European Directorate for the Quality of Medicines & HealthCare | Direction européenne de la qualité du médicament & soins de santé

Funded by the European Union and the Council of Europe

EUROPEAN UNION

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Implemented by the Council of Europe

General European OMCL Network (GEON)

GENERAL DOCUMENT

PA/PH/OMCL (19) 60 DEF

Benefits of Chemometrics for OMCLs

Full document title and reference	Benefits of Chemometrics for OMCLs PA/PH/OMCL (19) 60 DEF
Document type	Position Paper
Legislative basis	Council Directive 2001/83/EC and 2001/82/EC, as amended
Date of first adoption	June 2019
Date of original entry into force	June 2019
Date of entry into force of revised document	N/A
Previous titles/other references / last valid version	PA/PH/OMCL (19) 60 DEF
Custodian Organisation	The present document was elaborated by the OMCL Network / EDQM of the Council of Europe
Concerned Network	GEON

Benefits of chemometrics for OMCLs

1. Objectives

The aim of this paper is to introduce the basic principles of chemometrics, providing a list of applications suitable for OMCLs and giving examples of projects where the use of chemometrics was considered successful, as well as suggestions for further reading. It also outlines some chemometric methods currently applied for pharmaceutical analysis and provides a fundamental description of the figures and plots which can facilitate the evaluation of the quality of the corresponding models.

This paper may also help to raise OMCLs' awareness of possibilities to improve some aspects of their work with chemometrics. For example: development, optimisation and validation (robustness testing) of analytical methods, distinguishing differences within sets of data, verification of authenticity of suspect samples, assessment of analytical dossiers, implementing a quality by design (QbD) approach for pharmaceutical development, overall approach to implement a chemometric model, evaluation of chemometric results, etc.

2. Introduction

Quality control of pharmaceutical products is a very important topic which can be rather complex in certain situations. There are many analytical techniques (qualitative and quantitative) which produce a lot of information. Moreover, the use of modern analytical techniques has increased the amount and the complexity of experimental data. The situation becomes even more complex when, along with compliance testing, a broader analytical window needs to be applied, for example to identify characteristic analytical features of substances – in particular active pharmaceutical ingredients (APIs), to classify similar products, to detect the source of a substance/product, etc.

The analysis of a large set of analytical data for several samples is complex and is different from evaluating individual variables separately; the univariate approach cannot represent the true characteristics of a group of samples. The analysis of several variables for different samples makes it possible to understand the chemical characteristics associated with different groups. To do this, a multivariate analysis is conducted on a dataset composed of several different variables for all samples. This is the basic principle of the chemometric approach.

The main purpose of multivariate methods would be information extraction, particularly for large data sets with multiple dependent variables. This means that there is limited use to apply multivariate methods when the number of samples is low and/or just one or few parameters are measured. On the other hand, for large sample sets and/or many different measured parameters, multivariate methods will simplify data evaluation and possibly detect "hidden" characteristics like similarity between samples and/or analysis data.

Chemometric methods utilise algorithms that are able to analyse a whole set of data in order to extract the information of interest. Depending on the intended purpose, there are numerous chemometric methods available: exploratory methods (e.g. PCA), clustering methods (e.g. HCA), classification methods (e.g. k-NN, SIMCA, PLS-DA), regression methods (e.g. MLR, PLS), decomposition methods (e.g. MCR-ALS), etc. Chemometric methods are extensively described in the European Pharmacopoeia within the dedicated chapter *5.21. Chemometric Methods applied to analytical Data.*

Chemometrics should be seen as an analytical tool and the strategy selected for chemometric analysis should be based on the available data set. Different types of analytical data can be assessed using chemometrics (spectroscopic techniques (IR, NIR, Raman, XRF, NMR), separation techniques (HPLC, GC, CE), mass spectrometry, diffraction methods (XRD), etc.). Generally, any data that can be arranged in a table are suitable for chemometric analysis. Several dedicated commercially available

(e.g. Unscrumbler, SIMCA-P, R, MATLAB) and free-of-charge (e.g. R-Project, ChemFlow online, Excel Add-In) software packages are available[1]. These software packages should generate reproducible results.

The EDQM has been supporting the introduction of chemometrics in OMCLs by organising dedicated training sessions, applying chemometrics within hands-on training for the detection of falsified medicines, incorporating chemometrics in API Fingerprinting projects and conducting surveys in order to meet the needs of OMCLs on this matter. The majority of the participants from different OMCLs who completed such training acknowledged the possibilities to share knowledge and experience among the Network. Therefore, it was recognised that some additional information on the practical use of chemometrics in OMCLs is needed.

Considering the complexity of the subject, as well as the opportunity to obtain added value from the regular analyses performed by the OMCLs, a dedicated group was established in order to promote the use of chemometrics and to disseminate the knowledge established so far to facilitate its use.

3. Main areas of possible use of chemometrics within OMCLs

There are many possibilities to use chemometrics in OMCLs. Some of the applications which are already performed within the GEON, and which have been proven to be beneficial for the analysis and interpretation of analytical data, are described below. A more detailed discussion of each topic, as well as suggestions for further reading, are included in the Annexes.

3.1. Development, optimisation and validation (robustness verification) of analytical methods

This may be useful whenever an OMCL is developing *in house* methods or when they are involved in the evaluation of monograph methods. The chemometric methods applied for this purpose are called *Design of Experiments* (DoE). DoE allows the factors significantly influencing the performance of the analytical method to be selected, developed and optimised. In the same way, the influence on performance of small changes in these factors can be explored during validation, providing an objective approach for robustness testing.

3.2. Detection of illegal and falsified medical products

Several OMCLs have applied different *pattern recognition* methods to data generated using various techniques; this allows distinction among samples, verification of the authenticity of suspect samples and successful detection of falsified medicines.

Some pattern recognition techniques have also been used to successfully detect inter- and intra-batch composition variability of medicinal products.

3.3. Identification of chemical substances in medical products

Chemical imaging is a state-of-the-art technique that perform numerous measurements at the surface of a sample. The use of multivariate analysis on resulting spectra give access to the identification of chemical substances in a solid dosage form.

It was then possible to identify excipients of authorized drug products or to detect active ingredients in illegal tablets with an unknown composition.

(1) <https://www.camo.com/unscrambler/>; https://umetrics.com/>; https://www.mathworks.com/products/matlab.html>; <https://www.r-project.org/>; <https://galaxyproject.org/use/chemflow/>, respectively.

3.4. API Fingerprinting

API Fingerprinting MSS projects (MSS FP) represent a Network-wide strategy to detect illegal API sources. The aim of such projects is to develop a methodology that enables identification of characteristic chemical features of the API, thus collecting analytical data and providing a list of discriminating analytical techniques which could help to reveal possible falsification of these substances. With each MSS FP, lessons were learned that improved the design of subsequent projects, which now include well-organised sampling and testing processes which facilitate the chemometric analyses to follow. Reports from finalised MSS FP projects are available on the Extranet.

3.5. Herbal fingerprinting and analysis of plant-based food supplements

Herbal fingerprinting, especially based on chromatography, is becoming increasingly accepted for the identification and quality control of herbal components. These fingerprints are analytical profiles representing the complete composition of the samples. Chemometrics is used to extract information from these profiles about origin, quality or even active components, as well as to compare profiles of different herbal samples.

For the identification/detection of toxic or regulated plants in (suspicious) plant-based food supplements, chromatographic profiles can be used in conjunction with chemometrics.

3.6. Non-analytical data

Multivariate data analysis algorithms can be applied as part of a sampling plan for market surveillance study in order to limit the number of sample to be tested within a group of products. The recognition of common identifiers (e.g. formulation, form, strength, manufacturing site, etc.) allows to detect and discard similar samples. This could be of great importance for large market surveillance studies.

4. Chemometric methods

Generally, chemometrics is defined as “a chemical discipline that uses mathematics, statistics and formal logic to (a) design or select optimal experimental procedures; (b) to provide maximum relevant chemical information by analysing chemical data; and (c) to obtain knowledge about chemical systems.”

This definition is composed of three parts. The first (a) can, for example, be used to develop a new chromatographic method or to optimise a chromatographic separation with a minimal number of experiments. Indeed, chemometrics can be applied to select optimal parameters such as the pH range, buffer concentration, percentage of organic modifier or the column. This is generally known as experimental design (*Design of Experiments*, DoE).

The second and third parts, (b) and (c), are concerned with the extraction of information from chemical data. A simple example would be the decision on whether a sample is illegal or genuine, based on its infrared spectrum. The extraction of the information needed to come to this decision is the subject of part (b) of the definition. In part (c) the question would be asked: why is this sample illegal? Here the parts of the infrared spectrum relevant to the final decision and what they correspond to can be explored, e.g. a specific infrared region can correspond to the presence of a certain excipient, impurity or adulterant.

Chemometric methods use algorithms based on vector and matrix calculations. That is why manifest variables (e.g. spectra or chromatograms) are first transposed from graphical representations to a data table or matrix. Thus a data matrix contains samples arranged in rows and variables (e.g. the different wavelengths in a spectrum) sorted in columns.

4.1. General definitions

Some terminology used in the following discussion is defined below:

- *Manifest variables* represent what is measured or the observed data, on the basis of which a decision should be made. Examples include the absorbance values at different wavelengths across an infrared spectrum, the intensities at the different mass-to-charge ratios in a mass spectrum and the concentrations of impurities, as well as chemical or physical properties.

- A *Response variable* is what you want to know. It can be categorical or continuous. Examples of categorical variables are illegal vs genuine medicines and different API manufacturers. Examples of continuous variables are the dosage of API and the concentration of a certain impurity in samples.

- *Latent variables* are newly defined variables calculated from the manifest variables in the context of variable reduction.

- *Data space* is a multidimensional environment where all objects are plotted against their manifest variables.

- *Unsupervised techniques* are chemometric methods that use only manifest variables for data extraction. Information about manifest variables is not known. These techniques are often, but not exclusively, used for data exploration. A possible outcome here could be the presence of two distinct groups or clusters in an, at first sight, homogeneous group of samples.

- *Supervised techniques* are chemometric methods that relate manifest variables to a response variable. When the response variable is categorical, they are called classification or discrimination techniques. When the response is continuous, they are regression techniques.

4.2. Questions to be asked before starting

As for all method developments in analytical chemistry, some questions have to be answered before starting the analysis. The first question is about the goal of the chemometric analysis, followed by the work or measurements that will be performed and what information is needed.

In short:

(1) What do we want to know and what do we want to do?

- Develop and/or optimise a method? Plan experimental work in a rational way?
- Search for undefined differences? i.e. searching for differences within a group of samples, not knowing if there are any.
- Search for defined differences? i.e. searching for differences between groups of samples, e.g. APIs from different companies.
- Is the purpose to quantify something? e.g. the concentration of a certain impurity.
- Based on the chemometric analysis, are there new samples to be classified or quantified?

(2) Which manifest variables will be used or which measurements will be performed?

Manifest variables may be provided by spectroscopy, chromatography or other analytical techniques, and may be studied as such using analytical profiles or as data resulting from calculations.

(3) Is information on the manifest variables necessary?

Is information needed on which manifest variables are the most or least important in order to come to a decision (e.g. differentiation into groups, classification of a sample or quantification of an impurity).

Based on the answers to these questions, different approaches and chemometric methods will be used. This means the workflow of the analysis will be highly influenced by the purpose of the analysis, the type of manifest variables and the information needed from the analysis. The choice of chemometric techniques used is also strongly influenced by the knowledge and preferences of the analyst.

In the following section the general workflow used in the context of chemometrics is introduced. An analysis performed according to this workflow is applicable to the majority of the problems to be dealt with in an OMCL.

4.3. Workflow

Given the large number of methods available and the rapid scientific evolution in this field, only basic, frequently used chemometric methods are included in this document. They are gathered in a summary figure along with the preliminary questions to be asked before starting (Figure 1). A detailed explanation of the methodology and the methods is available in the respective annexes (Annex 1: Data pretreatment; Annex 2: Experimental design; Annex 3: Unsupervised methods; Annex 4: Supervised methods; Annex 5: Validation of models; Annex 6: Interpretation of the outcomes of chemometrics).

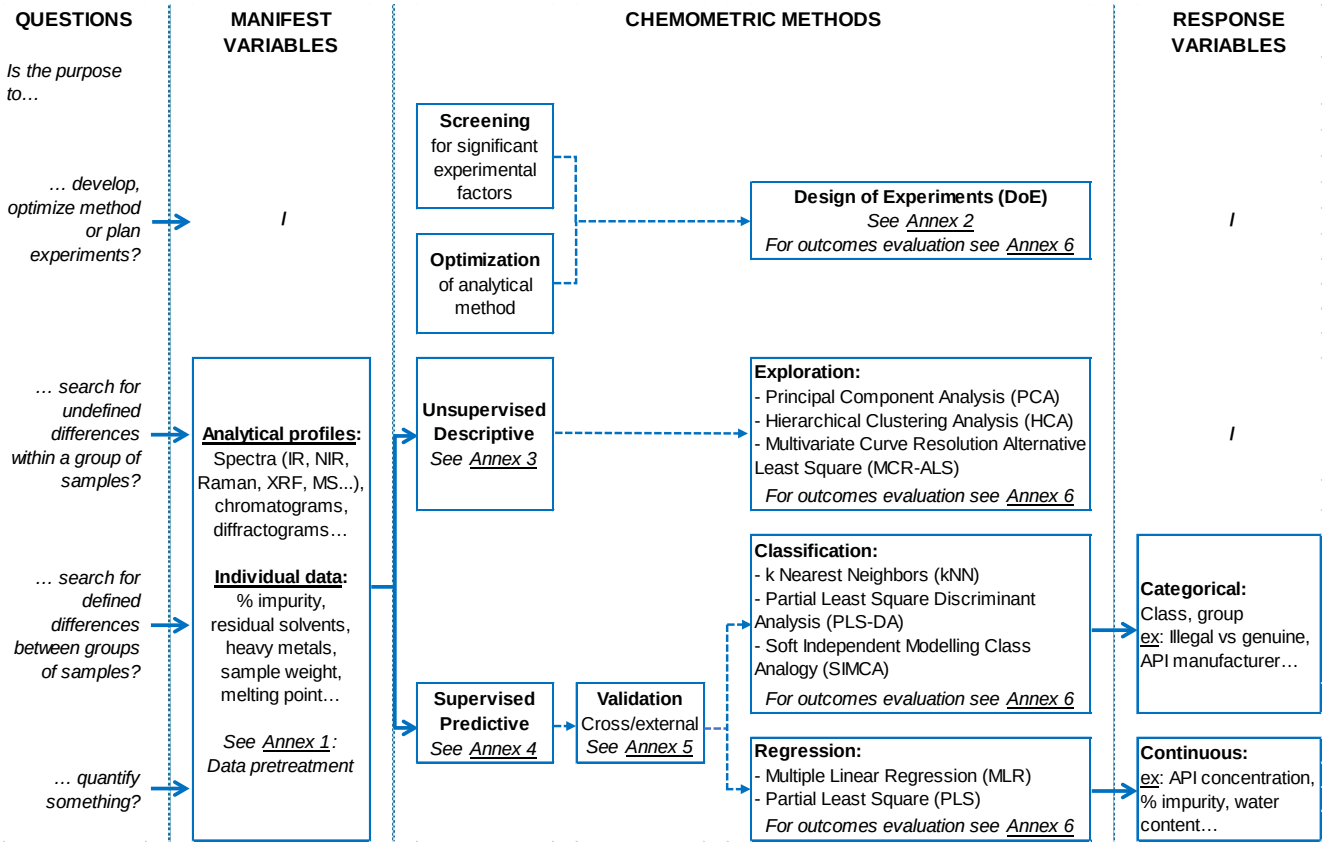


Figure 1: Workflow for chemometric analysis

5. Publishing of results obtained using chemometrics

Data dissemination is an important way to improve, increase and share scientific knowledge. The publication of survey results could also be an important step forward.

OMCLs are encouraged to discuss with their peer OMCLs the possibility of publishing all significant findings obtained using chemometrics, including those from combining or pooling data from multiple OMCL activities (via Annual Reports, Scientific Journals (including peer–reviewed journals) and Scientific Meetings, Internet sites). The diffusion of results could be important for inspections and the evaluation of dossiers.

Data can be published as prescribed in the document “Publicising the work of OMCLs and the GEON, Position Paper of the AdG GEON”.

6. Conclusions

- Chemometrics is a powerful analytical tool that can be used for the development and optimisation of analytical methods, fingerprinting, characterisation, contaminant profiling, falsified drug detection, development of prediction models, tracing the origin of suspect samples, etc.
- Chemometrics can be applied within different aspects of regular OMCL work, such as validation of analytical methods, investigation of reports of falsified products, use of Ph. Eur. general chapters, QbD approach within CTD.
- The GEON supports and promotes the use of chemometrics for API fingerprinting studies, authenticity testing, falsification detection and better sampling planning.
- OMCLs are encouraged to publish data where appropriate and suitable, and discuss with other OMCLs the possibility of publishing combined data.
- OMCLs are encouraged to share their ‘success stories’ with the Network and possibly beyond by sending abstracts of the work done and references for further reading to the Communication Group at the EDQM.

ANNEX 1: Data pretreatment

Before starting a chemometric analysis the data often has to be pretreated or transformed. Pretreatment of data has two purposes. The first is to make different manifest variables comparable, so that it is easier to interpret and compare the importance of the different manifest variables in solving the presented problem. The second is the elimination of noise in the data. This often occurs when using experimental data like spectra. Spectra are influenced by external parameters, such as temperature and fluctuations in the source. The inclusion of the noise data in the analysis can result in bad, unmeaningful or even faulty results and conclusions.

Different pretreatment techniques can be applied depending on the purpose and the type of data.

To make manifest variables comparable scaling techniques are used. Usually, scaling between 0 and 1 and mean centring are applied.

(1) To scale between 0 and 1, each column of the data table – or manifest variable – is scaled separately. From each value in column (x) the minimum value of the column is subtracted after which each value in the new column (x_1) is divided by the maximum value of this new column (x_1).

(2) In mean centring, each column of the data table – or manifest variable – is also scaled separately. Here the mean value of the column/manifest variable to be treated is subtracted from each individual value.

These techniques are often applied in projection methods such as principal component analysis (PCA) or partial least squares (PLS) regression in order to be able to compare the contribution of each variable to the obtained result. They are also applied in classical linear regression. In this case, after scaling the regression coefficients become a measure for the importance of a manifest variable in the regression. Without scaling or pretreatment the regression coefficients would be influenced by the difference in the range of values of the different manifest variables.

To eliminate noise in the data, especially when using spectral data, the most commonly used techniques are: standard normal variate (SNV) and the second derivative.

(1) SNV is applied to each spectrum separately and for each of the spectra the average and standard deviation of all the data points are calculated. The average value is subtracted from the absorbance for every data point and the result is divided by the standard deviation. In this way a part of the noise or scatter is removed from the spectra.

(2) The second derivative is also applied to each spectrum separately. This not only allows compensation for some noise but will also enlarge the visibility of differences between the spectra in the data set. This pre-treatment also has the advantage of correcting potential baseline drift.

A whole series of other more specific pretreatment techniques exist; these include warping when using chromatograms **to compensate for time shifts between chromatograms** due to column aging, temperature fluctuations or differences in mobile phase composition.

Data fusion techniques are applied **when different data blocks have to be combined** (e.g. chromatographic and spectroscopic data). Data fusion is particularly useful for fingerprinting of pharmaceutical samples. For example, this methodology improves the discrimination between APIs originating from different manufacturers.

ANNEX 2: Experimental design

One of the important applications of chemometrics is Design of Experiments (DoE) which is used for systematic and planned investigations and for solving the experimental problems that arise during the optimisation of analytical methods. The use of DoE enables the maximum utilisation of data from systematic and planned experiments. The mathematical criteria incorporated in the factorial design enable screening for significant experimental factors, and response surface methodology (RSM) is applied to optimise the experimental values of previously identified significant experimental factors.

- **DoE for screening and optimisation**

DoE involves the use of mathematical, statistical and graphical methods to design experiments, aiming for maximal usage of the information from experimental data. A major role of experimental design is screening. Most analytical processes are influenced by a wide variety of factors. Which factors are most significant is not always obvious. Initially, it is important to understand which factors are significant and then narrow down the final optimisation to three or four significant factors.

It is reasonable to assume that the outcome of an experiment is dependent on the experimental conditions. This means that the result can be described as a function based on the experimental variables (Equation 1):

$$y = f(x) \qquad \text{(Eq. 1)}$$

The function *f*(*x*) is approximated by a polynomial function (Equation 2) and represents the relationship between the experimental variables (marked as x) and the responses (marked as y) within a limited experimental domain.

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_{N-1}x_{N-1} + b_Nx_N \qquad \text{(Eq. 2)}$$

The terms marked as *b* (b_1, b_2, b_3, b_{N-1}, b_N) are coefficients that represent the estimated effect of the factors considered and b_0 is the average experimental response. The *simplest polynomial model* contains only linear terms and describes only the linear relationship between the experimental variables and the responses.

The next level of polynomial models contains additional terms (marked as b_{12}, b_{13}, b_{23}) that describe the interaction between different experimental variables. Thus, a *second order interaction model* is explained with the following polynomial (Equation 3):

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_{N-1}x_{N-1} + b_Nx_N + \\ + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + ... + b_{(N-1)N}x_{N-1}x_N \qquad \text{(Eq. 3)}$$

The two models above are mainly used to investigate the experimental system, i.e., screening. For this purpose full factorial, fractional factorial and star design are often used (Figure 2). The number of experiments depends on the type of full factorial design selected, represented by the equation:

$$N = 2^k + n$$

where k is the number of parameters studied and n is the number of central points included (n=3). It is obvious that the limit for the number of experiments it is possible to perform will easily be exceeded when the number of variables increases. In most investigations it is reasonable to assume that the influence of the interactions of the third order or higher are very small or negligible and can then be excluded from the polynomial model. Thus, experiments can be reduced by grouping the interaction terms into fractions. The number of experiments for fractional factorial design is defined with equation:

$$N = 2^{k-p}$$

where k is the number of variables and p the size of the fraction.

A star design is easy to carry out: all factors, except one, are set at level 0 (in coded units) and the levels of the studied factor are set at −1 and +1 (in coded units). It is easy to add a new factor, which is important at the end of the robustness study when the operator may want to verify whether a new factor is robust or not. Another advantage of star design is the introduction of a central point, an integral part of this type of design.

In most cases, it is not necessary to investigate the interactions between all of the included variables from the beginning. In the first screening it is recommended to evaluate the result and estimate the main effects according to a linear model. After this evaluation the variables that have the largest influence on the result are selected for new studies. Thus, a large number of experimental variables can be investigated without having to increase the number of experiments to the extreme.

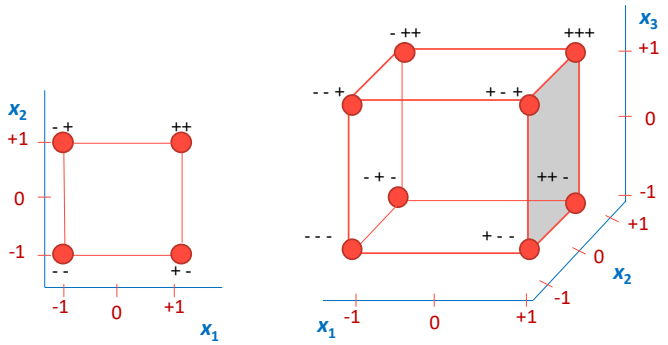


Figure 2a: Full factorial design of the experimental domain for two (left) and three variables (right). The levels of the factors are given by minus 1 for low level and plus 1 for the high level. A zero-level is also included, a centre, in which the mid-value of all variables is set.

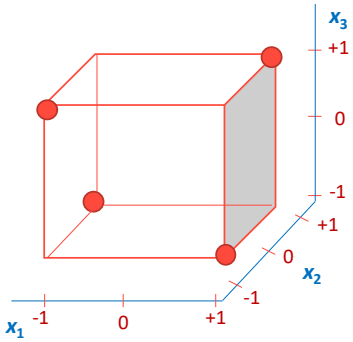


Figure 2b: Distribution of experiments for a 2^{3-1} fractional factorial design.

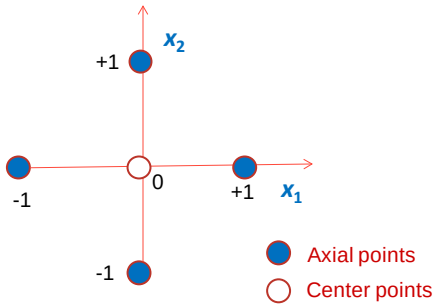


Figure 2c: Two-factor star design. The first factor has three levels (1, 0 and −1 in coded units). The second factor has only two new levels (−1 and +1) because the third level (0) is the same as factor 1. The other factors are similar to factor 2 and can be as numerous as the operator wishes.

Response surface designs

To be able to determine optimum (maximum or minimum) values for the variables, quadratic terms have to be introduced in the model. By introducing these terms, it is possible to determine non-linear relationships between the experimental variables and responses. The polynomial function below (Equation 4) describes a *quadratic model*:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_Nx_N + b_{12}x_1x_2 + b_{23}x_2x_3 + ... \\ ... + b_{(N-1)N}x_{N-1}x_N + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2 + ... + b_{NN}x_N^2 \quad \text{(Eq. 4)}$$

Response surface designs are useful for modelling a curved quadratic surface to continuous factors. A response surface model can pinpoint a minimum or maximum response, if one exists, inside the factor region. Three distinct values for each factor are necessary to fit a quadratic function, so the standard two-level designs cannot fit curved surfaces. Among the response surface DoE designs often mentioned are the Doehlert design, Box-Behnken design and the central composite design (Figure 3).

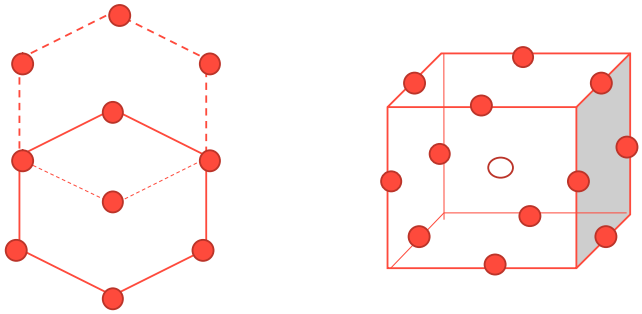


Figure 3a: Doehlert design with two variables (left) and Box-Behnken design (right)

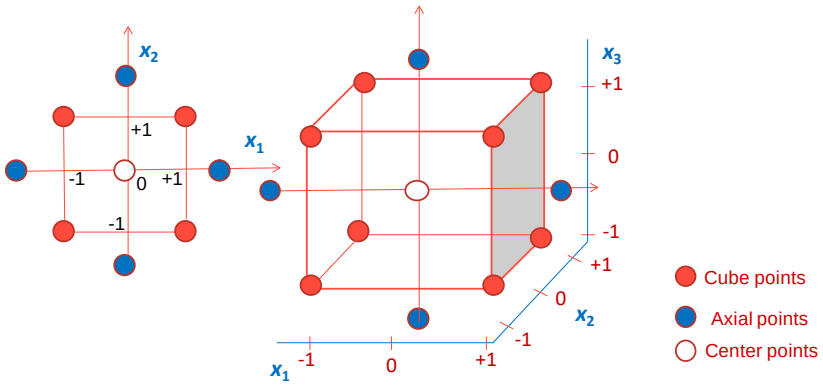


Figure 3b: Two- (left) and three-factor (right) Central Composite design

The central composite design combines a two-level fractional factorial and a star design. So, as well as the 2^k cube points that come from a *full factorial* design, there are also:

• Centre points, for which all the factor values are at the zero (or midrange) value and

• Axial (or star) points created by a *Screening Analysis*, for which all but one factor are set at zero (midrange) and that one factor is set at outer (axial) values.

The number of experiments (N) for central composite design is expressed by equation:

$$N = 2^k + 2k + n$$

where *k* is the number of parameters studied and *n* is the number of central points included (n=3). Here, the three repetitions at the central points are required to determine the experimental error variance and test the predictive validity of the model.

- **Response Surface Methodology**

So far, the experiments described have focused on: (a) identifying a few important variables from a large set of candidate variables, i.e., a screening experiment and (b) ascertaining how a small number of variables impact the response. In order to answer the question: “What specific levels of the important variables produce an optimum response?”, the goal is to find the optimum for a response (y) and to understand how the response changes in a given direction by adjusting the design variables.

Response surface methodology (RSM) is applied for optimisation of the experimental values of previously identified significant experimental factors. The graphical presentation of RSM results makes it easy to see how to reach the desired optimal range of the response (y) by controlling the experimental factors (x), knowing the intensity and direction of their effect on the response, as well the interaction among the factors. For interpretation of the RSM graphical representation refer to Annex 6.

ANNEX 3: Unsupervised methods

- **Principal Component Analysis (PCA)**

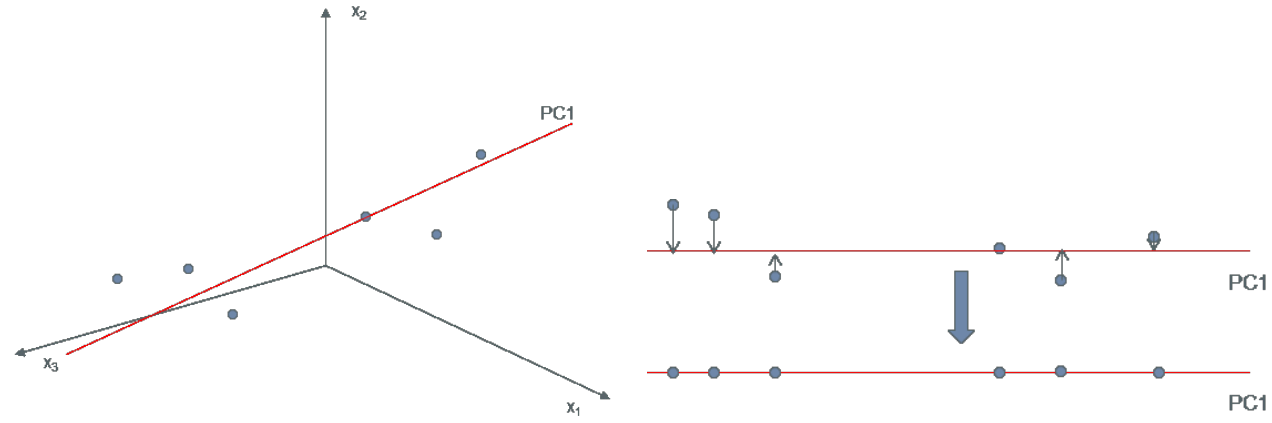
PCA is an unsupervised projection technique based on variable reduction through the definition of latent variables. In short, in PCA new variables (latent variables) are defined as linear combinations of the manifest variables according to the formula:

$$PC1 = a + b_1x_1 + b_2x_2 + \ldots + b_ix_i + \ldots + b_nx_n$$

in which PC1 is the first principal component, a is the residue, b_i the regression coefficient for the i^{th} manifest variable, x_i the i^{th} manifest variable and n the total number of manifest variables.

This new variable PC1 is defined in the direction of the highest variation in the data, e.g. in a data matrix composed of three manifest variables, PC1 will draw a new axis in the direction of the highest variation in the data, as shown in Figure 4a.

The samples can then be projected onto the new axis (Figure 4b), leading to a representation of the data in one dimension. The projected value of a sample on PC1 is called the score of the sample on PC1. The importance of the manifest variables in defining PC1 are called the loadings on PC1.



(a) (b)

Figure 4: (a) definition of PC1; (b) projection of samples on PC1

The use of PCA in a situation with only three manifest variables does not make sense. The power of PCA is the ability to represent multidimensional data (a high number of manifest variables) in two or three dimensions. As stated above PC1 is defined in the direction of the highest variation in the data. The second PC is defined in the direction of the highest remaining variation around PC1 and therefore is orthogonal to PC1 by definition. The third PC is further defined in the direction of the highest remaining variations around the plane PC1/PC2 and therefore as orthogonal to this plane. In theory as many PCs can be defined as there are samples in the data set, Although often the first three principal components are used to represent multidimensional data in a two or three dimensional plot, where PC1, PC2 and PC3 form the new axes. When the scores of the different samples are plotted, the result is called a score plot, while a loading plot shows the importance/loadings of the different manifest variables.

PCA is especially valuable for detecting trends in the data and for distinguishing different sample groups, e.g. APIs from different sources. Not only will PCA allow the detection of differences between groups of samples, but it can also define these differences based on the loading plots.

- **Hierarchical Cluster Analysis (HCA)**

HCA is the other unsupervised technique often used in OMCLs. HCA is a method used to evaluate the distance between samples and to group them in a tree-shaped chart called a dendrogram (Figure 5). In the data space, samples are considered as points and the distance between points will define the degree of (dis)similarity between samples. Therefore, HCA calculates a similarity measure between each pair of samples. Classical examples are the Euclidean distance and the correlation coefficient. In the former, HCA projects all samples in the multidimensional dataspace and calculates the Euclidean distance between each pair of samples, building the dendrogram according to these distances. When using the correlation coefficient, the correlation is calculated between the values of all manifest variables for each pair of samples. The dendrogram is created based on the obtained correlation values.

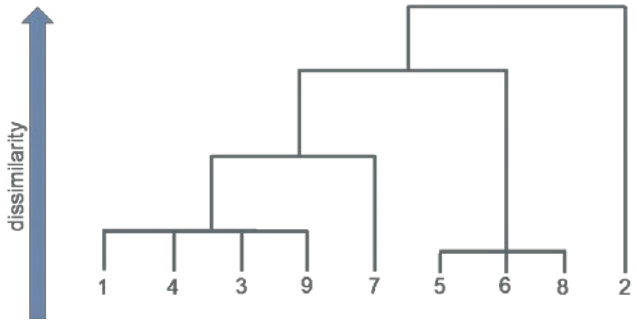


Figure 5: simplified example of a HCA dendrogram

In the dendrogram, samples are grouped according to their similarity, i.e. according to the Euclidean distance (the shorter the more similar) or the correlation (the higher the more similar). The length of the branches is a measure of the similarity between groups. For example, in Figure 5 samples 5, 6 and 8 are very similar since they are connected by the shortest branches. Samples 1, 4, 3 and 9 are also similar but this group is characterised by a higher heterogeneity than the group containing 5, 6 and 8. Sample 7 shows some similarity with group 1, 4, 3 and 9 and less with group 5, 6 and 8, while sample 2 is significantly different from all other samples.

HCA can be performed in a divisive or in an agglomerative way. With the divisive method the algorithm starts with all samples and splits them according to their similarity. With the agglomerative method individual samples are connected according to the similarity measure used. A wide variety of HCA algorithms exist, each using different ways to split/connect the samples and different similarity measures, but the general principle is always the same.

- **Multivariate Curve Resolution Alternating Least Squares (MCR-ALS).**

Hyperspectral images usually consist of hundreds of spectra gathered in a data cube, i.e. a three-dimensional matrix with two spatial dimensions (x, y) and one spectral dimension (λ). The data can be examined in a univariate way, as for conventional spectroscopy or, alternatively, chemometric methods may be used. MCR-ALS is a very popular method for the resolution of a spectral mixture without a priori knowledge of the chemical system. The assumption of a decomposition method is that the sample spectrum may be considered as the weighted sum of the spectra of pure chemical species. In a

chemical image, each pixel of the image contains the same pure spectra but their contribution is different from one pixel to the next. The method decomposes the spectral dataset X into the product of the matrix of pure spectra of species S^T and the matrix of their relative contribution C in the hyperspectral image according to the following formula (E is the residual error):

$$X = C.S^T + E$$

Studying the pure spectra delivered by the MCR-ALS outcome, it is possible to identify chemical compounds in a sample of unknown composition (use of spectral libraries), and to access the distribution map of substances in the sample (Figure 6).

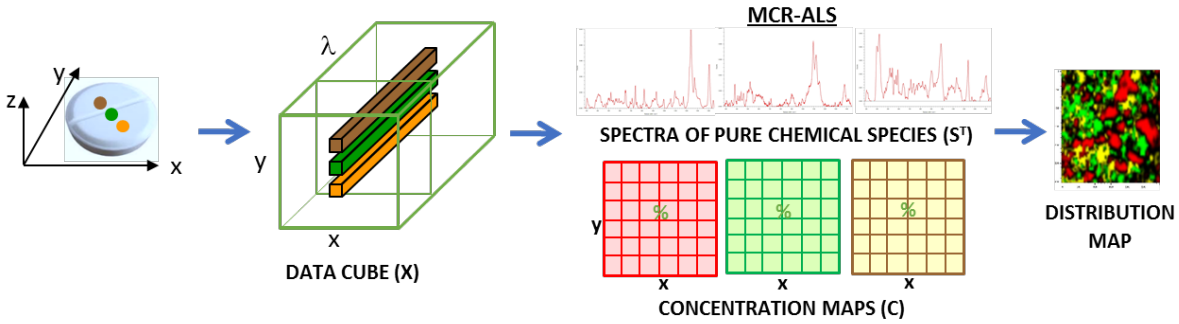


Figure 6: Basic principle of a decomposition method

ANNEX 4: Supervised methods

- **Classification/discrimination techniques**

- k-Nearest Neighbours (k-NN)

k-NN is probably the simplest chemometric classification technique. The algorithm is based on the class memberships of neighbouring objects. In short the samples of a sample set are projected into the multidimensional data space defined by their manifest variables. A new sample or a sample from an external test set is also projected and then the k-nearest neighbours are taken into account. Imagine a two class problem where k equals 3 (i.e. 3 neighbouring samples). From the three (k) closest samples, based on Euclidean distance or correlation similar to HCA, two belong to class 1 and one to class 2. In this case the new sample will be classified as belonging to class 1. In this algorithm k is the only value to be optimised. The choice of k can be done using cross-validation (see Annex 5). The method, being based on distance criteria and not on latent variables, does not give a graphical outcome.

k-NN is generally more efficient for binary classification problems, and k is often between 3 and 9. When k is uneven in a binary problem, samples will be classified. Where k is even, samples can also be unclassified. When a multiple-class problem is dealt with, unclassified samples can also occur with an uneven value of k.

- Soft Independent Modelling of Class Analogy (SIMCA)

SIMCA is a disjoint classification algorithm. In SIMCA each class of samples is modelled separately using PCA. A PCA analysis is performed on each class, represented by the samples in the training set belonging to this class. Based on this analysis, boundaries around the class are defined based on two spatial distances (Euclidean and Mahalanobis distances). The selection of the number of PCs to be taken into account to define the boundaries for each class is based on a cross-validation procedure (see Annex 5).

In the case of a classification problem with three classes, SIMCA will perform three PCA analyses (one for each class), and calculate limited spaces around the samples belonging to each class. When a new sample has to be classified, it is projected on the PC spaces of each class (Figure 7).

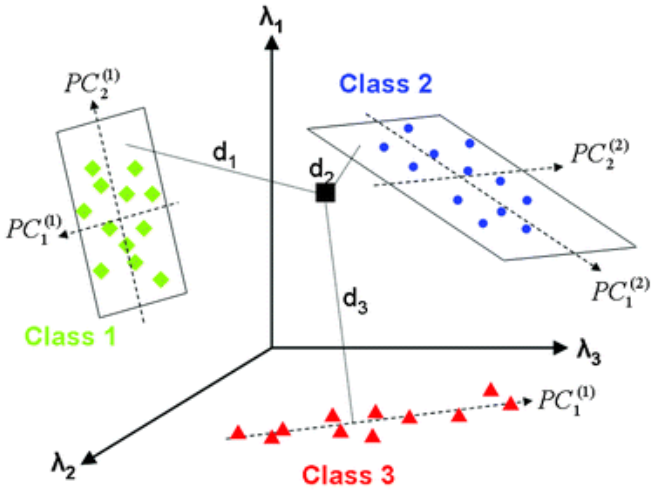


Figure 7: Graphical representation of a SIMCA analysis

When the sample falls within the limits around a certain class, the sample will be classified in this class. Since SIMCA is a disjoint modelling technique (each class is modelled separately), a new sample can be classified in more than one class or in none. This is also called a weak classifier.

SIMCA is one of the most popular classification techniques and often used when using spectroscopic data such as NIR or Raman spectroscopy. In this case SIMCA would allow identification of a sample based on the spectrum measured. This can be used in the identification of falsified samples, but also in the characterisation of e.g. party drugs.

- **Partial Least Squares-Discriminant Analysis (PLS-DA)**

PLS is by far the most commonly applied chemometric technique in a variety of domains. PLS is based on the same principles as PCA, i.e. on the definition of latent variables which are linear combinations of the manifest variables. The difference lies in the way these latent variables are defined. In PCA the latent variables are only based on the variation in the data since a response variable is not available. In PLS the latent variables are defined based on the co-variance of the data (the manifest variables) and the response variable(s), i.e. the model is built by trying to find a compromise between the best description of manifest variables and the best prediction of the response variable. Thus the first latent variable, called PLS factor 1 (PLS1), is defined in the direction of the highest co-variance with the response, the second in the direction of the highest remaining co-variance around PLS1, etc.

As in PCA, the latent variables are defined according to the following formula:

$$PLS1 = a + b_1x_1 + b_2x_2 + \ldots + b_ix_i + \ldots + b_nx_n$$

in which PLS1 is the first PLS factor, a is the residue, b_i the regression coefficient for the i^{th} manifest variable, x_i the i^{th} manifest variable and n the total number of manifest variables.

The PLS factors define new axes on which the samples are projected. As in PCA, the projection of the samples in the space defined by the PLS factors are called the scores, while the importance of each manifest variable in the definition of the respective PLS factors is called the loading. Visualisation of the dataspace becomes possible in two or three dimensions using a score and/or loading plot.

Once the latent variables are defined, they can be used to build models. When the response is categorical, PLS is combined with discriminant analysis (DA) to build a classification model. In short, all the samples of the training set are projected in the new space defined by the PLS factors. DA will now define some boundaries differentiating the different classes of the training set. A new sample will be projected into the PLS space and depending on its location relative to the boundaries, the sample will be assigned to a class.

PLS-DA can be used to solve the same problems as SIMCA. The advantage of PLS-DA is that it is a real classifier, assigning all samples to one exclusive class. Other advantages are that PLS-DA allows a visual interpretation of the data, taking into account the response variable, and that the model can be interpreted by evaluating the loadings. The latter allows identification of the most important variables, differentiation between the classes and assignation of new samples to a specific class.

- **Regression methods**

Regression techniques are used when the response variable is continuous. Although in their simplest form they are used daily in OMCL work to calculate calibration lines, regression methods are not widely used in the context of OMCL work.

The most basic regression technique is **Multiple Linear Regression (MLR)**, which is calculated according to the formula:

$$Y = a + b_1x_1 + b_2x_2 + \ldots + b_ix_i + \ldots + b_nx_n$$

in which Y is the response variable (e.g. concentration), a is the residue, b_i the regression coefficient for the i^{th} manifest variable, x_i the i^{th} manifest variable and n the total number of manifest variables. When the different manifest variables are scaled before analysis, the regression coefficient is an indication of the importance of the different manifest variables to model Y.

MLR is most often performed when data with a limited number of manifest variables is used. In case of a high number of manifest variables, e.g. when using spectroscopic data, MLR will generally result in bad modelling due to the fact that it is highly influenced by the presence of correlated manifest variables and noise in the data. To resolve this, two approaches can be followed: the first is variable selection and the second variable reduction.

In variable selection, algorithms are used to select only those manifest variables that contribute significantly to the modelling of the response variables. Different algorithms are available: forward selection in which the manifest variables are added to the model in a stepwise way, based on their contribution to the model, until no further improvement of the model can be obtained; backward selection in which all variables are included in the model and then in a stepwise way the variables contributing the least to the model are withdrawn. This continues until no further improvement of the model can be obtained; the third is stepwise selection in which forward and backward selection are used iteratively. These are the most basic algorithms, but a whole series of variable selection techniques exist, ranging from very simple to complex, e.g. genetic algorithms.

Variable reduction is based on the previously described techniques PCA and PLS and similar techniques. In short a PCA or PLS analysis is performed and the latent variables are calculated. The scores of the samples on the PCs or the PLS factors are then used in MLR instead of the manifest variables themselves. The selection of the number of PCs or PLS factors to be included is based on a cross-validation procedure (see Annex 5). When PCs are used this combination of PCA and MLR is called Principal Component Regression (PCR), when PLS factors are used it is simply called PLS.

ANNEX 5: Validation of models

When applied to chemometrics, "validation" does not refer to the regulatory validation supported by ICH-Q2, but it is instead considered as a procedure implemented in order to build, optimise and assess the prediction ability of a model.

The procedure for the elaboration of a chemometric model involves the use of many samples which response variable is perfectly known (e.g. identity, water content, API content…). A calibration set is dedicated to the building/optimisation of the model whilst a validation set is devoted to the evaluation of the model performance. When the number of sample is low a validation set may not be available, and then the calibration sample set must be split in 2 sub-sets (training set and test set).

A reliable chemometric model is usually elaborated in 3 steps:

- Building of the initial model with the training/calibration set,
- Optimisation of the model with the training/calibration set using cross-validation (also known as internal validation),
- Performance assessment of the final model with the test/validation set (also known as external validation).

- Test/validation set selection

To be able to correctly validate a model a representative set of samples not used during modelling, i.e. an external test set, has to be selected. In some cases, an external set of samples is available, but often the test set has to be selected from the calibration sample set. This selection can be randomly performed by the analyst (50-80% for the training set and 20-50% for the test set), although to avoid bias and to ensure that the test set is representative of the entire sample set, algorithms could be applied. A whole series of this kind of algorithms exist, ranging from very simple to very complex. Two regularly used algorithms which are efficient for resolving most of the problems facing an OMCL are Kennard-Stone and Duplex.

(1) Kennard-Stone projects the samples into the multidimensional data space defined by all manifest variables in the data set. The algorithm starts with the selection of the sample closest to the mean point in the data space for the test set, and continues with the sample farthest from the first sample, then the sample farthest from the first two samples…, until the predefined number of samples for the external test set is reached. In general, a test set is composed of 20 to 50% of the total sample set, depending on the sample set size. Alternatively Kennard-Stone can also be applied starting with the sample farthest from the mean point of the data space.

(2) Duplex also projects the samples into the data space and makes a pairwise selection of samples, starting with the selection of the two samples farthest away from each other for a first set. The second pair farthest away from each other is selected for a second set, while the third pair is again selected for the first set. This goes on until the predefined number of samples is reached in the second set, which will be the test set. The first set together with the remaining samples will be used for modelling.

- Cross-validation

A whole series of cross-validation algorithms exist, ranging from very simple to very complex, but they are all based on a resampling principle and only take the training set into account.

The simplest algorithm is leave-one-out cross-validation (LOOCV). Here one sample is taken out of the calibration set and a model is built with the remaining samples. The obtained model is then used to predict the selected sample. This procedure is repeated, each time with another sample. At the end a

prediction error is obtained for each sample of the calibration set. The prediction error for the whole calibration set is then calculated as correct classification rate (*ccr*, % samples correctly classified) for classification models or as the root mean squared error of cross-validation (RMSECV) for regression models. When cross-validation is used for selection of the optimal model, the model with the lowest RMSECV or the highest ccr is chosen. LOOCV has the disadvantage that it can lead to an over-fitted model, meaning that the model very closely describes the calibration set and is no longer able to predict/classify new samples correctly. This is often the case when LOOCV is used for the selection of the optimal model.

A more suitable alternative is n-fold cross-validation in which the calibration set is randomly divided into n parts. In each step one part is left out and the n-1 remaining parts are used for modelling. The obtained model is used to predict the sample in the part that was left out. This procedure is repeated n times, each time with another of the n-parts left out. At the end a prediction error is obtained for each sample of the calibration set and the *ccr* or RMSECV can be calculated. The most commonly used n-fold cross-validation is 10-fold cross-validation, where the training set is divided into 10 parts.

An example of a more complex, very efficient cross-validation algorithm is Monte Carlo cross-validation. Here you have a random selection of the cross-validation test sets, test sets can vary in size and samples can be selected more than once or not at all.

- **External validation**

External validation is the validation of the model using an external test set. This is the only correct way to validate a model, since in cross-validation the error is based on several slightly different models.

Once the model is obtained, it is used to classify or predict the samples of the external test set. For classification models the error is given as *ccr* for the test set and for regression models as the root mean squared error of prediction (RMSEP). It is up to the operator to decide the minimum *ccr* or maximum RMSEP acceptable for the intended use of the model.

ANNEX 6: Interpretation of the outcomes of chemometrics

A valuable feature of chemometric analysis is the production of graphical representations and numerical outcomes that allow evaluation and interpretation of analytical data. Depending on the chemometric methods implemented, several kinds of outcome are possible. Some of these are described in the following sections.

- **Response Surface Methodology (RSM)**

The response surface applied for optimisation of the experimental values of previously identified significant experimental factors can be visualised graphically as presented in Figure 8.

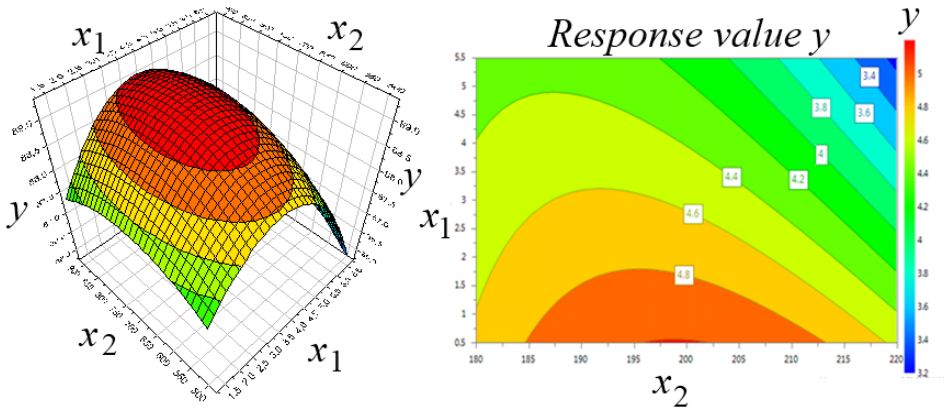


Figure 8: Response surface plot (left) and contour plot (right)

Graphs are useful to see the shape of a response surface: hills, valleys, and ridge lines. Hence, the function *f (x1, x2)* can be plotted against the levels of *x1* and *x2*. Three-dimensional graphs show the response surface from the side and are called *response surface plots*. Sometimes it is less complicated to view the response surface in two-dimensional graphs (*contour plots*) that show contour lines of *x1* and *x2* pairs which have the same response value *y.* In the case of the development of a chromatographic method, the diagram on the surface of the response of the chromatographic system explains the behaviour of the values of the chromatographic parameters/response when changing the experimental values of the studied factors. This diagram can be applied in order to predict the values of the chromatographic parameters at values of experimental factors beyond the test range.

- **Principal Component Analysis**

PCA is an exploratory method that highlights similarities and differences between analytical data. It is applied to reduce the dimensionality of the dataset while preserving as much as possible the observed variation. Original data are transformed into a new uncorrelated set of variables (principal components or PCs). Four plots are available for the interpretation of data from PCA.

The score plot is usually a 2- or 3-dimensional scatter plot where samples are projected in the PC space (Figure 9). First PCs model the relevant information in the chemical system whereas further PCs model the residual noise. A colour or a particular shape may be attributed to each sample, and it is then possible to observe the sample distribution, clusters of samples or atypical samples. The Hotelling T^2

ellipse is a graphical representation inside which 95% (for a risk of 5%) of the samples are expected to be located.

The loadings show the importance of the manifest variables to the different PCs. The loading plot is usually a 2- or 3-dimensional scatter plot where the loadings are projected in the PC space. When score and loading plots are combined, a bi-plot is obtained, showing the clustering of samples and the variables responsible or important for this clustering. Since loading and bi-plots may be difficult to interpret, loadings may be plotted as a curve for each principal component (Figure 10) showing the importance of the variables to a particular PC. Variables with high loadings (red boxes in Figure 10) are responsible for the greatest differences between samples. In certain cases (no use of derivative pre-treatment) loadings of individual PCs can be associated with analytical profiles (e.g. spectra). Thus the study of loadings may assist with the identification of variables linked to specific chemical information (e.g. vibration mode of H_2O for NIR spectroscopy). The study of loadings may also help in the selection of the number of PCs carrying relevant information, since high-level loadings exhibit random signals.

These 2 plots need to be examined in conjunction for a better understanding of which variable is responsible for the grouping of samples and which variable most contributes to the variability of the dataset.

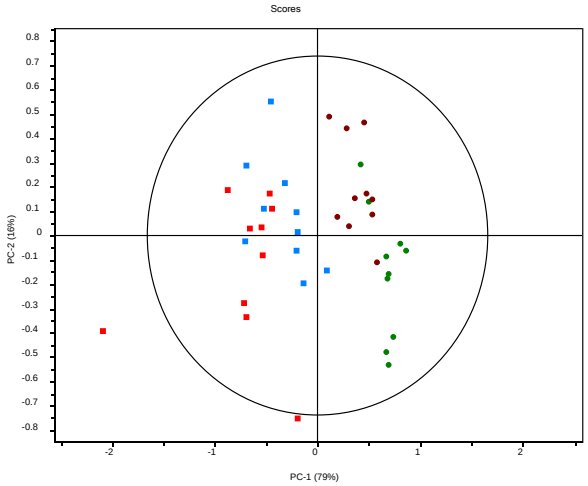


Figure 9: PCA score plot

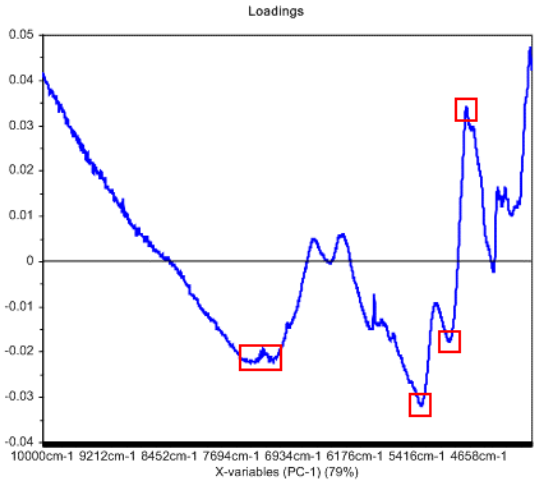


Figure 10: Loadings of the variables on PC1

The explained variance represents the ratio of the total variance of the dataset supported by the individual PCs. These numerical outcomes are visible in the graph of cumulative variances (Figure 11) and also in the axis labels of the score and loading plots. The graph of cumulative variances highlights the number of PCs needed to reliably model the relevant information in the dataset. In Figure 11, 95% of the variation in the data is described with the two first principal components.

An outlier is atypical data that may come, for example, from measurement issues or differences in the chemical or physical composition of the sample. These data have to be inspected in order to decide if they must be kept or discarded from the dataset. Outliers can be detected on the score plot and highlighted with 2 numerical outcomes: residual distance Q (orthogonal distance of the data to the model) and Hotelling distance T^2 (distance of the orthogonal projection of the data to the centre of the model). The influence plot compiles the 2 distances (Figure 12). Data with a high T^2 value (leverage on x-axis) act as a strong outlier since they have a leverage effect on the model. Removing such data strongly modifies the PCA model. Data with a high Q value (on y-axis) act as a low outlier with a moderate impact on the model.

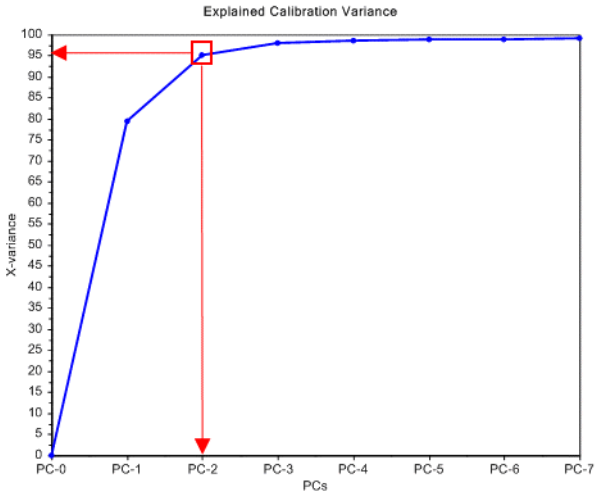


Figure 11: Plot of cumulative explained variance

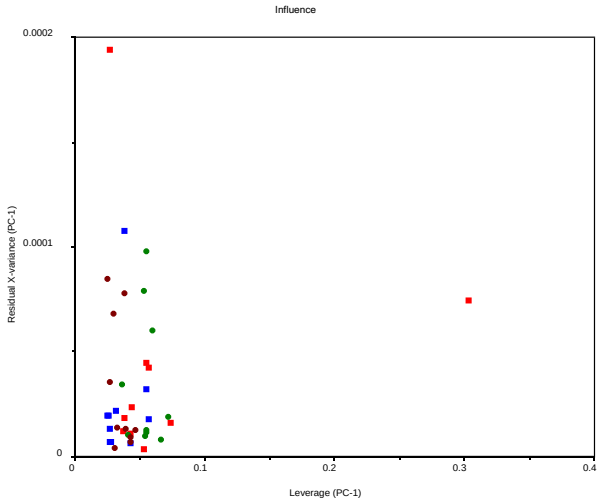


Figure 12: Influence plot of residual (Q) vs leverage (T^2)

- **Hierarchical Cluster Analysis (HCA)**

HCA is an exploratory/unsupervised method used to evaluate the distance between samples and to group them in a tree-shaped chart called a dendrogram (Figure 13). In the data space, samples are considered as points, and the distance between points will define the degree of (dis)similarity between samples. Geometric measurements are usually calculated with the Euclidean distance (using the coordinates of samples in the data space) or the Mahalanobis distance (including the correlation between variables). Different strategies and linkages (agglomerative, divisive, single, complete, average) may be tested in order to optimise the classification of samples.

In the dendrogram, samples are sorted and connected to each other with lines. The length of lines represents the distance between samples. The longer the line the more dissimilar the samples are. Clusters of samples may be defined using an arbitrary critical distance. For example, in Figure 13 if the critical distance is 5 then 3 clusters of samples are proposed, while if the critical distance decreases to 3 then 6 clusters of samples are suggested.

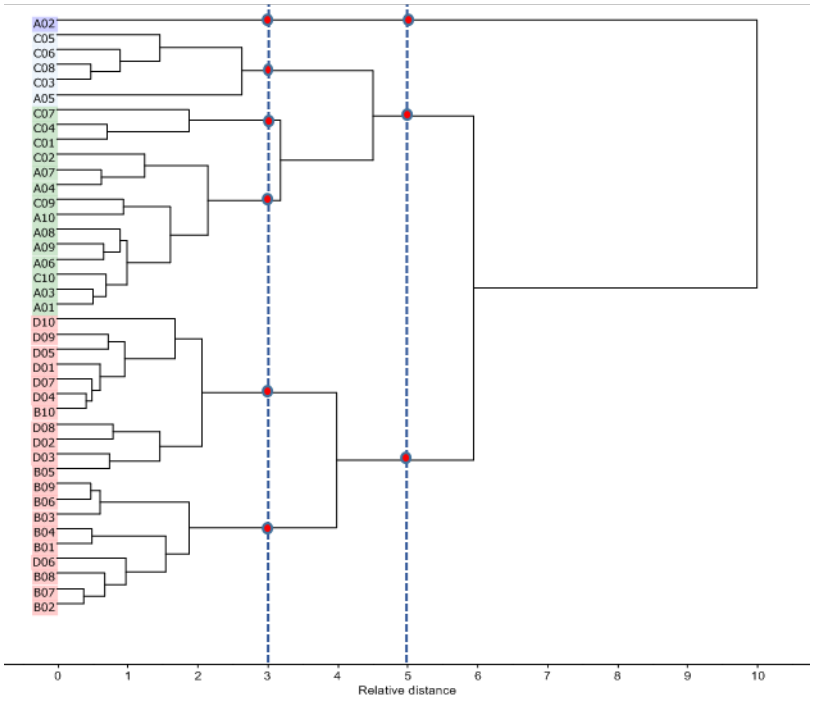


Figure 13: Dendrogram of 40 samples obtained with HCA

- **Soft Independent Modelling of Class Analogy (SIMCA)**

SIMCA is a data classification/discrimination method that requires the prior establishment and validation of disjoint PCA models for each membership class. Each PCA model is adjusted separately determining the optimal number of PCs. Score/loading plots are studied for each PCA. A cross-validation is performed to determine the acceptable *ccr*. New data is then projected in each of these PCA models, and its distance to the model is calculated. This data will be assigned to a class if its distance to the model (usually calculated with Q and T^2) is lower than the critical distance (based on the overall variation of each class and usually representing 95% confidence that particular data belongs to a class). With the SIMCA method, a new sample can belong simultaneously to several classes or to none.

The main outcome of SIMCA is the classification table that allocates samples to each class. Another plot using the sample distances to the model shows residual and leverage distances for a particular class. Figure 14 shows 4 samples predicted to be part of class A (residual and leverage distances lower than the critical distance in red) while 5 others are excluded.

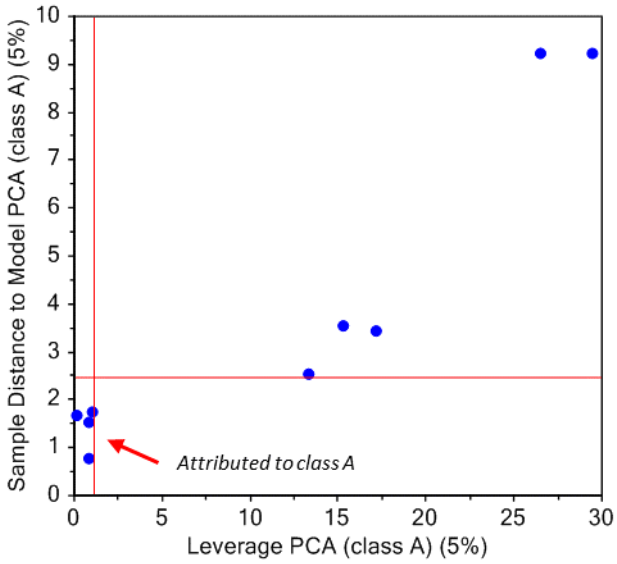


Figure 14: Distances of new samples to the model of class A

- **Partial Least Squares regression (PLS)**

PLS is a regression method used to find relationships between two blocks of data (see Annex 4). It is often implemented for the quantitative determination of the API content in a dosage form. A model is built taking into account the correlation between manifest variables (e.g. analytical data, spectra) and a continuous response variable (e.g. concentrations in mg/tablet).

The optimisation step of the model includes, for example, the selection of relevant variables, the appropriate pre-treatment of analytical data, the choice of the number of PLS factors and the detection of outlier samples. During this phase the operator aims to minimise the error of prediction (difference between predicted and reference values). A figure of merit is calculated testing known samples on the model built with the calibration set. The optimal number of PLS factors may be determined during the modelling calculating *Root Mean Square Error of Cross-Validation* (RMSECV) with an increasing number of factors entering the model. In figure 15, RMSECV decreases when the number of PLS factors increases, and 6 PLS factors are enough to properly describe the studied properties of the samples. RMSECV is expressed in the same unit than the concentration of samples, and then the performance of the model may be compared to the reference method. RMSECV between 1% and 2% of the reference value is considered acceptable.

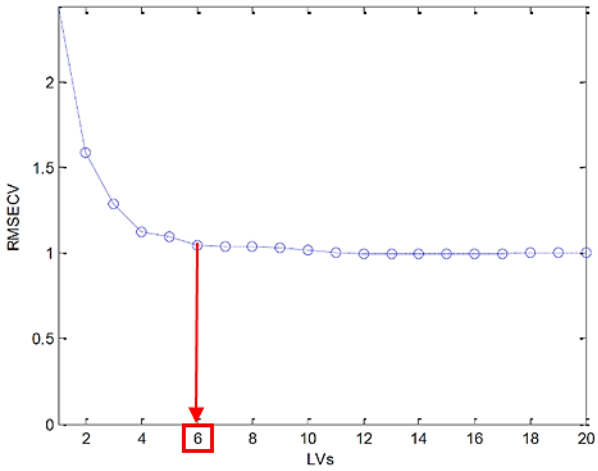


Figure 15: Standard error of cross-validation

The external validation step is implemented to demonstrate the compliance of the model with the intended purpose and evaluate its predictive ability. This phase is performed by using the test sample set and calculating the *root mean square error of prediction* (RMSEP). The model is considered reliable when the RMSEP is found to be close to RMSECV.

As PLS is a projection method, score and loadings plots are studied in the same way as for PCA.

The PLS model is assessed with the calibration plot that connects predicted and reference values (Figure 16). Performance criteria to be studied are: slope of the curve close to 1, intercept (*offset*) close to 0, RMSEC/RMSEP as low as possible and determination coefficient (R^2) close to 1 (RMSEC represents the deviation of the calibration samples from the model.)

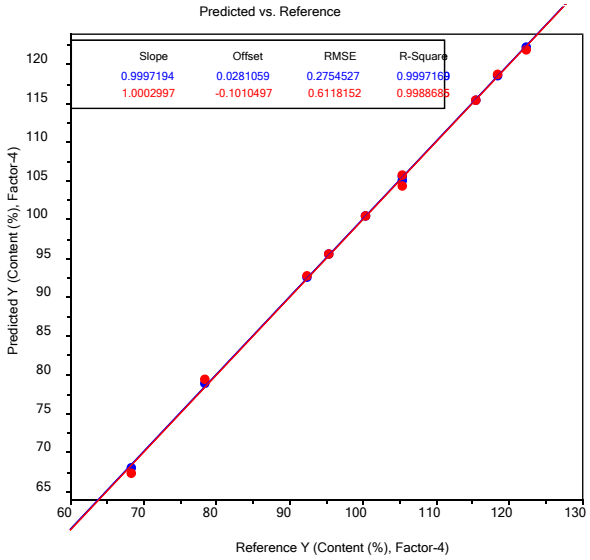


Figure 16: PLS calibration plot (calibration data in blue and prediction data in red)

ANNEX 7: Applications and scientific articles

The list provided here is not exhaustive.

- **Design of experiments**

GC/MS-HPLC/DAD

J. Acevska, G. Stefkov, R. Petkovska, S. Kulevanova, A Dimitrovska – *Chemometric approach for development, optimization, and validation of different chromatographic methods for separation of opium alkaloids,* Analytical and Bioanalytical Chemistry 403 (2012) 1117-1129.

- **API fingerprinting**

HPLC/MS-PCA-HCA

J. Acevska, G. Stefkov, I. Cvetkovikj, R. Petkovska, S. Kulevanova, J. Cho, A. Dimitrovska, *Fingerprinting of morphine using chromatographic purity profiling and multivariate data analysis*, Journal of Pharmaceutical and Biomedical Analysis 109 (2015) 18-27.

- **Detection of illegal and falsified products**

HPLC-SVM-SIMCA

E. Deconinck, P.Y. Sacré, P. Courselle, J.O. De Beer, *Chemometrics and chromatographic fingerprints to discriminate and classify counterfeit medicines containing PDE-5 inhibitors*, Talanta 100 (2012) 123-133.

IRTF-kNN

E. Deconinck, J.L. Bothy, B. Desmedt, P. Courselle, J.O. De Beer, *Detection of whitening agents in illegal cosmetics using attenuated total reflectance-infrared spectroscopy*, Journal of Pharmaceutical and Biomedical Analysis 98 (2014) 178-185.

NIR-SIMCA

I. Storme-Paris, H. Rebiere, M. Matoga, C. Civade, P.A. Bonnet, M.H. Tissier, P. Chaminade – *Challenging near infrared spectroscopy discriminating ability for counterfeit pharmaceuticals detection* - Analytica Chimica Acta 658 (2010) 163-174.

XRF-SIMCA

H. Rebiere, A. Kermaïdic, C. Ghyselinck, C. Brenier – *Inorganic analysis of falsified medical products using X-ray fluorescence spectroscopy and chemometrics*, Talanta 195 (2019) 490-496.

- **Inter-batch variability study**

NIR-PCA

H. Rebiere, C. Ghyselinck, L. Lempereur, C. Brenier – *Investigation of the composition of anabolic tablets using near infrared spectroscopy and Raman chemical imaging*, Drug Testing and Analysis 8 (2016) 370-377.

- **Chemical imaging**

Raman-MCRALS

H. Rebiere, M. Martin, C. Ghyselinck, P.A. Bonnet, C. Brenier – *Raman chemical imaging for spectroscopic screening and direct quantification of falsified drugs*, Journal of Pharmaceutical and Biomedical Analysis 148 (2018) 316-323.

- **Plant analysis**

E. Deconinck, M. Vanhamme, J.L. Bothy, P. Courselle – *A strategy based on fingerprinting and chemometrics for the detection of regulated plants in plant food supplements from the Belgian market: Two case studies*, Journal of Pharmaceutical and Biomedical analysis 166 (2019) 189-196.

E. Deconinck, C.A. Sokeng Djiogo, P. Courselle – *Chemometrics and chromatographic fingerprints to classify plant food supplements according to the content of regulated plants*, Journal of Pharmaceutical and Biomedical Analysis 143 (2017) 48-55.