for these batches was added to the average of the two original batches at 10 kg/h).
Hence, particularly interesting will be the comparison between the prediction results for this specific batch and the noisy ones. In Figure 4.8 the feed factor estimation for the validation dataset is reported together with the measured profile. The estimations is very good in all cases, with only some minimum discrepancies for the more noisy batches. However, even in that case the model is able to capture the general trend of the feed factor shape including some recurrent areas where the profile seems to drop more rapidly.

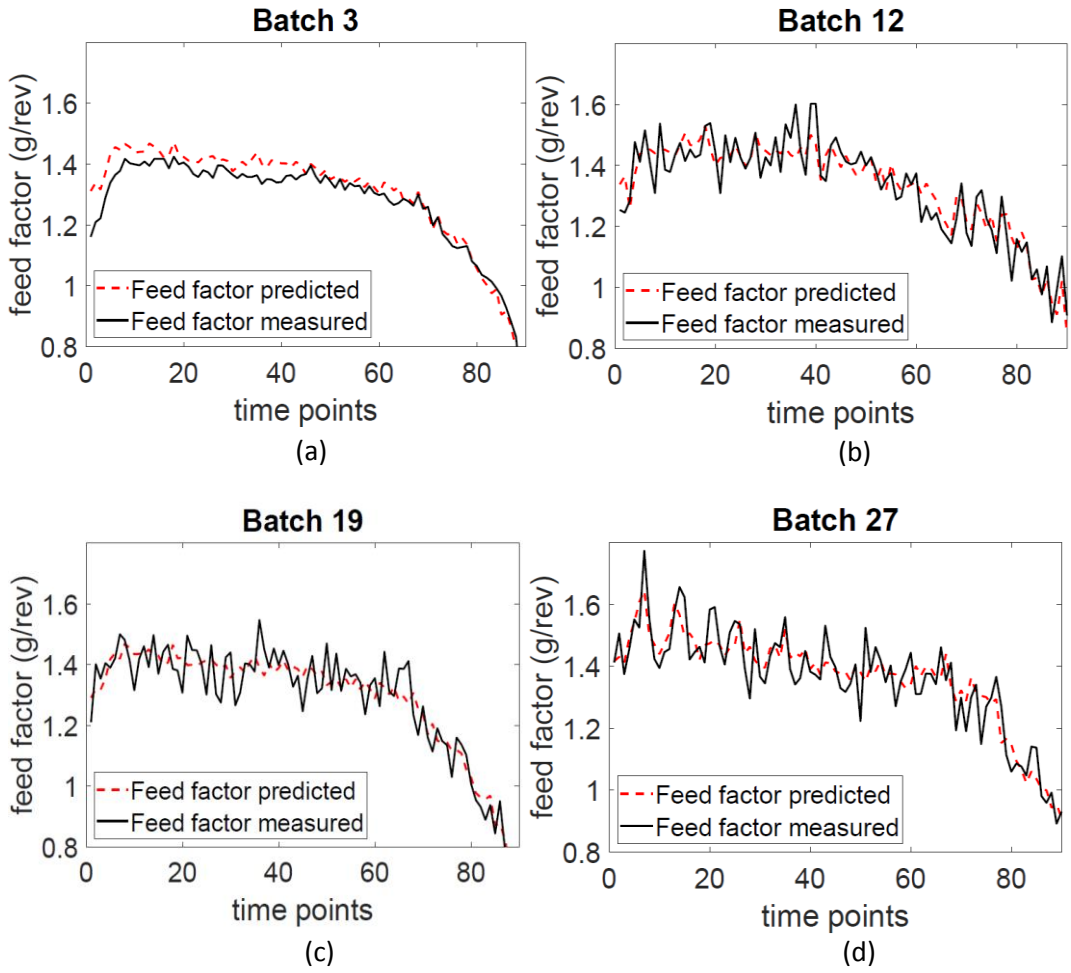


Figure 4.8 *MwPLS model results for the prediction of the feed factor of the validation dataset: (a) batch 3; (b) batch 12; (c) batch 19; (d) batch 27.*

The curve fitting performances for the validation dataset in terms of R^2 are reported in Table 4.7 and this time the average value is approximately equal to 0.87.

Table 4.7 *R^2 results for estimation of the feed factor profiles on the validation dataset.*

Batch number	Flowrate setpoint (kg/h)	R^2
#3	10	0.92
#12	5	0.85
#19	10	0.90
#27	20	0.83

Even in this case, the results can be considered appropriate to the purpose of the model and promising in terms of future development of more sophisticated models for more complex purposes, e.g. faults detection or on-line multivariate control.

4.6 Conclusions

In this Chapter, a new data-driven approach for powder feeding modelling has been presented from a forward-looking perspective that is promising both for development and for manufacturing operations. The idea of integrating multivariate statistical analysis in powder feeding modelling is something that has been sought for in the last years [8], but no references can be found about using a latent variable modelling approach for the prediction of the feed factor. The idea discussed in this Chapter is likely to arise further interest for manufacturing applications.

Two main concepts have been expanded in this Chapter from both a "static" and a "dynamic" perspective:

- the development of single models for single materials which have already been tested in the line through a systematic revision of the data collected for gravimetric experiments;
- the extension of the concept to multiple known materials models that are likely to produce good results in performance prediction for new unknown materials that cannot be tested in the line in the early stage of product development.

The proposed approach requires the combination of the knowledge about the equipment and process conditions, the properties of the powders utilized and the data collected by the equipment during the operations. The data are collected directly through the executions of dedicated gravimetric experiments and organised consistently in a "batch" manner.

Furthermore, the modelling methodologies require the use of a simple PLS extension for multiple response prediction for the "static" case and a more refined extension such as multi-way PLS for upgrading the model to a "dynamic" scenario. Both the methods can produce valuable results in prediction and address different purposes in the posteriori data-analysis.

The latter point on the extension to multiple materials model requires a good level of knowledge about the historical materials tested in the line and a structured approach to guide the input materials selection of the model. It is in this context that a systematic data-driven approach for a raw materials database investigation, as the one presented in Chapter 3, gains added value and can be integrated and personalized by the final users to achieve optimal results in the identification of materials surrogates with similar equipment performance.

The major drawback of this approach is the initial effort in constructing a various dataset of experiments for a relatively large number of powders. However, the efforts and the costs associated are estimated to be less than the cost of the actual manufacturing trials, both in terms of time, resources, and amount of powder required. Furthermore, the

estimation of feeder performance of a new drug formulation, based on materials surrogates tested in past, is a very powerful tool to aid and accelerate the drug development and design of new products for secondary continuous manufacturing line. The case study presented in this Chapter shows how it is possible to integrate a multivariate data-driven approach for the prediction of a target variables that is unlikely to be obtained in short time by the first principle understanding of the system. The good performance in estimation encourage a more frequent use of the available data to meet the modelling requirements of the manufacturing operations.

Most of the concepts reported in this Chapter have been successfully tested using several datasets from an experimental campaign of feeder trials. However, not all the results and the output of these analysis are presented in this Thesis project because of confidentiality issues. Nevertheless, no sufficient knowledge about the prediction of the behaviour of unknown materials have been collected and some more intense efforts in proving this concept are suggested.

In conclusion, addressing the generalised lack of first-principles understanding of powder flowability across a loss-in-weight feeder, a multivariate approach seems to be promising to speed up the development of feeding models for secondary continuous tableting lines, particularly with reference to the product and process design of new drug candidates.

Conclusions

In this Thesis, a data analytics approach for powder feeding modelling on continuous secondary pharmaceutical manufacturing processes has been proposed. In particular, this Thesis aims at providing an innovative data-driven approach to support the early stage of input materials selection in the product and process design phase of new continuous manufacturing lines.

The main contributions given by this study can be summarized in two aspects: the outline of a general procedure to aid the investigation of a raw materials database of API powders, and the proposal of a multivariate statistical approach for the development of powder feeding unit operation models.

In the first part, the proposed procedure aims at maximizing the amount of information that can be extracted from an available dataset of APIs measurements through a systematic methodology that: i) reorganizes the data, ii) explores the system, iii) identifies unknown patterns in the data, and iv) trains a classification model for class membership prediction of new future candidates. The procedure has been consolidated using an example of API powders candidates for a continuous tableting line. The industrial case study has shown that an integration between data-based knowledge of the measurements collected and system-based knowledge of materials experts is required to address the analytics approach in the right direction, following the four basic steps proposed. Firstly, the data need to be reorganized according to the limited availability and the purpose of the analysis, e.g. the identification of patterns of powder flowability in the presence of a restricted set of powder measurements. Secondly, the data can be modelled following a latent-variable approach to determine a reduced space of investigation and identify possible linear correlations between variables, e.g. PCA reduction of the original variables. Thirdly, the application of unsupervised clustering models is necessary for the identification of unknown patterns, and the choice can be restricted to a partitional clustering approach (if the number of clusters is known) or to an agglomerative clustering approach (if the number of clusters is unknown). In this case study, a hierarchical clustering method have been selected after a careful comparison of several different possible algorithms and four classes of powder flowability have been determined. The results have been confirmed by a cross-reading analysis of the models' outputs and the system-based knowledge of materials scientists. Lastly, a classifier must be designed to assign the correct class membership of new materials and a comparison between linear and non-linear methods is suggested in order to identify the correct training framework for the structure of the patterns analysed. In

this example, non-linear classifiers have shown superior performances than linear counterparts, since the patterns' boundaries were showing a non-linear shape. The outcomes of this analysis are likely to support the identification of surrogates input materials, and define a possible data-driven integration of powder characterization measurements into a following stage of data-driven modelling of the unit operations of the line.

The proposed methodology has been developed to be general and flexible, in such a way that it can be adapted and reproduced on any similar pharmaceutical raw materials database with none or minor changes in accord to the dataset structure, the users' expertise and the final goal of the analysis.

In the second part, the idea of using a data-driven approach for the prediction of a targeted quality variable for predicting feeding performance has been developed. The lack of first-principles understanding of powder flowability in feeding equipment has been addressed from a multivariate statistical approach, combining data from the equipment setup, materials properties data and process data from the feeder sensors in order to explain the correlation between these variables and the feed factor profile.

The PLS regression models are built both in a "static" scenario (where the feed factor response is a sampled average of ten consecutive hopper zones) and in a "dynamic" scenario (where the feed factor profile is evaluated along the entire duration of a gravimetric experiment). These models can provide useful information on the equipment setup, the manufacturing conditions that affects the powder processability across the feeder, the materials variability, the feeder performance and the estimation of the feeding behaviour of unknown materials.

Based on the promising results obtained, the future perspectives of this Thesis project are likely to be progressed in the next future. In particular, the first part on the raw materials database investigation can be extended to others downstream continuous processes. With the development of consistent and relatively large datasets of raw materials, some novel methodologies of unsupervised and supervised pattern recognition can be explored more into the details. In addition, many opportunities may arise from the extension of this general procedure to other raw materials databases, not only in the pharmaceutical industry. The second part contains several suggestions for the development of a multivariate modelling approach for powder feeding unit operations. Some possible areas for future research in this field are the integration of the database knowledge to predict new materials performances and the possibility of integrating a multivariate statistical tool for on-line control of the feeder operations, especially in the transition between the gravimetric and volumetric phases.

Ringraziamenti

È giunto il momento dei ringraziamenti alle persone che mi hanno sostenuto e accompagnato in questo percorso con passione e dedizione, dandomi la fiducia di mostrare le mie capacità.

Prof. Massimiliano Barolo

Per avermi trasmesso la passione per questa professione, ma soprattutto per aver creduto in me e avermi aiutato a concretizzare un mio sogno. Spero di aver contraccambiato la sua fiducia e di collaborare con il suo gruppo anche negli anni a venire.

Dr. Simeone Zomer

Per avermi accolto nel suo team con una speciale umanità e professionalità. Per spronarmi ogni giorno a fare il massimo, mettendoci tutta la passione per il proprio lavoro. In particolar modo, desidero ringraziarlo per formarmi dandomi la giusta responsabilità e l'opportunità di sbagliare e imparare dai miei errori.

Dr. Pierantonio Facco

Per i preziosi consigli e l'infinità disponibilità nell'insegnarmi in maniera semplice dei concetti difficili. Gran parte dei risultati che ho ottenuto in questo anno li devo a lui.

Dr. Jun Zhang

Per avermi insegnato ad applicare questi concetti, ma soprattutto per avermi spiegato passo dopo passo come inserirmi in una realtà enorme come GSK con semplicità e con la giusta dose di simpatia.

Ai miei colleghi tutti

Per avermi trattato sin dal primo giorno come un professionista, nonostante la mia giovane età e inesperienza.

A mamma e papà

Per avermi dato e per continuare a darmi tutto l'amore e la fiducia che hanno. Non è stata facile la mia scelta di trasferirmi, ma spero siano fieri dell'uomo che hanno cresciuto.

A mio fratello Diego

Perché nonostante l'apparenza sono fiero della responsabilità che ha sviluppato e di come non faccia sentire troppo la mia mancanza.

Nonni, zii, cugini e tutto il resto della mia bellissima famiglia

Per tutto l'amore che mi hanno trasmesso e per avermi cresciuto insieme a mamma e papà.

Alla compagnia

Per aver condiviso e continuare a condividere con me mille avventure. Per essere riusciti a rimanere un bellissimo gruppo e farmi sentire a casa anche quando sono distante. Vi considero parte integrante della mia famiglia e mi mancate ogni giorno.

A tutti i miei amici vecchi e nuovi e a tutte le persone che ho incontrato nei miei anni universitari e nelle mie esperienze all'estero

Non dimenticherò mai le bellissime esperienze che abbiamo condiviso. Un grazie speciale ad Aldo per aver condiviso ogni gioia e ogni difficoltà di questo percorso e per essersi dimostrato un grande amico oltre che un bravissimo collega.

Vi tengo tutti nel cuore e voglio condividere questo traguardo con voi perché ne siete parte integrante.

Antonio

Acknowledgements

Finally, it is time to thank all the people that support and walk with me in this experience with passion and dedication, giving me the opportunity to show my abilities.

Prof. Massimiliano Barolo

For passing me down the passion for this profession, but most of all for trusting and helping me in making my dream come true. I hope I had repaid his faith and I would be glad to work with his group for many years to come.

Dr. Simeone Zomer

For welcoming me into his team with a special kindness and professionalism. For pushing me every day at doing my best, with all his passion for his job. In particular, I would like to thank me for taking care of my development with the right dose of responsibility and giving me the opportunity of learning from my errors.

Dr. Pierantonio Facco

For all the precious advice and his infinite willingness in teaching me difficult concepts with an easy approach. Most of the results that I achieved in this last year are thank to him.

Dr. Jun Zhang

To teach me how to apply those concepts, but most importantly for helping me day by day in my integration progress in the enormous reality of GSK, with the right dose of simplicity and congeniality.

All my colleagues

For treating me as a professional from day one, despite my young age and inexperience.

Mum and Dad

For giving me all the love and the trust they have every day. My choice of moving was not easy, but I hope they are proud of the man they raised up.

My brother Diego

Because despite the appearance, I am proud of the responsibility he developed and how he is doing a great job in not making feel my absence.

Grandpas, grandmas, aunts, uncles, cousins and all the rest of my beautiful family

For all the love they give to me and for raising me up together with Mum and Dad.

The Crew

For sharing with me thousands of adventures. For staying together as a beautiful group and making me feel home when I am far away. I consider you as my family and I miss you every day.

To all my old and new friends that I have known in these academic years and all the people that I met during my experiences abroad

I will never forget all the beautiful experiences we have shared. A special thanks to Aldo for sharing with me every happiness and obstacles of this experience and for being a great friend and an excellent colleague.

I keep all of you in my heart and I want to share with you this achievement because you are well and truly an integral part of this.

Antonio

List of symbols

N	number of samples	[-]
V	number of variables	[-]
r	rank of the matrix	[-]
z	number of principal components	[-]
lv	latent variables	[-]
V_X	number of variables in the predictor variable matrix	[-]
V_Y	number of variables in the predicted variables matrix	[-]
Y	response variables in **Y**	[-]
K	number of time data points	[-]
J	number of process measurements or variables	[-]
I	number of batches	[-]
d	number of descriptors for a generic dataset	[-]
x	generic sample	[-]
y	generic sample	[-]
c_h	cophenetic correlation coefficient	[-]
b_{xy}	Euclidean distance between two generic samples x and y	[-]
t_{xy}	cophenetic distance between two generic samples x and y	[-]
$\bar{b}$	average of the b_{xy}	[-]
$\bar{t}$	average of the t_{xy}	[-]
$\%\ CC$	percentage of samples correctly classified	[-]
G	number of classes	[-]
d_{VC}	Vapnik-Chervonenkis dimension	[-]
C	cost parameter for SVM	[-]
ff	feed factor	[g/rev]
$\dot{m}$	mass flowrate	[g/s]
v_s	screw velocity	[rev/s]
PC1	principal component 1	[-]
PC2	principal component 2	[-]
PC3	principal component 3	[-]
ffc	flow function coefficient	[-]
w	weight	[g]
s	screw type	[-]
gb	gear box type	[-]
SSA	specific surface area	[g/m^2]
BD	bulk density	[g/cm^3]

TBD	tapped bulk density	[g/cm^3]
HR	Hausner ratio	[-]
CARR	Carr's compressibility index	[-]
WFA	wall friction angle	[°]

Greek letters

λ	eigenvectors	[-]
λ_i	eigenvalues	[-]
$\varphi(x)$	kernel function	[-]
γ	radial parameter of a radial basis kernel function	[-]
σ	standard deviation	[-]

Vectors and matrices

X	generic dataset or matrix of predictors (PLS)
Y	matrix of predicted variables
t	score vector
p	loading vector
T	matrix of the scores of **X**
P	matrix of the loadings of **X**
E	error or residual matrix
$\mathbf{X_{appr}}$	approximate reconstructed matrix
W	matrix of weights
J	matrix of the scores of **Y**
Q	matrix of the loadings of **Y**
$\mathbf{E_X}$	error or residual matrix of **X**
$\mathbf{E_Y}$	error or residual matrix of **Y**
$\overline{\mathbf{X}}$	three-dimensional matrix of predictors
$\overline{\mathbf{Y}}$	three-dimensional matrix of predicted variables

$\mathbf{x}$	vector of measurements for a generic sample *x*
$\mathbf{y}$	vector of measurements for a generic sample *y*
$\mathbf{Z}$	variance-covariance matrix
$\mathbf{X_1}$	matrix of the feeder setup data
$\mathbf{X_2}$	matrix of the material properties data
$\mathbf{X_3}$	matrix of the dynamic data collected by the feeder
$\mathbf{Y_1}$	matrix of the ten zones feed factor arrays
$\mathbf{Y_2}$	matrix of the continuous feed factor

Acronyms

k-NN	*k* nearest neighbours
PLS-DA	partial least-squares discriminant analysis
SVM	support vector machine
MSA	multivariate statistical analysis
LVs	latent variables
PCA	principal component analysis
PLS	partial least-squares
PAT	process analytical technology
MwPLS	multi-way partial least-squares
HCA	hierarchical clustering analysis
RBF	radial basis function
SRM	structural risk minimisation
ERM	empirical risk minimisation
OSD	oral solid dosage
FDA	food and drug administration
EMA	European medicines agency
QbD	quality-by-design
API	active pharmaceutical ingredient
NIR	near-infrared
PSD	particle size distribution
DEM	discrete element method

References

[1] Kleinebudde, P., Khinast, J., Ranten, J. (2017). *Continuous Manufacturing of Pharmaceuticals*, Wiley.

[2] Gernaey, K. V., Cervera-Padrell, A. E., & Woodley, J. M. (2012). A perspective on PSE in pharmaceutical process development and innovation. *Computers & Chemical Engineering*, 42, 15-29.

[3] McKenzie, P., Kiang, S., Tom, J., Rubin, A. E., & Futran, M. (2006). Can pharmaceutical process development become high tech?. *AIChE Journal*, *52*(12), 3990-3994.

[4] Scherer, F. M. (1993). Pricing, profits, and technological progress in the pharmaceutical industry. *Journal of Economic Perspectives*, *7*(3), 97-115.

[5] FDA. (2002). Guidance: Pharmaceutical cGMPs for the 21st century–A risk based approach.

[6] EMEA. (2004). ICH Q8, Q9, Q10 and Q11, Pharmaceutical Development.

[7] Lawrence, X. Y. (2008). Pharmaceutical quality by design: product and process development, understanding, and control. *Pharmaceutical research*, *25*(4), 781-791.

[8] Wang, Y., Li, T., Muzzio, F. J., & Glasser, B. J. (2017). Predicting feeder performance based on material flow properties. *Powder Technology*, *308*, 135-148.

[9] Koynov, S., & Muzzio, F. J. (2016). A quantitative approach to understand raw material variability. In *Process Simulation and Data Modeling in Solid Oral Drug Development and Manufacture* (pp. 85-104). Humana Press, New York, NY.

[10] Rietema, K. (1984). Powders, what are they?. *Powder Technology*, *37*(1), 5-23.

[11] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37-52.

[12] Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, *58*(2), 109-130.

[13] Gerlach, R. W., Kowalski, B. R., & Wold, H. O. (1979). Partial least-squares path modelling with latent variables. *Analytica Chimica Acta*, *112*(4), 417-421.

[14] Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, *185*, 1-17.

[15] Sandler, N., & Wilson, D. (2010). Prediction of granule packing and flow behavior based on particle size and shape analysis. *Journal of pharmaceutical sciences*, *99*(2), 958-968.

[16] Ferreira, A. P., & Tobyn, M. (2015). Multivariate analysis in the pharmaceutical industry: enabling process understanding and improvement in the PAT and QbD era. *Pharmaceutical development and technology*, *20*(5), 513-527.

[17] Tomba, E., De Martin, M., Facco, P., Robertson, J., Zomer, S., Bezzo, F., & Barolo, M. (2013). General procedure to aid the development of continuous pharmaceutical processes using multivariate statistical modeling–An industrial case study. *International journal of pharmaceutics*, *444*(1-2), 25-39.

[18] Muteki, K., Swaminathan, V., Sekulic, S. S., & Reid, G. L. (2011). De-risking pharmaceutical tablet manufacture through process understanding, latent variable modeling, and optimization technologies. *AAPS PharmSciTech*, *12*(4), 1324-1334.

[19] Largoni, M., Facco, P., Bernini, D., Bezzo, F., & Barolo, M. (2015). Quality-by-Design approach to monitor the operation of a batch bioreactor in an industrial avian vaccine manufacturing process. Journal of biotechnology, 211, 87-96.

[20] Sever, N. E., Warman, M., Mackey, S., Dziki, W., & Jiang, M. (2009). Process analytical technology in solid dosage development and manufacturing. In *Developing Solid Oral Dosage Forms* (pp. 827-841).

[21] Luypaert, J., Massart, D. L., & Vander Heyden, Y. (2007). Near-infrared spectroscopy applications in pharmaceutical analysis. *Talanta*, *72*(3), 865-883.

[22] Knop, K., & Kleinebudde, P. (2013). PAT-tools for process control in pharmaceutical film coating applications. *International journal of pharmaceutics*, *457*(2), 527-536.

[23] Jackson, J. E. (1981). Principal components and factor analysis: Part III—What is factor analysis?. *Journal of Quality Technology*, *13*(2), 125-130.

[24] Jackson, J. E. (2005). *A user's guide to principal components* (Vol. 587). John Wiley & Sons.

[25] Wise, B. M., & Gallagher, N. B. (1996). The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, *6*(6), 329-348.

[26] Wise, B. M., Gallagher, N. B., Bro, R., Shaver, J. M., Windig, W., & Koch, R. S. (2006). Chemometrics tutorial for PLS_Toolbox and Solo. *Eigenvector Research, Inc*, *3905*, 102-159.

[27] Valle, S., Li, W., & Qin, S. J. (1999). Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Industrial & Engineering Chemistry Research*, *38*(11), 4389-4401.

[28] Esbensen, K. H., & Geladi, P. (2009). Principal component analysis: concept, geometrical interpretation, mathematical background, algorithms, history, practice.

[29] Varmuza, K., & Filzmoser, P. (2016). Introduction to multivariate statistical analysis in chemometrics. CRC press.

[30] Wold, H. (1982). Soft modelling: the basic design and some extensions. *Systems under indirect observation, Part II*, 36-37.

[31] Wold, S., Ruhe, A., Wold, H., & Dunn, III, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, *5*(3), 735-743.

[32] Höskuldsson, A. (1988). PLS regression methods. *Journal of chemometrics*, *2*(3), 211-228.

[33] Camacho, J., Picó, J., & Ferrer, A. (2008). Bilinear modelling of batch processes. Part I: theoretical discussion. *Journal of Chemometrics*, *22*(5), 299-308.

[34] Nomikos, P., & MacGregor, J. F. (1995). Multi-way partial least squares in monitoring batch processes. *Chemometrics and intelligent laboratory systems*, *30*(1), 97-108.

[35] Shi, R., & MacGregor, J. F. (2000). Modeling of dynamic systems using latent variable and subspace methods. *Journal of Chemometrics*, *14*(5-6), 423-439.

[36] García-Muñoz, S., Kourti, T., MacGregor, J. F., Mateos, A. G., & Murphy, G. (2003). Troubleshooting of an industrial batch process using multivariate methods. Industrial & engineering chemistry research, 42(15), 3592-3601.

[37] Watanabe, S. (1985). *Pattern recognition: human and mechanical*. John Wiley & Sons, Inc..

[38] Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(1), 4-37.

[39] Gelbard, R., Goldman, O., & Spiegler, I. (2007). Investigating diversity of clustering methods: An empirical comparison. *Data & Knowledge Engineering*, *63*(1), 155-166.

[40] De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, *50*(1), 1-18.

[41] Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, *11*(2), 33-40.

[42] Brereton, R. G. (2003). *Chemometrics: data analysis for the laboratory and chemical plant*. John Wiley & Sons.

[43] Todeschini, R., Ballabio, D., Cassotti, M., & Consonni, V. (2015). N3 and BNN: Two new similarity based classification methods in comparison with other classifiers. *Journal of chemical information and modeling*, *55*(11), 2365-2374.

[44] Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

[45] Vapnik, V., Golowich, S.E., Smola , A.J. (1997). Support vector method for function approximation, regression estimation and signal processing. In *Advances in neural information processing systems* (pp. 281-287).

[46] Xu, Y., Zomer, S., & Brereton, R. G. (2006). Support vector machines: a recent method for classification in chemometrics. *Critical Reviews in Analytical Chemistry*, *36*(3-4), 177-188.

[47] Gunn, S. R. (1998). Support vector machines for classification and regression. *ISIS technical report*, *14*(1), 5-16.

[48] Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. *Analyst*, *135*(2), 230-267.

[49] Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of chemometrics*, *17*(3), 166-173.

[50] Górski, Ł., Sordoń, W., Ciepiela, F., Kubiak, W. W., & Jakubowska, M. (2016). Voltammetric classification of ciders with PLS-DA. *Talanta*, *146*, 231-236.

[51] Gibson, M. (Ed.). (2016). *Pharmaceutical preformulation and formulation: a practical guide from candidate drug selection to commercial dosage form*. CRC Press.

[52] Castle, B. C., & Forbes, R. A. (2013). Impact of quality by design in process development on the analytical control strategy for a small-molecule drug substance. *Journal of Pharmaceutical Innovation*, *8*(4), 247-264.

[53] Engisch, W. E., & Muzzio, F. J. (2015). Feedrate deviations caused by hopper refill of loss-in-weight feeders. *Powder Technology*, *283*, 389-400.

[54] K-Tron International, (2009). Smart Refill Technology in Loss-in-Weight Feeding, *Technical Paper*.

[55] Engisch, W. E., & Muzzio, F. J. (2012). Method for characterization of loss-in-weight feeder equipment. *Powder technology*, *228*, 395-403.

[56] Jenike, A. W. (1976). *Storage and flow of solids. Bulletin No. 123; Vol. 53, No. 26, November 1964* (No. NP-22770). Utah Univ., Salt Lake City (USA).

[57] Prescott, J. K., & Barnum, R. A. (2000). On powder flowability. *Pharmaceutical technology*, *24*(10), 60-85.

[58] Yu, Y. (1997). Theoretical modelling and experimental investigation of the performance of screw feeders.

[59] D.M.-S. Tim Freeman, R. Weinekotter, 2015. Predicting feeder performance from powder flow measurements, *Powder Bulk Solids*

[60] Cleary, P. W. (2007). DEM modelling of particulate flow in a screw feeder Model description. *Progress in Computational Fluid Dynamics, An International Journal*, *7*(2-4), 128-138.

[61] Hu, G., Chen, J., Jian, B., Wan, H., & Liu, L. (2010, June). Modeling and simulation of transportation system of screw conveyors by the Discrete Element Method. In *Mechanic Automation and Control Engineering (MACE), 2010 International Conference on* (pp. 927-930). IEEE.

[62] Owen, P. J., & Cleary, P. W. (2010). Screw conveyor performance: comparison of discrete element modelling with laboratory experiments. *Progress in Computational Fluid Dynamics, An International Journal*, *10*(5-6), 327-333.

[63] Owen, P. J., & Cleary, P. W. (2009). Prediction of screw conveyor performance using the Discrete Element Method (DEM). *Powder Technology*, *193*(3), 274-288.

[64] Stauffer, F., Vanhoorne, V., Pilcer, G., Chavez, P. F., Rome, S., Schubert, M. A., ... & De Beer, T. (2018). Raw material variability of an active pharmaceutical ingredient and its relevance for processability in secondary continuous pharmaceutical manufacturing. *European Journal of Pharmaceutics and Biopharmaceutics*, *127*, 92-103.

[65] Oka, S. S., Escotet-Espinoza, M. S., Singh, R., Scicolone, J. V., Hausner, D. B., Ierapetritou, M., & Muzzio, F. J. (2017). Design of an Integrated Continuous Manufacturing System. *Continuous Manufacturing of Pharmaceuticals*, 405-446.

[66] Dolnicar, S. (2002). A review of unquestioned standards in using cluster analysis for data-driven market segmentation.

[67] Yu, W., Muteki, K., Zhang, L., & Kim, G. (2011). Prediction of bulk powder flow performance using comprehensive particle size and particle shape distributions. *Journal of pharmaceutical sciences*, *100*(1), 284-293.

[68] Dray, S., & Josse, J. (2015). Principal component analysis with missing values: a comparative survey of methods. *Plant Ecology*, *216*(5), 657-667.

[70] Nomikos, P., & MacGregor, J. F. (1995). Multivariate SPC charts for monitoring batch processes. Technometrics, 37(1), 41-59.