Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

CHIMICA

Ciclo XXXI

Settore Concorsuale di afferenza: 03/A1

Settore Scientifico disciplinare: CHIM/01

CHEMOMETRICS APPLIED TO DIRECT MULTIVARIATE ANALYSIS

Presentata da: **Alessandro Zappi**

Coordinatore Dottorato

Prof. Aldo Roda

Relatore

Prof. Lucia Maini

Co-Relatore

Prof. Dora Melucci

Esame finale anno 2019

INDEX

ABSTRACT

The present Ph.D. Thesis is focused on applications and developments of chemometrics. After a short introduction about chemometrics (Chapter 1), the present work is divided in three Chapters, reflecting the research activities addressed during the three-year PhD work:

- Chapter 2 concerns the application of classification tools to food traceability (Chapter 2.1), plant metabolomics (Chapter 2.2), and food-frauds detection (Chapter 2.3) problems.
- Chapter 3 concerns the application of design of experiments for a bio-remediation research (Chapter 3.1) and for machine optimization (Chapter 3.2).
- Chapter 4 concerns the development of the net analyte signal (NAS) procedure and its application to several analytical problems. The main aim of this research is to face the matrix-effect problem using a multivariate approach.

Chemometrics is the science that extracts useful information from chemical data. The development of instruments and computers is bringing to analytical methodologies ever more sophisticated, and the consequence is that huge amounts of data are collected. In parallel with this rapid evolution, it is, therefore, important to develop chemometric methods able to handle and process the data. Moreover, the attention is also focusing on analytical techniques that do not destroy the analyzed samples. Chemometrics and its application to non-destructive analytical methods are the main topics of this research project.

Several analytical techniques have been used during this project: gas-chromatography (GC), bioluminescence, atomic absorption spectroscopy (AAS), liquid chromatography (HPLC), near-infrared spectroscopy, UV-Vis spectroscopy, Raman spectroscopy, X-ray powder diffraction (XRPD), attenuated total reflectance (ATR) spectroscopy.

Moreover, this research activity was carried out in collaboration with several external research groups and companies:

- University of Bologna: Department of Statistical Sciences (Prof. Giuliano Galimberti); Department of Biological, Geological and Environmental Sciences (Prof. Annalisa Tassoni); Department of Pharmacy and Biotechnology (Prof. Stefano Girotti)
- Cornell University (Ithaca, NY): Plant Biology and Horticulture Sections (Prof. Peter J. Davies)
- COOP Italia, cooperative society (Casalecchio di Reno, Bologna, Italy): quality control laboratory (Mr. Fernando Gottardi and Dr. Sonia Scaramagli)
- Industria Macchine Automatiche, IMA S.p.A. (Ozzano dell'Emilia, Bologna, Italy): IMA-Active division for solid-dose pharmaceutical formulations (Dr. Caterina Funaro, R&D Manager)
- Council for Agricultural Research and Economics (Bologna, Italy): Dr. Francesca Corvucci and Dr. Gian Luigi Marcazzan
- IZSLER, Zooprophylattic Experimental Institute of Lombardy and Emilia Romagna "Bruno Ubertini" (Brescia, Italy): Dr. Giorgio Fedrizzi

CHAPTER 1: INTRODUCTION TO CHEMOMETRICS

A general definition of chemometrics was given by Svante Wold [1]: "How to get chemically relevant information out of measured chemical data, how to represent and display this information, and how to get such information into data". In other words, the aim of chemometrics is to extract useful information from chemical data, which sometimes could mean obtaining a "single" number, or a single explicative graph, from a huge amount of data. Its purpose is to use mathematical, statistical, and numerical analyses applied to chemical problems in order to obtain mathematical models able to explain and, possibly, to "solve" it.

Svante Wold and Bruce Kowalski, the pioneers of this discipline, on June 10th, 1970 founded the informal Chemometrics Society, whose "primary function is communication" [2]. The aim of that Society was to give a landmark to chemist researchers that were starting to use statistical methods for their analytical problems. The new Society could help researchers in publishing their results in journals that, perhaps, were still not ready to accept the new point of view of chemometrics, even if most of the mathematical tools were already common in other areas such as econometrics and psychometrics. Anyway, the most important advice given by these pioneers was to use the mathematical tools, but maintaining always the point of view of the chemist. Since then, chemometrics has grown and evolved, and the contemporary evolution of computers has been a strong help. At now, there are several national Societies (for example in Great Britain, Spain, Sweden, Russia, Belgium, etc.) and there are also several Groups (for example, the Italian Group of Chemometrics, which is a section of the Italian Society of Chemistry). There are also two dedicated journals, *Journal of Chemometrics* (John Wiley & Sons, NJ) and *Chemometrics and Intelligent Laboratory Systems* (Elsevier, Netherland). Several companies have been developing software and tools for chemometrics (e.g. CAMO, Norway; R Core Team, Austria; Minitab Inc., USA; MathWorks, USA).

The use of chemometrics has involved many other scientific fields: besides analytical chemistry, from which it started and in which it is normally used, for instance, in spectroscopy [3] and chromatography [4], it is commonly used in metabolomics [5, 6], QSAR [7], bioengineering [8], polymer science [9], environmental chemistry [10], industrial chemistry [11], and so on. Up to now, a research on Scopus with the only keyword "chemometrics" gives more than 11000 results: this number did not reach 400 in 1990 (15 years after the first paper by Kowalski [2]), and reached more or less 2000 in 2000, with a peak of 954 papers published in 2017 only. Figure 1.1 shows the number-per-year published papers with "chemometric" as keyword, up to August 2018.

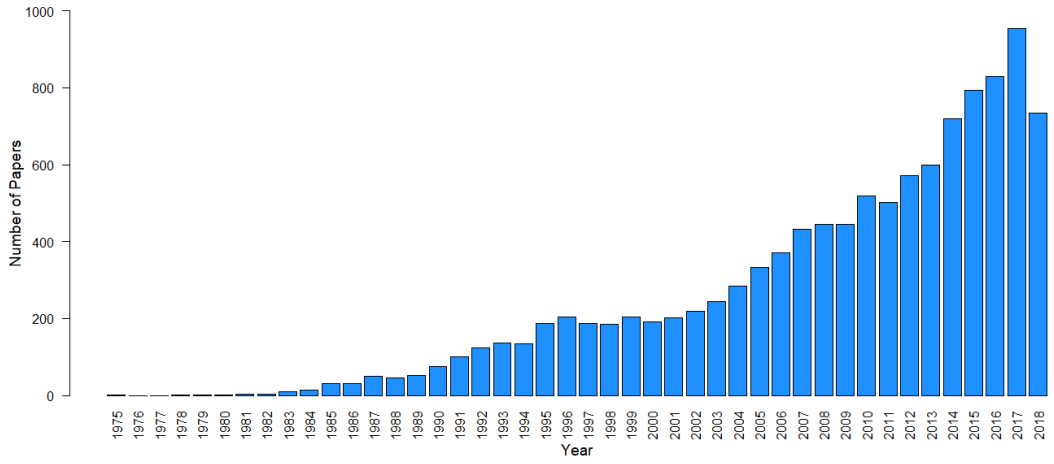


Figure 1.1: Number of papers published by year, concerning chemometrics (source: Scopus)

Wold and Kowalski were initially mainly interested in pattern recognition [2], which is still one of the main topics in chemometrics, but new frontiers have been opened.

For sure, the most known and used chemometric technique is principal component analysis (PCA) [12]. Looking for "PCA" on Scopus gives almost 88000 results (~200 of which dating back to before 1970). The role of PCA goes far beyond chemistry: it easily finds applications also in very different sciences, as social science [13] and psychology [14, 15]. PCA is the most typical example of explorative analysis. PCA does not give any quantitative or classification result. PCA's only goal is to visualize data, by simple 2D- or 3D-plots, starting from data having any number of dimensions (variables). By PCA plots, it is easy to visualize similarities and dissimilarities between samples and variables. The present Thesis will not discuss PCA in detail; however, some basic concepts could be useful to better understand what will be shown in the following chapters. PCA starts from a dataset matrix of dimensions n x v, where n is the number of samples, and v the number of variables. It computes a linear combination of the original variables in order to convert them into new variables, the principal components (PCs), which have the characteristics of being orthogonal to each other and ordered to have, in each PC, the maximum possible quantity of information, or explained variance (EV). In practice, PCA rotates the original space (spanned by the original variables) to the PC-space. The coordinates of samples in this new space are the *scores*, while the coordinates of variables are the *loadings*. Therefore, 2D- or 3D- scores and loadings plots often carry most of the EV, and are useful to examine several properties:

- Similarities between samples and variables: samples or variables that are close to each other in scores or loadings plot are considered similar
- Outliers: samples with scores far from all the others
- Relevant or not relevant variables: loadings far or close to the origin
- Role of variables in describing samples: scores and loadings that are in the same quadrant of the respective plot

The most important advantage of PCA is that, in general, PC1 vs PC2 plot is sufficient to describe the entire dataset, sacrificing only some percent of information. On the contrary, when proceeding in univariate mode starting from a dataset of *v* variables, all plots of each pair of variables should be studied to have a complete vision of the dataset. Figure 1.2 shows an example of scores and loadings plot.

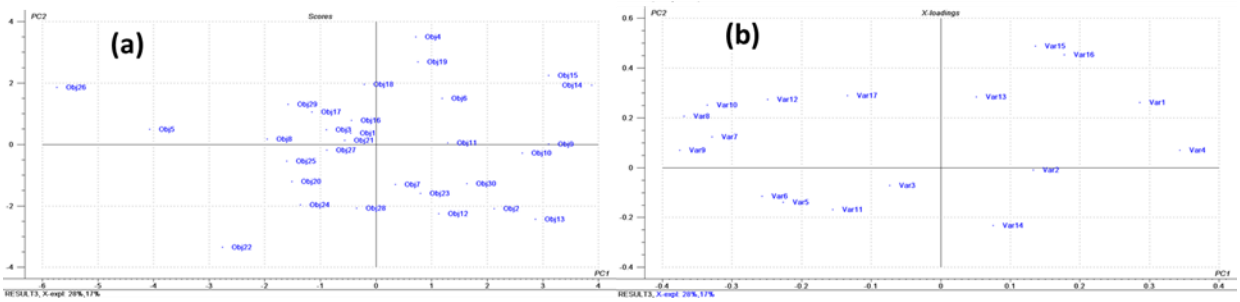


Figure 1.2: example of a) scores plot, and b) loadings plot

Hence, PCA scores and loadings may be used for further analyses. Chemometrics, in fact, has been developed in other areas, such as pattern recognition (or classification) [16], regression [17] and noise reduction [18].

Pattern recognition means to recognize an object as pertaining to a specific class by some chemical properties (for example, to classify an oil as Italian or not, based on its volatile fraction analysis [19]). For such analyses, a dataset containing samples from all classes under study is needed. "Classes" may be any characteristic joining groups of samples, such as geographical origin [19], commercial category [20], quality level [21], presence or absence of a disease [22] or of defects [23]. The mathematical model computed with such datasets has the first aim of discriminate the classes and recognize the training samples in the proper class. Then, new samples may be projected onto the model and assigned to a specific class. Several techniques are used with this aim, the most common of which is probably linear discriminant analysis (LDA) [19, 22]. However also artificial neural network (ANN) [24] received great interest, and other algorithms have been developed [25–27].

Also quantitative analysis is extremely important, and regression methods are largely used by chemometricians [28]. The classical regression method is ordinary least squares (OLS), that is the one used to compute common univariate calibration lines, which can somehow be extended to the multivariate field (multivariate linear regression, MLR). Among the others, OLS is very useful to perform computations for the design of experiments (DoE), a technique that will be discussed and applied in Chapter 3. Besides OLS, the two most important chemometric tools for regression are principal component regression (PCR) [29] and partial least squares (PLS) [17]. PLS will be discussed in detail Chapter 4.

Another important issue is noise reduction, which means removing, or at least reducing, the analytical noise present in every chemical analysis. It is a very complex issue, and there are dozens of chemometric tools aimed at noise reduction. These can be roughly divided into two categories: data pre-processing [30] (some of these techniques will be used further on in this Thesis) and variable selection [31].

Although all these analyses seem to have different goals, it is very easy to mix them to reach the desired result. For example, a noise reduction technique is often used to improve the results of a subsequent classification or regression analysis; the same PLS can be sometimes used for variables selection, and a technique combining PLS regression and classification has been developed: PLS-DA [32]. Moreover, all these methods can start from PCA scores and loadings, instead of original variables. For instance, PCR performs regression starting from PCA scores. Therefore, one can think chemometrics as a series of statistical and mathematical methods applied to chemical data, which can be independent of each other, but also inter-related. Sometimes, the solution of a specific problem could derive from a chemometric analysis completely different from the one decided at the beginning, or from several analyses used one after the other, starting from results of the previous one. Therefore, it often does not exist a specific chemometric way for solving a specific problem, and the obtained results can be probably always improved using other tools.

The present Ph.D. Thesis is divided into two parts. The first one (Chapters 2 and 3) shows the application of chemometric tools, in particular, LDA, PLS-DA, and design of experiments (DoE), to specific problems. Issues spanning from pure chemistry to biology, and also to an industrial process are discussed. For the industrial process, also a chemometric analysis for the on-line quality control will be shown. The second part (Chapter 4) will concern the development of net analyte signal (NAS) algorithm, in order to face up matrix effect and to develop a way for using standard addition method (SAM) in a multivariate field. In this second part, applications of the NASSAM algorithm will be shown for several different analytical techniques, again dealing with specific problems.

Besides chemometrics, the guiding principle of these two parts is, in general, direct analysis. Direct analysis means obtaining the desired result through chemical analyses that do not destroy the analyzed samples, possibly allowing some further chemical analyses. Although such principle has been not always followed (for example, for some DoE results a destructive analysis was necessary), most of the results obtained, and all of the NAS ones, derived from direct analyses.

References

1. Wold S, Sjöström M (1998) Chemometrics, present and future success. In: Chemometrics and Intelligent Laboratory Systems. pp 3–14
2. Kowalski BR (1975) Chemometrics: Views and Propositions. J Chem Inf Comput Sci 15:201–203 . doi: 10.1021/ci60004a002
3. Pasquini C (2003) Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. J Braz Chem Soc 14:198–219 . doi: 10.1590/S0103-50532003000200006
4. Stoll DR, Li X, Wang X, et al (2007) Fast, comprehensive two-dimensional liquid chromatography. J. Chromatogr. A 1168:3–43
5. Hollywood K, Brison DR, Goodacre R (2006) Metabolomics: Current technologies and future trends. Proteomics 6:4716–4723
6. Van Der Greef J, Smilde AK (2005) Symbiosis of chemometrics and metabolomics: Past, present, and future. J. Chemom. 19:376–386
7. Eriksson L, Johansson E (1996) Multivariate design and modeling in QSAR. Chemom. Intell. Lab. Syst. 34:1–19
8. Fukusaki E, Kobayashi A (2005) Plant metabolomics: potential for practical operation. J Biosci Bioeng 100:347–354 . doi: 10.1263/jbb.100.347
9. Nicholls IA, Andersson HS, Charlton C, et al (2009) Theoretical and computational strategies for rational molecularly imprinted polymer design. Biosens. Bioelectron. 25:543–552
10. Mas S, de Juan A, Tauler R, et al (2010) Application of chemometric methods to environmental analysis of organic pollutants: A review. Talanta 80:1052–1067
11. Russell EL, Chiang LH, Braatz RD (2000) Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. Chemom Intell Lab Syst 51:81–93 . doi: 10.1016/S0169-7439(00)00058-7
12. Bro R, Smilde AK (2014) Principal component analysis. Anal Methods 6:2812–2831 . doi: 10.1039/c3ay41907j
13. Vyas S, Kumaranayake L (2006) Constructing socio-economic status indices: How to use principal components analysis. Health Policy Plan 21:459–468 . doi: 10.1093/heapol/czl029
14. Lawrence EJ, Shaw P, Baker D, et al (2004) Measuring empathy: Reliability and validity of the Empathy Quotient. Psychol Med 34:911–919 . doi: 10.1017/S0033291703001624
15. Calder AJ, Burton AM, Miller P, et al (2001) A principal component analysis of facial expressions. Vision Res 41:1179–1208 . doi: 10.1016/S0042-6989(01)00002-5
16. Mutihac L, Mutihac R (2008) Mining in chemometrics. Anal. Chim. Acta 612:1–18
17. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: A basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130 . doi: 10.1016/S0169-7439(01)00155-1
18. Vaseghi S V. (2005) Advanced Digital Signal Processing and Noise Reduction
19. Melucci D, Bendini A, Tesini F, et al (2016) Rapid direct analysis to discriminate geographic origin

of extra virgin olive oils by flash gas chromatography electronic nose and chemometrics. Food Chem 204:263–273 . doi: 10.1016/j.foodchem.2016.02.131

20. Alcázar A, Ballesteros O, Jurado JM, et al (2007) Differentiation of green, white, black, Oolong, and Pu-erh teas according to their free amino acids content. J Agric Food Chem 55:5960–5965 . doi: 10.1021/jf070601a
21. Kowalkowski T, Zbytniewski R, Szpejna J, Buszewski B (2006) Application of chemometrics in river water classification. Water Res 40:744–752 . doi: 10.1016/j.watres.2005.11.042
22. Khanmohammadi M, Ansari MA, Garmarudi AB, et al (2007) Cancer diagnosis by discrimination between normal and malignant human blood samples using attenuated total reflectance-fourier transform infrared spectroscopy. Cancer Invest 25:397–404 . doi: 10.1080/02770900701512555
23. Custers D, Cauwenbergh T, Bothy JL, et al (2015) ATR-FTIR spectroscopy and chemometrics: An interesting tool to discriminate and characterize counterfeit medicines. J Pharm Biomed Anal 112:181–189 . doi: 10.1016/j.jpba.2014.11.007
24. Burns JA, Whitesides GM (1993) Feed-Forward Neural Networks in Chemistry: Mathematical Systems for Classification and Pattern Recognition. Chem. Rev. 93:2583–2601
25. Breiman L (2001) Random forests. Mach Learn 45:5–32 . doi: 10.1023/A:1010933404324
26. Huang G Bin, Zhu QY, Siew CK (2006) Extreme learning machine: Theory and applications. Neurocomputing 70:489–501 . doi: 10.1016/j.neucom.2005.12.126
27. Zhang ML, Zhou ZH (2007) ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognit 40:2038–2048 . doi: 10.1016/j.patcog.2006.12.019
28. Frank IE, Friedman JH (1993) A Statistical of Some Chemometrics View Regression Tools. Technometrics 35:109–135 . doi: 10.2307/1269656
29. Massy WF (1965) Principal Components Regression in Exploratory Statistical Research. J Am Stat Assoc 60:234–256 . doi: 10.1080/01621459.1965.10480787
30. Rinnan A, van den Berg F, Engelsen SB (2009) Review of the most common pre-processing techniques for near-infrared spectra. TRAC-Trends Anal Chem 28:1201–1222 . doi: DOI 10.1016/j.trac.2009.07.007
31. Guyon I, Elisseeff A (2003) An Introduction to Variable and Feature Selection. J Mach Learn Res 3:1157–1182 . doi: 10.1016/j.aca.2011.07.027
32. Ballabio D, Consonni V (2013) Classification tools in chemistry. Part 1: Linear models. PLS-DA. Anal. Methods 5:3790–3798

CHAPTER 2: PATTERN RECOGNITION

As already stated in the previous chapter, pattern recognition (or classification) was one of the main interests of Wold and Kowalski at the beginning of chemometrics.

The following Chapters 2.1, 2.2, and 2.3 will show some applications of pattern recognition.

The first one concerns the application of LDA to honeys analyzed by head-space gas-chromatography, with the aim of discriminating the twelve botanical classes from which samples came. This work was carried out in collaboration with Coop Italia (Casalecchio di Reno, Bologna, Italy) and the Council for Agricultural Research and Economics (CREA, Bologna, Italy). It was published on European Food Research and Technology (reference: 2018, 244(12):2149-2157).

The second one concerns LDA applied to a metabolomic profile of grapevines. The aim of that study was to use metabolites concentration as a fingerprint to discriminate between common plants and grapevines made resistant to illnesses. This work was carried out in collaboration with Prof. Annalisa Tassoni (Department of Biological, Geological, and Environmental Sciences, University of Bologna, Italy) and Prof. Peter J. Davies (Departments of Plant Biology and Horticulture, Cornell University, NY). It was published on Plant Physiology and Biochemistry (reference: 2019, 135:182-193).

The last work concerns again honey, but the aim was to discriminate natural honeys from adulterated ones. LDA was applied to variables whose analysis is mandatory for the Italian law, while PLS-DA was applied for an alternative method of discrimination, based on bio-luminescence of bacteria put in contact with honey samples. Also this work was carried out in collaboration with the Council for Agricultural Research and Economics (CREA, Bologna, Italy). It was published on European Food Research and Technology (reference: 2019, 245:315-324).

CHAPTER 2.1: BOTANICAL TRACEABILITY OF UNIFLORAL HONEYS BY CHEMOMETRICS BASED ON HEAD-SPACE GAS-CHROMATOGRAPHY

The present work was published by the journal “European Food Research and Technology” (reference: 2018, 244(12):2149-2157; https://doi.org/10.1007/s00217-018-3123-3). The experimental work was carried out in the laboratory of Coop Italia (Bologna, Italy), under the supervision of Mr. Fernando Gottardi and Dr. Sonia Scaramagli. Samples were collected both by Coop Italia and by Dr. Gian Luigi Marcazzan, from CREA institute (Bologna, Italy). Dr. Antonia Zelano prepared all samples for analysis. I analyzed the samples, performed all chemometric analyses and, together with my PhD co-Tutor (Prof. Dora Melucci), wrote the manuscript paper.

CHAPTER 2.2: SEASONAL CHANGES IN AMINO ACIDS AND PHENOLIC COMPOUNDS IN FRUITS FROM HYBRID CROSS POPULATIONS OF AMERICAN GRAPES DIFFERING IN DISEASE RESISTANCE

The present work was published by the journal “Plant Physiology and Biochemistry” (reference: 2019, 135:182-193; https://doi.org/10.1016/j.plaphy.2018.11.034). The experimental work was carried out in Cornell University (NY) by Prof. Peter J. Davies, Prof. Annalisa Tassoni, and Prof. Bruce I. Reisch. Data was collected by Prof. Tassoni and plant literature review was compiled by Prof. Davies. Prof. Melucci and I performed all chemometric analyses. All the cited authors contributed to write the manuscript paper, each one describing his part of the work.

CHAPTER 2.3: CHECKING SYRUP-ADULTERATION OF HONEY USING BIOLUMINESCENT BACTERIA AND CHEMOMETRICS

The present work was published by the journal “European Food Research and Technology” (reference: 2019, 245:315-324; https ://doi.org/10.1007/s0021 7-018-3163-8). All the experimental work was carried out in the Department of Pharmacy and Biotechnology of the University of Bologna (Bologna, Italy), in CREA-API (Bologna, Italy) and IZSLER (Brescia, Italy) institutes. Prof. Melucci and I performed all chemometric analyses. All cited authors contributed to write the manuscript paper, each one describing his part of the work.

CHAPTER 3: DESIGN OF EXPERIMENTS

The term “design of experiment” was created by the British statistician Sir Ronald Aylmer Fisher, who published a book with this title in 1935 [1] (in which, besides, he also introduced the concept of null hypothesis). His basic idea was simple: in order to get conclusive, or at least reliable, conclusions from the results of an experiment, this experiment has to be well designed and logically structured.

Over time, this idea of properly designing an experiment has been conjugated with the need of industry and laboratories of saving time and money (e.g. chemicals) while performing the experiment. This brought to the development of a series of methods which overall go under the name “design of experiments” (DoE). The main purpose of DoE is to obtain the maximum, or most robust, amount of information, using the lowest possible number of experiments or keeping a limit in terms of time or cost [2].

Like the other chemometric tools, also DoE has found several applications both in chemistry [3–6], but also in many other fields, such as: chemical engineering [7, 8], computer science [9], engineering [10, 11], medicine [12], psychology [13], and many others, such as health services organization [14].

The principle on which DoE is based is that the answer(s) generated by a system can be influenced by several parameters, called factors. For instance, an answer might be the yield of a reaction or the degree of satisfaction of a customer. In order to optimize the answer(s) of interest [2], a DoE can be built if some conditions are satisfied: factors (that is, variables) have to be controllable; these can assume at least two different levels; factors can be controlled each one independently of the others. In such a situation, the traditional strategy would be univariate, which means to vary the level of a single factor until the best solution has been achieved. Then, the first factor is kept constant at the optimal level, and a second one is varied, until a better solution is achieved. Then these two factors are left at their “optimal” level and the other ones are varied, one at a time. This situation is shown, for two factors, in Figure 3.0.1.

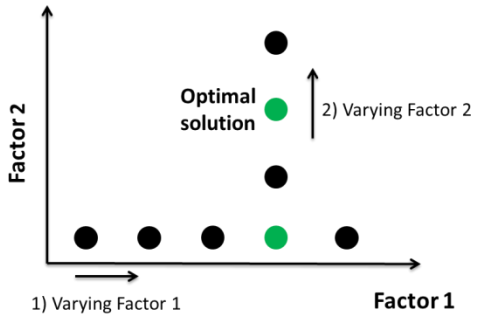


Figure 3.0.1: The “one at a time” strategy, each point is an experiment.
Green points indicate the optimal solutions for factors 1 and 2

However, this strategy suffers from two enormous drawbacks:

- As it can be seen from Figure 3.0.1, only a small portion of the experimental space is explored. If, for example, the overall best solution was situated above the second black point of factor 1, by this method it would never be achieved.

- In general, this strategy requires a higher number of experiments than a proper DoE, which means higher costs and much more time.

On the contrary, DoE fixes the lower and the higher limits of each factor, and executes only the experiments at the extreme points of the experimental space, as shown in Figure 3.0.2 (again for 2 factors)

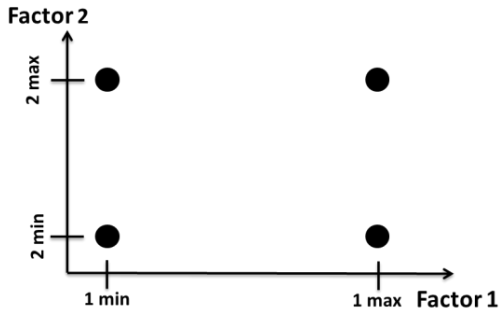


Figure 3.0.2: DoE strategy

By comparing Figure 3.0.1 and Figure 3.0.2, the main general consideration is that DoE requires a lower number of experiments, with the already cited advantages. However, from an analytical point of view, the most important advantage is that the experimental space not covered by the experiments can be "explored" by computing a mathematical linear model in the form

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 \qquad \text{(eq. 3.0.1)}$$

where y is the answer of interest, x_i are the factors, b_i are the regression coefficients, and b_0 is the intercept. Once created the model, any value of y can be calculated by inserting in eq. 3.0.1 the combination of interest of x_1 and x_2. Moreover, eq. 3.0.1 contains the term b_{12}, that is another interesting characteristic of DoE, which cannot be studied with the "one at a time" method. This term, indeed, takes into account the interaction between the two factors, which means that the two factors could influence *together* the behavior of y, making the mathematical model not perfectly addictive [15]. By adding levels to DoE, and degrees of freedom to the model, other effects can be studied. The most common is the square effect of factors, by adding a third level to the design, as shown in Figure 3.0.3

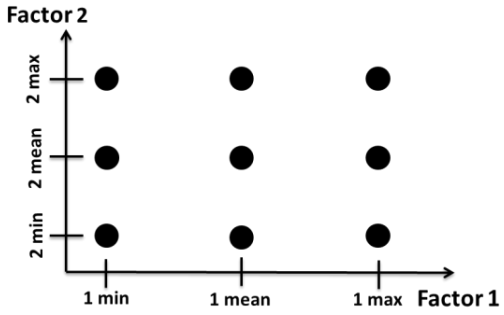


Figure 3.0.3: DoE strategy with three levels

In this case, the model becomes:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 + b_{11} x_1^2 + b_{22} x_2^2 \qquad \text{(eq. 3.0.2)}$$

The squared terms $b_{ii} x_i^2$ allow to study the curvature in the behavior of *y*, thus finding minima and maxima, if present. However, this obviously increases the required experiments.

The number *N* of experiments considered in a DoE ("cubic") scheme with *F* factors and *L* levels is $N=L^F$; thus, *N* increases rapidly with the number of factors. In general, the interaction between more than two factors or terms higher than the second order are considered not influential, therefore *L* commonly assumes value 2 or 3. If all the possible experiments are carried out, which means all the possible combinations of factors' levels, the resulting DoE is called "full factorial". However, several methods, that will not be discussed in detail, exist to reduce the number of experiments without losing the possibility of computing the model of interest with all the needed terms.

At this point, it is important to underline that the success of a DoE method strongly depends on its initial building phase. As stated before, DoE could be applied to whichever optimization problem in whichever field; however, the most important phase is that of selecting the correct variables that might be influential for the response to be optimized, and the proper levels. In such phase, it is very important to well know the studied problem and to take into account the available time and resources to carry out the experiments, which means also to decide how many experiments can be actually carried out in order to have enough information without exceeding the fixed limits. Moreover, it is important to decide how to properly collect the response(s), because it could require more analyses (which means again time and costs), and an erroneous procedure could make the entire project fail, giving misleading information. Therefore, this preparation part is the most important for the entire DoE procedure, also more than properly carrying out the experiments. In fact, in general, DoE is somehow robust, which means "resistant", to small errors in the experimental phase.

DoE is also very useful to understand which of the selected variables actually are the most important to describe the problem. Once created the linear model, each term is considered "important" if its regression coefficient *b* is significantly different from 0 [2]. In other words, a variable, or an interaction, can be considered influential for the answer if the *p*-value associated to its regression coefficient is lower than the selected significance level (usually 0.05 or 0.1). To create the linear model, the levels assumed by each factor are "scaled", so that the minimum has value -1 ("level -1"), the maximum is "level +1", and the central is "level 0". In this way, the computation of the model is simpler, but it also allows to eventually evaluate which variable is the most important for the problem, by considering the magnitude of the absolute values of *b* coefficients, together with the corresponding standard deviations.

Therefore, there are two main results that can be obtained by interpreting the DoE linear model:

- The calculated behavior of the response (*y*) variable in the entire experimental space. This can be visualized by a response surface plot, as the one reported in Figure 3.0.4. In this graph, the calculated value of the response is shown as a function of two factors. This allows to find the combination of

factor levels that optimize the response (if squared terms has been put in the model, maxima and minima could be visualized). Therefore, if the required result was to find such a combination, future experiments would be carried out at those levels, which will "guarantee" (in the limits of experimental error) the best result.

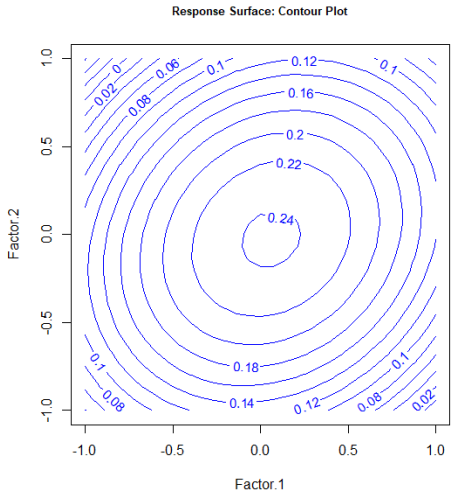


Figure 3.0.4: Example of response surface.
Each line represents a region with the same response value

- The most significant factors for describing the problem, which are the ones with low *p*-values. The ones with *p*-value higher than the significance level are probably not relevant for the problem; therefore, for an optimization problem, their level can be left at whichever value (generally, at the more convenient one).

Once terminated a DoE, if the results are not satisfactory, it can be decided to go on with a second DoE. In the subsequent DoE, the not relevant factors can be discarded or some new factors can be introduced, that were not taken into account before. Moreover, the experimental domain might be changed (for example, centering it around the optimal region found by the first DoE). Many strategies are available to perform a DoE. For example, some kind of design (as the Plackett-Burmann [16]) are considered only as preliminary, using few experiments to reduce the number of factors when it is too high at the beginning. Some other designs (as the Doehlert one [17]) allow to start with a low number of experiments, and then to simply extend the experimental domain by adding few more (this design is useful when it is not sure on which limits one should assume the factors). The choice again depends on available time and costs.

The following chapters will present two applications of DoE. In both cases, we had the possibility to perform a full factorial design. The first one concerns a study about the ability of a plant (*Polygonum aviculare* L.) to absorb some toxic metals while growing, with the aim of using it for bio-remediation of soils. This work is part of another PhD project in collaboration with the department of biological, geological, and environmental sciences of the University of Bologna. The second work concerns the optimization of work conditions of an industrial machine, and it is carried out in collaboration with the company IMA S.p.A. (Ozzano dell'Emilia, Bologna, Italy).

References

1. Fisher RA (1935) The design of experiments
2. Cela R, Claeys-Bruno M, Phan-Tan-Luu R (2010) Screening Strategies. In: Comprehensive Chemometrics. pp 251–300
3. Hibbert DB (2012) Experimental design in chromatography: A tutorial review. J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 910:2–13
4. Denmark SE, Butler CR (2008) Vinylation of aromatic halides using inexpensive organosilicon reagents. Illustration of design of experiment protocols. J Am Chem Soc 130:3690–3704 . doi: 10.1021/ja7100888
5. Carter CW, Baldwin ET, Frick L (1988) Statistical design of experiments for protein crystal growth and the use of a precrystallization assay. J Cryst Growth 90:60–73 . doi: 10.1016/0022-0248(88)90299-0
6. Roosta M, Ghaedi M, Daneshfar A, et al (2014) Optimization of the ultrasonic assisted removal of methylene blue by gold nanoparticles loaded on activated carbon using experimental design methodology. Ultrason Sonochem 21:242–252 . doi: 10.1016/j.ultsonch.2013.05.014
7. Gui MM, Lee KT, Bhatia S (2009) Supercritical ethanol technology for the production of biodiesel: Process optimization studies. J Supercrit Fluids 49:286–292 . doi: 10.1016/j.supflu.2008.12.014
8. Elsayed K (2015) Optimization of the cyclone separator geometry for minimum pressure drop using Co-Kriging. Powder Technol 269:409–424 . doi: 10.1016/j.powtec.2014.09.038
9. Simpson TW, Peplinski JD, Koch PN, Allen JK (2001) Metamodels for computer-based engineering design: Survey and recommendations. Eng. Comput. 17:129–150
10. Deloach R, Denver J (2000) The Modern Design of Experiments : A Technical and Marketing Framework 21st AIAA Advanced Measurement Technology and Ground Testing Conference. In: 21st Aerodynamic Measurement Technology and Ground Testing Conference
11. Wilson J, Sgondea A, Paxson DE, Rosenthal BN (2007) Parametric Investigation of Thrust Augmentation by Ejectors on a Pulsed Detonation Tube. J Propuls Power 23:108–115 . doi: 10.2514/1.19670
12. Qvist V, Stoltze K (1982) Identification of Significant Variables for Pulpal Reactions to Dental Materials. J Dent Res 61:20–24 . doi: 10.1177/00220345820610010401
13. Westfall J, Kenny DA, Judd CM (2014) Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. J Exp Psychol Gen 143:2020–2045 . doi: 10.1037/xge0000014
14. Syam SS, Côté MJ (2012) A comprehensive location-allocation method for specialized healthcare services. Oper Res Heal Care 1:73–83 . doi: 10.1016/j.orhc.2012.09.001
15. Pigeon JG (2006) Statistics for Experimenters: Design, Innovation and Discovery. Technometrics 48:303–304 . doi: 10.1198/tech.2006.s379
16. Plackett RL, Burman JP (1946) The Design of Optimum Multifactorial Experiments. Biometrika

33:305 . doi: 10.2307/2332195

17. Doehlert DH, Klee VL (1972) Experimental designs through level reduction of the d-dimensional cuboctahedron. Discrete Math 2:309–334 . doi: 10.1016/0012-365X(72)90011-8

CHAPTER 3.1: DoE FOR BIO-REMEDIATION

Introduction

The present chapter shows an application of DoE to a biological study. It is part of a larger research project, carried out in collaboration with the department of biological, geological, and environmental sciences of the University of Bologna, whose aim is to find new plants able to absorb high concentrations of dangerous metals from soils, thus resulting useful for bio-remediation purposes.

The ability of plants to absorb metals from soils is a topic of great interest for biologists [1]. Indeed, depending on their accumulation capacity, plants could be used for bio-indication [2, 3] (i.e. using them as analytical tool to monitor the presence of metals), or for bio-remediation [4, 5]. Bio-remediation (or phytoremediation if plants are used) means to use living organisms to accumulate pollutants, and then removing the pollutants from the area of interest by simply removing the organisms containing them. The advantages of using plants are that these can accumulate metals in several parts of their organism [6] (most of all in leaves [7, 8]), they can be easily removed, and they are adaptable to live in many different environments, even polluted ones [9]. Moreover, plants may be engineered to increase their accumulation capacity [10, 11]. It is under study also how metal accumulation influences plant life and growth [12].

Several plants have been studied for bio-monitoring and bio-remediation purposes, as for example dandelion (*Taraxacum officinale*) [6, 8], chicory (*Cichorium intybus* L.) [13], and *Plantago major* [14]. The present study focused its attention on common knotgrass (*Polygonum aviculare* L.), a plant that was already studied for its capacity of hyper-accumulating Hg [15], but never studied for the other metals.

In particular, we were interested in studying the bio-accumulation of Cu, Cd, Pb, Zn, and Cr, that are very common pollutants, in particular in urban areas [16, 17]. The previous part of the project, that will not be discussed in details here, proved that the *Polygonum aviculare* has the ability to absorb these metals; however, it seems not able to hyper-accumulate them. Anyway, it demonstrated to be useful for revealing actual levels of metals in the soil; thus, it could be considered as a bio-monitoring species.

The design of experiments (DoE) was applied to study how the presence of each metal at different concentrations affects the absorption of the others [18]. DoE was already applied for studying metal accumulation; however, in general, attention was focused on the effects of metals in biochemical development of the plant [19]. It has been already reported that metal accumulation depends on several characteristics of the soil, like pH, composition, and granulometry [18]. However, for the present study, the attention was focused only on metals; in fact, plants were cultivated in a growth chamber in hydroponic conditions (which means without the use of solid ground), with a standardized culture medium in which only metals in proper concentrations were added. Although the biology of plants by itself could be a great source of variability, it was therefore possible to “standardize” the plant cultivation. Hence, it was possible to check whether the different concentrations of Cd, Pb, and Cr (that were used as independent variables, while the concentrations of Cu and Zn were kept constant for all samples), and their interactions, can influence the

quantity of metals absorbed by *Polygonum aviculare*. Such a study can be useful to understand how this plant reacts in the presence of several metals in its environment, which is the situation of a polluted soil. In fact, a soil is not generally contaminated by a single chemical species. If performed also for other plant species, it could be useful also for finding the optimal plant for the bio-remediation of each specific soil.

Materials and Methods

Plants growing and DoE

As already stated, plants were grown in a growth chamber in hydroponic conditions. Hoagland medium [20] was used, with the following composition: KNO_3 (2 mM); $Ca(NO_3)_2{\cdot}4H_2O$ (2 mM); NH_4NO_3 (0.5 mM); $MgSO_4{\cdot}7H_2O$ (0.5 mM); KH_2PO_4 (0.25 mM); $Fe(Na_2)$-EDTA (40 μM); KCl (50 μM); H_3BO_3 (25 μM); $MnCl_4{\cdot}4H_2O$ (2 μM); $ZnSO_4{\cdot}7H_2O$ (2 μM); $CuSO_4{\cdot}5H_2O$ (0.5 μM); $(NH_4)_6Mo_7O_{24}{\cdot}4H_2O$ (0.075 μM); $CoCl_2{\cdot}6H_2O$ (0.15 μM); pH was adjusted to 5.8-6 with KOH. All chemicals were purchased by Sigma Aldrich (Merck, Darmstadt, Germany).

To this standard medium, the proper amount of metals was added, following the concentration decided for the experimental design. Constant concentrations of Cu (0.5 μM) and Zn (2 μM) were added to all samples.

Plants were grown for 4 weeks (1 for germination + 3 of treatment), and the liquid medium was replaced every week, in order to maintain "constant" characteristics. The first week (germination phase), seeds were put in Hoagland medium only, without metals, in order to let them sprout. Then, starting from the second week, metals were added to the medium. Some trials carried out by growing the plants for six weeks, instead of three, showed no significant differences in metal absorption; therefore, three weeks was chosen as growing time.

Sample preparation and analysis

After three weeks of treatment, plants were harvested, weighed, and frozen with liquid N_2, in order to grind them homogeneously. Samples were then dried and grinded again, in order to obtain a thin powder. 100 mg of that powder was put in a ceramic capsule, 1 ml of HNO_3 0.5 M was added and then put in a muffle furnace at 500°C for 5 h. This procedure was aimed at destroying and removing the organic matter (that should evaporate as H_2O and CO_2), leaving only the inorganic matter in the ashes. Ashes, which are soluble in water, were solved in 4 ml of HNO_3 0.5 M, and then analyzed by atomic absorption spectrometry (AAS). The instrument used is a Perkin Elmer AAnalyst 400 (Perkin Elmer, Waltham, MA, USA), controlled by the software WinLab 32, by Perkin Elmer. Table 3.1.1 shows the experimental conditions of AAS used for each analyzed metal.

Metal	Wavelength (nm)	Pyrolysis T (°C)	Analysis T (°C)
Cadmium	228.80	850	1650
Chromium	357.80	1650	2500
Copper	324.75	1000	2300
Lead	283.31	700	1800
Zinc	213.86	700	1800

Table 3.1.1: AAS experimental conditions for metals analyses

For each of the five analyzed metals (Cd, Cr, Cu, Pb, Zn), a calibration line was created. Standards for calibration lines were purchased by Merck (Darmstadt, Germany). Standard ranges were selected in order to stay in the linear range of each analyte, as tabulated in the software WinLab 32. Therefore, three standards were prepared for each metal, with the following concentrations:

- Cd: 2, 4, 6 ppb
- Cr: 2, 10, 25 ppb
- Cu: 2.5, 10, 25 ppb
- Pb: 5, 20, 40 ppb
- Zn: 0.1, 0.5, 0.8 ppm

Peak area was used as analytical signal, after verifying that peak height never overcame 0.6 AU, in order to always stay in the absorbance linear range. Before each analysis, a blank sample was analyzed, and the peak area of the sample was subtracted to the previous blank one. For each calibration line, the limit of detection (LoD) was computed, and it was verified that it never overcome the lowest standard concentration. When analyses had to be carried out in several days, every day three standards were analyzed and projected on the calibration line, to verify its validity. Three replicates were analyzed for each standard and sample. The injected volume was 20 μL for each analysis. Samples were properly diluted in order to obtain a signal in the calibration range, and the dilution factor was kept into account to calculate the metal concentration in the plant (expressed as ppm).

In order to analyze Cd, Cr, and Pb by AAS, some matrix modifiers are necessary. In particular $Mg(NO_3)_2$ (Perkin Elmer) for Cd and Cr, $PdCl_2$ (Fluka, Honeywell, Morris Planes, NJ, USA) for Cd, and $NH_4H_2PO_4$ (Sigma Aldrich) for Pb. 20 μL/mL of a solution containing all of them were added to each sample, final concentrations: 200 mg/L for $Mg(NO_3)_2$, 2.3 mg/L for $PdCl_2$, 4 mg/L for $NH_4H_2PO_4$. It was also verified that the presence of an unnecessary modifier did not influence the measurements of other metals (as Cu and Zn, that do not require any modifier).

Results and Discussion

A full factorial design was carried out with three factors (Cd, Pb, and Cr) and three levels, resulting in 27 total experiments. For each experiment, three plants were grown. This means that 81 plants were grown with different concentrations of Cd, Cr, and Pb in the culture medium. Nine of these samples were analyzed two times (without re-growing the plant but repeating the extraction procedure) due to uncertain results. The following Table 3.1.2 summarizes the DoE levels

Level	**-1**		**0**		**+1**	
Concentration	**μM**	**ppb**	**μM**	**ppb**	**μM**	**ppb**
Cd	0.01	1.6	0.07	8	0.14	16
Pb	1.83	380	14.48	3000	28.96	6000
Cr	6.92	360	23.08	1200	46.16	2400

Table 3.1.2: Metals concentrations according to DoE

DoE levels were chosen according to a previous study carried out by our group in the context of this project, which showed that urban soils, sampled in different places of Bologna and Milan, have, on average, the concentrations reported as level 0, while natural soils, sampled in countryside near the same two cities, have the concentrations reported as level -1. Level +1 was chosen as two times the concentrations of level 0, simulating a strongly polluted soil.

Appendix A shows the results obtained by the experiments carried out.

The execution of all 81 experiments (plus 9 replicates) made it possible to compute a linear model considering all the effects: individual variables, interactions between variables and square effects. Five linear models were computed, one for each dependent variable (the absorbed concentrations of Cd, Cr, Cu, Pb, Zn) The following Table 3.1.3 shows the linear coefficients and the corresponding standard deviations estimated by the linear models

	Cd		**Cr**		**Cu**		**Pb**		**Zn**	
	Coeff.	**Std. dev.**	**Coeff.**	**Std. dev.**	**Coeff.**	**Std. dev.**	**Coeff.**	**Std. dev.**	**Coeff.**	**Std. dev.**
b0	1.34	2.39	-0.0810	0.319	3.97	2.64	4.86	5.58	32.1	5.32
b1	2.25	1.14	-0.0479	0.153	0.560	1.26	-1.81	2.67	0.520	2.54
b2	1.34	1.16	0.0768	0.155	-0.0943	1.28	3.41	2.71	4.07	2.58
b3	-1.22	1.19	0.477	0.159	0.398	1.31	-2.56	2.77	0.903	2.64
b12	1.77	1.40	-0.0695	0.187	-2.78	1.55	-3.01	3.27	-0.00117	3.11
b13	-2.37	1.41	-0.172	0.188	0.509	1.56	5.10	3.29	-2.25	3.13
b23	-2.25	1.47	-0.124	0.196	-1.27	1.62	0.334	3.43	-4.91	3.27
b11	1.13	1.97	0.219	0.263	4.26	2.17	3.77	4.59	10.5	4.37
b22	1.17	1.97	0.501	0.263	4.51	2.17	2.50	4.59	12.0	4.37
b33	0.692	1.98	0.375	0.264	0.985	2.19	6.22	4.62	8.96	4.40

Table 3.1.3: coefficients and standard deviations estimated by linear models for plant DoE; b_0 is the intercept; 1: Cd; 2: Pb; 3: Cr

The following Table 3.1.4 shows, instead, the *p*-values of each regression coefficient

	Cd	**Cr**	**Cu**	**Pb**	**Zn**
b0	0.633	0.932	0.197	0.446	0.00
b1	0.0709	0.727	0.667	0.509	0.814
b2	0.285	0.750	0.944	0.233	0.0834
b3	0.339	0.00940	0.772	0.377	0.698
b12	0.235	0.674	0.0898	0.371	1.00
b13	0.119	0.347	0.751	0.138	0.410
b23	0.164	0.390	0.461	0.926	0.100
b11	0.607	0.332	0.0828	0.452	0.0146
b22	0.589	0.118	0.0626	0.610	0.00540
b33	0.747	0.247	0.672	0.209	0.0303

Table 3.1.4: *p*-values estimated by linear models for plant DoE; b_0 is the intercept; 1: Cd; 2: Pb; 3: Cr. The most significant p-values are highlighted in yellow

Table 3.1.4 shows which variables are the most significant for the absorption of each metal, that is, which one have a *p*-value lower than a significance level of 0.1 (or 0.05). From the observation of Tables 3.1.3 and 3.1.4, some remarks can be drawn for each metal (each linear model is independent from the others).

- *CADMIUM*

In Cd model, only the coefficient b_1 can be considered significant (p = 0.07), which is the coefficient corresponding to the independent variable Cd. Its coefficient is positive, which means that, in the explored space, *Polygonum aviculare* tends to absorb more Cd when its concentration in the medium increases, while the presence of other metals does not influence its absorption. This could mean that Cd may have some role in plant growth, therefore, it tends to accumulate this metal. The following Figure 3.1.1 represents the response surface for Cadmium with Cadmium and Chromium as independent variables. Cr has been chosen because its interaction with Cd (b_{13}) is slightly significant (p = 0.119). Figure 3.1.1 confirms that the absorption of Cd increases by increasing its presence in the medium.

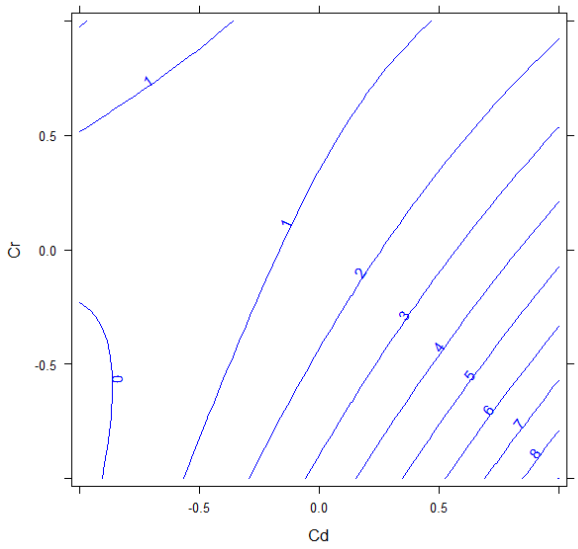


Figure 3.1.1: response surface for Cd dependent variable as a function of Cd (independent) and Cr

- *CHROMIUM*

Also in this case, the absorption of Cr is affected only by its presence in the medium (b_3, p = 0.00940), and again its absorption increases by augmenting the concentration in the medium, without any influence due to other metals. Cr (most of all Cr(VI), while Cr(III) was used for the present work) is known for its toxicity in plants [21], because it interferes with the photosynthetic path and causes other damages. However, it is absorbed by non-specific channels, while other essential ions pass through specific paths [22]; therefore, it can be easily absorbed by plants. This characteristic, however, makes the use of plants interesting for the bio-remediation of soil contaminated by this metal [23]. The following response surface (Figure 3.1.2) shows the calculated absorption of Chromium as a function of Chromium and Lead. Pb has been chosen because its square effect (b_{22}) is slightly significant for Cr absorption.

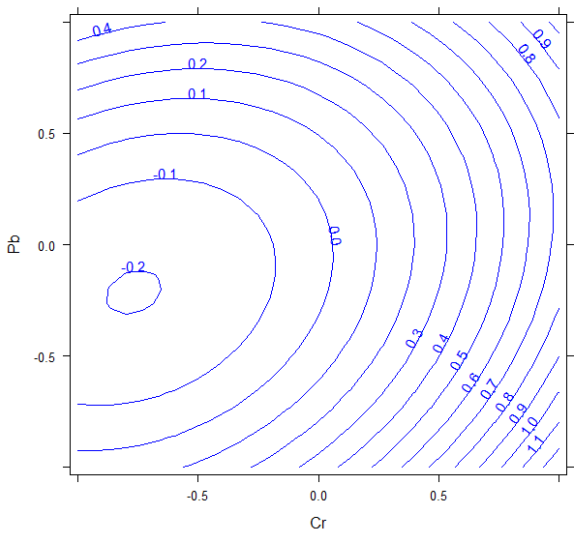


Figure 3.1.2: response surface for Cr dependent variable as a function of Cr and Pb

- *COPPER*

Copper, together with Zinc, is one of the two metals whose concentration was not varied in the present DoE, because their effects on plants have been widely studied [24] (and, until a certain level, these are also essential metals). Thus, we were interested only in studying how the presence of other metals influenced the absorption of these two. Table 3.1.4 shows that none of the three variables is significant by itself, however Cd and Pb influence Cu both as squares (b_{11} and b_{22}, p_{11} = 0.0828 and p_{22} = 0.0626) and with their interaction (b_{12}, p = 0.0898). Figure 3.1.3, therefore, represents the influence of these two metals on Copper.

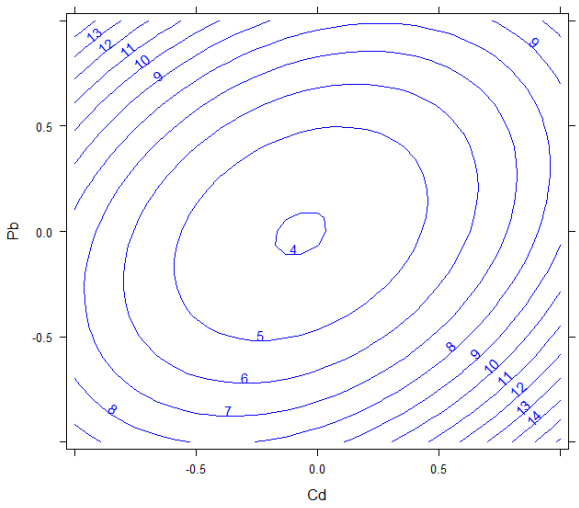


Figure 3.1.3: response surface for Cu dependent variable as a function of Cd and Pb

The response surface in Figure 3.1.3 shows that there is a minimum of absorption when both Cd and Pb are in their 0 levels, while maxima (for the explored space) are present when one of them is at the highest and the other one at the lowest level. Higher absorption is present also when both Cd and Pb are at the highest or lowest level. We do not have any explanation for such behavior; however, for maximizing the absorption of Cu, it seems only important that Cd and Pb do not stay at the urban (0) level.

- *LEAD*

In the case of Pb, none of the considered variables seem significant for its absorption. Interestingly, neither the Pb concentration in the medium is significant. It could mean that the studied plant is somehow "resistant" to the presence of Pb, and it never absorbs it beyond a certain limit. If such hypothesis was confirmed, it would mean that *Polygonum aviculare* would not be useful for bio-remediation of Pb pollution. Moreover, it is possible that Pb precipitates in the medium as $PbSO_4$ or $Pb_3(PO_4)_2$, or chelated by EDTA, being no more available for plant absorption, therefore it would be useful to study its behavior in other conditions (and, consequently, also its behavior with regard to the other metals, for which its concentration seems significant).

- *ZINC*

Zinc is the second metal that was not varied in the experiments. Table 3.1.4 shows that all the square effects are significant for its absorption (b_{11}, p = 0.0146; b_{22}, p = 0.00540; b_{33}, p = 0.0303). Figure 3.1.4, therefore, reports all the combinations of the three variables.

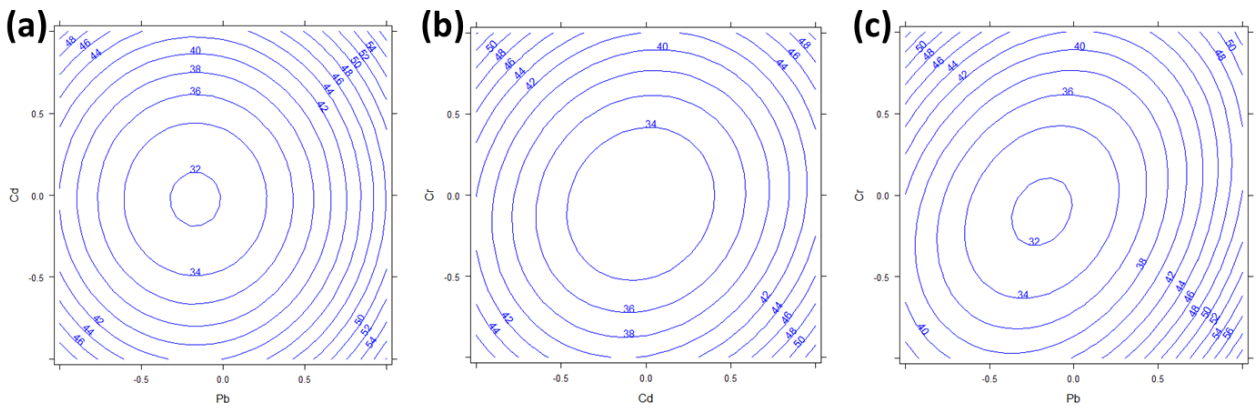


Figure 3.1.4: response surfaces for Zn dependent variable. a) Pb vs Cd; b) Cd vs Cr; c) Pb vs Cr

As in the case of Cu, also the absorption of Zn shows a minimum in the 0-levels of all variables, while it increases away from these central points. In particular, it has a local maximum when the concentration of Pb is highest.

Conclusions

A full factorial DoE with three factors and three levels was applied to the study of metal absorptions by *Polygonum aviculare*. Five metals were studied (Cd, Cr, Cu, Pb, and Zn), and the concentrations of three of them were varied in the growing medium. Some significant effects on the presence of one metal for the absorption of the others were observed.

This is only a preliminary study. As already stated, the biology of plants could have a great variability: therefore, replicates of the growths could be carried out to evaluate the repeatability of these results. Moreover, the toxicity and the effects of these metals on *Polygonum aviculare* have to be evaluated, as already studied for other plants [12]. The present study highlighted that, in general, high levels of metal

concentrations correspond to high absorbed concentrations. Therefore, a second DoE may be carried out starting from the concentrations here considered as high (+1), varying also Cu and Zn, and checking which concentrations would bring to plants death.

References

1. Broadley MR, Willey NJ, Wilkins JC, et al (2001) Phylogenetic variation in heavy metal accumulation in angiosperms. New Phytol 152:9–27 . doi: 10.1046/j.0028-646x.2001.00238.x
2. Malizia D, Giuliano A, Ortaggi G, Masotti A (2012) Common plants as alternative analytical tools to monitor heavy metals in soil. Chem Cent J 6: . doi: 10.1186/1752-153X-6-S2-S6
3. Dimitrova I, Yurukova L (2005) Bioindication of anthropogenic pollution with Plantago lanceolata (Plantaginaceae): metal accumulation, morphological and stomatal leaf characteristics. Phytol Balc 11:89–96
4. Pilon-Smits E (2005) Phytoremediation. Annu Rev Plant Biol 56:15–39 . doi: 10.1146/annurev.arplant.56.032604.144214
5. Salt DE, Blaylock M, Kumar NPBA, et al (1995) Phytoremediation: A novel strategy for the removal of toxic metals from the environment using plants. Bio/Technology 13:468–474 . doi: 10.1038/nbt0595-468
6. Normandin L, Kennedy G, Zayed J (1999) Potential of dandelion (Taraxacum officinale) as a bioindicator of manganese arising from the use of methylcyclopentadienyl manganese tricarbonyl in unleaded gasoline. Sci Total Environ 239:165–171 . doi: 10.1016/S0048-9697(99)00292-2
7. Aksoy A, Hale WHG, Dixon JM (1999) Capsella bursa-pastoris (L.) Medic. as a biomonitor of heavy metals. Sci Total Environ 226:177–186 . doi: 10.1016/S0048-9697(98)00391-X
8. Keane B, Collier MH, Shann JR, Rogstad SH (2001) Metal content of dandelion (Taraxacum officinale) leaves in relation to soil contamination and airborne particulate matter. Sci Total Environ 281:63–78 . doi: 10.1016/S0048-9697(01)00836-1
9. Shu WS, Ye ZH, Lan CY, et al (2002) Lead, zinc and copper accumulation and tolerance in populations of Paspalum distichum and Cynodon dactylon. Environ Pollut 120:445–453 . doi: 10.1016/S0269-7491(02)00110-0
10. Clemens S, Palmgren MG, Krämer U (2002) A long way ahead: Understanding and engineering plant metal accumulation. Trends Plant Sci. 7:309–315
11. Sheoran V, Sheoran AS, Poonia P (2016) Factors Affecting Phytoextraction: A Review. Pedosphere 26:148–166
12. Aggarwal A, Sharma I (2011) Metal toxicity and photosynthesis. Photosynth Overviews Recent Prog Futur Perspect 229–236
13. Simon L, Martin HW, Adriano DC (1996) Chicory (Cichorium intybus L.) and dandelion (Taraxacum officinale Web.) as phytoindicators of cadmium contamination. Water Air Soil Pollut 91:351–362 . doi: 10.1007/BF00666269
14. Galal TM, Shehata HS (2015) Bioaccumulation and translocation of heavy metals by Plantago major L. grown in contaminated soils under the effect of traffic pollution. Ecol Indic 48:244–251 . doi: 10.1016/j.ecolind.2014.08.013
15. Massa N, Andreucci F, Poli M, et al (2010) Screening for heavy metal accumulators amongst

autochtonous plants in a polluted site in Italy. Ecotoxicol Environ Saf 73:1988–1997 . doi: 10.1016/j.ecoenv.2010.08.032

16. Charlesworth S, Everett M, McCarthy R, et al (2003) A comparative study of heavy metal concentration and distribution in deposited street dusts in a large and a small urban area: Birmingham and Coventry, West Midlands, UK. Environ Int 29:563–573 . doi: 10.1016/S0160-4120(03)00015-1
17. Huber M, Welker A, Helmreich B (2016) Critical review of heavy metal pollution of traffic area runoff: Occurrence, influencing factors, and partitioning. Sci. Total Environ. 541:895–919
18. Orroño DI, Schindler V, Lavado RS (2012) Heavy metal availability in pelargonium hortorum rhizosphere: Interactions, uptake and plant accumulation. J Plant Nutr 35:1374–1386 . doi: 10.1080/01904167.2012.684129
19. Tkalec M, Štefanić PP, Cvjetko P, et al (2014) The effects of cadmium-zinc interactions on biochemical responses in tobacco seedlings and adult plants. PLoS One 9: . doi: 10.1371/journal.pone.0087582
20. Serrano-Cinca C, Fuertes-Callén Y, Mar-Molinero C (2005) Measuring DEA efficiency in Internet companies. Decis Support Syst 38:557–573 . doi: 10.1016/j.dss.2003.08.004
21. Rodriguez E, Santos C, Azevedo R, et al (2012) Chromium (VI) induces toxicity at different photosynthetic levels in pea. Plant Physiol Biochem 53:94–100 . doi: 10.1016/j.plaphy.2012.01.013
22. Shahid M, Shamshad S, Rafiq M, et al (2017) Chromium speciation, bioavailability, uptake, toxicity and detoxification in soil-plant system: A review. Chemosphere 178:513–533
23. Sinha V, Pakshirajan K, Chaturvedi R (2018) Chromium tolerance, bioaccumulation and localization in plants: An overview. J. Environ. Manage. 206:715–730
24. Ebbs SD, Kochian L V. (1997) Toxicity of Zinc and Copper to Brassica Species: Implications for Phytoremediation. J Environ Qual 26:776 . doi: 10.2134/jeq1997.00472425002600030026x

CHAPTER 3.2: DoE FOR MACHINE OPTIMIZATION

Introduction

The present chapter shows a second application of DoE for the optimization of working parameters of an industrial type rotary bin blender for powder mixtures. Moreover, as a second part of the same work, a chemometric approach for exploiting a NIR probe, connected to the machine for evaluating the mixing quality over time, will be discussed. This study was carried out in collaboration with IMA S.p.A. (Industria Macchine Automatiche, Ozzano dell'Emilia, Bologna, Italy), a company that designs and manufactures automatic machines for processing and packaging pharmaceutical, cosmetic, and food products.

Design of experiments finds, probably, its best application in the optimization of industrial processes [1, 2]. Indeed, most industrial machines are aimed to perform always the same procedure (for producing, mixing, packaging, etc. the same product), without changing the working conditions. DoE, therefore, can be used before starting the production to find the optimal machine settings giving the best possible product with a small number of preliminary experiments [3]. For this reason, the American FDA suggests the use of DoE to control quality attributes and process parameters [4].

In the literature, several applications of DoE to industrial processes are present, going from optimization of pharmaceutical formulations for drug delivery [5], to membranes for recycling wastewater [6], to cement [7]. Also a helicopter rotor that minimizes vibrations was optimized by DoE [8].

Powder blending is another topic of interest, most of all in the pharmaceutical industry [9, 10], where the mixing performances may have a direct influence on drug effectiveness [11]. In this field, besides using DoE for machine optimization, it is also interesting to have an analytical technique able to follow in-line (i.e. directly inside the machine) the blending progress. The alternative (that is in any case important) is to analyze the product only at the end of the process. However, most of all if the process requires several steps, if some critical issue is found on the final product, such method is not able to understand at which point of the process it emerged. Another possibility could be to get a sample from different points of the process and analyze all of them; however, in general, this requires a great amount of time and analyses, and, in the worst case, also to stop the process for sampling. Therefore, the in-line approach is simpler, cheaper, and non-destructive. The ability to analyze without altering the sample is another important characteristic for an industry. Obviously, the preferred analytical technique for such purpose is spectrophotometry, in particular near infra-red (NIR) spectroscopy [10]. Indeed, most of the molecules, in particular those of pharmaceutical interest, are somehow active to NIR radiation. The problem of using such in-line approach is that a huge amount of data may be collected for every single batch, and therefore chemometrics has to be applied to extract the required information.

These practices for optimizing all aspects of industrial processes goes under the name of "quality by design" (QbD) [12].

The main goal of a powder blending process is to obtain a homogeneous distribution of drug's components, with uniform properties in the entire batch. In particular, for a pharmaceutical formulation, the content of active pharmaceutical ingredients (APIs), the molecule(s) with the required pharmaceutical properties, is the most important. Indeed, an inhomogeneous powder would bring to dosages which are not compliant with the therapeutic window, with the possible consequence of being harmful or without effects on the patient. Such problems in mixing may imply, for the industry, expensive reworks or recalls of the product. Powder homogeneity depends on several factors [13], as, for example, quality of raw materials, geometry of blender, blending speed and time. Some of these factors may be difficult to control (as the geometry of blender and the quality of raw material, if provided from an external company), while others can be optimized by DoE.

In the present study, IMA company, the machine supplier, was interested in finding the optimal working conditions of its rotatory bin blender, with the aim of offering a better service to its customers. The purpose of DoE optimization was to obtain a powder mixture characterized by homogeneity and with good flowability. Moreover, a NIR probe was connected to the blender: the second request of IMA was to develop a general chemometric method to use such a probe for following the process in place. The aim was to in time decide whether the batch mixing was going fine or not, even without any other information about the mixture (because a supplier may not have the same instruments used in this work for the optimization section). Riboflavin (vitamin B2) was used as an example of API.

Materials and Methods

Machine settings and DoE

The machine under study, manufactured by IMA S.p.A. (Ozzano dell'Emilia, Bologna, Italy), is a laboratory type rotatory bin blender, Cyclops Lab™, with a total volume capacity of 15 L. The following Figure 3.2.1 shows the studied blender.

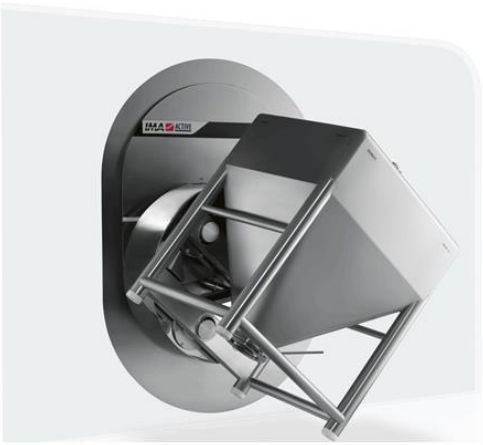


Figure 3.2.1: Cyclops Lab™ bin blender (image provided by IMA website https://ima.it/pharma/machine/cyclops/)

This machine rotates on its horizontal axis, and the micro-NIR spectrophotometer (MicroNIR spectrometer™, VIAVI Solutions inc., San Jose, CA, USA) is set on the top of the bin (at the top of the

"cubic" part, not visible in Figure 3.2.1). A NIR spectrum is registered for each rotation, every time the probe is at the bottom of the chamber and the powder "falls" over it by gravity.

The composition of the powder to be mixed was maintained constant for all experiments: riboflavin, 3%$_{w/w}$ (BASF, Ludwigshafen, Germany); pregelatinized starch, 48.5%$_{w/w}$ (Colorcon, Harleysville, PA, USA); microcrystalline cellulose (MCC), 48%$_{w/w}$ (Avicel® PH-101 and Avicel® PH-102, Merck, Darmstadt, Germany); sodium stearyl-fumarate, 0.5%$_{w/w}$ (PRUV®, JRS Pharma, Rosenberg, Germany). Powders were put into the machine bin following a sandwiching method [14]: first of all, half of the total starch, then half of the total MCC, riboflavin, sodium stearyl-fumarate (used as a lubricant), the second half of MCC, and, finally, the second half of starch. Working always with these conditions allowed to focus the study only on the effects of DoE variables, thus reducing other sources of variability. Sandwiching is useful for blending because it increases the contact area between API and excipients, facilitating the diffusion process of the former into the latter.

The independent variables and levels for DoE were decided in agreement with the machine experts of IMA company; these are: bin filling level, rotation speed, total mixing time, and particle size of MCC. Three levels were used for bin filling, rotation speed, and mixing time, while MCC was provided with two different particle sizes (PH-101, 50 μm, and PH-102, 100 μm); therefore, it is a qualitative variable with two levels. Therefore, a full factorial design was carried out with 58 experiments ($3^3 \cdot 2+4$ replicates of the central points, two for each level of particle size). Table 3.2.1 summarizes the DoE levels

Level	**-1**	**0**	**+1**
Bin filling (%)	30	52.5	75
Rotation speed (rpm)	10	15	20
Mixing time (min)	15	20	25
Particle size (μm)	50	-	100

Table 3.2.1: DoE factors and levels for blender optimization

As it can be seen from Table 3.2.1, bin filling never reaches 100%, as recommended by IMA manufacturers, and the number of rotations (thus the number of IR spectra collected for each experiment) can be obtained multiplying rotation speed and mixing time (ranging from 150 to 500).

The final mixture powders were characterized in terms of homogeneity and flowability.

Homogeneity was evaluated by HPLC analysis. For each experiment, one sample was taken from inside the bin, as soon as the rotation stopped, and two more were taken from random points when the powder was discharged (common guidelines suggest to take at least three sample from inside the machine in different positions; however, the geometry of the bin did not permit to take more than one). 50.0 ± 0.1 mg of each sample were solved in a 50 ml solution. The solvent was $H_2O_{\text{MilliQ grade}}$: Methanol : Acetonitrile = 40:30:30; sonication was employed to facilitate dissolution. 100 μL of this solution was diluted in 1 mL of mobile phase before HPLC analysis, in order to facilitate riboflavin elution. Therefore, the expected riboflavin concentration for an ideal 3%$_{w/w}$ powder is 3 ppm. Each solution was analyzed three times; thus, for each experiment, nine HPLC analyses were performed. Five standards of pure riboflavin were prepared with the

same solvents, ranging from 1 ppm to 5 ppm, and a calibration line was computed with peak areas as independent variables (R^2 = 0.998; slope = 14.0 ± 0.4; intercept = 0.8 ± 1.2, LoD = 0.3 ppm). HPLC analyses were carried out with an Agilent 1260 Quaternary Infinity LC (Agilent Technologies, Santa Clara, CA, USA) equipped with a C18 chromatographic column (Agilent ZORBAX Eclipse Plus C18). The mobile phase was composed by 70% H_2O buffered with 0.05 M formic acid (pH = 3.75) and 30% acetonitrile (isocratic conditions), with a flow rate of 1 mL/min. For each analysis, a 5-µL aliquot of solution was injected. HPLC detector was a UV-Vis spectrophotometer, set to 266.5 nm (where riboflavin has a maximum of absorption). Chromatographic peaks were integrated by the software controlling the instrument, Agilent ChemStation (Agilent Technologies). Homogeneity was evaluated by the mean concentration of the nine analyses carried out for each experiment (which has to be close to 3, indicating a powder sample with an average concentration of $3\%_{w/w}$), and by the relative standard deviation (RSD, which has to be minimized, indicating low variability between the replicates of the same experiment).

Flowability, instead, was evaluated by calculating the Carr Index (ICarr) [15] by:

$$ICarr = \frac{\rho_{tapped} - \rho_{bulk}}{\rho_{tapped}} \qquad \text{(eq. 3.2.1)}$$

where ρ_{bulk} is the bulk density. Bulk density is obtained by putting a known mass of powder in a 100-mL graduated cylinder and calculating mass/volume. The tapped density ρ_{tapped} is calculated in the same way as the bulk, but pressing the powder inside the cylinder before reading the volume [16] (pressing can be performed by different methods and goes on until no further changes in the volume are observed). Flowability is considered optimal when ICarr is low. In fact, the less compressible a material is, the more flowable it will be [15]. Carr put a borderline between free-flowing and non-free-flowing materials at 20-21% of ICarr, considering powders as non-free-flowing [15]. Thus, DoE has to look for a minimum of ICarr, which should be around 20-21%.

Chemometric approach for IR data analysis

As already stated, one NIR spectrum is collected for each rotation of the blender. Thus, for each experiment of the planned DoE, 150-500 spectra are collected, in the range 908-1676 nm, with steps of 6 nm. This brings to have 58 datasets (the number of DoE experiments) with dimensions (150-500) x 125. Those data, with a proper chemometric procedure, can be useful to follow the mixing process in-line, in particular if such chemometric process can be automated and performed directly during spectra acquisition.

The idea of the following method is to use the spectra by themselves to analyze every single experiment, but also in a combined manner to create a sort of control chart which could detect experiments that seem going well by themselves, but with an overall "strange" behavior.

The first step is to study the NIR spectra of each experiment. This can be done by the moving block method (MBM), a well-known method [9, 17] and very suited for in-line IR analyses [18]. MBM starts from a group of spectra (of a selectable number), called "window", and, for each variable, it computes mean and standard

deviation, reducing the window into two calculated spectra. Then a mean value is calculated for each of these two spectra, further reducing the window into two values. At this point, the window moves to a second group of spectra (for instance, the first window could take spectra from 1 to 5, the second from 2 to 6) and the computation is repeated. The window continues to move until all the spectra are considered. At the end of the computation, all the couple of values can be plotted in function of window number, and some information about the process can be obtained by the observation of such graphs, in particular from the values derived from the standard deviation spectra (MBSD). In general, if MBSD plot is flat and does not have discontinuity points, it means that no differences are present in the NIR spectra, thus the process is proceeding without trouble.

The second possibility considered for exploiting NIR spectra is computing a control chart, basing on an idea of Wold et al. [19]. As it is normal for a control chart, this procedure requires a training set of successful experiments; thus, a group of experiments-spectra is gathered in a single dataset. In this case, a progress over time has to be studied, therefore a PLS model is computed using NIR spectra as predictors ($\boldsymbol{X}$) and "time" as response ($\boldsymbol{y}$). The variable "time" can be considered as the experimental time at which the corresponding NIR spectrum was collected. However, if the lengths of the experiments are different (as in this specific case), timescales will be shifted from one experiment to the other (some experiments stop after 15 min, some other after 25 min, and also the spectra are collected with different scan time, due to different rotation speeds). So, the simple variable "time" would no longer be correct. Therefore, the "time" variable can be somehow scaled and converted to a progress percentage (PP%), in order to have all the NIR spectra in the range 0-100% of experimental progress. Then, PLS is computed and X-scores are stored. In general, only X-scores from the first PLS-factor may be used. However, depending on explained variance and RMSE, there might be cases in which also higher factors are useful, then also the corresponding X-scores can be stored. In any case, from now on the procedure works on a single factor at a time. Basing on PP% variable (or time, if possible), scores are divided in subsets of specific PP% intervals: in this case, it was chosen to divide X-scores into 10 intervals of 10% PP% each. In general, any interval is possible, and even different intervals for each subset. For each interval, then, mean and standard deviation of X-scores are computed, and the control chart is created by considering as the acceptability region mean ± standard deviation (or some multiple of standard deviation, Wold [19] suggests 3 standard deviations). The acceptability of a new sample can be evaluated by projecting its NIR spectra on the PLS model, and the calculated X-scores are divided into the same subsets. Then the mean of each subset is projected on the control chart, and it can be easily seen whether the new experiment is inside the acceptability region or not, and also whether there is a particular moment in which the experiment has a particular behavior.

Results and Discussion

DoE

Appendix B shows the results of the 58 experiments-full factorial design obtained for this work.

The execution of all experiments of the full factorial design made it possible to compute a linear model with all interaction and square effects, except for the square effect of particle size, because that variable has only two levels. The following Table 3.2.2 shows the regression coefficients and the corresponding standard deviations estimated by the linear models for all responses

	HPLC mean		**HPLC std dev**		**iCarr**	
	Coeff.	**std. dev.**	**Coeff.**	**std. dev.**	**Coeff.**	**std. dev.**
b0	3.26	0.0692	0.456	0.0726	22.3	0.284
b1	-0.0125	0.0320	0.0150	0.0336	0.0578	0.132
b2	0.00806	0.0320	-0.00806	0.0336	0.152	0.132
b3	-0.0647	0.0320	0.00583	0.0336	0.00806	0.132
b4	-0.0217	0.0261	0.0235	0.0274	-1.06	0.107
b12	0.0196	0.0392	0.0425	0.0411	0.0667	0.161
b13	0.00833	0.0392	-0.0217	0.0411	-0.366	0.161
b14	-0.0464	0.0320	-0.00333	0.0336	0.186	0.132
b23	0.0225	0.0392	-0.0517	0.0411	0.183	0.161
b24	0.0119	0.0320	-0.000278	0.0336	-0.100	0.132
b34	-0.0608	0.0320	0.0114	0.0336	-0.158	0.132
b11	0.00139	0.0554	-0.104	0.0582	0.266	0.228
b22	-0.00361	0.0554	-0.110	0.0582	-0.364	0.228
b33	-0.0419	0.0554	0.0486	0.0582	0.236	0.228

Table 3.2.2: Coefficients and standard deviations estimated by the linear models; b0 is the intercept; 1: bin filling; 2: rotation speed; 3: mixing time; 4: particle size

Table 3.2.3, instead, shows the estimated *p*-values for each regression coefficient

	HPLC mean	**HPLC std dev**	**ICarr**
b0	0.00	0.00	0.00
b1	0.698	0.658	0.663
b2	0.803	0.812	0.256
b3	0.0499	0.863	0.952
b4	0.412	0.396	0.00
b12	0.620	0.308	0.681
b13	0.833	0.601	0.0286
b14	0.155	0.922	0.165
b23	0.569	0.217	0.263
b24	0.711	0.993	0.452
b34	0.0646	0.736	0.238
b11	0.980	0.0818	0.251
b22	0.948	0.0666	0.118
b33	0.454	0.408	0.306

Table 3.2.3: *p*-values estimated by linear models for blender DoE; b0 is the intercept; 1: bin filling; 2: rotation speed; 3: mixing time; 4: particle size.
The most significant p-values are highlighted in yellow

Table 3.2.3 shows which predictors are significant for each response (the ones with *p*-value lower than the significance level 0.1 or 0.05). Some remarks can be drawn for each response.

- *HPLC MEAN*

HPLC mean is the mean value of the replicates of each experiment analyzed by HPLC. Considering that the samples concurring to this mean came from different point of the batch, HPLC mean is thought as a measure of the mixing ability of the machine, because each point should have a riboflavin concentration of 3%$_{w/w}$, which correspond to an HPLC concentration of 3 ppm. Mixing time (b_3) seems to be the only significant variable for this response (p = 0.0499), with a certain significance in correlation with particle size (b_{34}, p = 0.0646). Therefore, it is useful to look at the response surface showing the calculated behavior of HPLC mean as a function of these two variables (Figure 3.2.2).

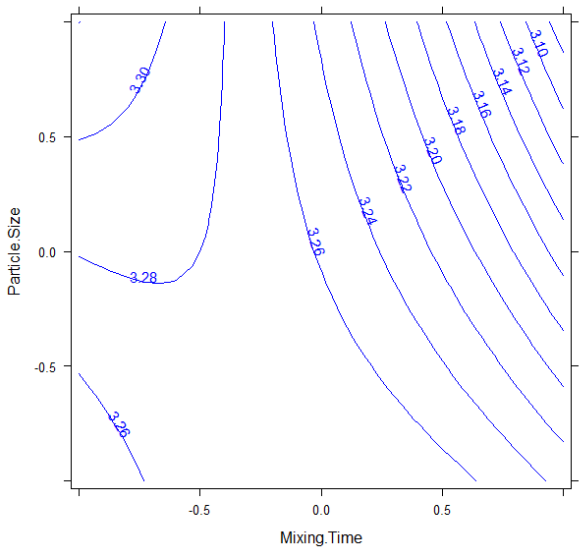


Figure 3.2.2: Response surface for HPLC mean as a function of mixing time and particle size

As it can also be seen from the table in Appendix B, HPLC mean is almost always higher than 3, and also in the response surface of Figure 3.2.2, the expected value 3 is never reached. This computation does not take into account the experimental error (one should take into account all the steps necessary to perform the HPLC analysis). However, Figure 3.2.2 clearly shows that better result can be obtained using the highest level of particle size of MCC (which means 100 μm instead of 50 μm) and increasing the mixing time (but this would mean to increase the experimental time).

- *HPLC STANDARD DEVIATION*

Also HPLC standard deviation derives from the replicates of the experiments, and it is thought as an estimation of the homogeneity of the powder: each HPLC sample should have more or less the same concentration value (3 ppm); thus, the overall standard deviation should be as low as possible. In this case, only the squared effects of bin filling (b_{11}) and mixing time (b_{22}) are slightly significant (p_{11} = 0.0818; p_{22} = 0.0666). The fact that few variables are significant is, nevertheless, encouraging: indeed, this could mean that the calculated standard deviations are mainly due to experimental errors occurring during HPLC analysis,

and this variability, that should be in any case low, is anyway enough to cover the variability due to the machine. Therefore, this means that the machine produces homogeneous mixing, no matter which the factor levels are. Figure 3.2.3 shows, anyway, the behavior of HPLC standard deviation as a function of bin filling and mixing time, and it can be seen that some improvement could be obtained by putting one of the two factors at the highest level and the other one at the lowest.

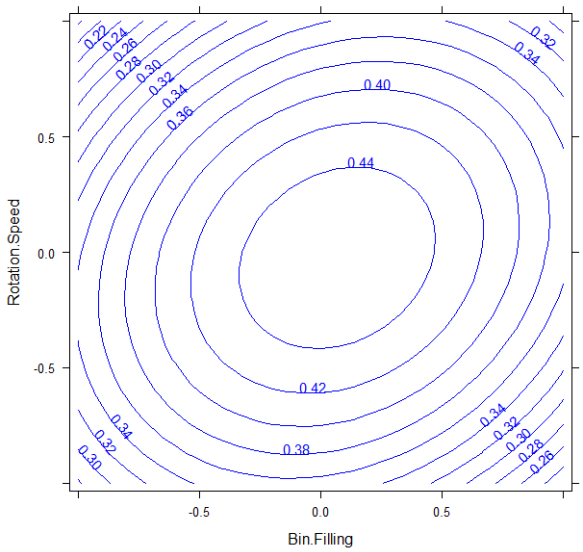


Figure 3.2.3: Response surface for HPLC standard deviation in function of bin filling and rotation speed

- *CARR INDEX*

As already stated, Carr index (ICarr) is a direct measure of powder flowability, a very important parameter for pharmaceutical powders: a high flowability (low ICarr) eases the tableting process. For ICarr, there is an enormous influence of particle size (b_4, p = 0), and also the interaction of bin filling and mixing time is significant (b_{13}, p = 0.0286). Figure 3.2.4 shows the response surface for ICarr as a function of particle size and bin filling.

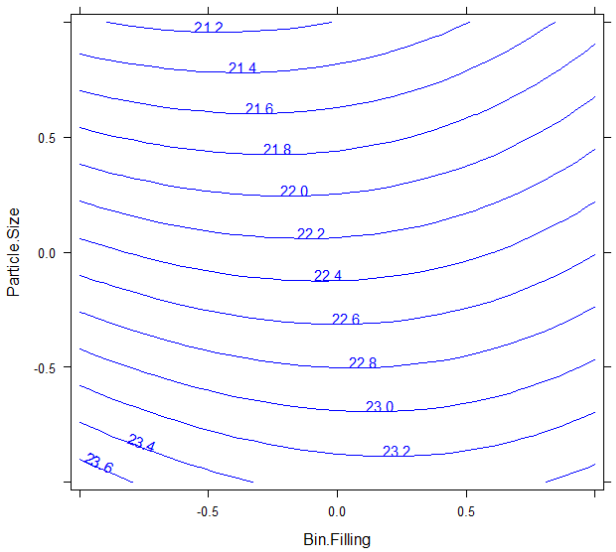


Figure 3.2.4: Response surface for ICarr in function of particle size and bin filling

From Figure 3.2.4, it is worth noting that the importance of particle size covers any possible effect of the other variables (in this case bin filling, although no other variable is significant by itself). It is clearly stated that the best level of particle size of MCC for minimizing ICarr is the highest (100 μm). Figure 3.2.5 shows the other significant effect: the interaction between bin filling and mixing time.

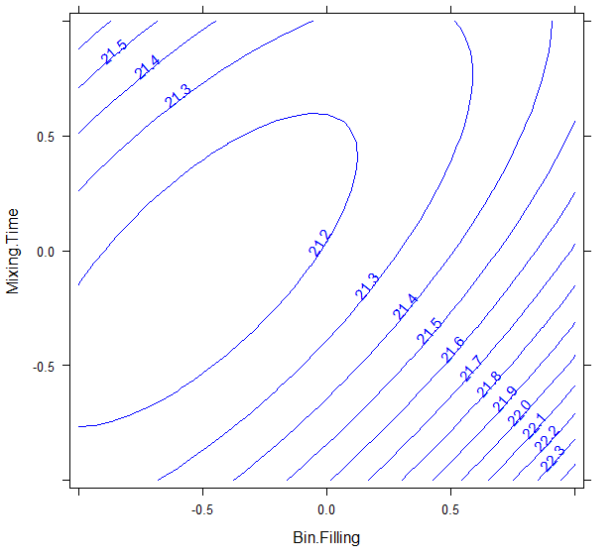


Figure 3.2.5: Response surface for ICarr as a function of bin filling and mixing time

Figure 3.2.5 (that is reported at level +1 of particle size), shows a minimum for ICarr in a region around the central level of mixing time (20 min) and comprised between the lowest and the central level of bin filling (30-52.5%). It also confirms manufacturers recommendation to not fill too much the bin blender for a better mixing.

NIR data

The following Figure 3.2.6 shows an example of NIR spectra pattern for one experiment (DOE 43, the first carried out).

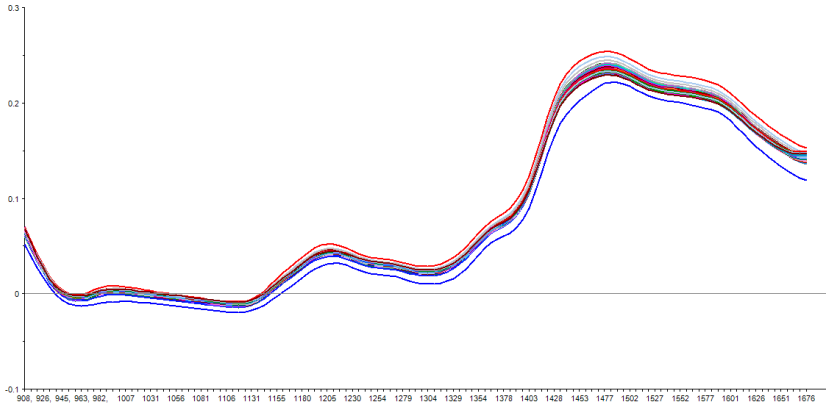


Figure 3.2.6: Example of NIR pattern

In Figure 3.2.6, it is interesting to note that the lower blue line and the higher red line correspond respectively to first and second rotation (125 total rotations were carried out in this experiment). The

difference from the other spectra, visible on first sight, is due to the fact that, at the beginning of mixing, powders are still separated ("sandwiched"), thus not totally comparable with the following rotations, in which the mixing begins. Although this is an obvious remark, it demonstrates that a visual inspection of data, also without any chemometric computation, could give some information about the dataset. Moreover, it confirms that NIR spectroscopy can be a good analytical tool for studying the mixing progress, even without knowing which species produce the NIR signal.

On these NIR patterns, MBM analysis was carried out. All MBM computations were carried out with a window of 10 spectra and a step of 1. Figure 3.2.7 shows moving block mean (MBM) and standard deviation (MBSD) for DOE 1 (because it has some interesting characteristics, but its spectra are a bit more confusing than that of DOE 43).

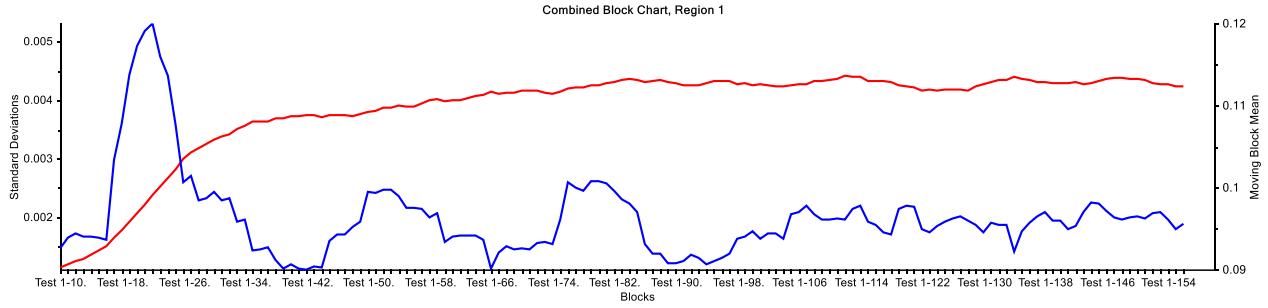


Figure 3.2.7: MB mean (red line) and MBSD (blue line) for DOE 1 experiment

As already stated before, MBSD is especially useful to understand the experimental behavior: if its graph is flat, and no discontinuity points are present, the experiment is going on without problems. However, the blue line of Figure 3.2.7, representing MBSD, shows three peaks on the left part of the graph. The first one, which is also the highest, is not a real problem: it can be easily interpreted by considering that it comes from the first rotations of the blender, when powders are starting to be mixed. It is not strange that, in the starting period, NIR spectra have a "great" variability. The other two peaks, however, may indicate some problem. In fact, the machine expert following the experiment said, even before looking at these results, that during that mixing some lumps were formed inside the blender. This is another demonstration that NIR is useful to in-line follow the mixing process (and that the chemometric analysis goes in the right way).

Before performing any computation, NIR spectra could be somehow pre-treated; however, there are no general guidelines about which pre-treatment should be used. Figure 3.2.8 shows the same MB mean and MBSD for DOE 1 experiment, but pre-treated by Savitzky-Golay (SG) first derivative, standard normal variate (SNV), and SG first derivative followed by SNV.

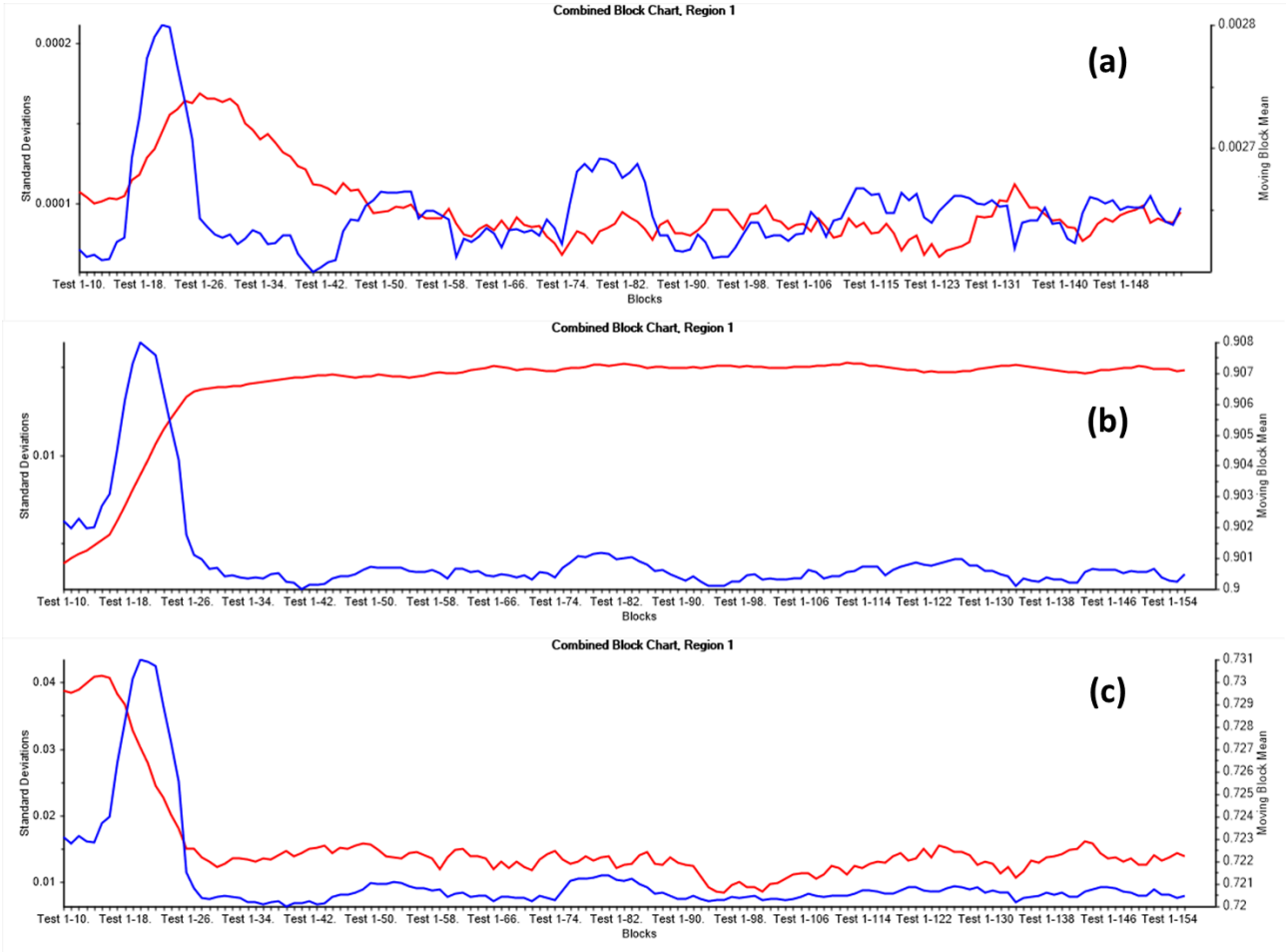


Figure 3.2.8: MBM (red line) and MBSD (blue line) for DOE 1 spectra pre-treated by a) SG first derivative; b) SNV; c) SG first derivative followed by SNV

Figure 3.2.8 shows that the use of SG first derivative enhances a bit the importance of MB mean, because here it also shows a peak at the beginning, while, with the original data, it only shows an increment at the beginning, and then it has a flat graph. SNV, instead, both with and without SG derivative, produced a too optimistic situation: in fact, the two peaks indicating lumps become very low, and it is difficult to distinguish them from the "baseline" of the graph. Therefore, it was decided to carry out all MBM analyses with original data, without any pre-treatment, because also the improvement given by SG derivative seems not so important to justify the addition of variability given by pre-treatment. In appendix B, MBM analyses for all DoE experiments are reported.

Some analyses (as for example DOE 28, reported in appendix B) show strong discontinuity points, generally due to a single biased spectrum, which has an effect on ten windows, producing a sort of square peak. Such a problem may arise, for instance, if a lump of material remains attached to the NIR probe from the previous rotation (which could indicate in any case a mixing problem). However, removing the biased spectrum, which, in general, is detectable also by a visual inspection, brings MBSD and MBM back to a normal behavior. From the graphs reported in appendix B, it is also interesting to note that the processes that seem to have some problems are mainly DOE 19, DOE 28, and DOE 37. All these samples (together with the already

cited DOE 1) have in common the bin filling and the rotation speed at the lowest levels -1. This could indicate that such combination of levels may bring, in general, to some mixing problems.

MBM analysis, in a daily routine, might also be useful to understand when to stop the mixing: if a discontinuity point is present after a certain time, it could mean that the blending was too long and has to be stopped before. Powders, indeed, may undergo a segregation process, if blending time is too long or experimental conditions are not optimal [20].

A control chart should be built with a set of experiments carried out with the same conditions and that gave good and replicable results. During this work, no particular criticisms were observed in any of the final products, and HPLC analyses showed that all of them were well mixed by the blender. However, experimental conditions were decided based on a DoE, so these are all different. Thus, in order to have a "homogeneous" dataset, experiments were divided into six subsets based on their level of particle size and bin filling, which seemed two of the most significant factors for DoE. Moreover, in this case, NIR spectra were pre-treated by SG first derivative followed by SNV, in order to further homogenize them, enhancing at the same time (by deriving) the differences. Several control charts may be created from each dataset, some examples are reported below, with the aim of explaining the method and to give some details.

A typical control chart created with NIR spectra obtained in the present work is shown in Figure 3.2.9. Samples with bin filling at level +1 and particle size +1 were used (nine samples in total, 2759 NIR spectra). As already stated, control charts are created independently from scores of each PLS-factor, Figure 3.2.9 reports the first three.

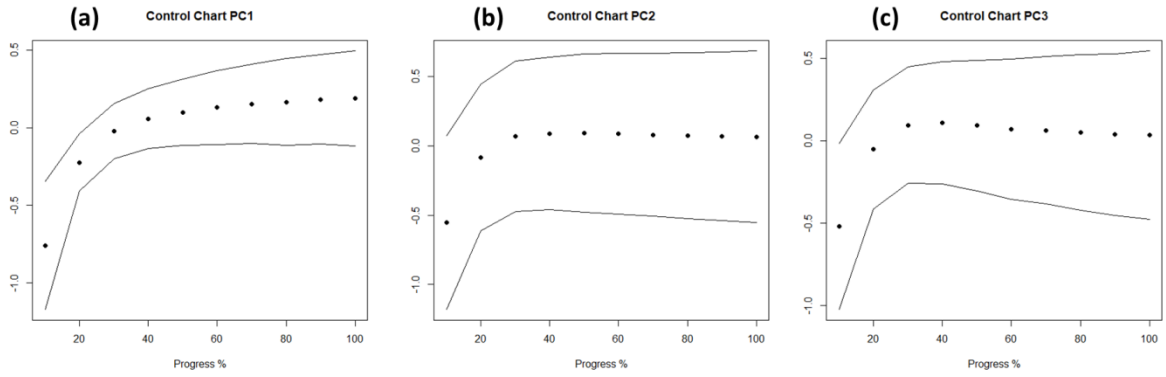


Figure 3.2.9: Example of control chart for a) Factor 1; b) Factor 2; c) Factor 3. Points are mean values of scores-blocks, lines represent mean ± standard deviation, the limits of acceptability region

The main difference with respect to univariate control chart is that the limits of acceptability are dynamic and change for each considered point, also because points represent different PP%, thus different analysis times, while, in general, a control chart is used for evaluating a final product. In this case, PP% was divided in subsets of 10% each, so ten points are present in these charts, but points may be augmented by reducing PP%-steps (increasing also the chart complexity), in order to have a more detailed vision of the process. At the same time, the acceptability region may be enlarged by using multiples of standard deviation (in this case, one standard deviation was used because there is already much variability in the original data coming

from a DoE). On these control charts, experiments external to the dataset can be projected. Figure 3.2.10 shows the projection of five batches with different combinations of levels of bin filling and particle size, but at 0 level of rotation speed and mixing time.

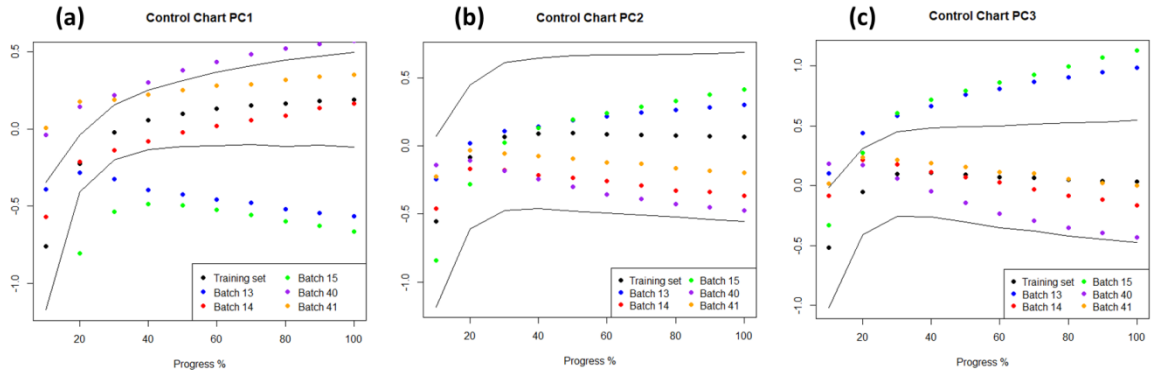


Figure 3.2.10: External samples projection on control charts. a) Factor 1; b) Factor 2; c) Factor 3

As it can be seen from Figure 3.2.10, samples have different behavior on different PCs: batches 13 and 15 (blue and green points), for instance, are outside acceptability for factor 1 and factor 3, except at the beginning of experiments, while they are totally inside for factor 2. Moreover, batch 14 (red points), which, compared to training set samples, has a totally different combination of levels of bin filling and particle size (0 and -1 respectively), so it was expected to be considered outside acceptability, is, instead, always inside. This is to demonstrate that, for such analysis, it is very important to have a trustable training set (and probably a group of experiments derived from DoE is not suitable) and to properly evaluate the PLS model to decide which PLS-factor is the more useful for the control chart. In this specific case, the cumulative explained variance reaches 70% only at factor 7 (at factor 3 it is 49%), so maybe it would be better to show that control chart. However, here the aim was simply to show the applicability of the proposed method to NIR spectra obtained by the blender probe, without pretending to be a conclusive work.

Conclusions

A full factorial design of experiments was carried out for the optimization of the mixing process in an industrial blender, and interesting results were obtained about the optimal parameters to improve homogeneity and flowability of the mixed powder.

Moreover, a possible chemometric approach for exploiting a NIR probe connected to the blender was shown. Although, particularly as for control chart, this work is not conclusive, it has been demonstrated that NIR spectra may be used to in-line control the mixing process, detecting possible faults in real-time. In fact, MBM analysis may be computed while the process is running, because it requires a window of ten (or also less) spectra at a time, and the step forward can be calculated while new spectra are acquired. Control charts may also be used to project new spectra in the proper PP% region during their acquisition. The hardest problem in this sense could arise from PLS computation, because the computational cost may be high if the

dataset becomes too large (the addition of one experiment brings to add 150-500 NIR spectra); however it has to be carried out only one time (or periodically, if the control chart has to be updated).

Therefore, a blender user may choose its operating conditions by DoE, particularly for flowability optimization. Then, a series of products with optimal characteristic may be produced, and their mixing may be monitored by MBM applied to NIR spectra registered in-line. Control charts can be created, while "traditional" control techniques are also applied for comparison and validation purposes. Then the blending process might be controlled in real-time by MBM and control chart projection.

From a chemometric point of view, the NIR data pre-treatment (for both MBM and control chart) and the PLS model interpretation still have to be optimized, in order to be able to create the best possible control chart. However, such optimizations can be made only with a proper set of trustable samples.

References

1. Rantanen J, Khinast J (2015) The Future of Pharmaceutical Manufacturing Sciences. J Pharm Sci 104:3612–3638 . doi: 10.1002/jps.24594
2. Jones B, Johnson RT (2007) Design and Analysis for the Gaussian Process Model. Qual Reliab Eng Int 23:517–543 . doi: 10.1002/qre
3. Vivancos J, Luis CJ, Costa L, Ortíz JA (2004) Optimal machining parameters selection in high speed milling of hardened steels for injection moulds. J Mater Process Technol 155–156:1505–1512 . doi: 10.1016/j.jmatprotec.2004.04.260
4. Fda (2012) Q8, Q9, & Q10 Questions and Answers -- Appendix: Q&As from Training Sessions (Q8, Q9, & Q10 Points to Consider). 301–827
5. Singh B, Kapil R, Nandi M, Ahuja N (2011) Developing oral drug delivery systems using formulation by design: vital precepts, retrospect and prospects. Expert Opin Drug Deliv 8:1341–1360 . doi: 10.1517/17425247.2011.605120
6. Khayet M, Zahrim AY, Hilal N (2011) Modelling and optimization of coagulation of highly concentrated industrial grade leather dye by response surface methodology. Chem Eng J 167:77–83 . doi: 10.1016/j.cej.2010.11.108
7. Bentz DP, Hansen AS, Guynn JM (2011) Optimization of cement and fly ash particle sizes to produce sustainable concretes. Cem Concr Compos 33:824–831 . doi: 10.1016/j.cemconcomp.2011.04.008
8. Ganguli R (2002) Optimum design of a helicopter rotor for low vibration using aeroelastic analysis and response surface methods. J Sound Vib 258:327–344 . doi: 10.1006/jsvi.2002.5179
9. Wu H, Khan MA (2009) Quality-by-Design (QbD): An integrated approach for evaluation of powder blending process kinetics and determination of powder blending end-point. J Pharm Sci 98:2784–2798 . doi: 10.1002/jps.21646
10. Shi Z, Cogdill RP, Short SM, Anderson CA (2008) Process characterization of powder blending by near-infrared spectroscopy: Blend end-points and beyond. J Pharm Biomed Anal 47:738–745 . doi: 10.1016/j.jpba.2008.03.013
11. Kim, KM, Kim, GT, Kang JS (2016) Design of experiments for wet granulation of valsartan and pravastatin fixed-dose combination tablet. Asian J Chem 28:2759–2763 . doi: https://doi.org/10.14233/ajchem.2016.20115
12. Patil AS, Pethe AM (2013) Quality by design (QbD): A new concept for development of quality pharmaceuticals. Int. J. Pharm. Qual. Assur. 4:13–19
13. Sudah OS, Coffin-Beach D, Muzzio FJ (2002) Effects of blender rotational speed and discharge on the homogeneity of cohesive and free-flowing mixtures. Int J Pharm. doi: 10.1016/S0378-5173(02)00377-0
14. Mittal B (2017) How to Develop Robust Solid Oral Dosage Forms from Conception to Post-Approval. How to Dev Robust Solid Oral Dos Forms from Concept to Post-Approval. doi: 10.1016/B978-0-12-804731-6.00004-2

15. Carr RL (1965) Evaluating flow properties of solids. Chem Eng. doi: 10.1016/j.jaerosci.2007.10.003
16. Emery E, Oliver J, Pugsley T, et al (2009) Flowability of moist pharmaceutical powders. Powder Technol 189:409–415 . doi: 10.1016/j.powtec.2008.06.017
17. Wu Y, Jin Y, Li Y, et al (2012) NIR spectroscopy as a process analytical technology (PAT) tool for on-line and real-time monitoring of an extraction process. Vib Spectrosc 58:109–118 . doi: 10.1016/j.vibspec.2011.10.006
18. Rosas JG, Blanco M, Santamaría F, Alcalà M (2013) Assessment of chemometric methods for the non-invasive monitoring of solid blending processes using wireless near infrared spectroscopy. J Near Infrared Spectrosc 21:97–106 . doi: 10.1255/jnirs.1041
19. Wold S, Kettaneh N, Fridén H, Holmberg A (1998) Modelling and diagnostics of batch processes and analogous kinetic experiments. Chemom Intell Lab Syst 44:331–340 . doi: 10.1016/S0169-7439(98)00162-2
20. Chaudhuri B, Mehrotra A, Muzzio FJ, Tomassone MS (2006) Cohesive effects in powder mixing in a tumbling blender. Powder Technol 165:105–114 . doi: 10.1016/j.powtec.2006.04.001

CHAPTER 4: NET ANALYTE SIGNAL

Introduction

The idea of Net Analyte Signal (NAS) was firstly proposed in 1980 by Boumans [1] in order to solve a problem of spectral interferences in ICP-AES. However, the mathematical development of NAS is due to Lorber [2], who in 1986 introduced the concept of NAS as "the part of the signal that is orthogonal to the spectra of the other components" [2], and proposed a way to compute figures of merit for this field. Since then, up to 2018, 240 papers have been published about further developments or applications of NAS.

The basic idea of NAS is extracting that part of the analytical signal (which can be whatever type of spectral signal) that is directly related and only due to the analyte of interest [3]. Therefore, in general, the NAS of the *k*th analyte of interest in a mixture may be computed as the part of the spectrum orthogonal to the contribution of the other coexisting species [3]. Mathematically speaking, "orthogonal" means also "independent", thus the NAS of the *k*th analyte is also independent from the signal of the other species.

Therefore, the NAS, due to its dependence only on the analyte of interest, can be used for quantification problems. In fact as the original signal from which it is calculated, NAS should increase linearly with the concentration of the analyte. It was again Lorber [3] who developed this idea in 1997. He also solved another important problem: in most of the cases, the signal due to the other species (orthogonal to NAS) is partially or totally unknown; therefore, it also has to be calculated.

A simple graphical representation of how the NAS algorithm works is represented in Figure 4.0.1, which is simplified because it does not take error vectors into account. In this figure, an example spectrum is reported as the black diagonal vector ($\boldsymbol{s_i}$). This vector is mathematically decomposed into the orthogonal blue (horizontal) and red (vertical) vectors. These represent the calculated signal due to interfering species and the net signal due to the analyte of interest only ($\boldsymbol{s_i}^*$) respectively.

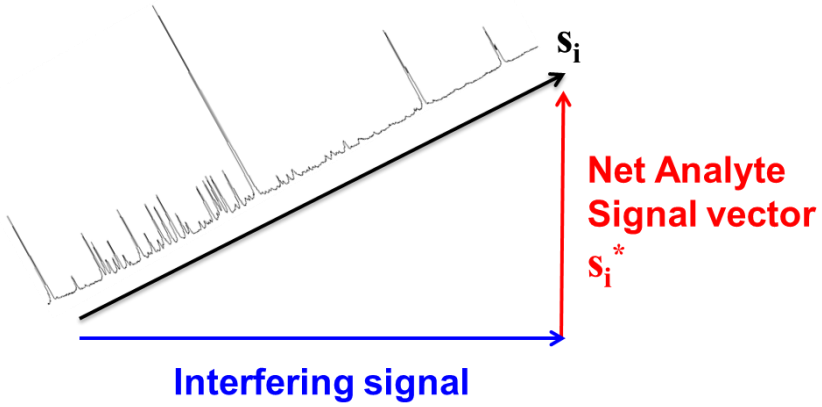


Figure 4.0.1: graphical representation of the NAS algorithm

Once obtained the NAS vector, its Euclidean norm can be used as a pseudo-univariate signal for quantification purposes [4].

The NAS procedure for quantification was firstly derived for the external-standard method. However, an interesting field of application is that of the standard addition method (SAM). The advantage of SAM, with respect to external standard, is that the matrix in which the analyte is dissolved can be totally unknown. Possible problems of overlapping signals between the analyte of interest and the other species in the matrix can be solved by the application of NAS (NASSAM). The first study which applied this procedure was that of Hemmateenejad [5], who also developed the mathematical treatment.

It is also interesting that, since its first development for ICP-AES [1], the NAS procedure has been successfully applied to several analytical techniques: UV-Vis spectroscopy [6], NMR [7], NIR spectroscopy [8, 9], spectrofluorimetry [10]. The NASSAM method has also been more recently applied to several fields: UV-Vis spectroscopy [5], polarography [11], spectrofluorimetry [12].

However, most of the application reported in the literature concern the analysis of solutions, where it is easier, with respect to solid samples, to apply the SAM. In fact, in preparing the liquid standard-added samples, there is not the problem of mass balance and, above all, if the analyte is soluble in the chosen solvent, the samples are always homogeneous. The main topic of the present work is to develop and apply the NASSAM method to solid samples, with the aim of quantifying both "artificial" analytes (in samples entirely prepared in laboratory, as feasibility studies), and real analytes in real samples. Hence, the problem of strong matrix effect in solids is faced. Application to different analytical techniques will be presented, showing the applicability of the developed method in different analytical fields.

The work is organized as follow. First of all, the developed algorithm for the present work will be explained. Next, an application of NASSAM to liquid home-made samples, analyzed by UV-Vis spectroscopy, will be presented. These samples, although not relevant for the topic of this work, are used to link this work with what is already present in the literature, and to show the applicability of the developed NAS procedure. Then, an application to home-made samples of beeswax, analyzed by Raman spectroscopy, and another one applied to home-made saffron samples, analyzed by gas-chromatography, will be shown. Next, two applications to pharmaceutical samples analyzed by X-ray powder diffraction, one home-made and one real, will be presented. Finally, an application to sediments (real samples) analyzed by NIR spectroscopy in ATR mode will be presented.

NAS Algorithm

The algorithm used in the present work starts from the ones of Lorber [3] and Hemmateenejad [5].

Obviously, the starting point is to prepare and analyze the standard added samples. Preparation and analysis of samples will be shown case by case for each type of sample in the proper following section. In general, once analyzed all the samples, a data-matrix is obtained, called $\boldsymbol{S}$, in which the rows (or vectors $\boldsymbol{s}_i$) represent the samples and the columns represent the spectral variables (e.g. wavelengths). The matrix $\boldsymbol{S}$ will have dimensions n x v, where n is the total number of i samples ($1 \leq i \leq n$), and v is the number of spectral

variables. Besides, to each sample an added concentration value is associated, and all these values form the $\boldsymbol{c_{add}}$ vector.

The first step of the mathematical procedure is to perform a partial least square (PLS) regression [13]. PLS is a well-known chemometric modeling method that computes a regression between a matrix of independent variables (in this case $\boldsymbol{S}$) and a matrix, or vector, of dependent variables ($\boldsymbol{c_{add}}$). Put simply, PLS searches the directions of maximum variance of the $\boldsymbol{S}$ matrix (which means that it computes principal components, PCs, as PCA) and corrects (rotate) them to achieve the maximum possible correlation with the dependent variable. In such a way, the final regression model takes into account both the variance included in independent and dependent variables, and their covariance, aiming to describe the original data as better as possible. The advantages of PLS over other regression methods [13] (as for example the traditional multiple linear regression, MLR) is that it gives good results even when the number of variables exceeds the number of samples ($v >> n$) and when the variables are strongly co-linear. Both of these problems are present when spectral variables are used as independent variables.

PLS computation [14] starts by performing a singular value decomposition (SVD) [15] on the cross-product matrix $\boldsymbol{Q} = \boldsymbol{S}^t \boldsymbol{c_{add}}$ (where superscript *t* indicates matrix transpose). The first left singular vector of SVD can be seen as the direction of maximal variance in the cross-product $\boldsymbol{Q}$ and is indicated as "weight", $\boldsymbol{w}$. This vector is used to calculate the scores ($\boldsymbol{t}$) of the first PLS-factor (the equivalent of the first PC in PCA)

$$\boldsymbol{t} = \boldsymbol{S}\boldsymbol{w} \qquad \text{(eq. 4.0.1)}$$

This vector is used to calculate the loadings for both independent ($\boldsymbol{p}$) and dependent ($\boldsymbol{q}$) variables

$$\boldsymbol{p} = \boldsymbol{S}^t\boldsymbol{t}/(\boldsymbol{t}^t\boldsymbol{t}) \qquad \text{(eq. 4.0.2)}$$

$$\boldsymbol{q} = \boldsymbol{c_{add}}^t\boldsymbol{t}/(\boldsymbol{t}^t\boldsymbol{t}) \qquad \text{(eq. 4.0.3)}$$

Finally, the original data matrix is deflated: this means that the information included in this first factor is subtracted from the original data

$$\boldsymbol{E} = \boldsymbol{S} - \boldsymbol{t}\boldsymbol{p}^t \qquad \text{(eq. 4.0.4)}$$

$$\boldsymbol{f} = \boldsymbol{c_{add}} - \boldsymbol{t}\boldsymbol{q}^t \qquad \text{(eq. 4.0.5)}$$

$\boldsymbol{E}$ and $\boldsymbol{f}$ are then used iteratively in place of $\boldsymbol{S}$ and $\boldsymbol{c_{add}}$: the PLS cycle starts again from eq. 4.0.1 to compute the second PLS-factor, and, when it arrives at eq. 4.0.4 and 4.0.5, it calculates new $\boldsymbol{E}$ and $\boldsymbol{f}$ matrices that are used to compute the third factor. The cycle goes on until the operator decides that the number of PLS-factors is sufficient, or until all the possible factors are computed. The vectors $\boldsymbol{w}$, $\boldsymbol{t}$, and $\boldsymbol{p}$ are collected during the computation into the weights ($\boldsymbol{W}$), scores ($\boldsymbol{T}$), and loadings ($\boldsymbol{P}$) matrices.

The regression coefficients are calculated starting from the scores, firstly by performing a regression between these and the independent variables (to obtain $\boldsymbol{\beta}$), then by converting them to the original variables ($\boldsymbol{B}$):

$$\boldsymbol{\beta} = (\boldsymbol{T}^t\boldsymbol{T})^{-1}\boldsymbol{T}^t\boldsymbol{c_{add}} \quad \text{(eq. 4.0.6)}$$

$$\boldsymbol{B} = \boldsymbol{S\beta} \quad \text{(eq. 4.0.7)}$$

PLS can be used to remove part of the analytical noise from data. In fact, it can be decided to keep a number *A* of PLS-factors and discard the others, in order to keep only the information retained by these first factors. In general, the higher factors report spurious information or noise, so these can be discarded. The optimal number of PLS-factors depends on the dataset; a general rule is to stop to that factor that minimizes the root mean squared error (RMSE). However, sometimes RMSE continuously decreases by adding factors; thus, the choice for the present work has been to compute the NAS procedure with all the possible components and to select the optimal component only at the end, taking into account not only RMSE.

Matrices $\boldsymbol{T}$ and $\boldsymbol{P}$ from PLS are then used to start the NAS procedure [5]. The first step consists in rebuilding the $\boldsymbol{S}$ matrix ($\boldsymbol{S_{reb}}$) from scores and loadings

$$\boldsymbol{S_{reb}} = \boldsymbol{T}_A\boldsymbol{P}_A^t + \boldsymbol{m} \quad \text{(eq. 4.0.8)}$$

$\boldsymbol{m}$ is the mean vector of the original $\boldsymbol{S}$ matrix and it is added to the product to remove the automatic mean centering performed by the PLS algorithm. Subscripts "*A*" indicate that this computation can be performed for each *A* number of PLS-principal components. The advantage of using $\boldsymbol{S_{reb}}$ instead of $\boldsymbol{S}$ is that part of the analytical noise has been removed by PLS. $\boldsymbol{S_{reb}}$ is then used to mathematically calculate the matrix signal, which is that part of the analytical signal due to all the species present in the mixture, except the *k*th analyte of interest (this step is called "rank annihilation")

$$\boldsymbol{S_{-k}} = \boldsymbol{S_{reb}} - \alpha\boldsymbol{c_k}\boldsymbol{s}^t \quad \text{(eq. 4.0.9)}$$

where

$$\boldsymbol{c_k} = \boldsymbol{S_{reb}}\boldsymbol{S_{reb}^+}\boldsymbol{c_{add}} \quad \text{(eq. 4.0.10)}$$

$$\alpha = 1/(\boldsymbol{s}^t\boldsymbol{S_{reb}}\boldsymbol{c_k}) \quad \text{(eq. 4.0.11)}$$

In equations 4.0.9, 4.0.10, and 4.0.11 $\boldsymbol{s}$ is a vector which is recommended to contain information about the pure analyte. It can be calculated as a linear combination of the row of $\boldsymbol{S}$. However the simpler way to obtain this vector is to use the spectrum (or its equivalent signal) of the pure analyte of interest (or a mean of some replicates of such signal). $\boldsymbol{c_k}$ is the added concentration vector projected onto the PLS space, and α is a scalar that is used as a correction factor; superscript "+" indicates the Moore-Penrose pseudo-inverse of a matrix [16].

What happens in practice in equation 4.0.9 is that to each spectrum a quantity is subtracted, which is calculated as the signal of the pure analyte "corrected" by multiplying it for its added concentration in the sample and for α. In such a way, the signal of the analyte of interest (or at least most of it) is removed from

the total signal, and what remains in $\boldsymbol{S_{-k}}$ is only that part of the signal due to the other constituents of the sample (i.e. the matrix or the interfering signal, the blue line in Fig. 4.0.1).

At this point, a problem arises regarding the zero-added sample. Although the $\boldsymbol{c_{add}}$ vector is modified to $\boldsymbol{c_k}$, the differences between the two vectors are slight. Therefore, the value of $\boldsymbol{c_k}$ for the zero-added sample is not so different from 0. This brings the product $\alpha \boldsymbol{c_k s^t}$ to be negligible, and so $\boldsymbol{s_{-k,0}} \cong \boldsymbol{s_{reb,0}}$. The effect of this is that the zero-added samples have, sometimes, an anomalous behavior with respect to the other samples. This anomaly affects the final standard addition line (by reducing R^2, even dramatically). Hence, in some of the cases shown in the present work, the zero-added sample will be removed from the computation from the beginning.

Anyway, $\boldsymbol{S_{-k}}$ can be used to compute a projection matrix ($\boldsymbol{H}$) that will be used to calculate that part of the original signal that is orthogonal to the interfering signal, as shown in Fig. 4.0.1

$$\boldsymbol{H = I - S_{-k} S^{+}_{-k}} \qquad \text{(eq. 4.0.12)}$$

where $\boldsymbol{I}$ is an identity matrix with proper dimensions (in this case *v* x *v*).

At this point, the original spectra can be projected on the NAS space, spanned by $\boldsymbol{H}$, by simply multiplying the *i*th spectrum ($\boldsymbol{s_i}$) for $\boldsymbol{H}$

$$\boldsymbol{s^*_i = H s_i} \qquad \text{(eq. 4.0.13)}$$

The $\boldsymbol{s^*_i}$ vectors are the net analyte signal ones. The information reported by them is orthogonal to the interfering signals, so it is only related to the analyte of interest. Therefore, the Euclidean norm of these vectors ($\|\boldsymbol{s^*_i}\|$) can be used as a pseudo-univariate signals to create a classical standard addition line, with general equation $\|\boldsymbol{s}^*\| = a \cdot \boldsymbol{c_{add}} + b$, from which the concentration of the analyte of interest in the original sample can be directly extrapolated by:

$$c_E = b/a \qquad \text{(eq. 4.0.14)}$$

As already stated before, a different solution (c_E) is obtained for each *A*th PLS-factor. The selection of the optimal factor can be made upstream by considering only RMSE, or it can be made at this point by taking that factor that optimizes both RMSE and the determination coefficient (R^2) of the standard addition line. If a reliable reference value for c_E is present, it could be chosen also the PLS-component that optimizes c_E. However, in most of the real cases, such a reference is not present, thus it is better to try to optimize only RMSE and R^2. In any case, in order to not include too much noise in the model, it is preferable to use a value of *A* as low as possible, even if it often happens that RMSE decreases and R^2 increases continuously by adding factors without any minimum and maximum. There are also cases in which RMSE and R^2 are in conflict. For example at some *A*th component there may be a minimum (often relative) of RMSE with a bad R^2 (< 0.9). In such cases it is often preferable to optimize R^2 "sacrificing" a bit the optimization of RMSE.

Hence, and a greater value of *A* is taken; alternatively, if possible, the best compromise between the two parameters (again with the lowest possible *A*) has to be found.

Improvements to NAS Algorithm

Depending on the specific case, the so far presented algorithm requires some improvements in its computation, in order to optimize the final results. First of all, as it will be shown later, the results are strongly dependent on the data-pretreatment carried out before the application of NAS. Therefore, for each specific case presented, the required data-pretreatment will be shown.

In this chapter, some specific improvements for the NAS algorithm will be presented.

The most important was developed by Ferré [17], and regards a correction applied to matrix $\boldsymbol{S}$ before projecting the spectra on the NAS space. This correction has been introduced to include also in $\boldsymbol{S}$ the information derived only from the *A* selected latent variables. This is obtained by converting $\boldsymbol{s_i}$ to $\boldsymbol{s_{i,A}}$ by:

$$\boldsymbol{s_{i,A}} = \boldsymbol{P_A S_{W,A}^t s_i} \qquad \text{(eq. 4.0.15)}$$

The matrix $\boldsymbol{S_{W,A}}$ is calculated by [18]:

$$\boldsymbol{S_{W,A}} = \boldsymbol{W_A}(\boldsymbol{P_A W_A})^+ \qquad \text{(eq. 4.0.16)}$$

Now $\boldsymbol{s_{i,A}}$ belongs to the space spanned by the columns of $\boldsymbol{P_A}$, therefore eq. 4.0.13 can be modified to include this information also in the NAS vectors

$$\boldsymbol{s_{i,A}^*} = \boldsymbol{H s_{i,A}} \qquad \text{(eq. 4.0.17)}$$

Another improvement to the algorithm was developed by Bro [19] who derived a method that does not require assumptions about the structure of data and does not require the computation of the matrix $\boldsymbol{S_{-k}}$. Therefore, Bro proposed to compute the $\boldsymbol{H}$ matrix by:

$$\boldsymbol{H} = \boldsymbol{b_A}(\boldsymbol{b_A^t b_A})^{-1}\boldsymbol{b_A^t} \qquad \text{(eq. 4.0.18)}$$

where superscript "-1" indicates matrix inversion and $\boldsymbol{b_A}$ is the PLS-coefficients vector for *A* factors. This method skips the steps from eq. 4.0.8 to eq. 4.0.11 and substitutes eq. 4.0.12 passing directly from the PLS model to the computation of the projection matrix. Then, NAS vectors are calculated as in equation 4.0.13.

It is worth noting that the original algorithm of Lorber with the addition of Ferré variant, and the Bro method give exactly the same results (for *A* PLS-factors). Thus, in the present work two possible ways to obtain NAS have been studied: by the original Lorber procedure, and by the Ferré-Bro method. Most of the results that will be shown afterward have been obtained by the Ferré-Bro procedure, but the method used will be made explicit in any presented case.

Hemmateenejad [5] proposed also another improvement that has been demonstrated to be useful for liquid samples only, because it is based on the Beer-Lambert law, which probably is not totally applicable to the analytical methods applied for solid samples. It is a correction applied to the $\boldsymbol{S}$ matrix and c_{add} vector at the

beginning of the procedure. The basic idea is that the spectra in $\boldsymbol{S}$ ($\boldsymbol{s_i}$) are the sum of the signal of the zero-added sample (the original one, $\boldsymbol{s_0}$) and that of the standards added to the sample ($\boldsymbol{S_{sm}}$): $\boldsymbol{S = s_0 + S_{sm}}$; therefore, the signal of the standard can be obtained by:

$$\boldsymbol{S_{sm} = S - s_0} \qquad \text{(eq. 4.0.19)}$$

Moreover, considering that also the matrix signal is addictive and that the addition of pure analyte (in solution) does not influence it, in $\boldsymbol{S_{sm}}$ the matrix signal has been removed. Thus, $\boldsymbol{S_{sm}}$ can be decomposed according to the Beer-Lambert law by multiplying the signal of the pure analyte for a proper concentration vector:

$$\boldsymbol{S_{sm} = c_s s} \qquad \text{(eq. 4.0.20)}$$

This vector $\boldsymbol{c_s}$ is the correction applied to the original $\boldsymbol{c_{add}}$ for this procedure, and it can be obtained by inversion of eq. 4.0.20

$$\boldsymbol{c_s = S_{sm} s (s^t s)^{-1}} \qquad \text{(eq. 4.0.21)}$$

Once obtained $\boldsymbol{S_{sm}}$ and $\boldsymbol{c_s}$, these are used instead of $\boldsymbol{S}$ and $\boldsymbol{c_{add}}$ in the NAS procedure, already for computing the PLS model. It is obvious that such a correction can be used only when the signal of all the present species may be considered addictive, which can be a possible approximation for solutions analyzed by UV-Vis spectroscopy (and only in the linear range of Beer-Lambert law). This correction cannot be considered for solid samples, where the addition of standards also modifies the concentrations of the other species (unless specific precautions are used), especially when the analytical techniques that do not follow the Beer-Lambert law (e.g. X-ray powder diffraction) are employed. Therefore, the "Hemmateenejad-correction" will be used in the present work only in the application of NASSAM to liquid home-made samples.

Figures of Merit

The figures of merit are those parameters that indicate if a chemometric model is reliable. For the NASSAM model, two of these figures of merit are the already cited RMSE of the PLS model for *A* factors, and the determination coefficient, R^2, of the NAS standard addition line. The former has to be minimized, while the latter has to be as close as possible to 1.

Another important figure of merit that has to be taken into account is the standard deviation of the extrapolated value because it indicates the precision of that value: a low standard deviation indicates a high precision. If a reference value is present, the standard deviation is also useful to evaluate the accuracy, that is the "closeness" of the calculated to the expected value. In the present work, two possible ways of estimating the standard deviation of c_E ($s_{C,E}$) have been used. The first possibility is simply to compute the standard deviation in the final univariate NAS standard addition line [20] by:

$$s_{C,E} = \frac{s_{y/x}}{b}\sqrt{\frac{1}{n} + \frac{\overline{NAS}^2}{b^2 \sum_{i=1}^{n}(c_{add,i} - \overline{c_{add}})^2}} \qquad \text{(eq. 4.0.22)}$$

where $s_{y/x}$ is the RMSE of the standard addition line and $\overline{NAS}$ and $\overline{c_{add}}$ are the mean value of the calculated NAS Euclidean norms and of the added concentrations respectively. Such a computation, however, does not take into account all the multivariate analysis used to arrive at the NAS values: it uses them as these were simple univariate signals, losing most of their complexity. Therefore, this kind of estimation is not satisfactory.

A second way for estimating $s_{C,E}$ is by using the jackknife method [21]. The jackknife method performs the computation of the NAS algorithm by leaving one sample at a time out from the computation. In practice, a first NAS model is computed (with all the samples, either with Lorber or Ferré-Bro method), in order to find the optimal number *A* of factors and the c_E value. In a further step, all the computation is repeated starting from PLS and keeping the chosen *A* as the optimal number of components, but leaving out from the computation the first sample. Then the computation is repeated in the same way by reintegrating the first sample and leaving out the second, and so on. At the end of "jackknifing", as many c_E values as the number of samples are obtained (each representing a dataset in which the *i*th sample has been left out). The estimation of the standard deviation of the NAS extrapolated value is obtained as the standard deviation of this vector of c_E values. In this way, $s_{C,E}$ estimation is obtained by keeping into account at least part of the variability due to NAS computation, because changing the original dataset, also by adding or subtracting a single sample, changes (sometimes also dramatically) the $\boldsymbol{H}$ matrix, hence strong differences in $\boldsymbol{s}^*_{i,A}$ and their Euclidean norms may be observed. The jackknife method may also be used by leaving out more than one sample at a time; however, due to the relatively small number of samples, in this work the jackknife has been applied in "leave-one-out" mode. From the mean of the jackknife c_E values, also an estimation of the extrapolated value can be derived; however, numerically, no significant differences have ever been observed between this mean value and the original extrapolated one, so it has been preferred to show the latter as the result of NAS computation.

Two other important figures of merit are sensitivity (*Sn*) and limit of detection (*LoD*).

Sn of a calibration line is, in general, its slope; therefore, the *Sn* of the NAS standard addition line could be considered simply as its *b*. However, Bro [19] suggested that *Sn* can be computed from the vector of regression coefficients of PLS at the chosen *A* component as:

$$Sn = \frac{1}{\|\boldsymbol{b}_A\|} \qquad \text{(eq. 4.0.23)}$$

As for $s_{C,E}$, such computation takes into account the multivariate nature of the NAS procedure, also because, when using the Ferré-Bro method, the vector $\boldsymbol{b}_A$ is the only one used to calculate the $\boldsymbol{H}$ matrix. Therefore, when presenting NAS results, both the slope of the final standard addition line and *Sn* computed by eq.

4.0.23 will be shown. It will be interesting to note that very slight differences are in general obtained for these two values.

The limit of detection is defined as "the minimum quantity of analyte that shows a signal significantly different from the blank" or "the analyte concentration giving a signal equal to the blank signal, plus three standard deviations of the blank" [22]. In practice, it is the lower detectable concentration by any method. For a calibration line in the univariate field, estimation of *LoD* is in general obtained by

$$LoD = 3\frac{s_{y/x}}{b} \qquad \text{(eq. 4.0.24)}$$

However, it is still not so well established how to compute *LoD* for a multivariate model, and several methods to extend eq. 4.0.24 to the multivariate case have been proposed [23, 24].

Even for the NAS procedure, there is not an univocal way to estimate *LoD*. Lorber [3] proposed a method that requires the numerical computation of a further statistical parameter. A simpler way to estimate *LoD* (and the one chosen for the present work) has been proposed again by Hemmateenejad [5], who partially extended eq. 4.0.24 to the NAS procedure:

$$LoD = 3\frac{\|\boldsymbol{\varepsilon}\|}{Sn} \qquad \text{(eq. 4.0.25)}$$

The $\boldsymbol{\varepsilon}$ vector (of which the Euclidean norm is calculated) is obtained by registering spectra of several blank samples (e.g. empty sample holders) and by projecting them onto the NAS space (by multiplying these spectra for the $\boldsymbol{H}$ matrix). $\boldsymbol{\varepsilon}$ is the mean of the so obtained NAS vectors. In this procedure, the blanks are treated as unknown samples and projected onto the NAS space spanned by the samples of interest. Their signals are used as the minimum signal detectable by the technique. The multiplication by 3 (and dividing by *Sn*) brings to the minimum concentration that gives a signal significantly different from the blank, which is the definition of *LoD*. For the external standard method, the *LoD* should be lower than the lower standard concentration analyzed. In SAM case, the lower concentration is the extrapolated one, so the *LoD* should be lower than c_E.

All the computations for the present work were performed with the software "R" version 3.4.3 (R Core Team, Vienna, Austria) [25]. PLS was computed with the package "pls" [26]. Moore-Penrose pseudo-inverse was computed with the package "MASS" [16]. NAS algorithm and its extensions were written by the author, the R code is reported in Appendix C. Any signal pre-processing was carried out with the software "The Unscrambler" version 10.4 (CAMO, Oslo, Norway) and the resulting signals were imported in "R".

Further Considerations

Considering what has been said until now, there are two main critical points in using NAS procedure: the selection of both the method (between the original Lorber's or the Ferré-Bro variant, and sometimes also the Hemmateenejad variant could be introduced) and the number of PLS principal components. The R algorithm developed for the present work, reported in Appendix C, is thought for performing NAS with both

procedures, and with each PLS-factor for each procedure, in order to obtain all the possible results. Therefore, the selection of the best result is made *a posteriori* following the previously cited rules of optimization of RMSE and R^2.

As an example of the differences in the possible results, the following Table 4.0.1 reports the ones obtained by the two procedures as a function of the number of PLS-components in a specific case. This work regards the quantification of a home-made adulteration of paraffin in a beeswax sample and will be discussed in detail in Chapter 4.2. What is important for the present discussion is that the expected value is $1.5\%_{w/w}$. Standard deviations are not presented here.

		Lorber procedure		**Ferré-Bro procedure**	
PLS-factor	**RMSE**	**c_E ($\%_{w/w}$)**	**R^2**	**c_E ($\%_{w/w}$)**	**R^2**
1	4.50	172	-0.020	116	-0.060
2	4.20	12.9	0.108	60.8	0.523
3	2.70	7.72	0.408	52.1	0.757
4	1.31	3.86	0.878	42.3	0.956
5	0.713	2.58	0.976	50.5	0.997
6	0.646	2.26	0.979	42.2	1.000
7	0.645	2.09	0.985	41.7	1.000
8	0.646	1.45	0.984	41.8	1.000
9	0.646	0.871	0.997	42.2	1.000
10	0.646	0.775	0.997	42.2	1.000
11	0.646	0.483	0.997	42.2	1.000
12	0.646	0.379	0.997	42.2	1.000

Table 4.0.1: Lorber and Ferré-Bro procedures comparison for the example case. R^2 are referred to the final NAS regression line

Some comments can be made based on Table 4.0.1. First of all, the first factor has the highest RMSE and the worst predicted value for both procedures. A minimum of RMSE can be found at factor 7, a good result is found at this factor for the Lorber procedure, while it is still poor for Ferré-Bro (although with a better R^2). The extrapolated values decreases at higher components for Lorber, while they are close to each other (at least starting from factor 4) for Ferré-Bro. R^2 are always higher for Ferré-Bro.

All these trends can be observed in general when working with these NAS procedures in solid samples: the lower is RMSE, the better is the extrapolated concentration. Lorber c_E decreases by augmenting the number of factors, while Ferré-Bro values are "more stable". Usually Ferré-Bro procedure shows better R^2.

The stability of results and the better R^2 of Ferré-Bro method are probably due to the fact that (thinking to the Bro algorithm), it uses only the PLS-regression coefficients to calculate the ***H*** matrix and, then, NAS vectors. Therefore, in this situation, it is like that the "linearization" of samples, that brings to R^2 close to 1, is already made by PLS, and the following procedure takes advantage of the PLS work. Lorber procedure, instead, starts from scores and loadings; consequently, it cannot take advantage of the work of PLS. However, if the starting spectra do not have a good linear behavior (which means, for example, that peak intensities do not increase linearly with added concentration), like in this case, PLS model could not be optimal, and, therefore, Ferré-Bro procedure could bring to totally unexpected and erroneous results, while

Lorber method may be better. This is the main drawback found in the present work in using the NAS procedure (and the most problematic for validation): the choice of the optimal procedure is case-dependent, and, up to now, it is difficult to provide a general rule to *a-priori* understand which of the two procedures should be used.

Anyway, in the presented example case, the Lorber procedure is chosen because, in general, it gives results closer to the expected one. Then, the optimal PLS-factor is chosen "blindly" by looking at the minimum RMSE and the maximum R^2 (or the better compromise between the two), with the minimum number of factors, in order not to include noise. In this specific case, 7 factors were chosen, because the absolute minimum of RMSE (0.645) and a relative maximum of R^2 (0.985) is obtained. This brings to $c_E = 2.09\%_{w/w}$, which is in good agreement with the expected value. However, c_E could be improved by taking 8 components (1.45 $\%_{w/w}$), sacrificing a bit both RMSE and R^2. In general, for the present work, the "blind" selection of PLS-factors is preferred, in order to not add too much subjectivity to the final results.

References

1. Boumans PWJM (1980) ICP: D.c. arc in a new jacket? Spectrochim Acta Part B At Spectrosc 35:57–71 . doi: 10.1016/0584-8547(80)80053-X
2. Lorber A (1986) Error Propagation and Figures of Merit for Quantification by Solving Matrix Equations. Anal Chem 58:1167–1172 . doi: 10.1021/ac00297a042
3. Lorber a, Faber K, Kowalski BR (1997) Net Analyte Signal Calculation in Multivariate\rCalibration. Anal Chem 69:1620–1626 . doi: 10.1021/ac960862b
4. Faber NM (1998) Mean centering and computation of scalar net analyte signal in multivariate calibration. J Chemom 12:405–409 . doi: 10.1002/(SICI)1099-128X(199811/12)12:6<405::AID-CEM520>3.0.CO;2-8
5. Hemmateenejad B, Yousefinejad S (2009) Multivariate standard addition method solved by net analyte signal calculation and rank annihilation factor analysis. Anal Bioanal Chem 394:1965–1975 . doi: 10.1007/s00216-009-2870-1
6. Navarro-Villoslada F, Pérez-Arribas L V., León-González ME, Polo-Díez LM (1999) Matrix effect modelling in multivariate determination of priority pollutant chlorophenols in urine samples. Anal Chim Acta 381:93–102 . doi: 10.1016/S0003-2670(98)00701-6
7. Alam MK, Alam TM (2000) Multivariate-Analysis and Quantitation of O-17-Nuclear Magnetic-Resonance in Primary Alcohol Mixtures. SpectrochimActa Part A 56:729–738
8. Olesberg JT, Arnold M a, Hu SY, Wiencek JM (2000) Temperature-insensitive near-infrared method for determination of protein concentration during protein crystal growth. Anal Chem 72:4985–90
9. Boschetti CE, Olivieri AC (2001) Net analyte preprocessing: A new and versatile multivariate calibration technique. Analysis of mixtures of rubber antioxidants by NIR spectroscopy. J Near Infrared Spectrosc 9:245–254 . doi: 10.1255/jnirs.310
10. Espinosa-Mansilla A, Durán-Merás I, Galian R (2001) Simultaneous fluorimetric determination of pteridin derivatives: Comparison between synchronous, partial least-squares, and hybrid linear analysis methods. Appl Spectrosc 55:701–707 . doi: 10.1366/0003702011952622
11. Asadpour-Zeynali K, Majidi MR, Tahmasebpour M (2009) Net analyte signal standard addition method for the simultaneous determination of cadmium and nickel. J Serbian Chem Soc 74: . doi: 10.2298/JSC0907789A
12. Asadpour-Zeynali K, Bastami M (2010) Net analyte signal standard addition method (NASSAM) as a novel spectrofluorimetric and spectrophotometric technique for simultaneous determination, application to assay of melatonin and pyridoxine. Spectrochim Acta - Part A Mol Biomol Spectrosc 75:589–597 . doi: 10.1016/j.saa.2009.11.023
13. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: A basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130 . doi: 10.1016/S0169-7439(01)00155-1
14. Wehrens R (2011) Chemometrics with R. Multivariate Data Analysis in the Natural Sciences and Life Sciences

15. Golub GH, Reinsch C (1970) Singular value decomposition and least squares solutions. Numer Math 14:403–420 . doi: 10.1007/BF02163027
16. Venables WN, Ripley BD (2002) Modern Applied Statistics with S Fourth edition by
17. Ferré J, Faber NM (2003) Net analyte signal calculation for multivariate calibration. Chemom Intell Lab Syst 69:123–136 . doi: 10.1016/S0169-7439(03)00118-7
18. de Jong S (1993) SIMPLS: An alternative approach to partial least squares regression. Chemom Intell Lab Syst 18:251–263 . doi: 10.1016/0169-7439(93)85002-X
19. Bro R, Andersen CM (2003) Theory of net analyte signal vectors in inverse regression. J Chemom 17:646–652 . doi: 10.1002/cem.832
20. Tellinghuisen J (2005) Simple algorithms for nonlinear calibration by the classical and standard additions methods. Analyst 130:370–378 . doi: 10.1039/b411054d
21. Stute W (1996) The jackknife estimate of variance of a Kaplan-Meier integral. Ann Stat 24:2679–2704 . doi: 10.1093/biomet/81.3.602
22. Miller JN, Miller JC (2010) Statistics and Chemometrics for Analytical Chemistry, Sixth. Edinburgh Gate, Harlow, England
23. Ostra M, Ubide C, Vidal M, Zuriarrain J (2008) Detection limit estimator for multivariate calibration by an extension of the IUPAC recommendations for univariate methods. Analyst 133:532–539 . doi: 10.1039/b716965p
24. Allegrini F, Olivieri AC (2014) IUPAC-consistent approach to the limit of detection in partial least-squares calibration. Anal Chem 86:7858–7866 . doi: 10.1021/ac501786u
25. R Foundation for Statistical Computing (2018) R: A language and environment for statistical computing. R A Lang. Environ. Stat. Comput.
26. Mevik BH, Wehrens R, Liland KH (2015) pls: Partial Least Squares and Principal Component Regression. R package version 2.5-0. J Stat Softw. doi: 10.1159/000323281

CHAPTER 4.1: NAS APPLIED TO UV-VIS SPECTROSCOPY

Introduction

The net analyte signal procedure has been already successfully applied several times to samples analyzed in solution by UV-Vis spectroscopy [1–4]. Therefore, the work that will be shown in this chapter is not intended to offer a new development of NAS but is useful to understand how the NAS algorithm works, which applications already existing in the literature can be the starting point for the ones developed in the present work, and which new applications and results can be obtained. For the following work, the Ferré-Bro [5, 6] algorithm will be used with the addition of the Hemmateenejad variant [2], as explained in the former chapter. Hemmateenejad variant can be used because, as it will be shown later, the work is developed to stay in the linear range of the Beer-Lambert law.

The basic idea is very simple: to prepare a solution of an "analyte" of interest with the addition of a second substance playing the role of interfering species, and to quantify the analyte of interest. In order to have a better control on this study, the chosen compounds were colored inorganic salts, which gave colored solutions. Thus, a "visual" control is also possible. Three cases were studied:

- For the first one (the simpler) the chosen analyte of interest was copper sulfate, $CuSO_4$ (blue), and the interfering was potassium dichromate, $K_2Cr_2O_7$ (yellow). In this case, the concentration of $K_2Cr_2O_7$ was maintained constant in the added samples, what generally happens in the SAM method in solution. Moreover, the different colors of these salts give different and well distinguishable signals. This case will be called from now on "Cu-1"
- For the second case, the same $CuSO_4$ and $K_2Cr_2O_7$ were chosen as the analyte of interest and interfering, but, in this case, the concentration of $K_2Cr_2O_7$ was reduced while $CuSO_4$ increased with the additions. This simulates what happens in solid samples, where the addition of the analyte of interest, in general, reduces the relative concentration of the interfering substances. This case will be called "Cu-2"
- For the third case, again $K_2Cr_2O_7$ was used as interfering, but the analyte of interest was chromium chloride, $CrCl_3$ (green). The green color of $CrCl_3$ gives a signal overlapping with that of $K_2Cr_2O_7$, which is useful to stress the algorithm and check if it is able to remove the interfering signal. This case will be called "Cr"

Materials and Methods

$CuSO_4 \cdot 5H_2O$ (MW = 294.6 g/mol), $K_2Cr_2O_7$ (MW = 249.19 g/mol), and $CrCl_3 \cdot 6H_2O$ (MW = 266.45 g/mol) were purchased by Merck (Darmstadt, Germany). All these salts were weighted with a 0.1 mg precision weight scale. Mother solutions of all salts were prepared in 50 mL flasks, while the starting solutions of the analytes of interest and the standard added solutions were prepared in 10 mL flasks. All solutions were prepared with distilled water. All analyses were carried out with a Cary 60 UV-Vis Spectrophotometer

(Agilent, Santa Clara, CA), in plastic cuvettes, the wavelength range was 1000-300 nm with 1 nm step. All samples were analyzed three times.

For "Cu-1" case, 0.9988 g of $CuSO_4$•$5H_2O$ were solved in 10 mL of distilled water (C = 0.0800 mol/L). This is the "unknown" solution, which concentration has to be evaluated by NAS. Then, two 50 mL-solutions were prepared, the first one with 1.2512 g of $CuSO_4$•$5H_2O$ (C = 0.501 mol/L), and the second one with 0.0120 g of $K_2Cr_2O_7$ (C = $9.63 \cdot 10^{-4}$ mol/L). These were used to prepare the standard added solutions. The standard added solutions were prepared by putting 5 ml of the unknown solution, and 1 ml of the $K_2Cr_2O_7$ solution, plus an added volume of the second $CuSO_4$ solution as reported in Table 4.1.1. Distilled water was then added to these solutions to reach 10 ml.

	Added volume of $CuSO_4$ (mL)	Added concentration (mol/L)
add.0	0	0
add.1	0.2	0.0100
add.2	0.4	0.0150
add.3	0.6	0.0200
add.4	0.8	0.0250

Table 4.1.1: standard added solutions for "Cu-1" case

Therefore, the concentration of $CuSO_4$ in the add.0 solution, which is also the expected value from NASSAM, is 0.0400 mol/L.

For "Cu-2" case, 0.9960 g of $CuSO_4$•$5H_2O$ were solved in 10 mL of distilled water (C = 0.0798 mol/L). Again, this is the "unknown" solution. For the standard added solutions, two 50 mL-solutions were prepared, the first one with 1.2240 g of $CuSO_4$•$5H_2O$ (C = 0.490 mol/L), and the second one with 0.0146 g of $K_2Cr_2O_7$ (C = 0.00117 mol/L). The standard added solutions were prepared as reported in Table 4.1.2, by 5 mL of the unknown solution mixed with increasing concentrations of $CuSO_4$ and decreasing concentration of $K_2Cr_2O_7$. Also in this case, distilled water was used to reach 10 mL.

	Added volume of $CuSO_4$ (mL)	Added concentration (mol/L)	Volume of $K_2Cr_2O_7$ (mL)	Concentration of $K_2Cr_2O_7$ (mol/L)
add.0	0	0	1	$1.17 \cdot 10^{-4}$
add.1	0.2	0.0100	0.8	$9.37 \cdot 10^{-5}$
add.2	0.4	0.0150	0.6	$7.03 \cdot 10^{-5}$
add.3	0.6	0.0200	0.4	$4.69 \cdot 10^{-5}$
add.4	0.8	0.0250	0.2	$2.34 \cdot 10^{-5}$

Table 4.1.2: standard added solutions for "Cu-2" case

In this case, the expected value in add.0 sample is 0.0399 mol/L.

For the "Cr" case, the "unknown" solution was prepared by 0.4020 g of $CrCl_3$•$6H_2O$ solved in 10 mL (C = 0.0301 mol/L). For the standard added solution, two 50 mL-solutions were prepared, the first one with 0.2665 g of $CrCl_3$•$6H_2O$ (C = 0.100 mol/L), the second one with 0.0125 g of $K_2Cr_2O_7$ (C = 0.00100 mol/L).

The standard added solutions were prepared by putting 5 ml of the unknown solution mixed with 1 mL of $K_2Cr_2O_7$ solution and the added volume of the second $CrCl_3 \cdot 6H_2O$ solution as reported in Table 4.1.3.

	Added volume of $CrCl_3$ (mL)	Added concentration (mol/L)
add.0	0	0
add.1	10	0.0100
add.2	15	0.0150
add.3	20	0.0200
add.4	25	0.0250

Table 4.1.3: standard added solutions for "Cr" case

The expected extrapolated concentration from NASSAM in this case is 0.0151 mol/L.

Results and Discussion

- **"Cu-1" and "Cu-2" cases**

As already stated, Cu-1 is the simpler case that will be shown. Five samples were prepared as reported in Table 4.1.1 in which the concentration of $CuSO_4$ was increased, while the concentration of the interfering $K_2Cr_2O_7$ was maintained constant. The following Figure 4.1.1 shows the UV-Vis spectra obtained by these samples.

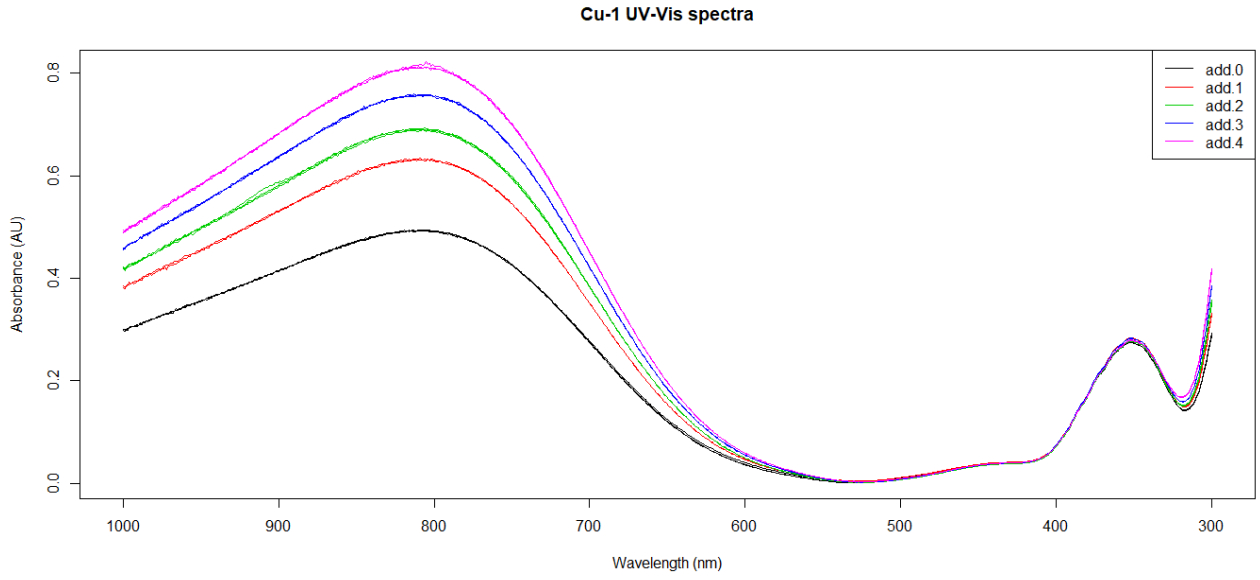


Figure 4.1.1: UV-Vis spectra of "Cu-1" samples, the legend is according to Table 4.1.1

The same procedure was adopted for the "Cu-2" samples: five samples increasing the concentration of $CuSO_4$ as reported in Table 4.1.2. The difference is that the concentration of the interfering $K_2Cr_2O_7$ decreases in each sample, simulating what happens in a solid sample. Figure 4.1.2 shows the original spectra of these samples.

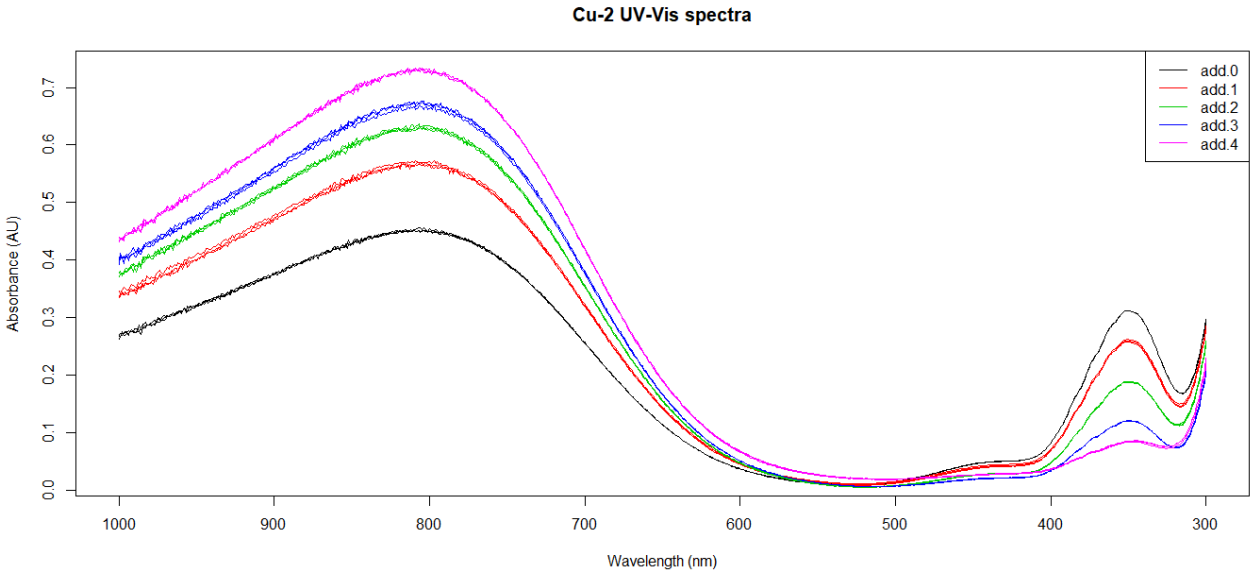


Figure 4.1.2: UV-Vis spectra of "Cu-2" samples, the legend is according to Table 4.1.2

$CuSO_4$ and $K_2Cr_2O_7$ have two well distinct absorption peaks: the former (blue) with a maximum at ~807 nm, the latter (yellow) with a maximum at ~352 nm. Therefore, there is no signal-overlap and the standard addition problem could be solved also in a univariate mode by using the absorbance of the $CuSO_4$ maximum, ignoring the presence of $K_2Cr_2O_7$. Therefore, these cases are most of all useful to show how the NAS procedure works because the results of the passages shown in the previous chapter are calculated "spectra", which are displayable and helpful to understand what is happening to data.

As already stated for the liquid case, the Hemmateenejad variant [2] can be used. It consists in subtracting the zero-added sample signal from the others, according to eq. 4.0.19. As a result, for "Cu-1" the signal of $K_2Cr_2O_7$ is removed in all samples (because it has the same magnitude for all samples), and the signal of $CuSO_4$ is reduced, while, for "Cu-2", the interfering signal becomes negative in the added samples, as shown in Figure 4.1.3

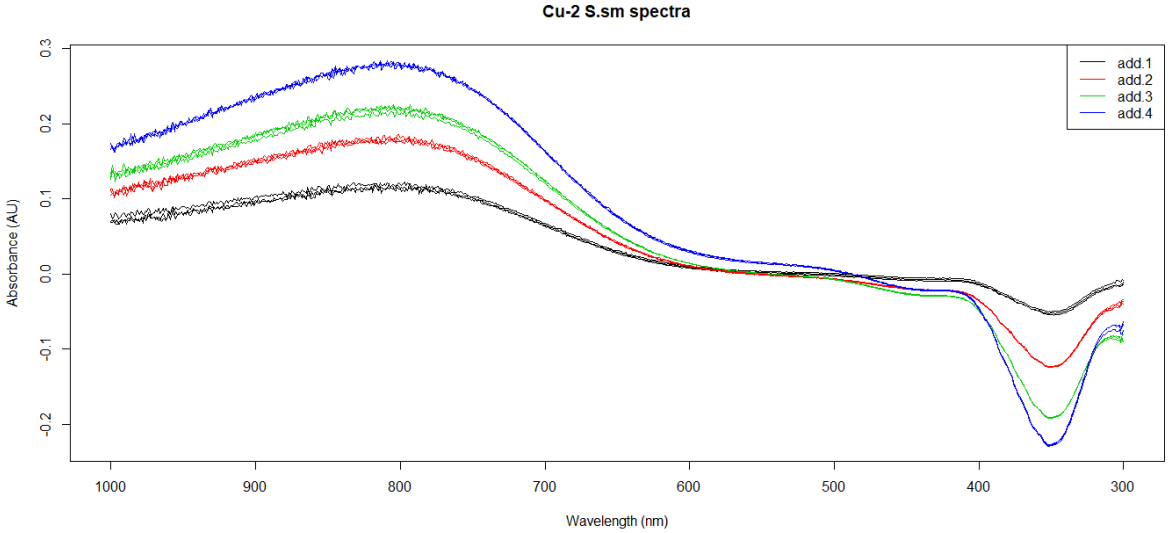


Figure 4.1.3: Calculated S_{sm} spectra for "Cu-2" samples, according to eq. 4.0.19

The application of Hemmateenejad variant gives slight differences on the final results. Therefore, it is more useful to show the NAS procedure on the original spectra, in order to better follow the computations. The differences in results between the two possible modes (with or without Hemmateenejad variant) are shown only at the end. In any case, the Ferré-Bro method was chosen for both "Cu-1" and "Cu-2" for its better results, and the zero-added sample was removed from the dataset.

The first step is always to compute a PLS regression by using the spectra as independent variables, and the added concentrations as dependent ones. For both "Cu-1" and "Cu-2" data, a minimum in RMSE was observed at the third PLS-component; thus, 3 factors were used for the following computations. From scores and loadings, the original spectra are recalculated and the result is shown in Figure 4.1.4

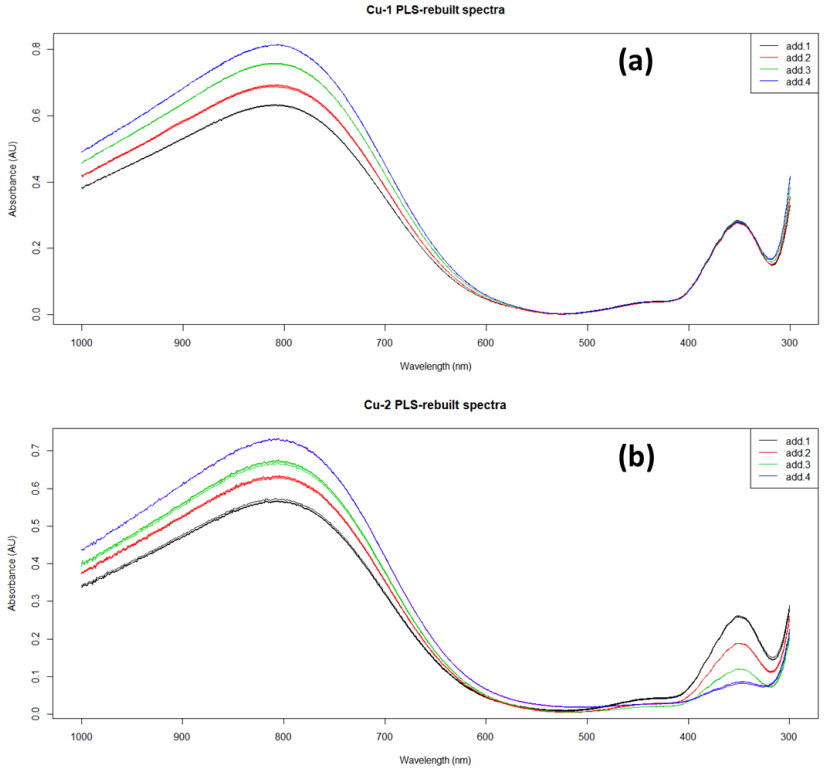


Figure 4.1.4: PLS-rebuilt spectra for a) "Cu-1" and b) "Cu-2" cases

As it can be observed by comparing Figure 4.1.4 with Figures 4.1.1 and 4.1.2, no significant differences are displayable between the original and the rebuilt spectra. For this reason, in some particular cases, the PLS regression may be also skipped, or substituted with a simpler PCA or PCR. However, for the present work, the PLS regression was always applied, because the differences between original and rebuilt spectra, even if difficult to be seen with the naked eye, may be dramatic for the final result. Moreover, if the Bro method [6] has to be used, according to eq. 4.0.18, the regression coefficient vector is required, thus PLS is necessary.

According to eq. 4.0.9, the matrix spectrum has to be calculated. For this purpose, a pure spectrum is necessary. For solid samples, this pure spectrum is simply obtained by analyzing the pure analyte added to the other samples. Similarly, in the liquid cases, as a "pure" spectrum it was used the one obtained from the solution used for the additions; a different solution was prepared for each case, but the concentration was

always ~0.08 mol/L, so the spectrum was similar for both cases. Figure 4.1.5 shows the second right term of eq. 4.0.9, which is the pure spectrum multiplied for c_k and α ($\alpha \boldsymbol{c_k s^t}$)

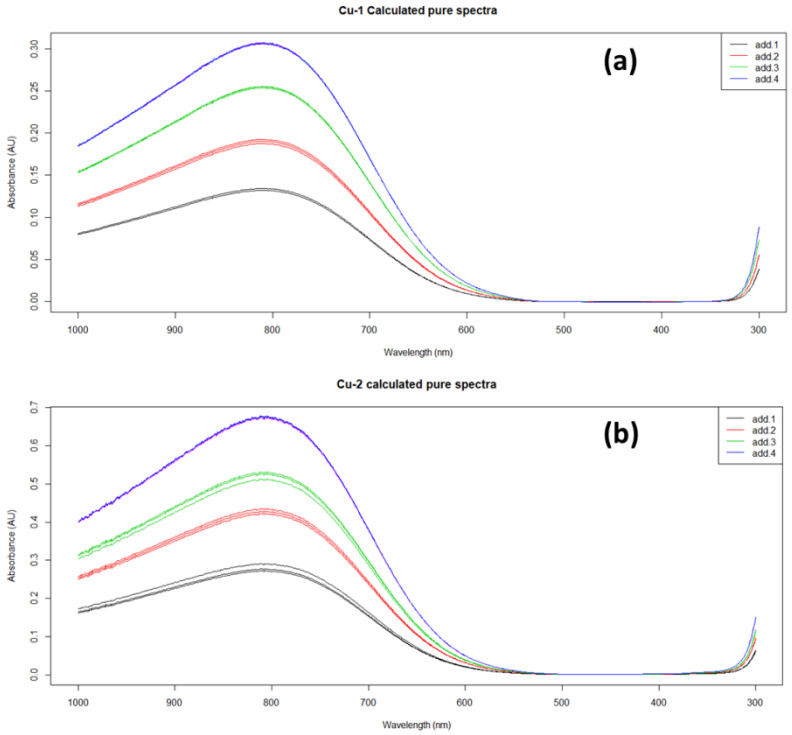


Figure 4.1.5: Calculated pure spectra according to eq. 4.0.9 for a) "Cu-1" and b) "Cu-2" cases

A different "calculated pure" spectrum is present for each addition because the value of c_{add} and, consequently, the value of c_k changes (eq. 4.0.10). These pure spectra are subtracted from the rebuilt one, according to eq. 4.0.9. The results of these subtractions are shown in Figure 4.1.6

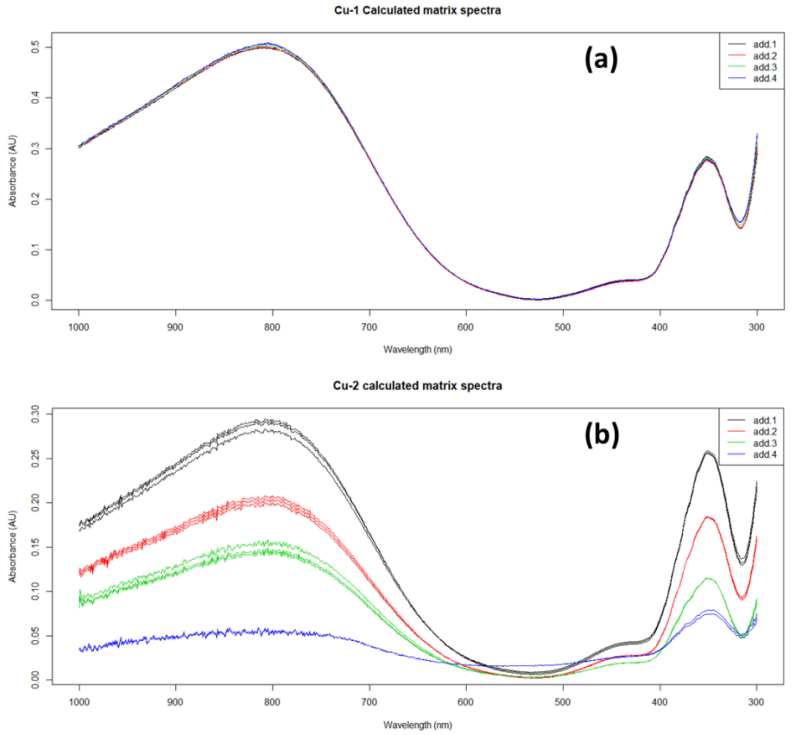


Figure 4.1.6: Matrix calculated spectra according to eq. 4.0.9 for a) "Cu-1" and b) "Cu-2" cases

The theory of NAS states that, at this point, in the matrix spectra only the signal due to interfering species should be present, and the one due to the analyte has been removed. However, in this case, the added concentration has been used to compute the matrix spectrum, and not the total one (and also the "pure" spectrum is not totally pure, but it has its own concentration). Therefore, what should be seen in Figure 4.1.6 are spectra in which the signal due to added concentration has been removed, which means spectra close to the zero-added one. This is the case of "Cu-1" (Figure 4.1.6a) where the calculated spectra are close to each other and similar to the zero-added one of Figure 4.1.1. For "Cu-2", however, the situation is totally different: matrix spectra are not close to each other, they are different from the zero-added one and there is also an inversion of the $CuSO_4$-peak trend (the lower addition has the higher peak). This is probably due to the presence of the interfering peak at different concentration: the difference between "Cu-1" and "Cu-2" is in the values of α (0.932 and 2.41 respectively) that are calculated from the rebuilt matrix (eq. 4.0.11). Therefore, differences in the product $\alpha \boldsymbol{c_k} \boldsymbol{s}^t$ are amplified in "Cu-2". This is a general behavior, also for solid samples: it is very difficult to find a case in which the matrix spectra are sharp as that of "Cu-1"; there is always some effect that makes them deviate from "ideality".

The matrix spectra are then used to compute the projection matrix $\boldsymbol{H}$ (eq. 4.0.12) and, finally, the original spectra are projected on the NAS space (eq. 4.0.13). The results of this projection are the net spectra shown in Figure 4.1.7

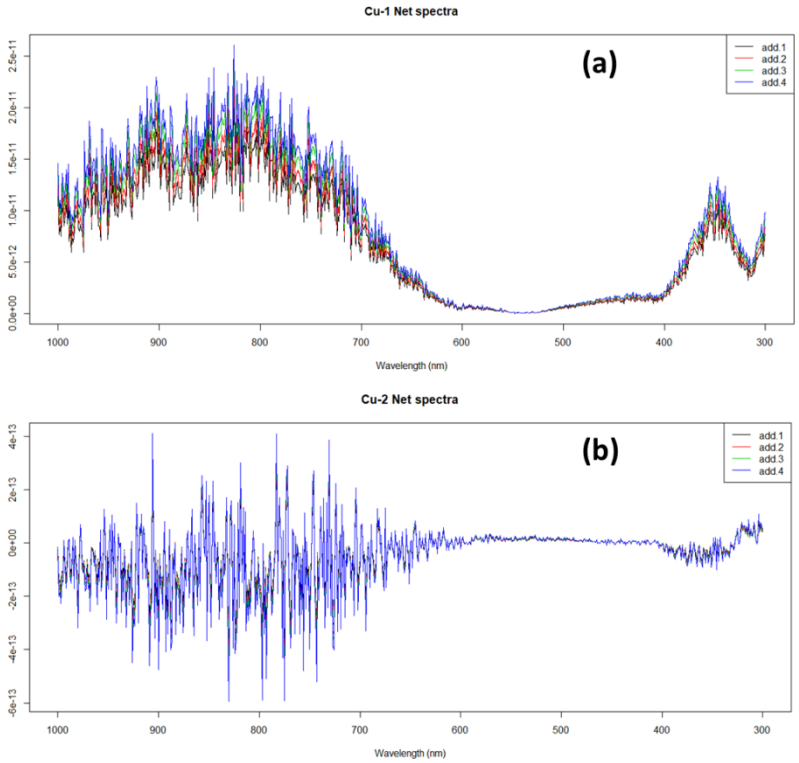


Figure 4.1.7: Net spectra for a) "Cu-1" and b) "Cu-2" cases

What is important of these spectra is their Euclidean norm, which means that it is not important if the reported signal is positive or negative, because a sum of squares will be computed. It is interesting, from

Figure 4.1.7, that "Cu-1" shows both a signal in the $CuSO_4$ region and a second signal in $K_2Cr_2O_7$ region. "Cu-2", instead, shows an intense signal, no matter how confused it may be, in $CuSO_4$ region and a much less intense signal in $K_2Cr_2O_7$ region. This means that, for "Cu-2", the net signal is more reliable than for "Cu-1", because the signal of the interfering $K_2Cr_2O_7$ has been more effectively removed, which is the objective of NAS. In both cases, anyway, the net signal increases while increasing the added concentration (for "Cu-2", it is the absolute value of each point that increases).

At this point, the Euclidean norms of each net spectrum were computed and used as a pseudo-univariate signal to create a pseudo-univariate standard added line, from which the concentration of interest was extrapolated.

The following Figure 4.1.7 and Table 4.1.4 show the NAS-standard addition line and the corresponding figures of merit for "Cu-1" and "Cu-2" cases.

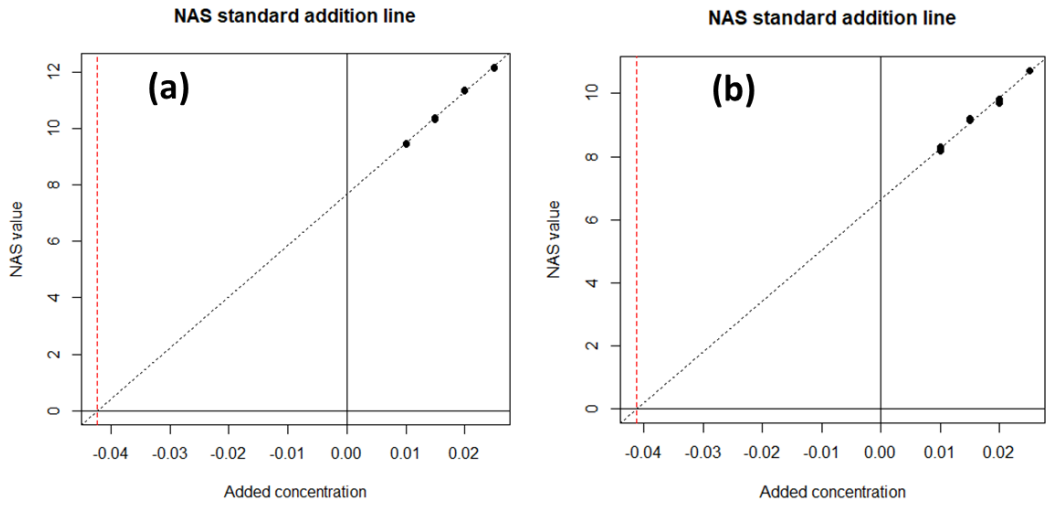


Figure 4.1.8: NAS-standard addition line for a) "Cu-1" and b) "Cu-2" cases. The red vertical line indicates the extrapolated concentration

	"Cu-1"	**"Cu-2"**
Expected concentration	0.04	0.04
$\mathbf{c_E}$	0.0423	0.0413
$\mathbf{s_{C,E}}$	$2.32 \cdot 10^{-4}$	$4.92 \cdot 10^{-4}$
b	7.66	6.65
$\mathbf{s_b}$	0.0551	0.0909
a	181	161
$\mathbf{s_a}$	2.91	4.95
$\mathbf{R^2}$	0.997	0.990
PLS-factor	3	3
RMSE	$1.39 \cdot 10^{-5}$	$1.31 \cdot 10^{-4}$
Sn	15.0	13.9
LoD	0.02	0.03

Table 4.1.4: Results and figures of merit for "Cu-1" and "Cu-2" cases

Table 4.1.4 shows a very good agreement between the expected concentration value and the one obtained by extrapolation from the NAS-standard addition line for both "Cu-1" and "Cu-2". Moreover, R^2 of the two lines are very close to the ideal value of 1 indicating good models and *LoD*s are lower than the extrapolated values (even if in the same order of magnitude). In general, it can be stated that, although the examined cases are very simple, and in some steps the behavior of NAS procedure is different for the two cases, the NAS method successfully predicted the concentrations of the analytes of interest. These results are encouraging for applying NAS to more complex and real cases.

- **"Cr" case**

The complication present in "Cr" case, with respect to "Cu-1" and "Cu-2", is that the analyte of interest, $CrCl_3$, has two UV-Vis maxima: the first one at ~615 nm, the second one at ~430 nm. This second peak partially overlaps the absorption band of the interfering $K_2Cr_2O_7$ (maximum at ~352 nm), as shown in Figure 4.1.9

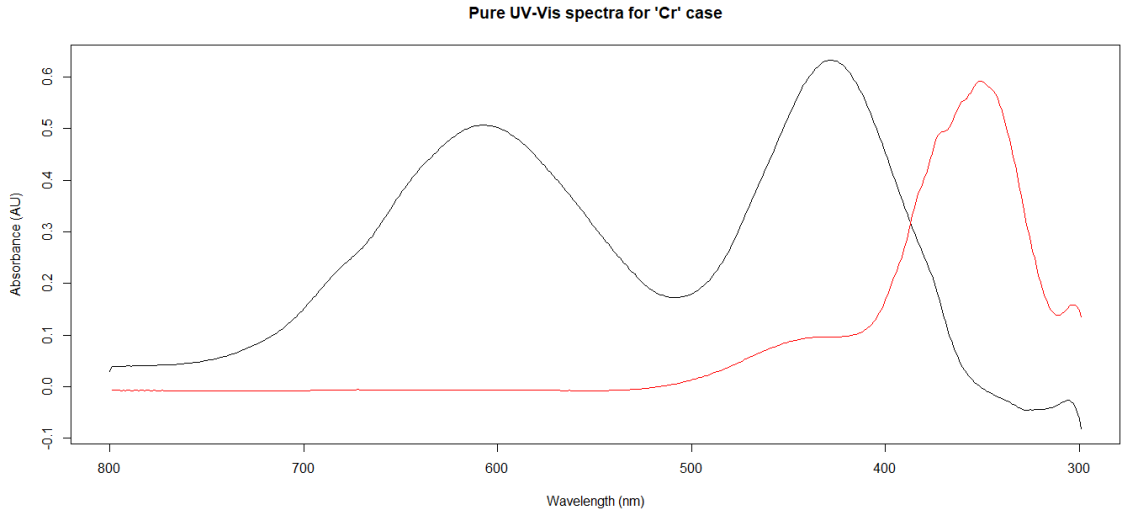


Figure 4.1.9: UV-Vis spectra of $CrCl_3$ (black) and $K_2Cr_2O_7$ (red)

For univariate purposes, the first-peak maximum could be taken to compute a standard addition line. However, in order to stress the NAS algorithm, both the entire spectrum and only the second $CrCl_3$-peak were used to compute a NAS standard addition line. This case is interesting because, for the second analyte-peak, the Beer-Lambert law, due to the overlap, is not fully respected; therefore, the Hemmateenejad variant should not be applicable. The following Table 4.1.5, indeed, shows the results of NAS procedure on the entire spectra with and without Hemmateenejad variant.

	With Hemm. variant	Without Hemm. variant
Expected concentration	0.015	
$\mathbf{c_E}$	0.0189	0.0133
$\mathbf{s_{C,E}}$	$3.63 \cdot 10^{-4}$	$7.36 \cdot 10^{-4}$
b	4.48	1.99
$\mathbf{s_b}$	0.193	0.0605
a	218	150
$\mathbf{s_a}$	10.5	3.43
$\mathbf{R^2}$	0.975	0.995
PLS-factor	3	2
RMSE	$3.38 \cdot 10^{-4}$	$5.52 \cdot 10^{-4}$
Sn	7.10	151
LoD	0.3	0.004

Table 4.1.5: Results and figures of merit for "Cr" case using full spectra

It is interesting to note that, without Hemmateenejad variant, the extrapolated concentration is closer to the expected value, the R^2 is higher, and, most of all, *LoD* is much lower and it is lower than the extrapolated value, while, with Hemmateenejad variant, it is much higher. This result can be considered as a demonstration that Hemmateenejad variant is useful for the liquid case (it gave very good results for "Cu-1 and "Cu-2", and even in this case the extrapolated concentration is close to the expected one), but its results are reliable only when the Beer-Lambert law is completely respected.

Table 4.1.6, instead, shows the results obtained using only the second peak of $CrCl_3$ (the one with the maximum at ~352 nm)

	With Hemm. variant	Without Hemm. variant
Expected concentration	0.015	
$\mathbf{c_E}$	0.0213	0.0144
$\mathbf{s_{C,E}}$	$4.96 \cdot 10^{-4}$	$1.08 \cdot 10^{-3}$
b	3.35	1.53
$\mathbf{s_b}$	0.107	0.0491
a	157	106
$\mathbf{s_a}$	6.06	2.78
$\mathbf{R^2}$	0.985	0.993
PLS-factor	2	2
RMSE	$7.35 \cdot 10^{-4}$	$5.66 \cdot 10^{-4}$
Sn	5.07	107
LoD	0.2	0.003

Table 4.1.5: Results and figures of merit for "Cr" case using the second peak of $CrCl_3$

Besides the confirmation of what already stated before about c_E, R^2, and *LoD*, what is interesting here is that, without the variant, the extrapolated value is closer to the expected one with respect to that obtained before.

Figure 4.1.10 shows the passage from the original to the net spectra in “Cr” case (second peak only). It should be noted that the original spectra are noisier than the ones reported for “Cu” cases, but the NAS is efficient in managing also this noise. Moreover, net spectra show a negative peak where the interfering signal is present in the original ones; it is a general behavior of NAS: for managing the interfering signal, sometimes it appears as a negative peak in the net spectra.

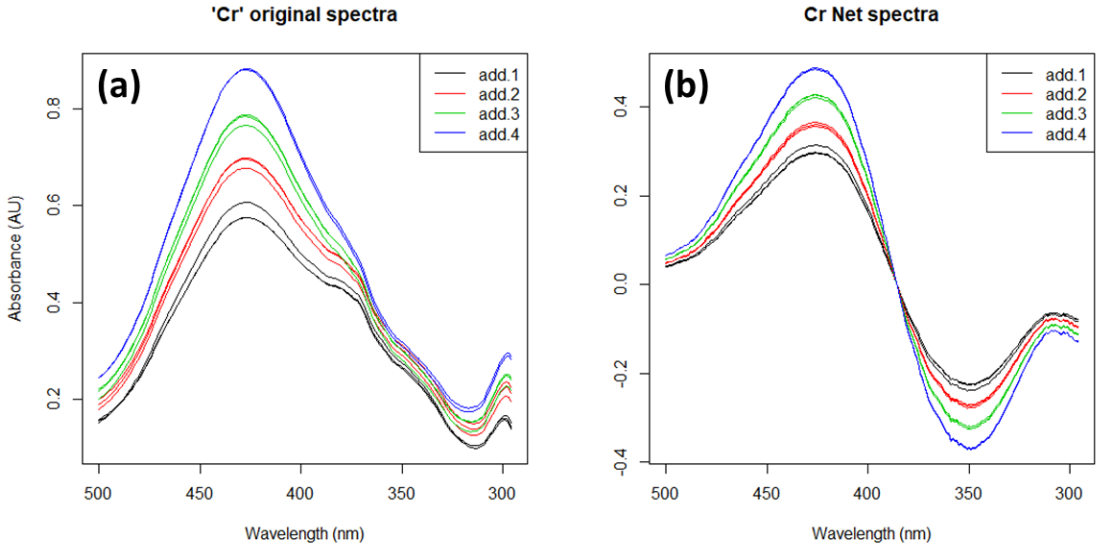


Figure 4.1.10: Original (a) and Net (b) spectra for the “Cr” case using only the second $CrCl_3$ peak

Conclusions

The present chapter showed an application of NAS to simple home-made solutions analyzed by UV-Vis spectroscopy. This was simply an exercise, however, it is useful from a theoretical point of view, to better understand how NAS algorithm works, and how the signal is managed through the procedure. Such graphic visualization can be obtained also for the other cases, which will be shown in further chapters. However their meaning would be more difficult to explain and understand, thus, in general, it will not be shown.

References

1. Navarro-Villoslada F, Pérez-Arribas L V., León-González ME, Polo-Díez LM (1999) Matrix effect modelling in multivariate determination of priority pollutant chlorophenols in urine samples. Anal Chim Acta 381:93–102 . doi: 10.1016/S0003-2670(98)00701-6
2. Hemmateenejad B, Yousefinejad S (2009) Multivariate standard addition method solved by net analyte signal calculation and rank annihilation factor analysis. Anal Bioanal Chem 394:1965–1975 . doi: 10.1007/s00216-009-2870-1
3. Shayanfar A, Asadpour-Zeynali K, Jouyban A (2013) Solubility and dissolution rate of a carbamazepine-cinnamic acid cocrystal. J Mol Liq 187:171–176 . doi: 10.1016/j.molliq.2013.06.015
4. Asadpour-Zeynali K, Bastami M (2010) Net analyte signal standard addition method (NASSAM) as a novel spectrofluorimetric and spectrophotometric technique for simultaneous determination, application to assay of melatonin and pyridoxine. Spectrochim Acta - Part A Mol Biomol Spectrosc 75:589–597 . doi: 10.1016/j.saa.2009.11.023
5. Ferré J, Faber NM (2003) Net analyte signal calculation for multivariate calibration. Chemom Intell Lab Syst 69:123–136 . doi: 10.1016/S0169-7439(03)00118-7
6. Bro R, Andersen CM (2003) Theory of net analyte signal vectors in inverse regression. J Chemom 17:646–652 . doi: 10.1002/cem.832

CHAPTER 4.2: NAS APPLIED TO RAMAN SPECTROSCOPY

Introduction

Regarding the application of net analyte signal procedure to Raman spectroscopy, several studies are reported in the literature, where it was used for the computation of figures of merit [1–4]. There is also one study in which the combination Raman spectroscopy – NAS was used to create a control chart for quality control purposes [5]. However, at the author knowledge, only one study was carried out in which NAS was applied to Raman spectroscopy for quantification [6]: in that work, a NAS-calibration line was created by analyzing several ethanol aqueous solutions in order to quantify the concentration of ethanol in beverages as, for instance, wine.

The aim of the present work is to show an application of NAS procedure for quantification purposes in solid samples analyzed by Raman spectroscopy: the specific study concerns the quantification of paraffin in beeswax samples.

Beeswax is one of the products of bees (*Apis Mellifera*) and, besides honey, probably the most important. Bees use beeswax as a structural material for their hives: they use it for building honeycombs (hexagonal cell structure into which pollen, honey, and larvae are placed), and to close any hole in the structure (both internal and external). Bees use beeswax also for defense: if some insect enters in the hive, bees cover it with beeswax until it suffocates. Humans are very interested in this product, too. Although the most known beeswax products are candles [7], humans use beeswax since ancient times for art [8, 9], but also for coating [10]; ancient Egyptians used beeswax also for their mummies [11]. Besides candles, also today beeswax has several applications, for example in food packaging [12], food preservation [13], and also in drug and cosmetic formulations [14]. Moreover, a recent study [15] used the composition of beeswax as bio-monitor for the presence of pesticide residuals in the environment. These uses of beeswax, in particular for food and drugs, makes it important to study and characterize this product. Therefore, several studies were carried out about its biosynthesis [16, 17] and composition [18–20].

Composition of beeswax, analyzed by GC-MS [18] and Raman spectroscopy [21], has been found to be a mixture of hydrocarbons (~15%), esters (mono-, ~35%, di-, ~14%, tri-, ~3%, and hydroxy-, ~12%), free acids (~12%), acid esters (~3%) and other compounds (~7%). All these hydrocarbons and esters are characterized by medium-length linear carbon chains (up to 34-35 atoms). Such a composition makes beeswax a white (or yellow-brown, depending on impurities of pollen or honey) solid with a characteristic texture soft and plastic (but sometimes also hard and crumbly), a low melting point (~40°C), and low viscosity when melted.

The large employment of beeswax makes this product subject to frauds, the principal of which is by the addition of paraffin [22]. Paraffin is a class of linear alkanes, with carbon chains that can have different length, determining different properties [23], derived from petroleum processing. Its physical properties are

similar to beeswax ones: it is white, soft and with a low melting point. Therefore, it is easy to mix paraffin with beeswax, obtaining a product with characteristics very similar to the pure beeswax.

Mostly due to the food and pharmaceutical uses of beeswax, the adulteration control is important. Although also the use of pure paraffin in these fields (most of all in cosmetics) is studied and controlled [24], there is mainly an economic problem: paraffin is cheaper than pure beeswax, so mixing it to beeswax and sell the product as "pure beeswax" brings to an illicit profit for the producer. The official analytical method for this kind of control is gas-chromatography coupled with mass spectrometry (GC-MS) [22, 25]. Some alternatives are the study of physical and chemical parameters [26] or the use of FTIR spectroscopy [27, 28]. However, the biggest problem of GC-MS [25] is the variability of paraffin composition. The presence of paraffin-adulteration can be detected by finding hydrocarbons with an even number of carbons (occurring in trace amounts in beeswax) or with more than 35 atoms of carbon in the structure (not present in beeswax [20]). Moreover, GC-MS is a destructive technique, that requires some sample pre-treatment (at least, it has to be solved in heptane) and long analysis time.

As already stated, the alternative approach for quantification of paraffin-adulteration in beeswax is based on Raman spectroscopy and NASSAM method. The advantage of Raman spectroscopy, over the official GC-MS method, is that it is not-destructive for the sample and the analysis is more rapid and cheaper. For the present work, a beeswax sample was adulterated in laboratory with paraffin and the amount of the adulterant was quantified by NAS.

Materials And Methods

Beeswax and paraffin samples were furnished by beekeepers to the CREA-API institute (Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria – Unità di Ricerca in Apicoltura e Bachicoltura, Bologna, Italy). To obtain pure beeswax, old and broken honeycomb (entirely produced by bees) were melted.

Standard-added samples were prepared by melting a proper quantity of beeswax and paraffin and mixing them once in the liquid state. In melting phase, the temperature was controlled to not exceed 62-64°C in order not to decompose beeswax (at ~85°C it tends to lose color, while at 120°C it burns). Five samples were prepared with an increasing concentration ($\%_{w/w}$) of paraffin, according to Table 4.2.1. The sample with lower concentration was used as the zero-added one, and the added concentrations of the others were calculated according to this one. Samples were prepared in beakers covered with an aluminum sheet, previously washed with hexane. The total weight of each sample is 2.00 ± 0.05 g.

	Paraffin total concentration ($\%_{w/w}$)	Added concentration ($\%_{w/w}$)
add.0	1.5	0.0
add.1	3.9	2.4
add.2	6.4	4.9
add.3	9.0	7.5
add.4	12.9	11.4

Table 4.2.1: Paraffin-adulterated standard added samples

Each sample, plus a pure paraffin one, was analyzed three times with Raman spectroscopy. In order to carry out Raman analyses, a portion of the sample was melted on a watch-glass and placed on a hot plate. Once the sample is liquefied, a drop is placed on a glass slide (18 x 18 mm^2) and covered with another slide. In order to avoid the solidification of the sample in contact with the cold surface of the slides, it was continuously heated and an extremely thin layer was melted between the two slides to avoid beeswax stratification. Instrumental analyses were carried out by a DXR Raman Microscope (Thermo Fisher Scientific, Waltham, MA) with a laser source (wavelength 532 nm). OMNIC for Dispersive Raman software (Thermo Fisher Scientific, Waltham, MA) was used to handle all the acquisition parameters specified in Table 4.2.2.

Parameter	Value
Laser wavelength	532 nm
Laser power (max 10 mW)	10.0
Aperture	25 µm pinhole
Grating	900 lines/mm
Estimated resolution	2.7-4.2 cm^{-1}
Estimated spot size	1.3 µm
Min range limit	50 cm^{-1}
Max range limit	3500 cm^{-1}
Accessory	Microscope
Objective	20x

Table 4.2.2: Raman operative conditions

Raman spectra were collected in the interval 100-3500 cm^{-1}, with steps of 1 cm^{-1}.

Results and Discussion

The starting dataset is composed of 18 rows (3 replicates of samples in Tab. 4.2.1 and 3 replicates of a pure paraffin sample) and 3527 columns (Raman shifts). Figure 4.2.1 shows the Raman spectra of a pure sample of beeswax and a pure sample of paraffin.

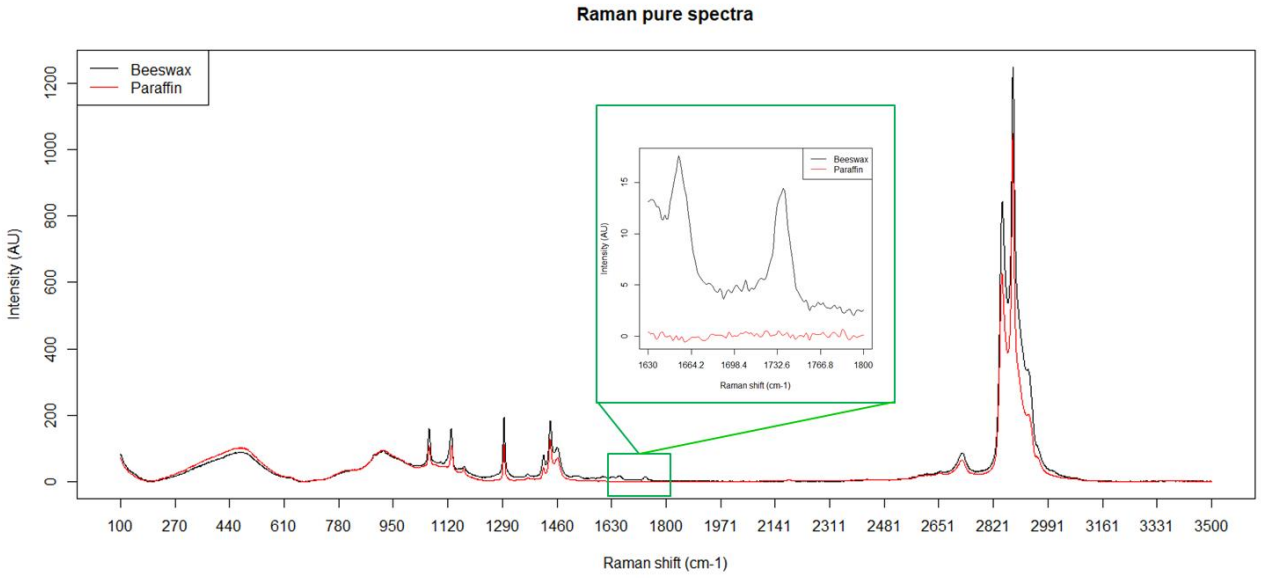


Figure 4.2.1: Raman spectra of pure beeswax and paraffin

What can be stated from figure 4.2.1 is that there are very slight differences between the spectra of beeswax and paraffin. This is not surprising: indeed, most of the peaks present in these spectra are due to stretching and bending of C-H and C-C bonds, which are predominant in both beeswax and paraffin. From a qualitative point of view, the peaks at 1735 cm^{-1} and 1656 cm^{-1} are very important. These are due to the stretching of C=O bond in, respectively, esters and amides. Although not intense, these peaks are present in beeswax and totally absent in paraffin, that is composed of alkanes only [28]. These peaks are highlighted in Figure 4.2.1 in the green rectangle. Figure 4.2.2, instead, shows the original Raman spectra for the added samples.

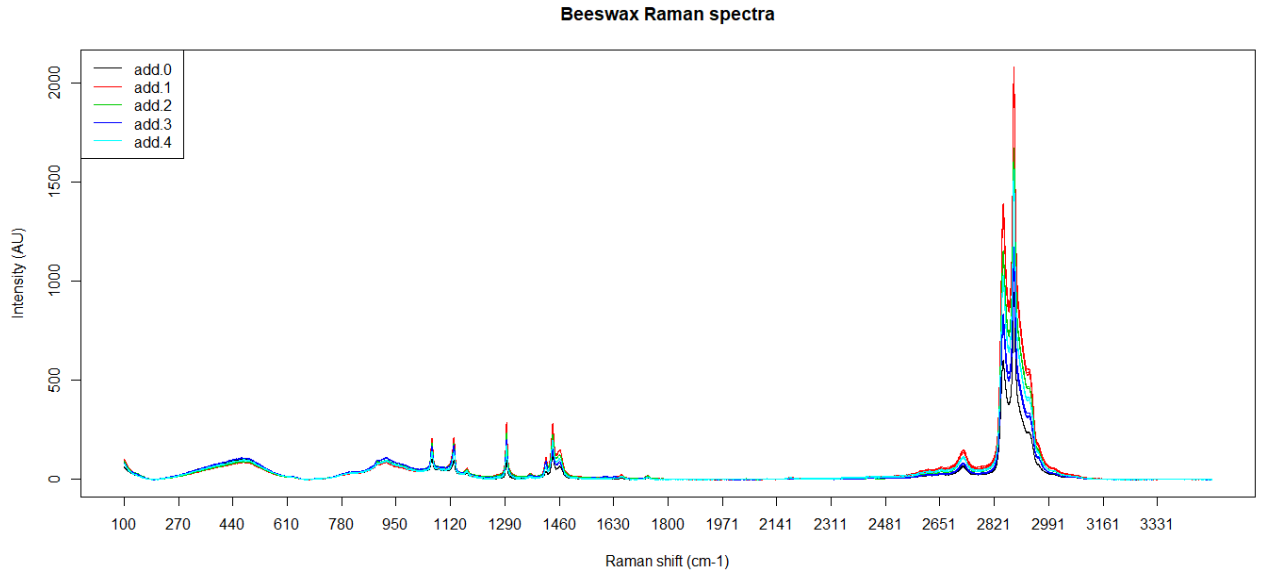


Figure 4.2.2: Raman spectra of the added samples, the legend is according to table 4.2.1

Looking at Figure 4.2.2, it is interesting to note that the (evident) peaks are not linearly increasing with the concentration of paraffin: add.0 (black) has the less intense signal (for almost all peaks), then, in ascending

order, there are add.3 (blue), add.4 (light blue), add.2 (green), and add.1 (red) is the most intense. Therefore, a univariate quantification seems impossible (at least without pre-treating data). However, the NAS procedure was applied to the raw data, and the best results were obtained by the original Lorber procedure, without Ferré-Bro variant. In this case, also the zero-added sample was kept in the computation. Table 4.2.3 reports the results obtained for this case.

Expected concentration	1.5
$\mathbf{c_E}$	2.09
$\mathbf{s_{C,E}}$	0.258
b	20.8
$\mathbf{s_b}$	2.17
a	9.94
$\mathbf{s_a}$	0.331
$\mathbf{R^2}$	0.985
PLS-factor	7
RMSE	0.645
Sn	12.5
LoD	--

Table 4.2.3: Results and figures of merit for the beeswax-Raman case

In this case, a blank sample was not available for the computation of *LoD*. The extrapolated concentration (2.09%$_{w/w}$), also considering the relative standard deviation (0.258%$_{w/w}$), is not significantly different from the expected value (1.5%$_{w/w}$, the total concentration of paraffin present in add.0 sample, as reported in Table 4.2.1), and the R^2 (0.985) is close to its ideal value. Figure 4.2.3 shows the NAS standard addition line obtained in this study, Figure 4.2.4 shows the corresponding net spectra.

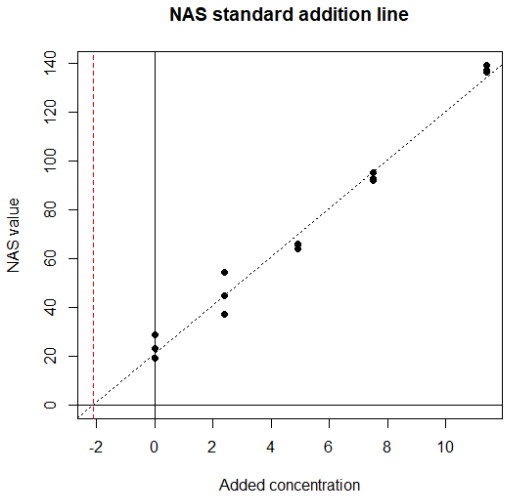


Figure 4.2.3: NAS-standard addition line for the beeswax-Raman case.
The red vertical line indicates the extrapolated concentration

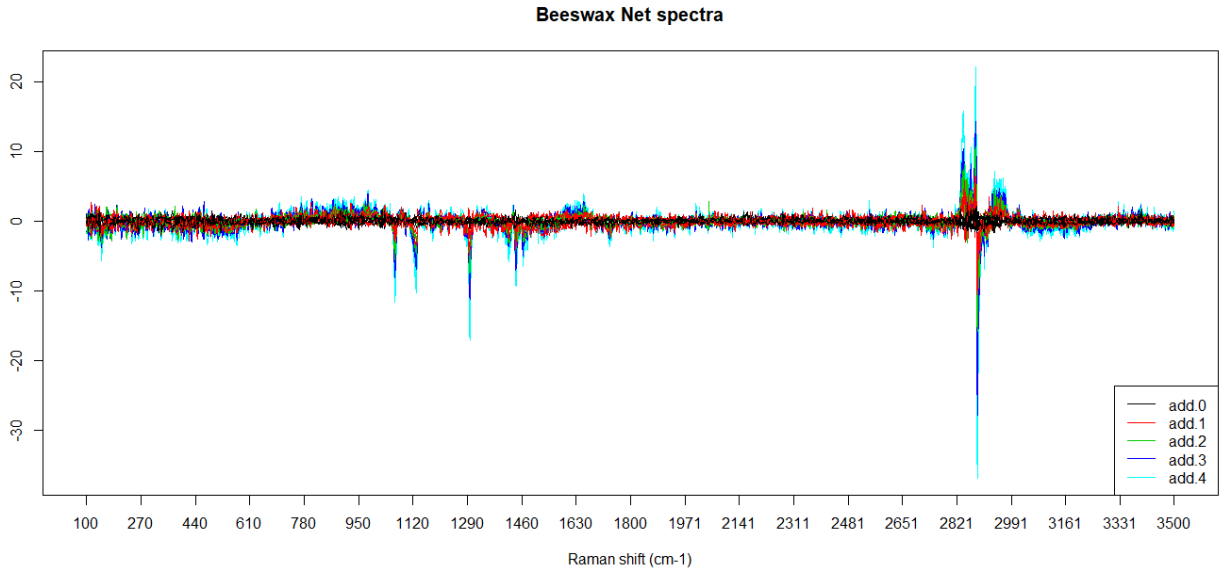


Figure 4.2.4: Net spectra for the beeswax-Raman case

Net spectra in Figure 4.2.4 (as the standard addition line in Figure 4.2.2) show that the NAS procedure has been able to obtain a linear behavior of samples, even if the starting spectra seemed not to have it. Looking at the peaks (no matter if positive or negative), the increasing order is now correct, the lowest signal (with almost no peaks) is the one of add.0, and the highest one is add.4.

Conclusions

NAS procedure was applied to samples of beeswax artificially adulterated with paraffin and analyzed by Raman spectroscopy. Although from the original Raman spectra it was difficult to qualitatively detect the presence of paraffin, the concentration extrapolated from NAS procedure is in very good agreement with the expected value.

References

1. Ren M, Arnold MA (2007) Comparison of multivariate calibration models for glucose, urea, and lactate from near-infrared and Raman spectra. Anal Bioanal Chem 387:879–888 . doi: 10.1007/s00216-006-1047-4
2. Short SM, Cogdill RP, Anderson C a (2007) Determination of figures of merit for near-infrared and Raman spectrometry by net analyte signal analysis for a 4-component solid dosage system. AAPS PharmSciTech 8:E96
3. Kang G, Lee K, Park H, et al (2010) Quantitative analysis of mixed hydrofluoric and nitric acids using Raman spectroscopy with partial least squares regression. Talanta 81:1413–1417 . doi: 10.1016/j.talanta.2010.02.045
4. Palermo RN, Short SM, Anderson CA, et al (2012) Determination of figures of merit for near-infrared, Raman and powder X-ray diffraction by net analyte signal analysis for a compacted amorphous dispersion with spiked crystallinity. J Pharm Innov 7:56–68 . doi: 10.1007/s12247-012-9127-9
5. Rocha WF de C, Poppi RJ (2011) Multivariate control charts based on net analyte signal (NAS) and Raman spectroscopy for quality control of carbamazepine. Anal Chim Acta 705:35–40 . doi: 10.1016/j.aca.2011.03.024
6. Li QB, Yu C, Zhang QX (2013) [Research on Raman spectra quantitative analysis model of ethanol aqueous solution based on net analyte signal]. Guang Pu Xue Yu Guang Pu Fen Xi 33:390–394 . doi: 10.3964/j.issn.1000-0593(2013)02-0390-05
7. Fine PM, Cass GR, Simoneit BRT (1999) Characterization of fine particle emissions from burning church candles. Environ Sci Technol 33:2352–2362 . doi: 10.1021/es981039v
8. d'Errico F, Backwell L, Villa P, et al (2012) Early evidence of San material culture represented by organic artifacts from Border Cave, South Africa. Proc Natl Acad Sci 109:13214–13219 . doi: 10.1073/pnas.1204213109
9. Andreotti A, Bonaduce M, Colombini MP, et al (2006) Combined GC/MS analytical procedure for the characterization of glycerolipid, waxy, resinous, and proteinaceous materials in a unique paint microsample. Anal Chem 78:4490–4500 . doi: 10.1021/ac0519615
10. Garnier N, Cren-Olivé C, Rolando C, Regert M (2002) Characterization of archaeological beeswax by electron ionization and electrospray ionization mass spectrometry. Anal Chem 74:4868–4877 . doi: 10.1021/ac025637a
11. Buckley SA, Evershed RP (2001) Organic chemistry of embalming agents in Pharaonic and Graeco-Roman mummies. Nature 413:837–841 . doi: 10.1038/35101588
12. I.K. Greener OF (1989) Barrier Properties and Surface Characteristics of Edible, Bilayer Films. J Food Sci 54:1393–1399
13. McHugh TH, Senesi E (2000) Apple Wraps: A Novel Method to Improve the Quality and Extend the Shelf Life of Fresh-cut Apples. J Food Sci 65:480–485 . doi: 10.1111/j.1365-2621.2000.tb16032.x

14. Strickley RG (2004) Solubilizing Excipients in Oral and Injectable Formulations. Pharm. Res. 21:201–230
15. Chauzat MP, Martel AC, Cougoule N, et al (2011) An assessment of honeybee colony matrices, Apis mellifera (Hymenoptera: Apidae) to monitor pesticide presence in continental France. Environ Toxicol Chem 30:103–111 . doi: 10.1002/etc.361
16. Blomquist GJ, Chu AJ, Remaley S (1980) Biosynthesis of wax in the honeybee, Apis mellifera L. Insect Biochem 10:313–321 . doi: 10.1016/0020-1790(80)90026-8
17. Piek T (1964) Synthesis of wax in the honeybee (Apis mellifera L.). J Insect Physiol 10:563–572 . doi: 10.1016/0022-1910(64)90027-7
18. Tulloch AP (1971) Beeswax: Structure of the esters and their component hydroxy acids and diols. Chem Phys Lipids 6:235–265 . doi: 10.1016/0009-3084(71)90063-6
19. Tulloch AP (1970) The composition of beeswax and other waxes secreted by insects. Lipids 5:247–258 . doi: 10.1007/BF02532476
20. Tulloch AP, Hoffman LL (1972) Canadian beeswax: Analytical values and composition of hydrocarbons, free acids and long chain esters. J Am Oil Chem Soc 49:696–699 . doi: 10.1007/BF02609202
21. Edwards HGM, Farwell DW, Daffner L (1996) Fourier-transform Raman spectroscopic study of natural waxes and resins. I. Spectrochim Acta - Part A Mol Spectrosc 52:1639–1648 . doi: 10.1016/0584-8539(96)01730-8
22. Maia M, Nunes FM (2013) Authentication of beeswax (Apis mellifera) by high-temperature gas chromatography and chemometric analysis. Food Chem 136:961–968 . doi: 10.1016/j.foodchem.2012.09.003
23. Dirand M, Chevallier V, Provost E, et al (1998) Multicomponent paraffin waxes and petroleum solid deposits: Structural and thermodynamic state. Fuel 77:1253–1260 . doi: 10.1016/S0016-2361(98)00032-5
24. Johnson W, Bergfeld WF, Belsito D V., et al (2012) Safety Assessment of Isoparaffins as Used in Cosmetics. Int J Toxicol 31:269S–295S . doi: 10.1177/1091581812463087
25. Waś E, Szczęsna T, Rybak-Chmielewska H (2016) Efficiency of GC-MS method in detection of beeswax adulterated with paraffin. J Apic Sci 60:145–161 . doi: 10.1515/JAS-2016-0012
26. Serra Bonvehi J, Orantes Bermejo FJ (2012) Detection of adulterated commercial Spanish beeswax. Food Chem 132:642–648 . doi: 10.1016/j.foodchem.2011.10.104
27. Maia M, Barros AIRNA, Nunes FM (2013) A novel, direct, reagent-free method for the detection of beeswax adulteration by single-reflection attenuated total reflectance mid-infrared spectroscopy. Talanta 107:74–80 . doi: 10.1016/j.talanta.2012.09.052
28. Svečnjak L, Baranović G, Vinceković M, et al (2015) N approach for routine analytical detection of beeswax adulteration using ftir-atr spectroscopy. J Apic Sci 59:37–49 . doi: 10.1515/JAS-2015-0018

CHAPTER 4.3: NAS APPLIED TO GAS-CHROMATOGRAPHY

Introduction

The present chapter will show an application of NAS procedure to gas-chromatography in head-space mode. Although gas chromatography (GC) is often used as a quantitative method, in particular when coupled with mass spectrometry (MS) [1, 2], at the author knowledge it is the first time that NAS procedure is applied to GC for quantification purposes. Moreover, in general, GC is used to quantify a single analyte at a time, while, in the present work, the aim is to quantify a product, which produces several signals in the chromatogram. In fact, the specific case will regard the quantification of a home-made turmeric impurity in a saffron sample.

Several studies demonstrated that both saffron and turmeric (as other spices) may be used to fight diseases as cancer [3–5], diabetes [3], Alzheimer [6], depression [7], or as antibacterial [8, 9]. Therefore, probably, no health problems are associated with the addition of one of these spices to the other.

However, saffron is obtained by the dried red stigmas of the flower of *Crocus sativus L.*, which is cultivated only in some regions of Asia (Kashmir, northern Iran) and Europe (Castilla la Mancha, Spain, Kozani, Greece, Abruzzo, Sardinia, and Sicily, Italy) [10]. The limited areas of production and the laborious process required to obtain this spice makes it one of the most expensive in the world. This motivates the interest in economical frauds by adding less expensive spices to the market product [11, 12]. The authenticity of saffron is protected by the Standard ISO 3632-1 (2011) and by the presence of several PDO designations (as, for example, the Italian "Zafferano dell'Aquila", the major in terms of production and global exports) [10].

Due to their yellow color (even if with a totally different flavor), three of the most important adulterants for saffron are turmeric (*Curcuma longa*), safflower (*Carthamus tinctorius*), and marigold (*Calendula officinalis*) [13]. Several studies were carried out concerning saffron adulteration with different analytical techniques. The most important in this field is DNA analysis [10] to detect botanical species different from *Crocus sativus*. However, such technique is time-consuming and requires specialized personnel to produce and interpret reliable results. Therefore, chemical analyses are developing: FTIR spectroscopy [14], nuclear magnetic resonance [15], laser-induce breakdown spectroscopy [13], and liquid chromatography coupled with mass spectroscopy [16].

GC has been widely used for studying saffron. Coupled with MS, it allowed to identify the components of saffron volatile fraction [17, 18]. All studies identified safranal (2,6,6-trimethyl-1,3-cyclohexadiene-1-carboxaldehyde) as the most powerful aroma-active and abundant compound. Amanpour et al. [18] quantified safranal with a concentration of ~2200 μg/g over a total aromatic concentration estimated in ~8700 μg/g, as the sum of all the identified aromatic compounds. Therefore, safranal is considered as the molecule which mostly contribute to the typical saffron aroma. Other aroma-active compounds are most of all aldehydes, alcohols, and ketones. Two of the most important of these are isophorone (3,5,5-trimethyl-2-cyclohexene-1-one) and 4-ketoisophorone [17, 18]. However, the volatile fraction of saffron strongly

depends on its geographical origin. Volatile fraction, indeed, has been used as a fingerprint for the geographical discrimination of saffron samples by GC-MS [19, 20]. In all these cases, chemometrics was used to manage GC data, in particular LDA [19], in a work analogous to that showed in chapter 2.1 for honey discrimination.

Another work [21] used chemometrics, in particular partial least squares discriminant analysis (PLS-DA) for the same purpose of geographical discrimination. Moreover, they used other chemometric tools to select the most important compounds for such discrimination (finding, again, safranal, isophorone, and 4-ketoisophorone, plus 20 other compounds). However, they carried out these analyses on GC data collected with a flame ionization detector (FID). GC-FID has the disadvantage, over GC-MS, of not being able to directly identify molecules, unless comparing the retention times with certified standards. Therefore, GC-FID is useful for collecting a fingerprint of the volatile fraction, but not to identify the specific molecules.

The aim of the present chapter is to use the GC-FID fingerprint of some artificially adulterated saffron samples, with increasing amounts of turmeric, to quantify, with the NASSAM procedure, the turmeric concentration of the less adulterated one, used as zero-added concentration sample. GC data were obtained in head-space mode, by the same ultra-fast gas chromatograph (flash-GC), Heracles II (AlphaMOS, Toulouse, France), already used in chapter 2.1 for honey. The present work was carried out in collaboration with Coop Italia (Casalecchio di Reno, Bologna, Italy), that provided samples and the analytical instrumentation.

Materials and Methods

Saffron and turmeric samples were collected in a Coop Italia supermarket. The problem for such study is that no pure standards of the two spices exist, most of all pulverized (powders are necessary to properly mix the two species). Therefore, two commercial samples were collected and their purity was confirmed before sample preparation and GC-analysis by a DNA analysis, again carried out in Coop Italia laboratory. The trademark of the saffron sample is Zaffy® (Aromatica S.r.l., Milan, Italy), while that of turmeric one is Cannamela (Cannamela S.r.l., Bologna, Italy).

Four aliquots of saffron were manually adulterated with a proper quantity of turmeric according to table 4.3.1. All samples had a total weight of 100 ± 1 mg, thus, total concentration values reported in table 4.3.1 are also the milligrams of turmeric weighted and mixed to saffron.

	Turmeric total concentration ($\%_{w/w}$)	**Added concentration ($\%_{w/w}$)**
add.0	5	0
add.1	10	5
add.2	15	10
add.3	20	15

Table 4.3.1: turmeric-adulterated standard added samples

Powders were mixed by a Vortex mixer (Thermo Fisher Scientific, Milan, Italy) keeping them under agitation for 3 min, in order to homogenize samples. Each added sample was prepared two times.

30 ± 1 mg of each prepared sample was then put in a 20 mL vial and sealed with a magnetic cap. Two replicates of each sample were analyzed; thus, four replicates of each added sample were available for NAS analysis. Moreover, four replicates of pure saffron and four of pure turmeric were analyzed. The description of the flash-GC Heracles II (AlphaMOS, Toulouse, France) has been already provided in Chapter 2.1 about botanical discrimination of honey. Table 4.3.2 shows Heracles' operating conditions.

Parameter	Value
Oven conditions	20 min at 50 °C, 500 rpm
Syringe temperature	60 °C
Injected volume	5 mL
Trap temperature	40 °C
Trapping time	65 s
Trap desorption	240 °C
Columns temperature program	40°C (2 s) to 270 °C (21 s) by 3°C s^{-1}
FID temperature	270 °C
Acquisition duration	100 s
Digitalization of the signal	0.01 s

Table 4.3.2: Heracles II operating conditions for saffron analysis

Results and Discussion

The starting dataset matrix has dimensions 16 x 20000. The number of variables, however, is too high and is not easy to manage for a common personal computer. Therefore, in order to reduce data dimensions maintaining a good representability, one variable every ten was taken, and the remaining were discarded. This brought to a 16 x 2000 matrix. Figure 4.3.1 shows the chromatograms of pure saffron and turmeric, appending the signal of the second column subsequent to the first one (as it will be used for NAS computations). Figure 4.3.2, instead, represents samples chromatograms; the signal from the two columns are separated, for the sake of readability.

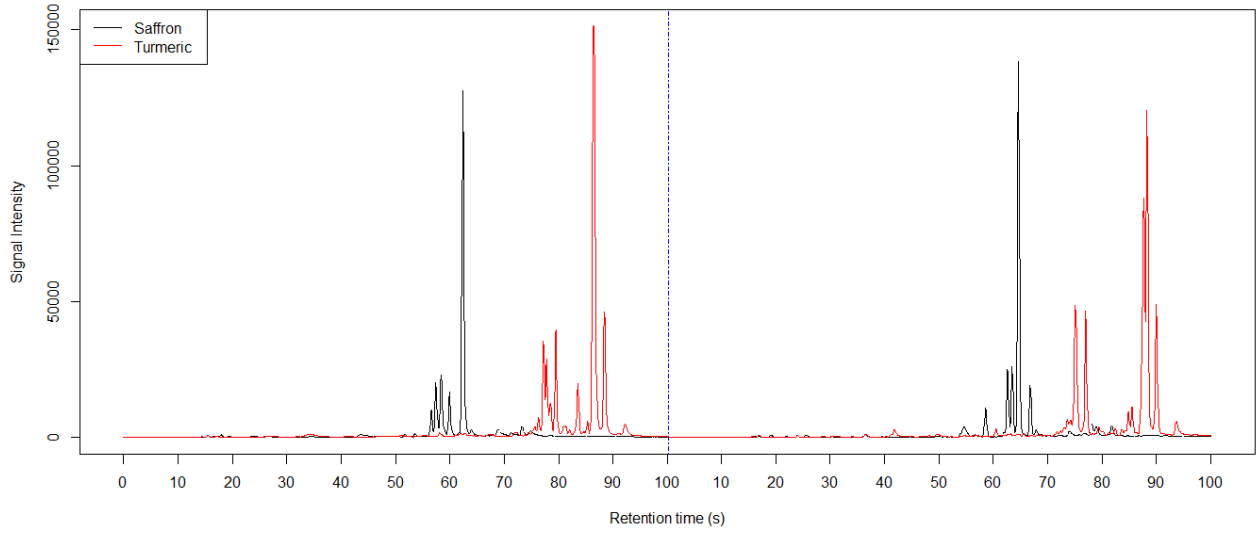


Figure 4.3.1: Pure saffron and turmeric chromatograms.
The vertical blue line divides column 1 (on the left) from column 2 (on the right)

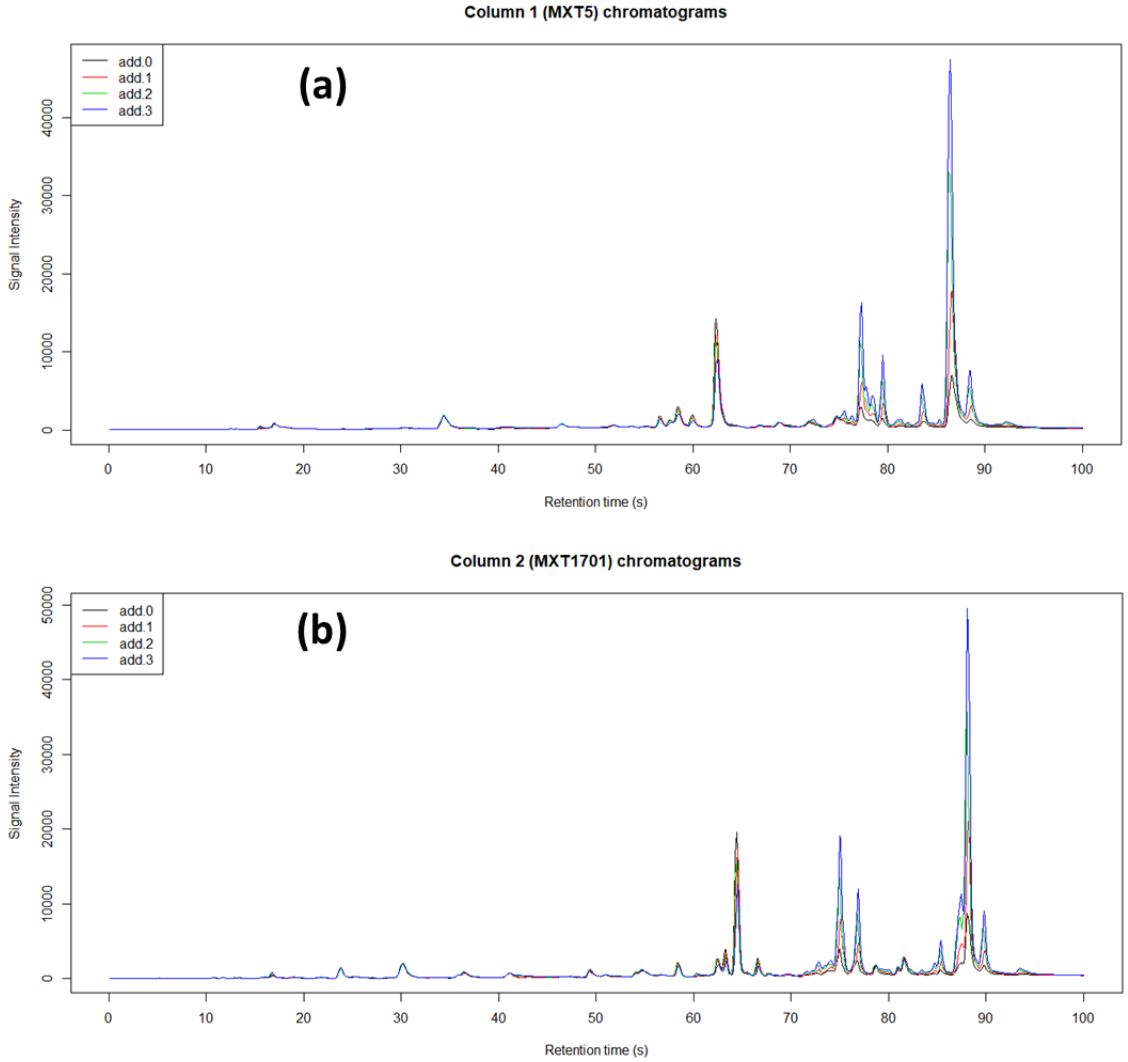


Figure 4.3.2: chromatograms of the added samples for a) MXT5 and b) MXT1701 columns. Legends are according to table 4.3.1

Some consideration can be drawn from Figures 4.3.1 and 4.3.2. In Heracles, the head-space of the same sample is split into the two columns before the analysis. Therefore, the two columns analyze the same molecules. Moreover, the two columns have a slight difference of polarity: this is why the two chromatograms may seem similar at first sight. However, in this study, no peak recognition was carried out (also because the FID detector makes it possible with good reliability only by comparison with standards). Therefore it is not so simple to recognize the correspondence of a peak from the first column to one on the second column. Moreover, Figure 4.3.1 shows that, at least for the higher peaks, there is a slight overlap between chromatograms of saffron and turmeric. Thus, a quantitative analysis might take advantage of such characteristic. However, in general, it is easy to have overlaps also in chromatography, and a pure sample may also not be available. Therefore the decision was to use the entire chromatograms, the second one following the first, for NAS quantification of turmeric. Another interesting information, that can be derived from a visual inspection, is that the intensity of peaks related to turmeric (starting at about 70 s for both columns) increases with the added concentration of the analyte, while the peaks related to saffron decrease correspondingly.

To these data, NAS was applied using Ferré-Bro method, without removing the zero-added sample. Table 4.3.3 and Figure 4.3.3 report the results and figures of merit obtained by NAS procedure.

Expected concentration	5
$\mathbf{c_E}$	3.58
$\mathbf{s_{C,E}}$	0.726
b	7603
$\mathbf{s_b}$	69.9
a	2125
$\mathbf{s_a}$	7.47
$\mathbf{R^2}$	0.999
PLS-factor	9
RMSE	1.06
Sn	2031
LoD	3.27

Table 4.3.3: Results and figures of merit for saffron-GC case

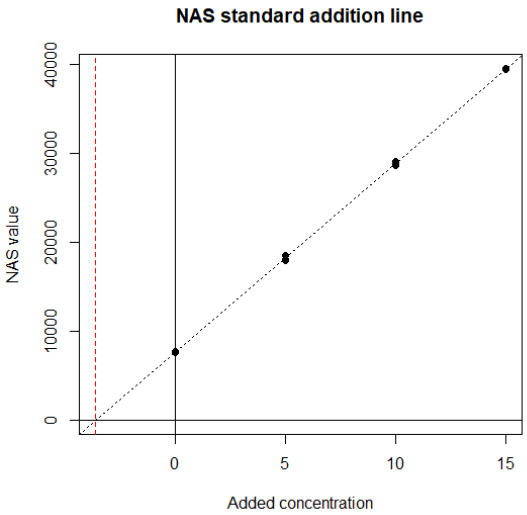


Figure 4.3.3: NAS-standard addition line for saffron-GC case. The red vertical line indicates the extrapolated concentration

From Table 4.3.3, it can be seen that the extrapolated concentration ($3.58\%_{w/w}$) is underestimated compared to the expected value ($5\%_{w/w}$). However, the relative standard deviation ($0.726\%_{w/w}$) makes the extrapolated value not significantly different from the expected one, with *p*-value = 0.0687 (15 degrees of freedom). One problem that could be relevant is the limit of detection: in this case, it is very close to the extrapolated concentration ($3.27\%_{w/w}$). This is probably due to the fact that the chromatograms of blank samples (empty vials) are not flat, but show some peaks due to laboratory air. These peaks, although of low intensity, are summed to sample peaks and may have some influence on the projection (***H***) matrix of NAS. Then, when the blank chromatograms are projected on the ***H*** matrix, the effect of blank signals may be in some way amplified, increasing *LoD* value.

Conclusions

NAS method was applied to saffron samples adulterated on purpose by turmeric, and analyzed in head-space mode by flash-GC. The extrapolated concentration of turmeric is a bit underestimated, but not significantly different from the expected value. Therefore, the concentration of the analyte of interest was obtained by the analysis of its volatile fraction, and not by the direct analysis of the powder, as it happens for the other cases presented in this Thesis.

References

1. Biniecka M, Caroli S (2011) Analytical methods for the quantification of volatile aromatic compounds. TrAC - Trends Anal Chem 30:1756–1770 . doi: 10.1016/j.trac.2011.06.015
2. Koek MM, Jellema RH, van der Greef J, et al (2011) Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. Metabolomics 7:307–328 . doi: 10.1007/s11306-010-0254-3 [doi]\r254 [pii]
3. Imran M, Nadeem M, Saeed F, et al (2017) Immunomodulatory perspectives of potential biological spices with special reference to cancer and diabetes. Food Agric. Immunol. 28:543–572
4. Bhagat N, Chaturvedi A (2016) Spices as an alternative therapy for cancer treatment. Syst Rev Pharm 7:46–56 . doi: 10.5530/srp.2016.7.7
5. Butt MS, Naz A, Sultan MT, Qayyum MMN (2013) Anti-oncogenic perspectives of spices/herbs: A comprehensive review. EXCLI J. 12:1043–1065
6. Mirmosayyeb O, Tanhaei A, Sohrabi HR, et al (2017) Possible role of common spices as a preventive and therapeutic agent for Alzheimer's disease. Int. J. Prev. Med. 2017
7. Sarris J (2018) Herbal medicines in the treatment of psychiatric disorders: 10-year updated review. Phyther. Res. 32:1147–1162
8. Priyanka R, Vasundhara M, Ashwini J, et al (2014) Screening fresh, dry and processed turmeric (Curcuma longa L.) extract against pathogenic bacteria. Res J Pharm Biol Chem Sci 5:1041–1046
9. Shahmoradi Ghaheh F, Mortazavi SM, Alihosseini F, et al (2014) Assessment of antibacterial activity of wool fabrics dyed with natural dyes. J Clean Prod 72:139–145 . doi: 10.1016/j.jclepro.2014.02.050
10. Bosmali I, Ordoudi SA, Tsimidou MZ, Madesis P (2017) Greek PDO saffron authentication studies using species specific molecular markers. Food Res Int 100:899–907 . doi: 10.1016/j.foodres.2017.08.001
11. Moore JC, Spink J, Lipp M (2012) Development and Application of a Database of Food Ingredient Fraud and Economically Motivated Adulteration from 1980 to 2010. J. Food Sci. 77
12. Johnson R (2014) Food Fraud and " Economically Motivated Adulteration " of Food and Food Ingredients. Congr Res Serv Rep January:1–40
13. Varliklioz Er S, Eksi-Kocak H, Yetim H, Boyaci IH (2017) Novel Spectroscopic Method for Determination and Quantification of Saffron Adulteration. Food Anal Methods 10:1547–1555 . doi: 10.1007/s12161-016-0710-4
14. Petrakis EA, Polissiou MG (2017) Assessing saffron (Crocus sativus L.) adulteration with plant-derived adulterants by diffuse reflectance infrared Fourier transform spectroscopy coupled with chemometrics. Talanta 162:558–566 . doi: 10.1016/j.talanta.2016.10.072
15. Petrakis EA, Cagliani LR, Polissiou MG, Consonni R (2015) Evaluation of saffron (Crocus sativus L.) adulteration with plant adulterants by1H NMR metabolite fingerprinting. Food Chem 173:890–896 . doi: 10.1016/j.foodchem.2014.10.107
16. Han J, Wanrooij J, van Bommel M, Quye A (2017) Characterisation of chemical components for

identifying historical Chinese textile dyes by ultra high performance liquid chromatography – photodiode array – electrospray ionisation mass spectrometer. J Chromatogr A 1479:87–96 . doi: 10.1016/j.chroma.2016.11.044

17. Urbani E, Blasi F, Chiesi C, et al (2015) Characterization of volatile fraction of saffron from central Italy (Cascia, Umbria). Int J Food Prop 18:2223–2230 . doi: 10.1080/10942912.2014.968787
18. Amanpour A, Sonmezdag AS, Kelebek H, Selli S (2015) GC-MS-olfactometric characterization of the most aroma-active components in a representative aromatic extract from Iranian saffron (Crocus sativus L.). Food Chem 182:251–256 . doi: 10.1016/j.foodchem.2015.03.005
19. Karabagias IK, Koutsoumpou M, Liakou V, et al (2017) Characterization and geographical discrimination of saffron from Greece, Spain, Iran, and Morocco based on volatile and bioactivity markers, using chemometrics. Eur Food Res Technol 243:1577–1591 . doi: 10.1007/s00217-017-2866-6
20. Anastasaki E, Kanakis C, Pappas C, et al (2009) Geographical differentiation of saffron by GC-MS/FID and chemometrics. Eur Food Res Technol 229:899–905 . doi: 10.1007/s00217-009-1125-x
21. Aliakbarzadeh G, Parastar H, Sereshti H (2016) Classification of gas chromatographic fingerprints of saffron using partial least squares discriminant analysis together with different variable selection methods. Chemom Intell Lab Syst 158:165–173 . doi: 10.1016/j.chemolab.2016.09.002

CHAPTER 4.4: QUANTIFYING API POLYMORPHS IN FORMULATIONS USING X-RAY POWDER DIFFRACTION AND MULTIVARIATE STANDARD ADDITION METHOD COMBINED WITH NET ANALYTE SIGNAL ANALYSIS

The present paper was published by the journal “European Journal of Pharmaceutical Sciences” (reference: 2019, 130:36-43; https://doi.org/10.1016/j.ejps.2019.01.014). It presents an application of NASSAM to pharmaceutical ingredients analyzed by X-ray powder diffraction. The experimental work was carried out in our laboratories, under the supervision of Prof. Lucia Maini. Prof. Giuliano Galimberti (here and in the other chapters concerning NAS), helped with computations and development of NAS, while Dr. Rocco Caliandro developed and applied to these data the algorithm RootProf, that will be presented in the paper. Prof. Dora Melucci and I performed all NAS analyses. All the cited authors contributed to write the manuscript paper, each one describing his part of the work.

CHAPTER 4.5: NAS APPLIED TO IR-ATR

Introduction

The present chapter shows the application of NAS to data acquired by infrared spectroscopy in attenuated total reflection (IR-ATR) mode. The aim of this study is the quantification of biogenic silica in marine sediments. Thus, in this case, the matrix was natural and totally unknown, making almost impossible the use of a quantitative not-destructive method different from standard addition method. Some problems with NAS arose in this case; therefore, also an alternative method was applied to reach the goal.

Biogenic silica (BSi) is produced by diatoms, unicellular microalgae present in oceans, waterways, and soils [1], that use silica for their cell wall, called frustule [2]. At diatoms death, frustules may resist and settle on the ocean floor (or in diatom living place) [3]. The interest in diatoms lays in the fact that these act as a "pump" for CO_2, transferring it from air into the ocean [4, 5], and fix it to produce oxygen [6]. It has been estimated that diatoms produce 20% of the world oxygen by photosynthesis. Therefore, the study of diatoms may help to understand CO_2 cycle in order also to understand (and forecast) climate changes [5]. To evaluate the presence of diatoms in past periods, most of all in oceans, the quantification of BSi in sediments has been thought as a good indicator. However, it suffers from some drawbacks [5]: most part of the frustules dissolves in water and, while "falling" from surface waters to sea floor, BSi is subjected to marine currents. As a consequence, it can be moved from one part of the ocean to another, and the sediments may not reflect anymore the diatom activity at the surface [7]. This is why other indicators of diatoms activity have been studied, such as organic carbon and $CaCO_3$ [8], ^{210}Pb [9], and barium [10]. Anyway, settled BSi stratifies over time and it can be collected and studied by coring sea (or lake) floor. Therefore, the quantification of BSi at different core-depth may give information about diatoms activity over time.

The study of BSi is a challenge also from the analytical point of view. The greatest problem is the presence of lithogenic silica (aluminosilicates) in sediments [11, 12], derived from volcanic activity or transported by rivers, that is difficult to distinguish from BSi. The main difference between aluminosilicates and BSi is that the former is crystalline, while the latter is amorphous. Therefore, the official technique that was developed for BSi study is the so-called "wet method" [11]. The wet method is based on a different rate of dissolution of BSi and aluminosilicates in an alkaline solution, BSi being more rapid in solving. Thus, the sediment is solved in an alkaline solution, BSi is separated by aluminosilicates and it is collected in the supernatant of such solution. Finally, it can be analyzed without the interference of lithogenic silica. The wet method is economic because it only requires an alkaline medium. However, it is very time-consuming, because it requires several steps of dissolution and filtration to be sure to solve all the BSi. Moreover, results have shown a strong dependence on experimental conditions, and some lack of reproducibility [13]. These drawbacks obviously aggravate the destructivity of the technique.

Therefore, some alternative techniques have been proposed for the quantification of BSi in sediments. For example, the direct analysis of sediments by X-ray powder diffraction [14]. The problem, in this case, is that BSi (amorphous) "signal" in XRPD is a "bulge" that, in general, is considered a non-signal. A possible

solution would be to heat the sample in order to convert BSi to crystalline cristobalite [15]. Another idea was to prepare a set of synthetic sediment mixtures and to calculate a PLS model in order to quantify BSi by interpolation [16]. It is, however, clear that such a method requires a prior knowledge of all chemical species present in the matrix.

Besides XRPD, the two analytical techniques mostly used for this kind of research are X-ray fluorescence [8] and FTIR spectroscopy [8, 17]. In particular, the latter has gained attention because of its simplicity. FTIR has been used for quantification purposes, coupled with chemometrics and using the entire spectra [16], but also exploiting the Beer-Lambert law to quantify BSi and other minerals using their peculiar absorption bands [18]. Also in this case, however, the problem is signal overlapping. In fact, the main BSi IR absorption bands are [16]: ~470 cm^{-1}, caused by asymmetric vibrations of SiO_4 tetrahedron, ~800 cm^{-1}, due to stretching vibrations of Si-O-Si group, ~945 cm^{-1}, assigned to Si-OH vibration, and ~1100 cm^{-1}, caused by asymmetric stretching of SiO_4 tetrahedron. The problem is that only the band at ~945 cm^{-1} may be considered characteristic of BSi, because the others are shared with lithogenic quartz. Due to its crystallinity, quartz has a low presence of Si-OH groups, thus its absorption at ~945 cm^{-1} is low (quartz has also a characteristic band of absorption at ~695 cm^{-1}, due to symmetric bending of SiO_4 tetrahedron).

In the present study, BSi was quantified in several sediment samples by the use of NASSAM method applied to FTIR spectra collected in ATR mode. Another problem of this kind of analysis is, in general, the scarcity of samples, due also to the difficulty in collecting them. For the NASSAM procedure, part of the sample is lost for producing the standard-added samples, but IR-ATR is totally non-destructive because it neither requires the preparation of a KBr tablet as in traditional FTIR analysis. Therefore, the prepared samples are preserved for possible further analyses. As already stated, NAS quantification showed some problems for some samples; therefore, also an alternative chemometric method was tried, based on orthogonal signal correction (OSC) [19], a pre-processing technique that seems very suitable for IR spectra [20].

Materials and Methods

Sample Collection

Sediment samples come from the Gondwana Station, the so-called "Site D" (Figure 4.5.1), an Italian research center located in the western sector of the Ross Sea, Antarctica, at 75° 06' S and 164° 28' 5" E.

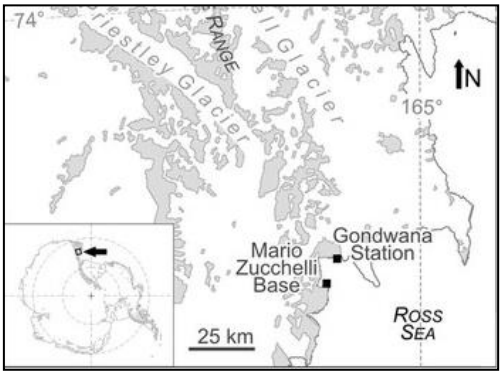


Figure 4.5.1: Location of Gondwana Station, Site D

The sediment core from which samples come was collected during 2003-2004 PNRA (Italian National Program for Research in Antarctica) Campaign at a depth of 972 m. To collect the sediment, a box corer was used, then a plastic cylinder (22 cm long, 12 cm of diameter) was inserted into the corer to obtain one core. Immediately after collection, the core was sub-sampled to obtain sections of 1 cm thickness, stored at -20°C in precleaned polycarbonate Petri capsules. Sub-samples were named with a two digits code: a letter indicating the sampling place, "D", and a number, from 0 to 21, indicating the core height (0 indicating the top). Part of these samples were analyzed *in situ*, while another part was sent to CNR (Consiglio Nazionale delle Ricerche, Bologna, Italy) for BSi analysis. On five of these samples, a wet analysis was carried out by a CNR expert, giving some expected values for NASSAM method (due to the high time necessary, the wet analysis was not carried out on all samples). Another artificial sample was prepared and analyzed by a ring test, certifying its BSi content at 53%$_{w/w}$.

Samples Preparation and Analysis

20 natural samples (from D0 to D21, D3 was not present) and the "53%" standard were prepared for the application of NASSAM method and analyzed always in the same way. For standard additions, diatomaceous flour (or diatomite) was used as a proxy of the analyte of interest, BSi. Diatomite was purchased by Sigma Aldrich (Merck, Darmstadt, Germany) under the commercial name Celite®.

Each sample was manually ground in an agate mortar for two minutes before preparation, in order to have a powder as homogeneous as possible and to remove small pebbles. Then, it was heated at 105°C for 1 h in a ventilated oven to remove water traces. The same procedure was carried out for diatomite. Four aliquots of each sample were prepared by adding diatomite according to Table 4.5.1, in order to have a total weight of 200 mg and added concentrations of 5%, 10%, and 15%$_{w/w}$. Powders were weighted with a five decimals weight scale.

	Sediment weight (mg)	**Diatomite weight (mg)**	**Added concentration (%$_{w/w}$)**
add.0	200	0	0
add.1	190	10	5
add.2	180	20	10
add.3	170	30	15

Table 4.5.1: Biosilica standard added samples preparation

Added samples were then homogenized with an MM200 ball mill (Retsch, Düsseldorf, Germany). Each sample was put in a 1.5 ml volume stainless steel cylinder and mixed for 1 h at 20 Hz. Then, samples were kept in a dryer until analysis. Some samples and standards were analyzed with a scanning electron microscope (SEM) to look for the presence of frustules. Figure 4.5.2 reports some of the obtained images, that demonstrate that the diatomite used and the 53% standard present frustules similar to the ones observed in sediment samples.

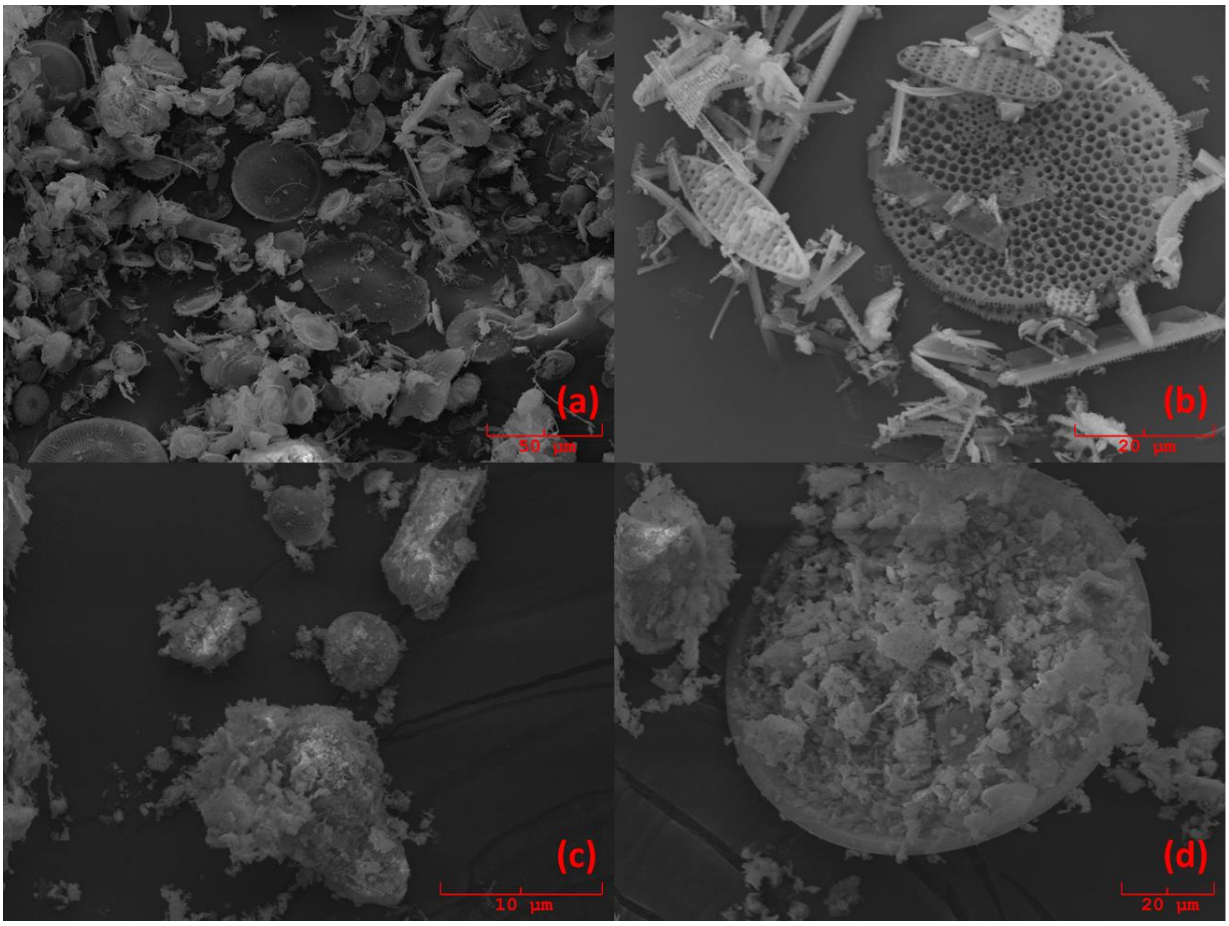


Figure 4.5.2: SEM images of samples: a) pure diatomite; b) 53% standard; c) D1 sample; d) D4 sample

The spectrophotometer used for the present work was a Bruker ALPHA FT-IR (Bruker, Billerica, MA, USA) equipped with a Bruker Platinum ATR (Bruker) accessory. The ATR probe had a mercury-cadmium-tellurium (MCT) detector of dimensions 0.6 x 0.6 cm and a single reflection diamond crystal. In order to improve spectra reproducibility (that is important for a quantitative analysis), powder samples were put inside a ring (1 cm diameter) and weighted before the analysis, with the aim of creating a sort of "tablet" of the same weight and dimensions for all samples. Samples were also subjected always to the same pressure given by the mechanical press of the ATR probe. IR spectra were collected in the range 4000-400 cm^{-1}, with a resolution of 4 cm^{-1} and 64 scans for each analysis. Analyses were carried out at room conditions and a blank spectrum (air) was collected and automatically subtracted before each analysis. 4-6 replicates were analyzed for each sample, mixing the powder inside the ring between the replicates.

Results and Discussion

The IR range from 4000 to 1300 cm^{-1} does not contain any signal due to the samples, there is only a residual of the water signal, thus, it was discarded from all analyses.

Previous to NAS computations, spectra were pre-processed by multiplicative scatter correction (MSC) [21]. MSC is a common pre-processing method for IR spectra, and has the aim of removing the scatter present in

spectra. The computation for this method is divided into two parts [22]. The first one regards the estimation of correction coefficients by:

$$\boldsymbol{x_{org}} = b_0 + b_{ref,1} \cdot \boldsymbol{x_{ref}} + \boldsymbol{e} \qquad \text{(eq. 4.5.1)}$$

where $\boldsymbol{x_{org}}$ is one original sample spectrum, $\boldsymbol{x_{ref}}$ is a reference spectrum, and $\boldsymbol{e}$ is the residual. In most of the applications, as a reference spectrum the average of all involved spectra is used. Then each spectrum is corrected using the correction coefficients *b* (scalars):

$$\boldsymbol{x_{corr}} = \frac{\boldsymbol{x_{org}} - b_0}{b_{ref,1}} = \boldsymbol{x_{ref}} + \frac{\boldsymbol{e}}{b_{ref,1}} \qquad \text{(eq. 4.5.2)}$$

The result is that $\boldsymbol{x_{corr}}$ spectra are much more reproducible than the original ones. MSC was applied to the replicates of each added sample in order to obtain a correction as that reported in Figure 4.5.3.

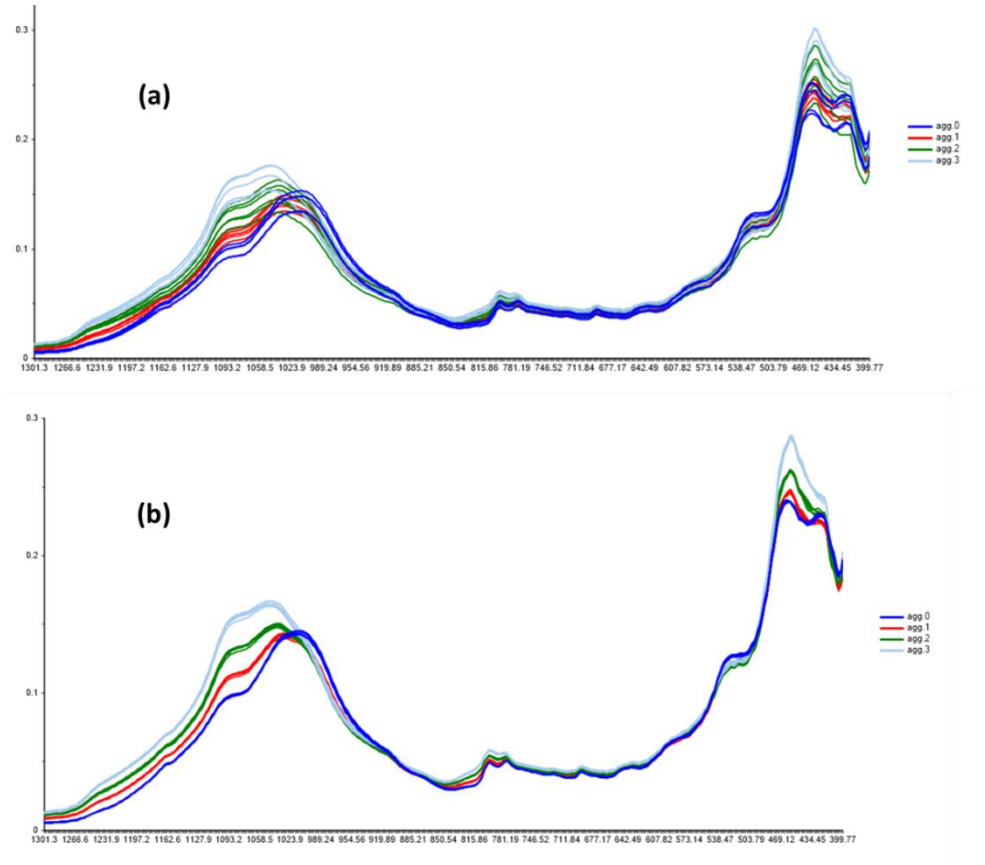


Figure 4.5.3: Example of sediment spectra a) original and b) after MSC pre-treatment

Figure 4.5.3, shows that the original spectra have already a good reproducibility, obtained by the use of the ring during ATR analyses. However, MCS pre-treatment improves reproducibility and partially enhances the increase of signal intensity with the increasing of BSi concentration. Figure 4.5.4 shows the same spectra reported in Figure 4.5.3 (D10 sample) with the addition of diatomite spectra (after MSC pre-treatment). Figure 4.5.4 shows that the signal increasing in added samples is most visible where the diatomite has its absorbance peaks, that are the ones reported in the literature [16]. However, the band that should be

characteristic of BSi (~945 cm^{-1}) is very flat also for the pure diatomite. Therefore, it neither could be used for qualitative purposes.

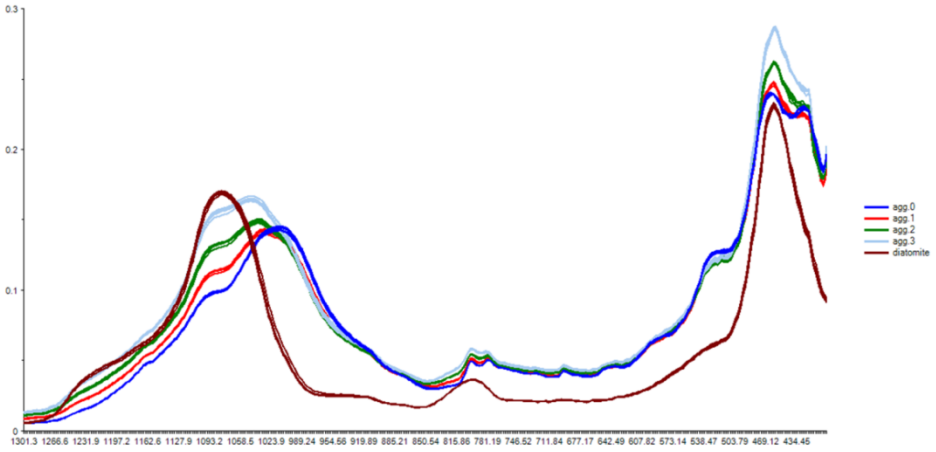


Figure 4.5.4: D10 and pure diatomite spectra

On MSC-pretreated spectra of all samples, the NAS method was applied, always with Ferré-Bro variant. The following Table 4.5.2 shows extrapolated concentrations, standard deviations, LoD, NAS R^2, and expected values (if present) for all sediment samples. For all samples, the 2nd PLS-factor was used.

Sample	C_E ($\%_{w/w}$)	$s_{C,E}$ ($\%_{w/w}$)	LoD	R^2	Expected Concentration ($\%_{w/w}$)
Std 53%	53.6	6.02	126	0.992	53
D0	3.23	0.903	10	0.996	
D1	3.24	1.92	30	0.993	
D2	10.4	1.85	24	0.994	
D4	12.0	1.16	18	0.988	13.9 ± 2.1
D5	5.41	1.10	18	0.993	
D6	14.2	1.54	8	0.989	14 ± 2
D7	5.87	1.27	7	0.999	
D8	9.45	0.671	13	0.993	
D9	< 0	-	-	-	9.0 ± 1.4
D10	9.44	0.885	1	0.997	
D11	2.94	0.646	9	0.994	
D12	< 0	-	-	-	
D13	< 0	-	-	-	
D14	< 0	-	-	-	
D15	3.25	0.184	4	0.993	
D16	< 0	-	-	-	
D17	2.71	0.535	4	0.994	
D18	4.80	1.67	5	0.996	4.27 ± 0.65
D19	4.04	0.332	5	0.998	
D20	2.31	0.579	14	0.999	
D21	3.12	1.86	9	0.998	3.48 ± 0.53

Table 4.5.2: NAS results for sediment samples

As it can be observed from Table 4.5.2, some drawbacks appeared in these analyses. The most evident is that some samples (D9, D12, D13, D14, and D16) have been reported with extrapolated concentration "< 0". These samples have two distinct behaviors: some of them (D9 and D11) have an R^2 of the final NAS standard addition line always lower than 0.7, for any PLS-factor, therefore their results are not reliable at all; the others have, for all factors, good R^2, but extrapolated values always negative. This means that, in the final NAS standard addition line, either slope or intercept was negative. As an example, in Figure 4.5.5 it is reported the NAS standard addition line for sample D16.

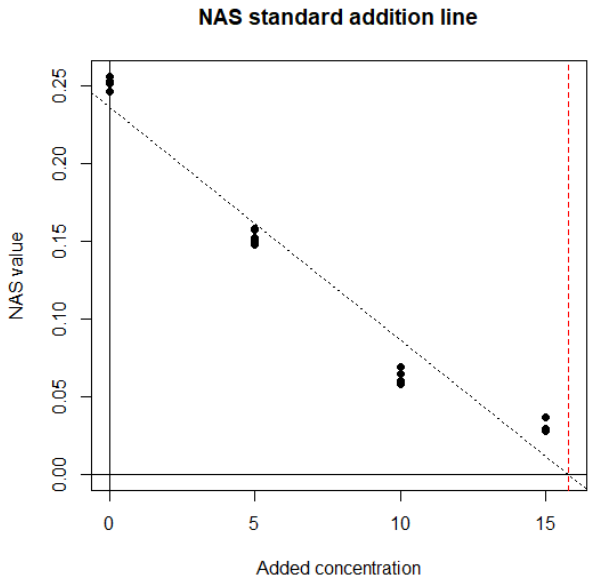


Figure 4.5.5: NAS standard addition line for sample D16, with negative slope

Such a behavior, with negative slope (or, sometimes, intercept), was already shown by NAS in other cases (although not cited), as for example when dealing with X-ray data. In those cases, however, the problem was solved by pre-treating data or by removing the zero-added sample (and this is one of the reasons that brought to those choices in some cases). In this case, however, no satisfactory solution was found for the five samples here reported as "< 0": the removal of zero-added sample did not change the situation for D16-like samples and brought the others to the negative slope case. Further pre-treatments (SNV, area normalization) were in some cases even deleterious, because also the linearity was eliminated. Moreover, one of the goals of this work is to find a chemometric method that could be used for every sediment sample analyzed by IR-ATR, thus the use of a further pre-treatment (besides MSC) would be considered satisfactory if good results would be obtained for all sediment samples, which is not the case for any method used till now. A possible and simple explanation for the behavior of these samples could be a mistake occurred during sample measurements or the absence of signal linearity. However, the NAS method has already demonstrated being somehow "resistant" to such drawbacks (as in the case of Raman data). Therefore, the solution to this problem is still under study.

Another problem emerging from Table 4.5.2 is that of *LoD*s, that, except for D6 and D10, are always higher than the extrapolated values. This is probably due to the fact that the blank signal, although always lower than all sample signals, is not flat, thus it gives a NAS vector with a significant intensity. This problem could probably be solved by removing the baseline from all spectra, including blank ones, that in IR spectroscopy is due to scattering effects. However, the baseline-correction pre-treatment has shown some problems when used together with MSC (most of the results were in the order of $10^{-2}\%_{w/w}$). Thus, also in this case, a solution is still under study.

Anyway, there are also positive and encouraging remarks shown by Table 4.5.2, the most important of which is that the extrapolated values for those samples for which an expected one was present, are close to that (except for the already cited D9). It is an important result, because the expected values were obtained by the destructive and time-consuming wet method, while the NASSAM one is more rapid and does not destroy the analyzed samples, thus it could be considered a valid alternative to the official method. Standard deviations of the extrapolated values are, in some cases, quite high, but this is due to the low reproducibility of IR-ATR analyses in terms of signal intensities. This is why it is important to use MSC as a pre-treatment, otherwise standard deviations would be probably even higher. Therefore, other pre-treatment methods will be explored in further works, as for example EMSC [21], an extension of MSC taking into account the second-order polynomial fitting to the reference spectrum. Moreover, it is reported in the literature [23] that BSi concentration in superficial sediments is relatively low ($< 10\%_{w/w}$), and this is in agreement with, at least, the results obtained for the two most superficial samples, D0 and D1.

Even if not conclusive, a second approach for the quantification of BSi in sediments by IR-ATR is here proposed. The starting point is pre-processing original spectra with orthogonal signal correction (OSC) [20], instead of MSC. OSC in general works as a PLS regression (thus, unlike most of the other pre-processing methods, requires also a response, *y*, variable), but, unlike PLS, it calculates the weight vectors, $\boldsymbol{w}$, in order to minimize, and not maximize, the covariance between $\boldsymbol{X}$ (the matrix of spectra) and $\boldsymbol{y}$ (the vector of added concentrations, $\boldsymbol{c_{add}}$). As for PLS, "principal components" are calculated iteratively, and the computation is stopped when a satisfactory result has been reached. The final goal of this procedure is to obtain a corrected $\boldsymbol{X}$ matrix that is (almost) orthogonal to $\boldsymbol{y}$. This brings to a corrected matrix in which as much of the information related to $\boldsymbol{y}$ has been retained, while the discarded information is that not related to $\boldsymbol{y}$ (orthogonality guarantees it). In fact, the greatest problem of other pre-treatment methods is that there is no control over which part of the information is discarded from $\boldsymbol{X}$. Thus, it may happen that also part of (if not all) the information related to the response(s) variable(s) is removed from data, making a further regression model unstable. OSC, instead, guarantees a corrected $\boldsymbol{X}$ matrix very suitable for a regression model.

Therefore, OSC was applied to all samples. The total number of components used for all samples is 4. Figure 4.5.6 shows a sample (D2) matrix corrected by OSC.

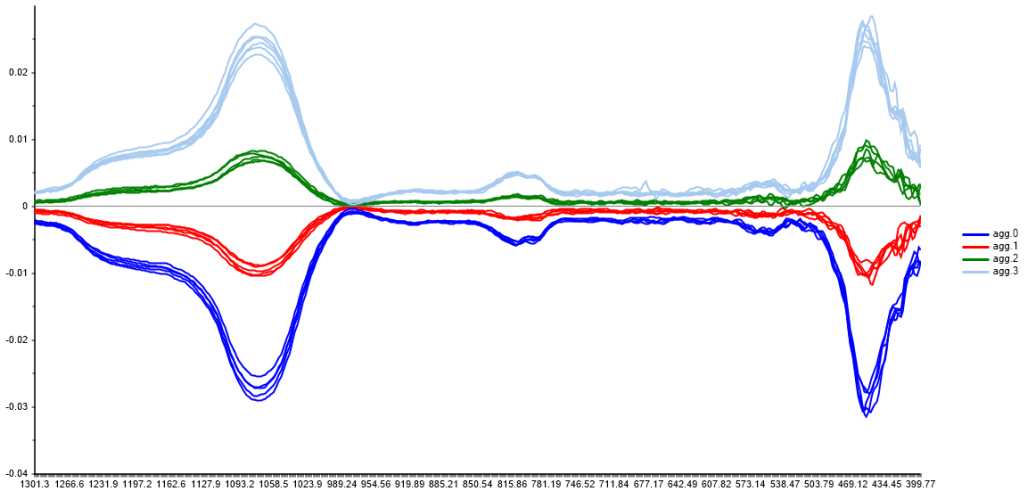


Figure 4.5.6: D2 sample spectra pre-treated by OSC

What is interesting to note from Figure 4.5.6 is that the profiles obtained by OSC are very close to those of pure diatomite reported in Figure 4.5.4 (brown profile), but the diatomite spectra were not used for the computation. Thus, OSC seems to actually extract the signal due only to the analyte of interest, as NAS does. Sometimes, as reported in Figure 4.5.7 (for sample D10), there are some intersections of sample signals, but the peaks are, anyway, in the same positions.

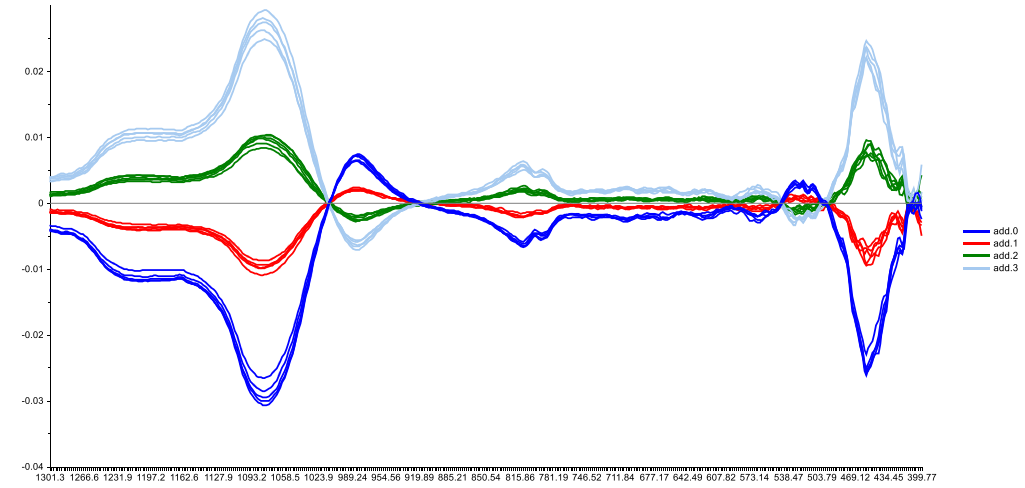


Figure 4.5.7: D10 sample spectra pre-treated by OSC

The problem is how to use such data for BSi quantification. OSC, as presented here, is incompatible with NAS: the specularity of OSC-spectra (add.0 with add.3 and add.1 with add.2) is maintained at the end of any NAS computation, thus there is no linearity on the final standard addition line and it cannot be used for extrapolation. Any other pre-treatment does not solve this problem (and it is likely to lose the advantage gained by the use of OSC). Therefore, a possible correct procedure would be to use a blank sample (i.e. a sample devoid of the analyte of interest): analyze it by IR-ATR, project its spectrum on the OSC model (and not use it to compute the pre-treatment), then project the OSC-spectrum on a PLS model computed with the

standard added samples. However, the original problem comes back: it is very difficult to obtain such real blank sample, therefore such a solution is hardly practicable.

As a possible use of these data, a tentative was made to compute a PLS model with the three added samples (add.1, add.2, and add.3), and then predict the concentration value of the zero-added sample. The idea behind this method is that OSC removes the signal due to the matrix: thus, on the zero-added sample (the sediment), only the signal due to BSi naturally present in the sample remains, thus it becomes a sort of "standard" which can be used for extrapolation. The following Table 4.5.3 reports the calculated concentrations for such method and the differences (in absolute values) between NAS-values and OSC-values. For all samples, the extrapolation was performed on the 4th PLS-component, the same number used for OSC pre-treatment (R^2 both in calibration and in validation were always higher than 0.95 for that component). The values reported are the mean of the calculated ones for the 4-6 replicates of the zero-added sample.

	Std 53%	**D0**	**D1**	**D2**	**D4**	**D5**	**D6**
OSC	31.5	0.300	5.25	10.9	8.30	1.50	3.27
|OSC-NAS|	22.2	2.9	2.0	0.5	3.7	3.9	10.9

	D7	**D8**	**D9**	**D10**	**D11**	**D12**	**D13**	**D14**
OSC	6.18	19.7	7.24	9.31	14.6	13.0	4.48	1.71
|OSC-NAS|	0.3	10.2	--	0.1	11.6	--	--	--

	D15	**D16**	**D17**	**D18**	**D19**	**D20**	**D21**
OSC	6.09	11.5	8.25	3.50	4.82	2.75	7.50
|OSC-NAS|	2.8	--	5.5	1.3	0.8	0.4	4.4

Table 4.5.3: Predicted values for spectra pre-treated by OSC and difference with NAS extrapolated values

As it can be seen from Table 4.5.3, the differences between concentrations obtained in this case and NAS-extrapolated values are, in most of the cases, not so high: OSC values may be considered inside the confidence interval of NAS ones. However, there are also cases for which OSC values are strongly different from NAS (also of the double, as for D8, or more, as for D6 and D11). Moreover, the concentration calculated for 53% standard is significantly different from the expected value, and also the other samples for which an expected value is present are not so good (except for D9, expected 9.0 ± 1.4 and D18, expected 4.27 ± 0.65). Another problem is that the standard deviation, calculated automatically by the software, are always in the degree of magnitude of the calculated concentration, making it not significantly different from zero (that's why standard deviations were not reported). However, these results are not totally unreliable: it has been reported in the literature [23, 24] that the BSi content in Ross sea hardly reaches 20%$_{w/w}$ (due to dissolution and water fluxes), and OSC values never reach this limit (as NAS ones). Therefore, the problem might be simply to find a way for using these data in a trustable way, for example finding a proper "blank" sample on which to perform extrapolation or studying a way for removing the specularity of spectra when computing OSC.

Conclusions

NAS procedure was applied to sediment samples analyzed by IR-ATR spectroscopy. Although some critical issues emerged, some encouraging results were obtained. In fact, when a reference value was present, NAS results were in agreement with it, and the results are also in agreement with what reported in literature for BSi concentration in Ross Sea [23, 24]. This demonstrates that NASSAM methodology may be a valid alternative to the traditional wet method for the quantification of biogenic silica in sediments. Most of the problems may be probably solved by further improving the computations, without even the need of re-prepare and re-analyze the samples.

Moreover, a second approach to the problem has been proposed, considering a different data pre-treatment as a possible solution for the NAS problems. Also in this case, some encouraging results have been obtained, but the method still needs to be developed.

References

1. Guillard RRL, Ryther JH (1962) Studies of Marine Planktonic Diatoms: I. Cyclotella Nana Hustedt, and Detonula Confervacea (Cleve) Gran. Can J Microbiol 8:229–239 . doi: 10.1139/m62-029
2. Hamm CE, Merkel R, Springer O, et al (2003) Architecture and material properties of diatom shells provide effective mechanical protection. Nature 421:841–843 . doi: 10.1038/nature01416
3. Smetacek VS (1985) Role of sinking in diatom life-history cycles: ecological, evolutionary and geological significance. Mar Biol 84:239–251 . doi: 10.1007/BF00392493
4. Schubert CJ, Villanueva J, Calvert SE, et al (1998) Stable phytoplankton community structure in the Arabian sea over the past 200,000 years. Nature 394:563–566 . doi: 10.1038/29047
5. Ragueneau O, Tréguer P, Leynaert A, et al (2000) A review of the Si cycle in the modern ocean: Recent progress and missing gaps in the application of biogenic opal as a paleoproductivity proxy. Glob Planet Change 26:317–365 . doi: 10.1016/S0921-8181(00)00052-7
6. Reinfelder JR, Milligan AJ, Morel FMM (2004) The Role of the C 4 Pathway in Carbon Accumulation and Fixation in a Marine Diatom. Plant Physiol 135:2106–2111 . doi: 10.1104/pp.104.041319.2106
7. Kumar N, Anderson RF, Mortlock RA, et al (1995) Increased biological productivity and export production in the glacial southern Ocean. Nature 378:675–680 . doi: 10.1038/378675a0
8. Liu X, Colman SM, Brown ET, et al (2013) Estimation of carbonate, total organic carbon, and biogenic silica content by FTIR and XRF techniques in lacustrine sediments. J Paleolimnol 50:387–398 . doi: 10.1007/s10933-013-9733-7
9. Langone L, Frignani M, Labbrozzi L, Ravaioli M (1998) Present-day biosiliceous sedimentation in the Northwestern Ross Sea, Antarctica. J Mar Syst 17:459–470 . doi: 10.1016/S0924-7963(98)00058-X
10. Dymond J, Suess E, Lyle M (1992) Barium in Deep-Sea Sediment: A Geochemical Proxy for Paleoproductivity. Paleoceanography 7:163–181 . doi: 10.1029/92PA00181
11. DeMaster DJ (1981) The supply and accumulation of silica in the marine environment. Geochim Cosmochim Acta 45:1715–1732 . doi: 10.1016/0016-7037(81)90006-5
12. Kamatani A, Oku O (2000) Measuring biogenic silica in marine sediments. Mar Chem 68:219–229 . doi: 10.1016/S0304-4203(99)00079-1
13. Conley DJ (1998) An interlaboratory comparison for the measurement of biogenic silica in sediments. Mar Chem 63:39–48 . doi: 10.1016/S0304-4203(98)00049-8
14. Eisma D, Van Der Gaast SJ (1971) Determination of opal in marine sediments by X-ray diffraction. Netherlands J Sea Res 5:382–389 . doi: 10.1016/0077-7579(71)90019-6
15. Calvert SE (1966) Accumulation of diatomaceous silica in the sediment of the Gulf of California. Geol Soc Am Bull 77:569–572 . doi: 10.1130/0016-7606(1966)77[569:AODSIT]2.0.CO;2
16. Meyer-Jacob C, Vogel H, Boxberg F, et al (2014) Independent measurement of biogenic silica in sediments by FTIR spectroscopy and PLS regression. J Paleolimnol 52:245–255 . doi:

10.1007/s10933-014-9791-5

17. Vogel H, Rosén P, Wagner B, et al (2008) Fourier transform infrared spectroscopy, a new cost-effective tool for quantitative analysis of biogeochemical properties in long sediment records. J Paleolimnol 40:689–702 . doi: 10.1007/s10933-008-9193-7
18. Bertaux J, Froehlich F, Ildefonse P (1998) Multicomponent analysis of FTIR spectra; quantification of amorphous and crystallized mineral phases in synthetic and natural sediments. J Sediment Res 68:440–447 . doi: 10.2110/jsr.68.440
19. Fearn T (2000) On orthogonal signal correction. Chemom Intell Lab Syst 50:47–52 . doi: 10.1016/S0169-7439(99)00045-3
20. Wold S, Antti H, Lindgren F, Öhman J (1998) Orthogonal signal correction of near-infrared spectra. Chemom Intell Lab Syst 44:175–185 . doi: 10.1016/S0169-7439(98)00109-9
21. Rinnan Å, Berg F van den, Engelsen SB (2009) Review of the most common pre-processing techniques for near-infrared spectra. TrAC - Trends Anal Chem 28:1201–1222 . doi: 10.1016/j.trac.2009.07.007
22. Geladi P, MacDougall D, Martens H (1985) Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. Appl Spectrosc 39:491–500 . doi: 10.1366/0003702854248656
23. Frignani M, Giglio F, Accornero A, et al (2003) Sediment characteristics at selected sites of the Ross Sea continental shelf: Does the sedimentary record reflect water column fluxes? Antarct Sci. doi: 10.1017/S0954102003001123
24. Nelson DM, DeMaster DJ, Dunbar RB, Smith WO (1996) Cycling of organic carbon and biogenic silica in the Southern Ocean: Estimates of water-column and sedimentary fluxes on the Ross Sea continental shelf. J Geophys Res C Ocean 101:18519–18532 . doi: 10.1029/96JC01573

APPENDIX A: DoE TABLE FOR PLANTS PROBLEM

DoE table referred to the plant work, shown in chapter 3.1. Values of the Y block are expressed in ppm (metal weight over plant weight).

	X block			Y block				
	Cd	**Pb**	**Cr**	**Cd_{plant}**	**Cr_{plant}**	**Cu_{plant}**	**Pb_{plant}**	**Zn_{plant}**
DOE 1	0	0	0	2.00	0.33	9.6	18.6	71
DOE 2	-1	-1	-1	0.54	0.14	10.6	3.4	65
DOE 3	1	-1	-1	2.37	0.29	9.5	3.7	59
DOE 4	-1	1	-1	0.56	0.23	22.2	29.2	73
DOE 5	1	1	-1	37.84	1.7	15.9	26.8	86
DOE 6	-1	-1	1	0.64	2.3	15.9	3.7	80
DOE 7	1	-1	1	3.46	2.9	17.7	39.6	74
DOE 8	-1	1	1	0.49	1.8	22.7	33.0	63
DOE 9	1	1	1	2.46	1.0	16.7	27.8	59
DOE 10	0	-1	-1	0.9	0.13	8	36	40
DOE 11	-1	0	-1	0.66	0.17	7	47	39
DOE 12	0	0	-1	3.3	0.14	7	3.8	46
DOE 13	1	0	-1	3.4	0.18	9	2.1	36
DOE 14	0	1	-1	2.5	0.18	8.2	10	53
DOE 15	-1	-1	0	2.2	0.84	7.6	6	41
DOE 16	0	-1	0	0.7	0.28	9	13	41
DOE 17	1	-1	0	3.2	0.15	37	0.7	64
DOE 18	-1	0	0	0.35	0.26	8	12	38
DOE 19	1	0	0	1.1	0.29	8	24	49
DOE 20	-1	1	0	2.9	0.15	8.7	5	57
DOE 21	0	1	0	1.5	0.95	7	9	52
DOE 22	1	1	0	2.6	0.30	8.3	8	54
DOE 23	0	-1	1	1.4	0.24	7.6	2	34
DOE 24	-1	0	1	1.5	0.36	8.8	10	51
DOE 25	0	0	1	1.9	1.2	8	3.6	43
DOE 26	1	0	1	2.8	0.68	8.6	2.7	47
DOE 27	0	1	1	2.3	0.55	7.5	4.7	60
DOE 1bis	0	0	0	5.3	0.99	9	16.3	49
DOE 5bis	1	1	-1	6.6	0.30	8	11.4	73
DOE 8bis	-1	1	1	2.0	3.4	8	23.3	59
DOE 15bis	-1	-1	0	2.3	0.58	7	4.4	41
DOE 16bis	0	-1	0	4.6	0.21	6	1.6	36
DOE 17bis	1	-1	0	4.9	0.06	7	0.7	42
DOE 24bis	-1	0	1	3.5	0.18	8	3.0	66
DOE 26bis	1	0	1	3.1	0.36	5	3.7	46
DOE 27bis	0	1	1	2.5	0.33	4	3.7	61

APPENDIX B: DoE TABLE AND MB RESULTS FOR MACHINE OPTIMIZATION PROBLEM

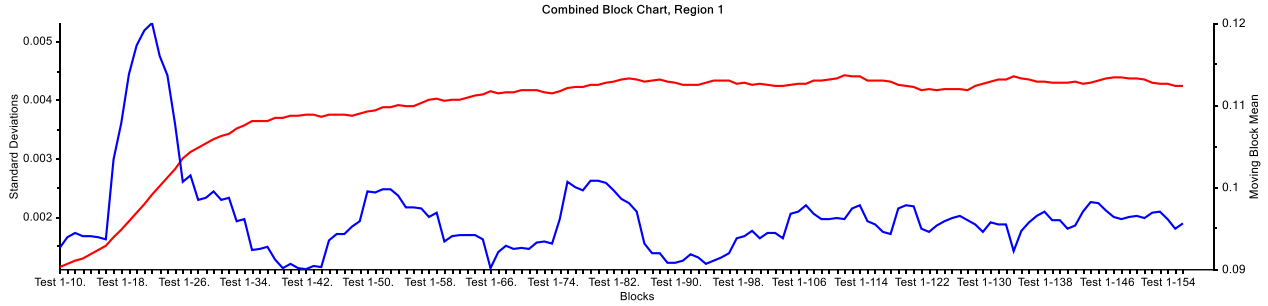
DoE table referred to the bin blender work, shown in chapter 3.2.

	X Block				Y Block		
	Bin Filling (%)	**Rotation Speed (rpm)**	**Mixing Time (min)**	**Particle Size (µm)**	**Mean HPLC Concentrations (ppm)**	**Standard deviation HPLC (ppm)**	**iCarr (%)**
DOE 1	-1	-1	-1	-1	3.14	0.22	22.48
DOE 2	0	-1	-1	-1	3.25	0.45	22.44
DOE 3	1	-1	-1	-1	3.18	0.22	24.03
DOE 4	-1	0	-1	-1	3.26	0.34	23.98
DOE 5	0	0	-1	-1	3.34	0.29	22.47
DOE 6	1	0	-1	-1	3.33	0.68	23.52
DOE 7	-1	1	-1	-1	3.16	0.26	23.49
DOE 8	0	1	-1	-1	3.25	0.46	23.50
DOE 9	1	1	-1	-1	3.37	0.24	23.02
DOE 10	-1	-1	0	-1	3.46	0.40	23.48
DOE 11	0	-1	0	-1	3.23	0.42	23.53
DOE 12	1	-1	0	-1	3.28	0.11	22.05
DOE 13	-1	0	0	-1	3.22	0.07	23.55
DOE 14	0	0	0	-1	3.18	0.12	23.96
DOE 15	1	0	0	-1	3.37	0.67	23.48
DOE 16	-1	1	0	-1	3.14	0.29	24.02
DOE 17	0	1	0	-1	3.13	0.22	23.99
DOE 18	1	1	0	-1	3.26	0.18	23.48
DOE 19	-1	-1	1	-1	3.14	0.20	23.95
DOE 20	0	-1	1	-1	2.91	0.32	22.46
DOE 21	1	-1	1	-1	3.21	0.25	23.54
DOE 22	-1	0	1	-1	3.35	0.50	23.99
DOE 23	0	0	1	-1	3.75	0.84	23.51
DOE 24	1	0	1	-1	3.43	0.15	23.47
DOE 25	-1	1	1	-1	3.13	0.27	23.47
DOE 26	0	1	1	-1	3.11	0.15	24.01
DOE 27	1	1	1	-1	3.18	0.38	23.51
DOE 28	-1	-1	-1	1	3.57	0.45	20.00
DOE 29	0	-1	-1	1	3.38	0.24	21.47
DOE 30	1	-1	-1	1	3.19	0.18	23.52
DOE 31	-1	0	-1	1	3.07	0.18	21.95
DOE 32	0	0	-1	1	3.16	0.31	20.96
DOE 33	1	0	-1	1	3.27	0.45	23.48
DOE 34	-1	1	-1	1	3.56	0.40	20.90
DOE 35	0	1	-1	1	3.51	0.70	20.00
DOE 36	1	1	-1	1	3.04	0.34	22.53
DOE 37	-1	-1	0	1	3.33	0.35	21.98
DOE 38	0	-1	0	1	3.38	0.20	19.97
DOE 39	1	-1	0	1	3.35	0.43	20.53
DOE 40	-1	0	0	1	3.04	0.32	22.02
DOE 41	0	0	0	1	3.22	0.41	20.47
DOE 42	1	0	0	1	3.14	0.56	21.00
DOE 43	-1	1	0	1	3.39	0.31	20.19
DOE 44	0	1	0	1	3.21	0.25	20.97
DOE 45	1	1	0	1	3.29	0.33	20.96
DOE 46	-1	-1	1	1	3.08	0.49	20.53
DOE 47	0	-1	1	1	3.23	0.72	21.56
DOE 48	1	-1	1	1	2.64	0.06	20.03
DOE 49	-1	0	1	1	3.09	0.22	22.51
DOE 50	0	0	1	1	2.63	1.07	21.55
DOE 51	1	0	1	1	3.31	0.36	20.97
DOE 52	-1	1	1	1	3.22	0.06	20.53
DOE 53	0	1	1	1	3.23	0.30	22.46
DOE 54	1	1	1	1	3.06	0.28	21.98
DOE 55	0	0	0	-1	3.56	0.76	24.59
DOE 56	0	0	0	-1	3.34	0.80	22.49
DOE 57	0	0	0	1	2.97	0.09	20.53
DOE 58	0	0	0	1	3.13	0.52	20.96

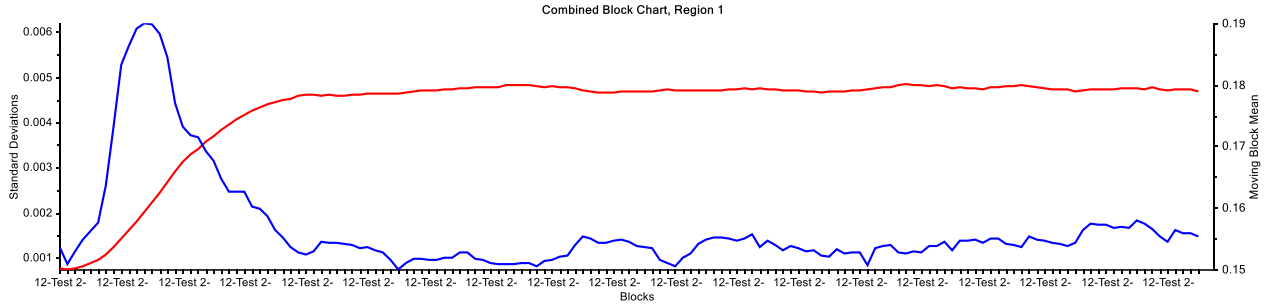
MB ANALYSES GRAPHS OF DoE EXPERIMENTS

For all graphs, red line represents the MB mean, blue line represents MB standard deviation

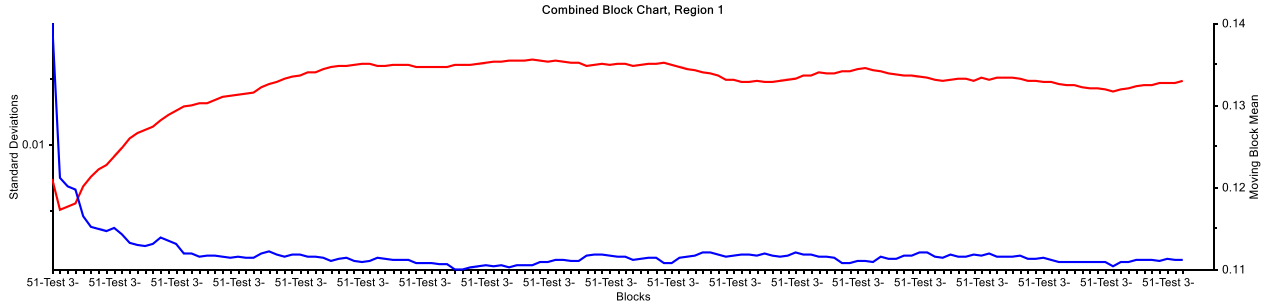
DOE 1



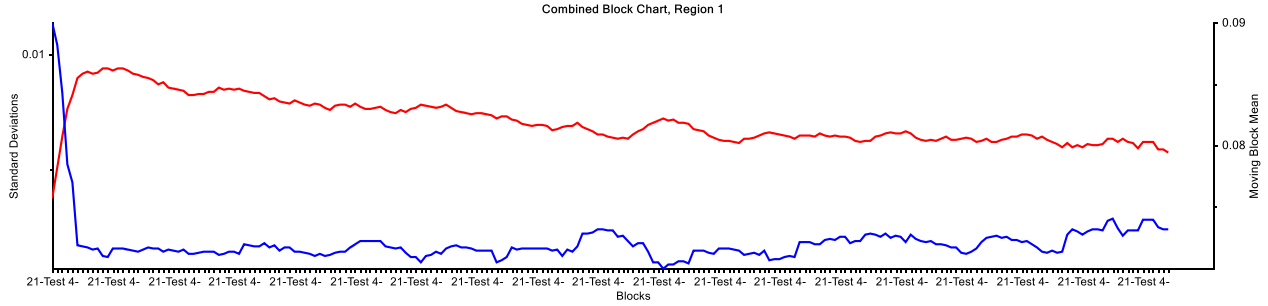
DOE 2



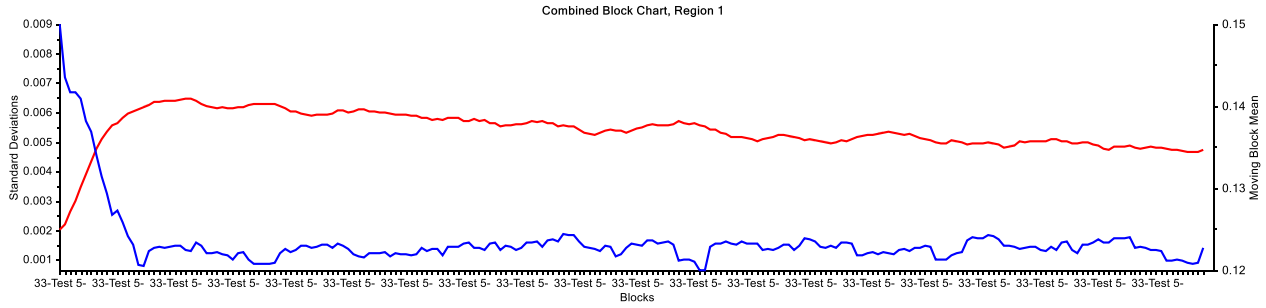
DOE 3



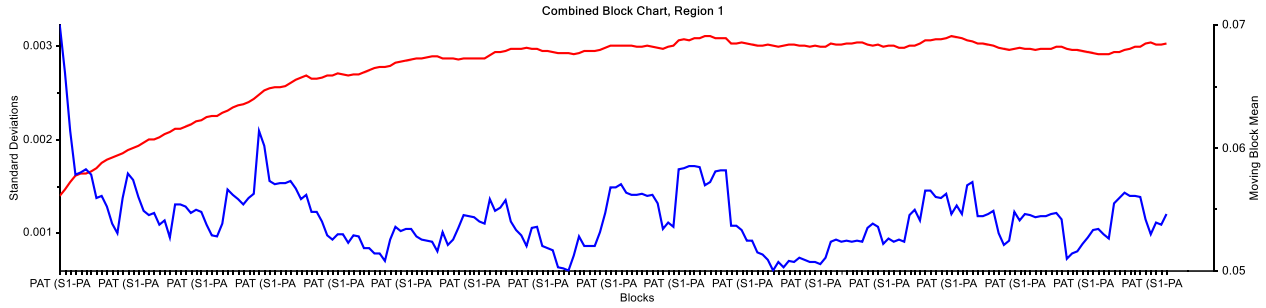
DOE 4



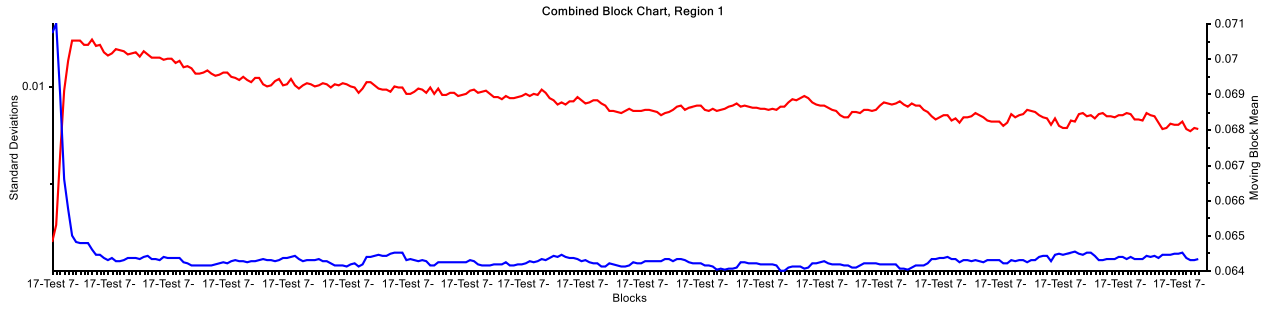
DOE 5



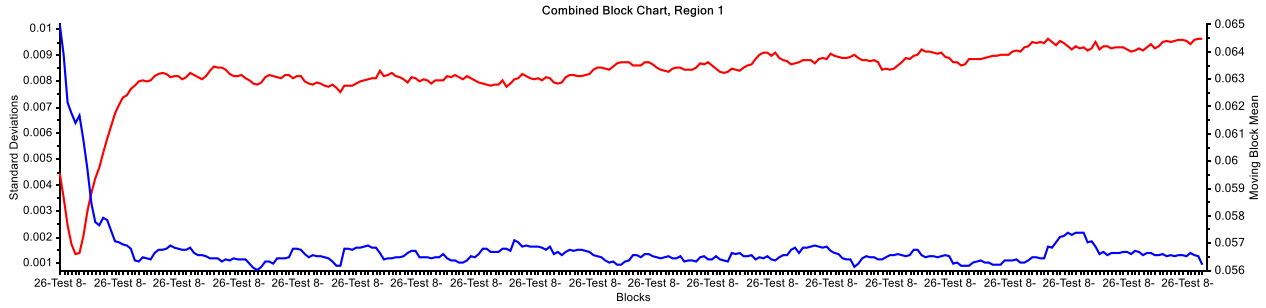
DOE 6



DOE 7

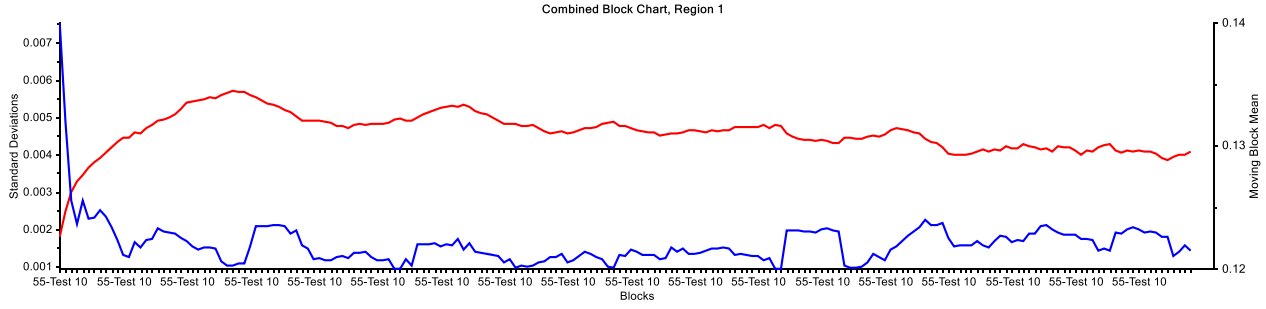


DOE 8

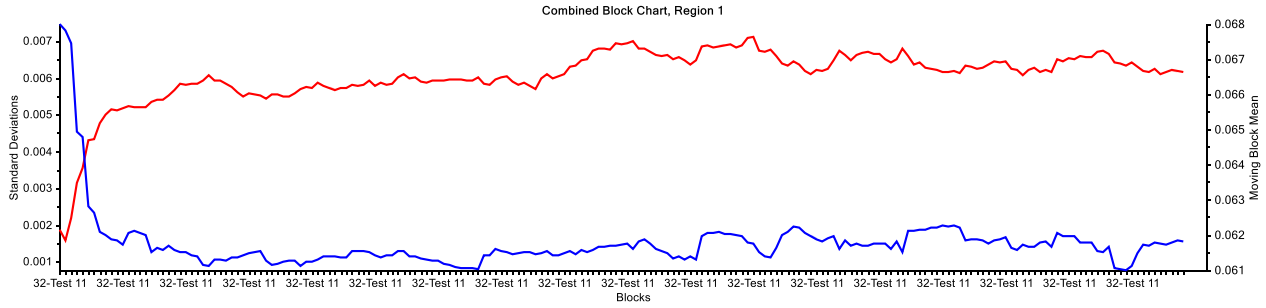


DOE 9 not present due to technical problems

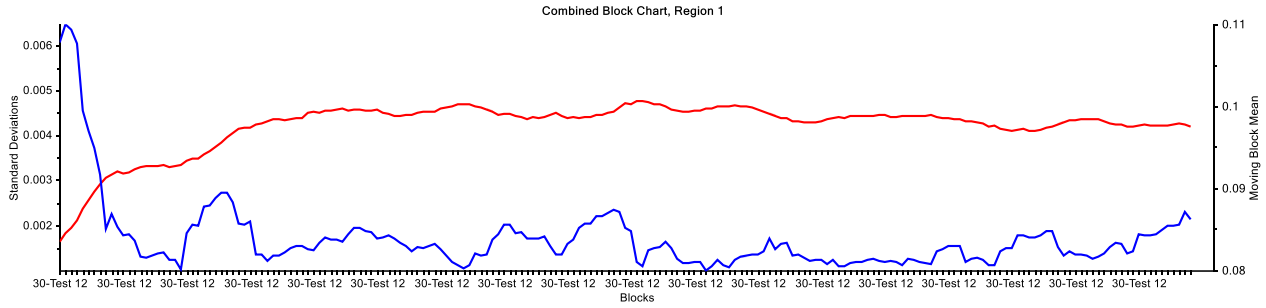
DOE 10



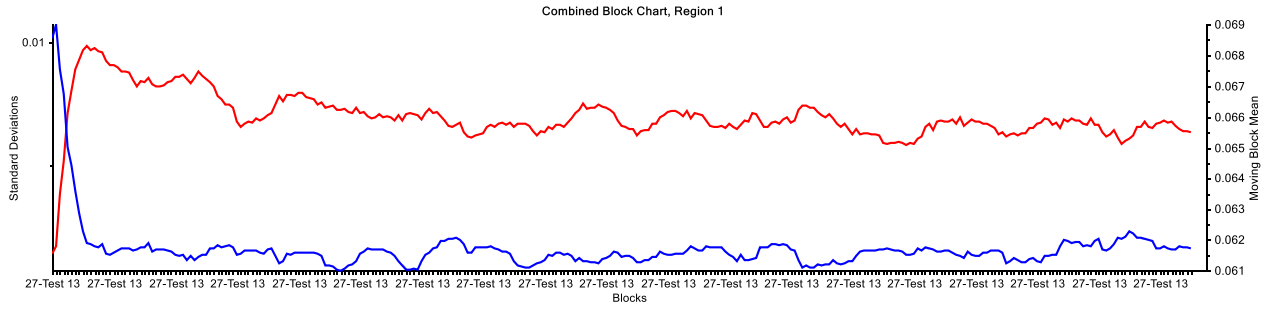
DOE 11



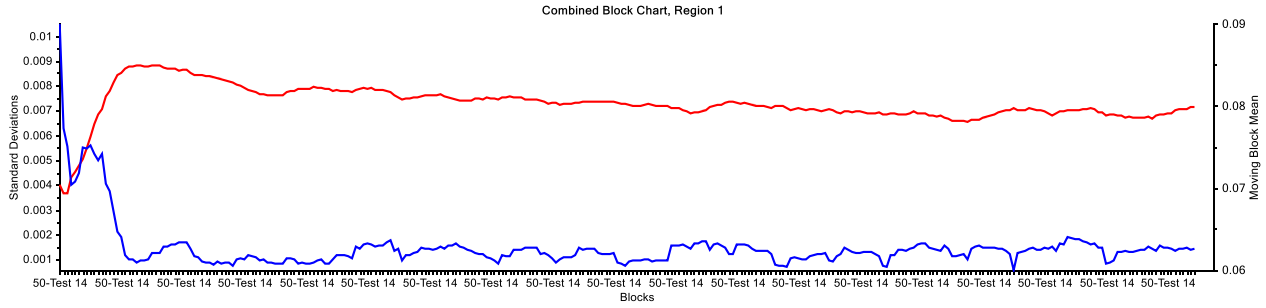
DOE 12



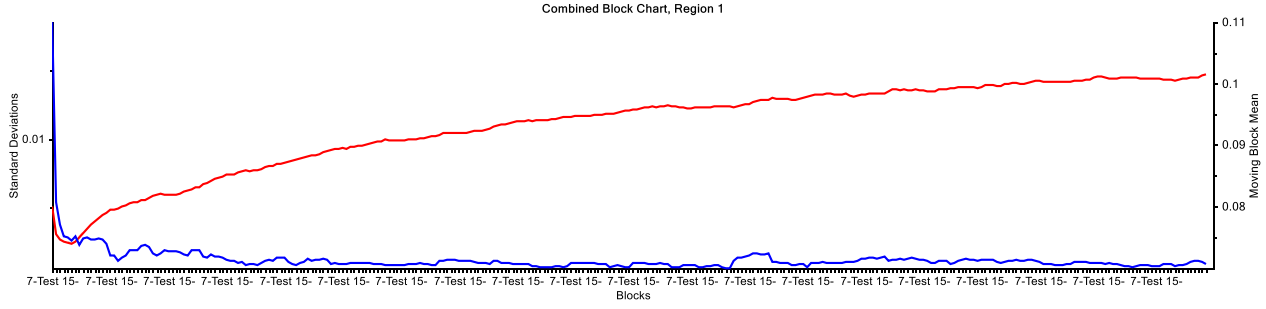
DOE 13



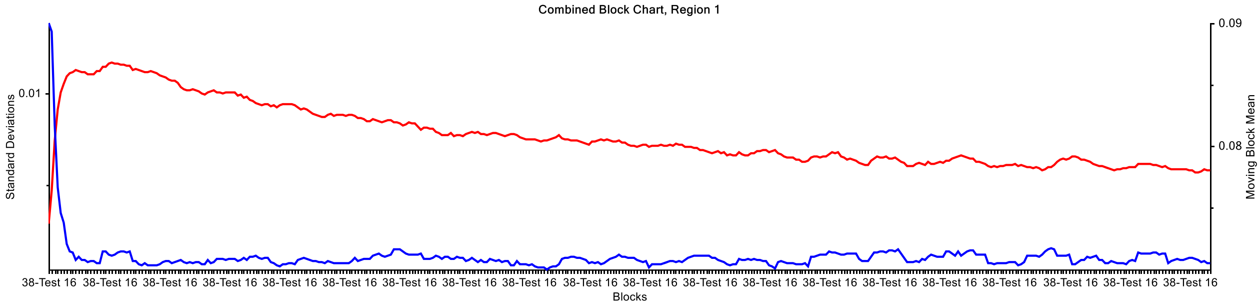
DOE 14



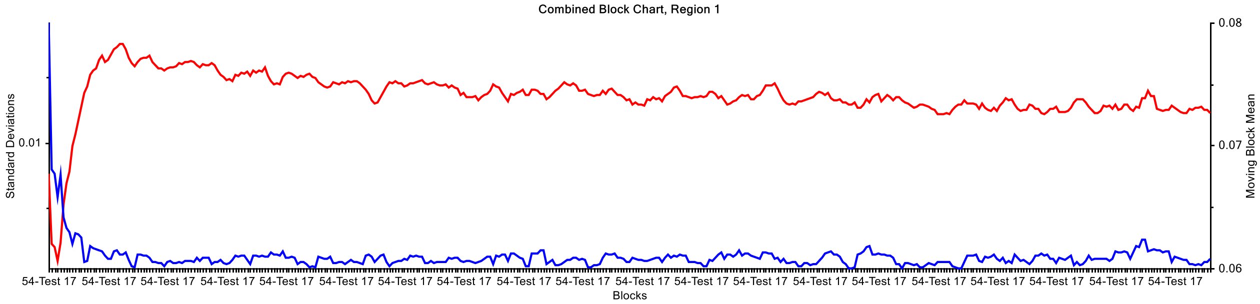
DOE 15



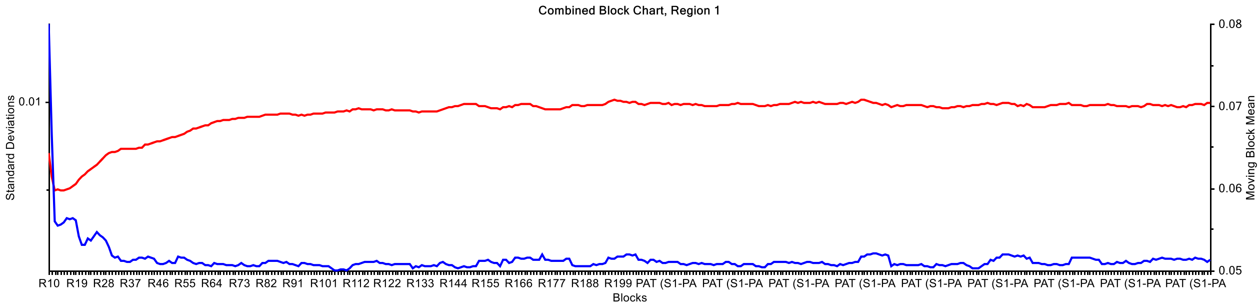
DOE 16



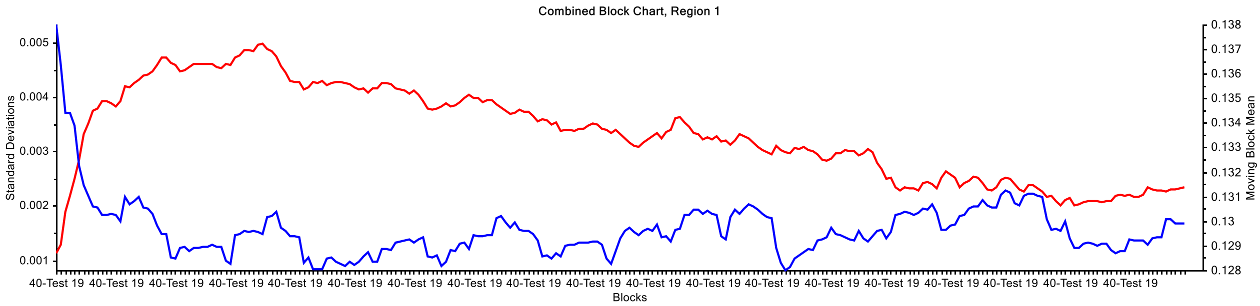
DOE 17



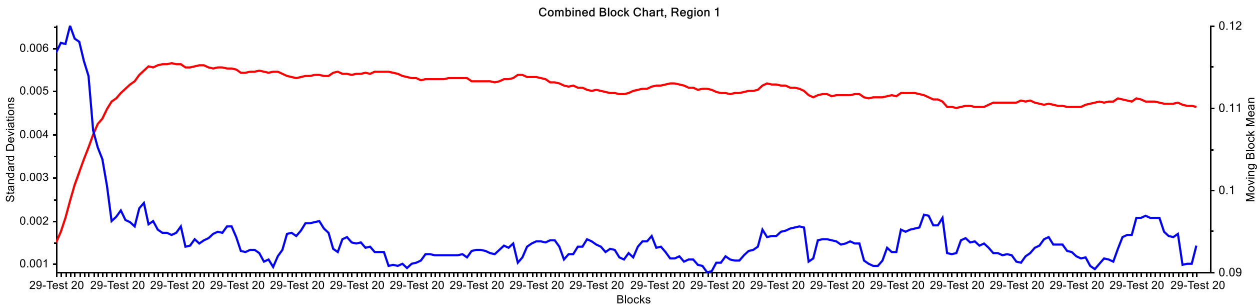
DOE 18



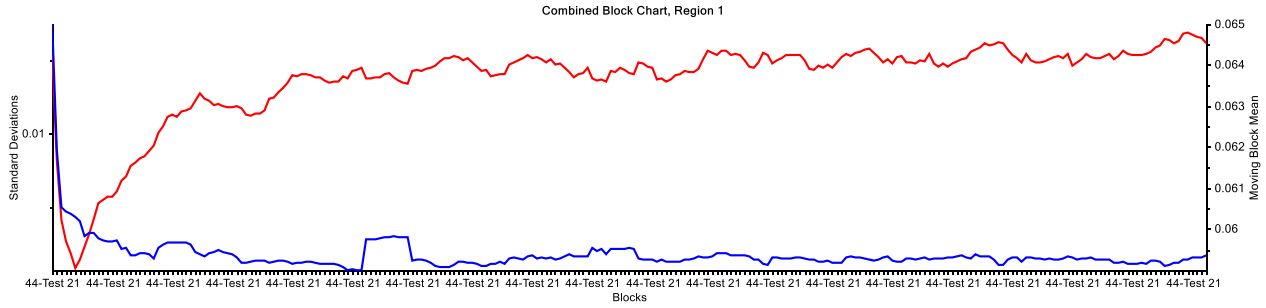
DOE 19



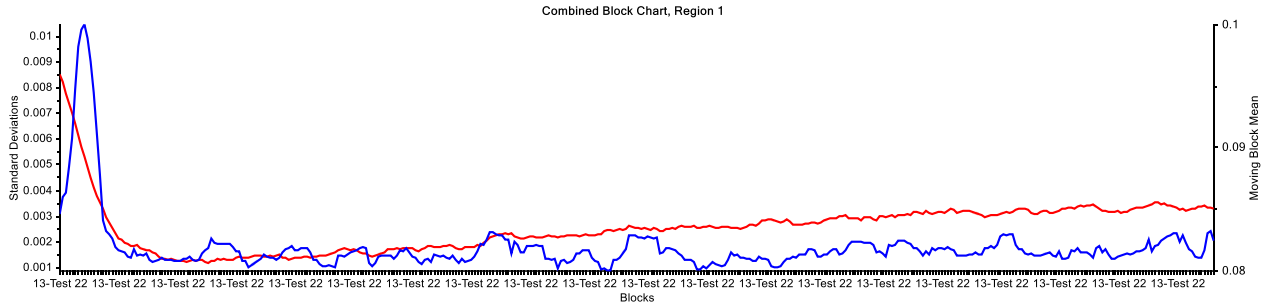
DOE 20



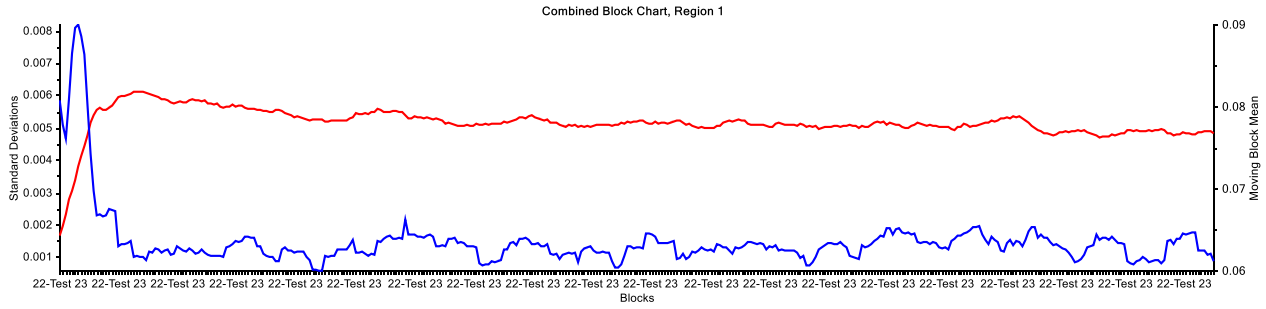
DOE 21



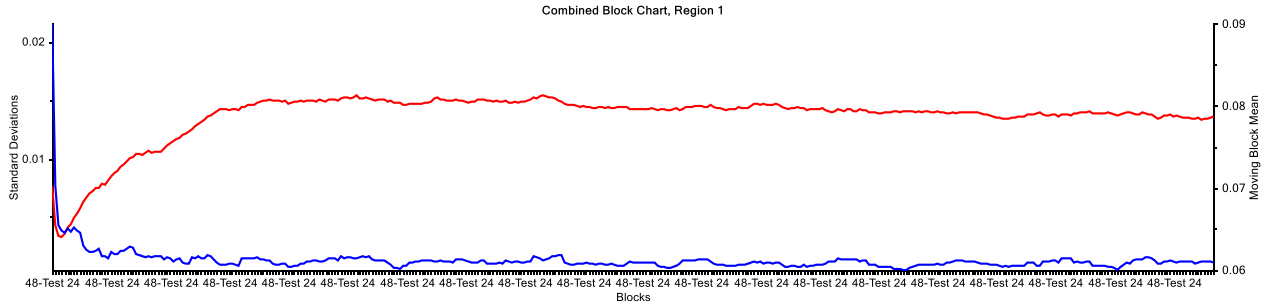
DOE 22



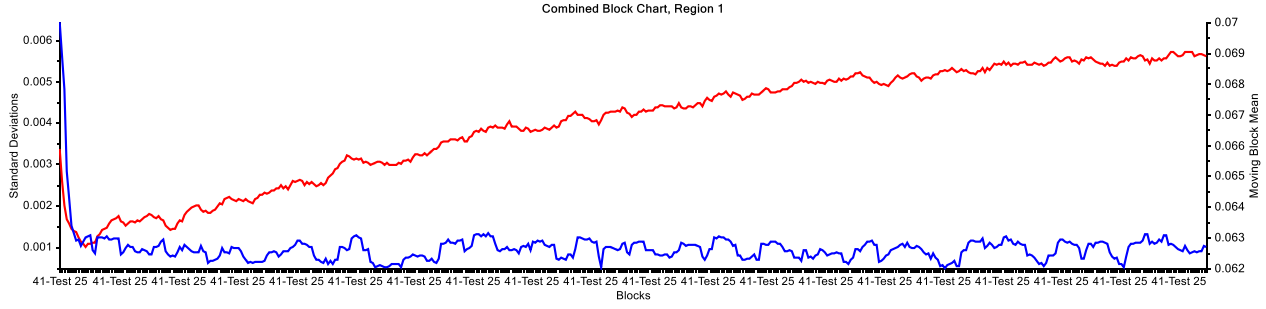
DOE 23



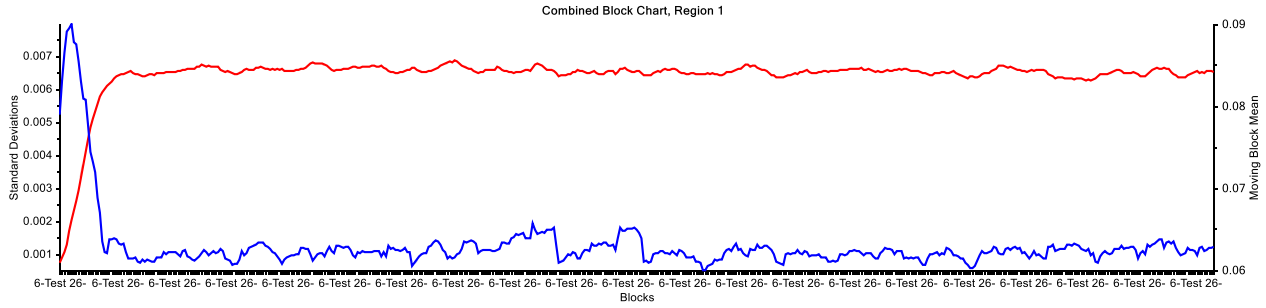
DOE 24



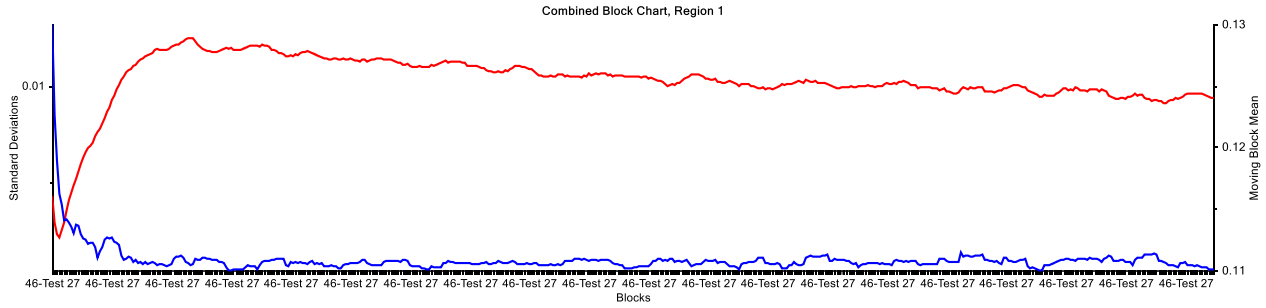
DOE 25



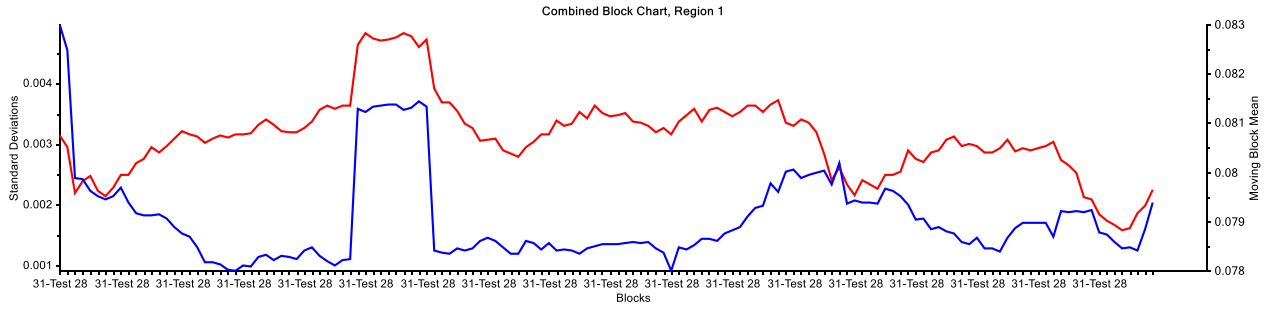
DOE 26



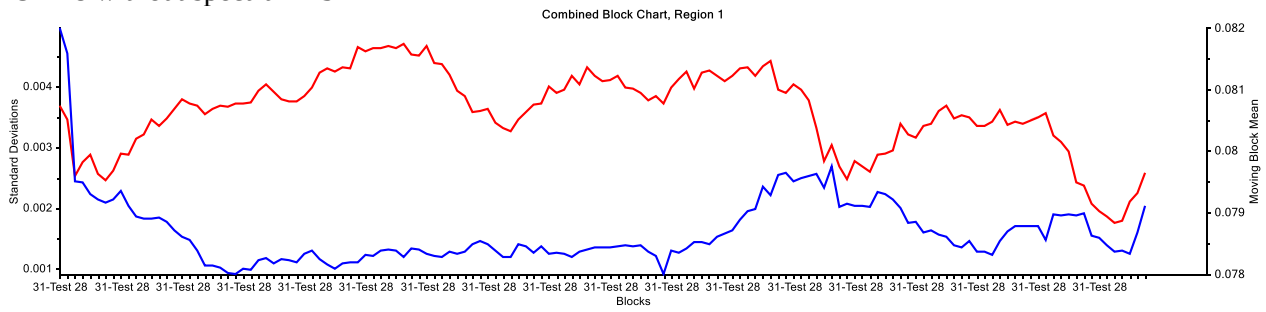
DOE 27



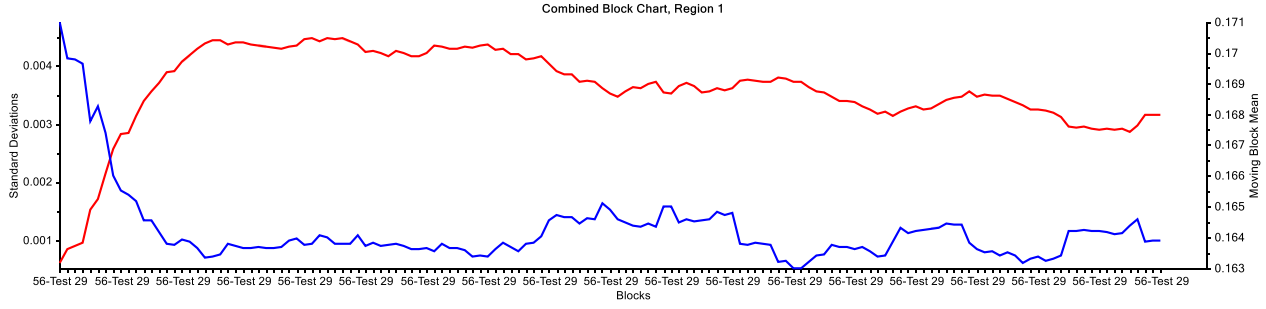
DOE 28



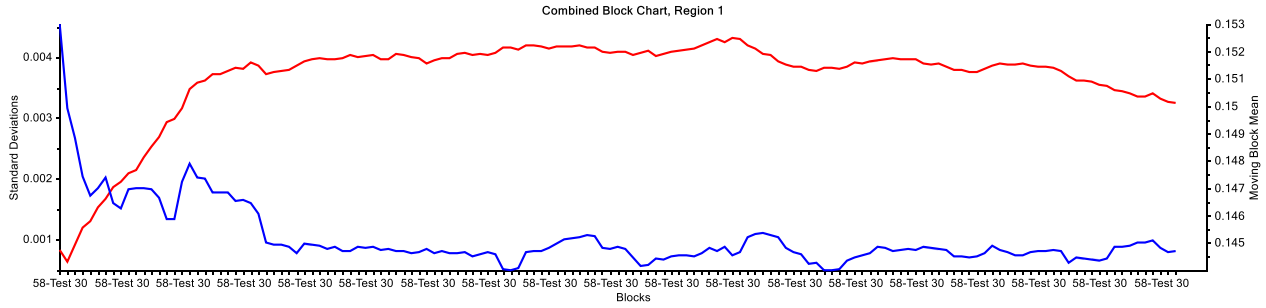
DOE 28 without spectrum 49



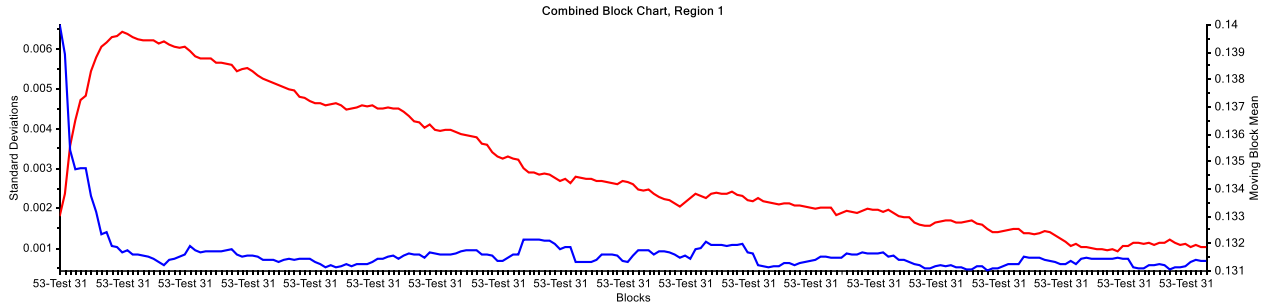
DOE 29



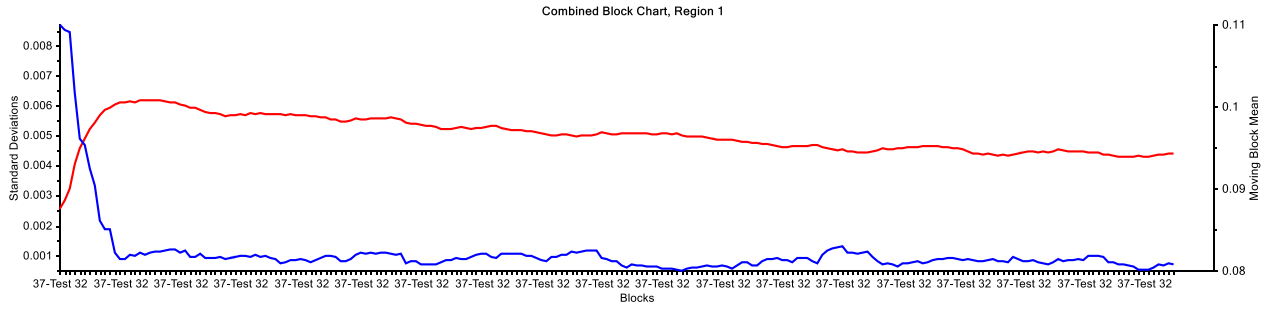
DOE 30



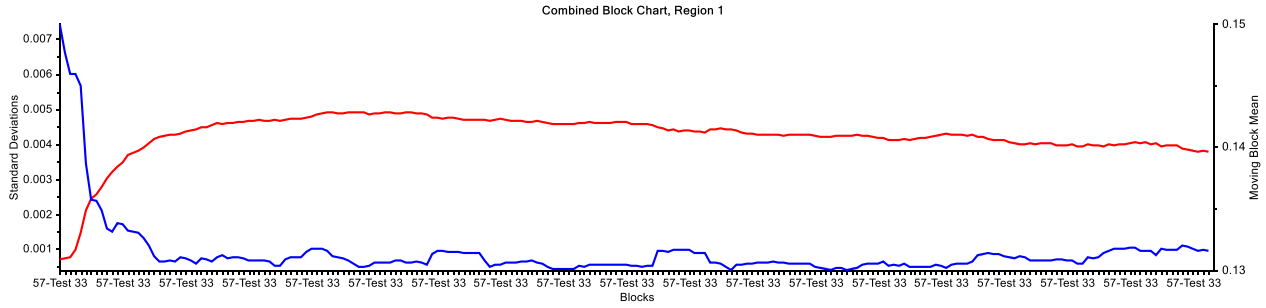
DOE 31



DOE 32

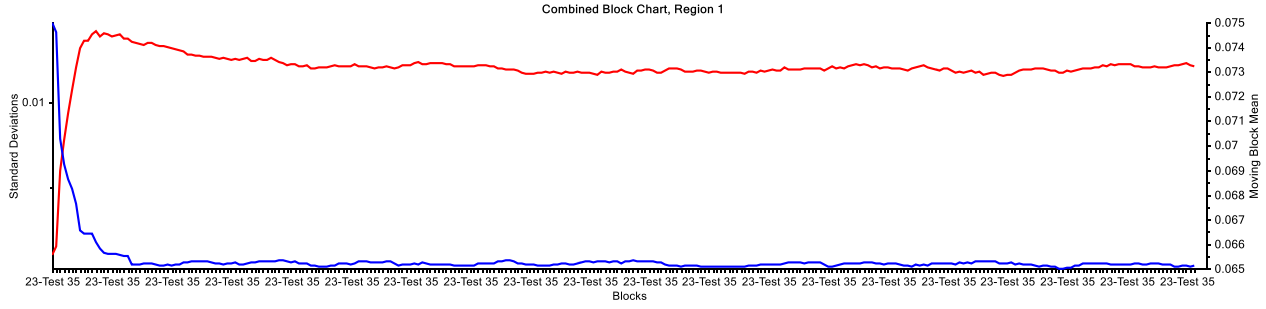


DOE 33

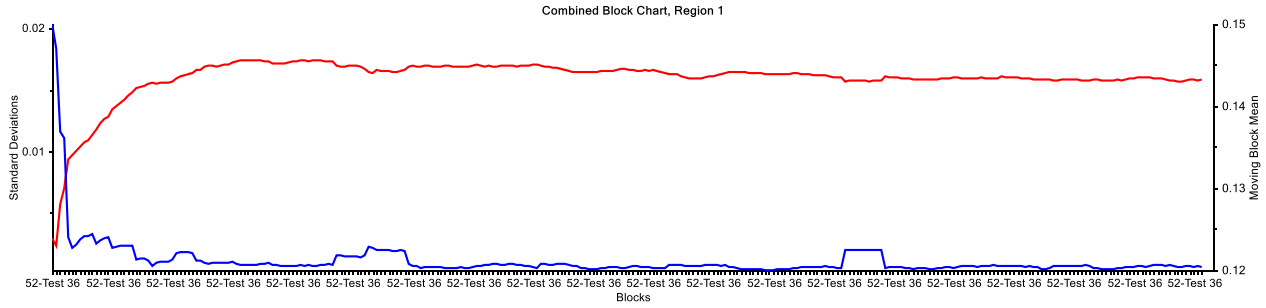


DOE 34 not present due to technical problems

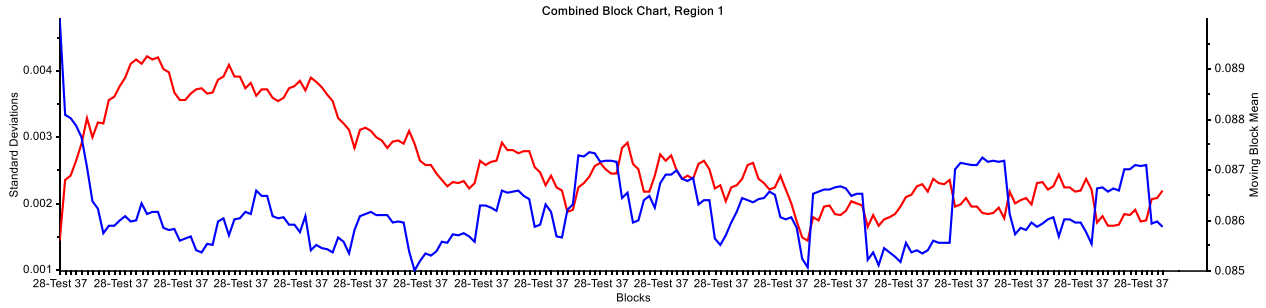
DOE 35



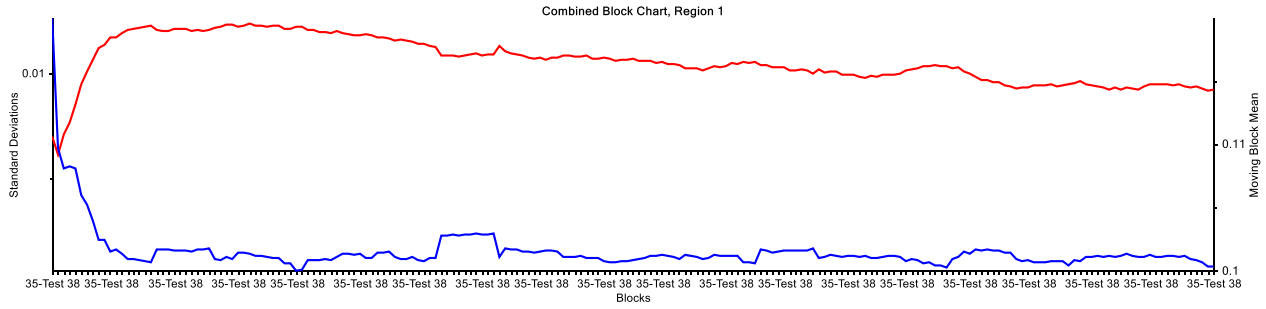
DOE 36



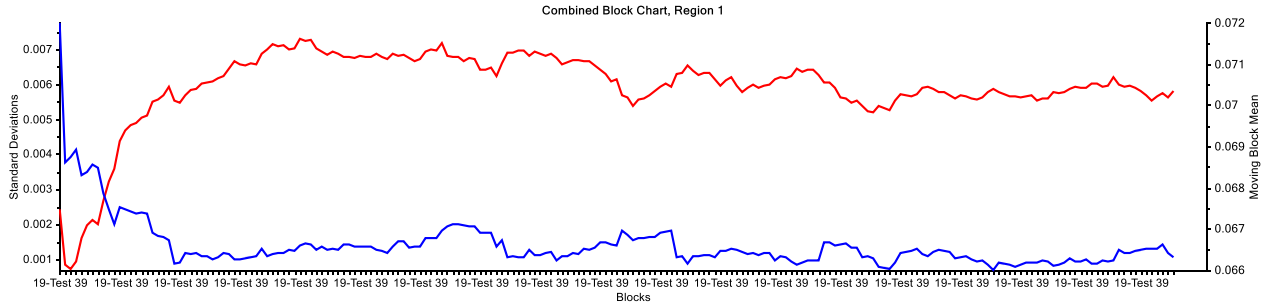
DOE 37



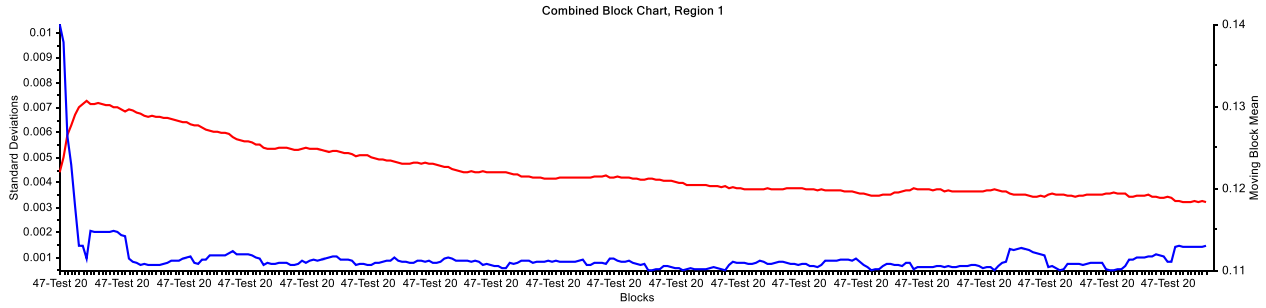
DOE 38



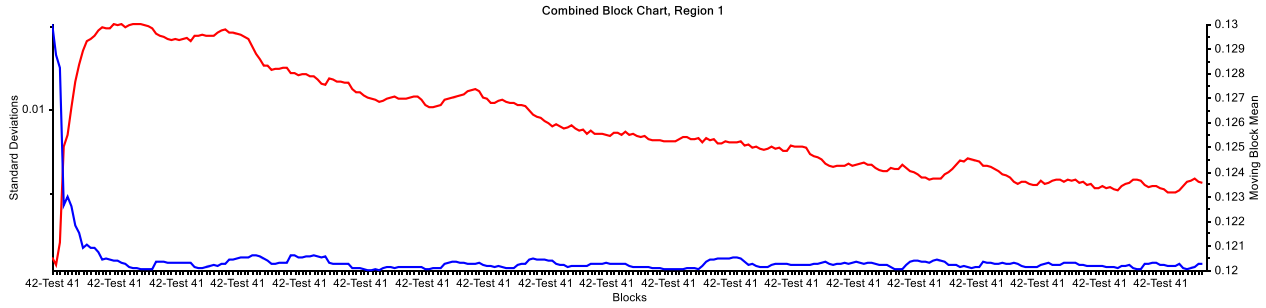
DOE 39



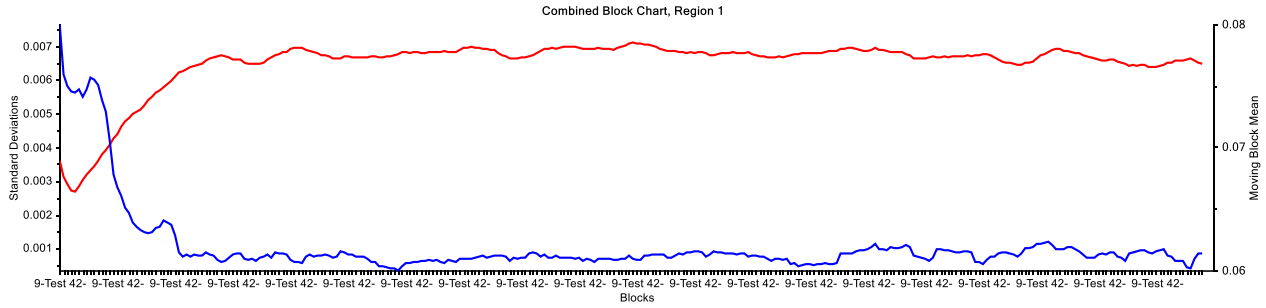
DOE 40



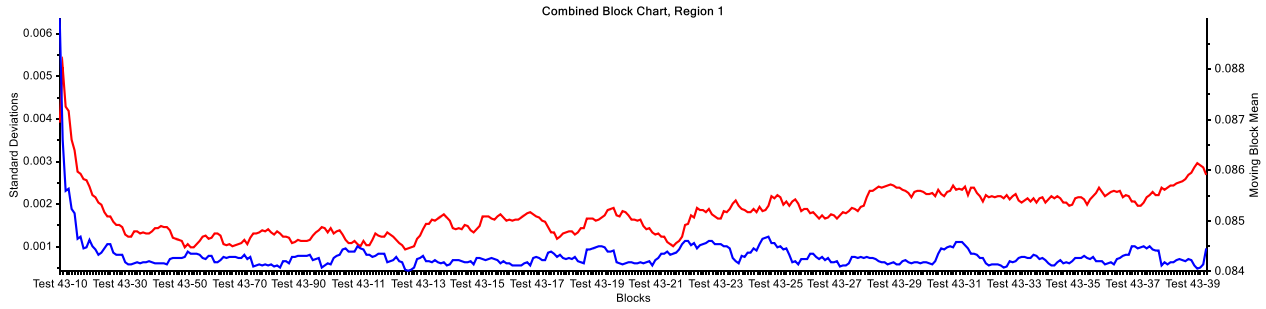
DOE 41



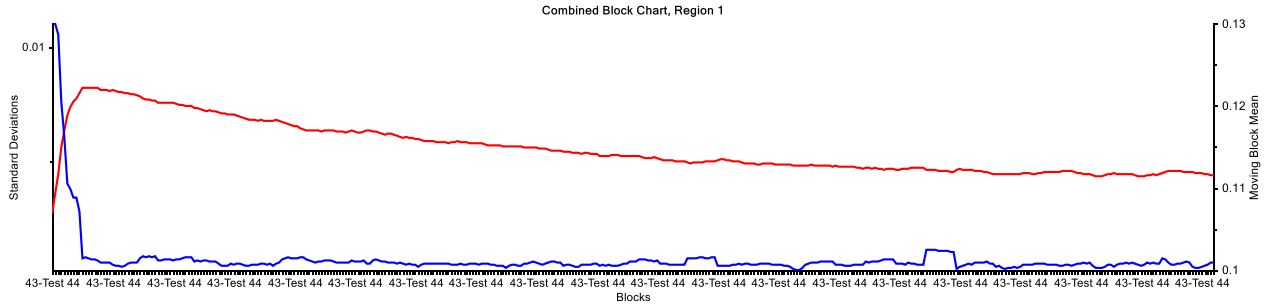
DOE 42



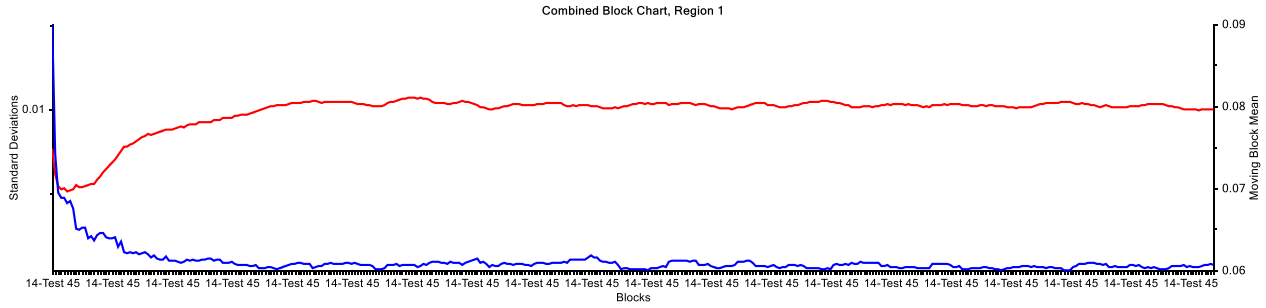
DOE 43



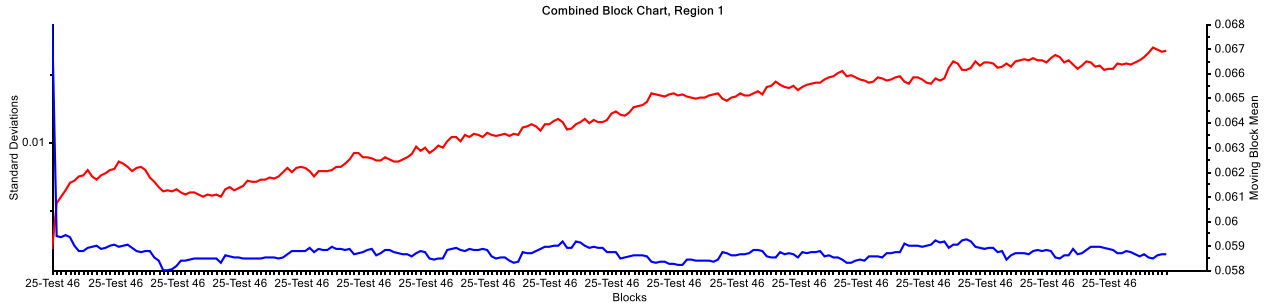
DOE 44



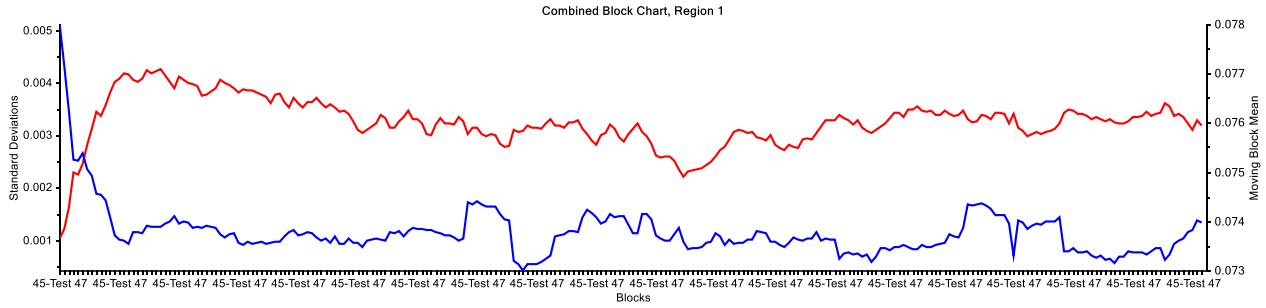
DOE 45



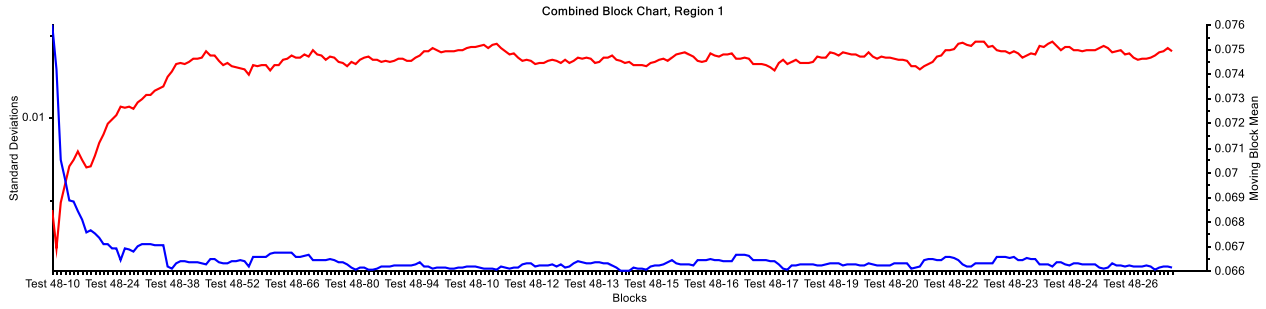
DOE 46



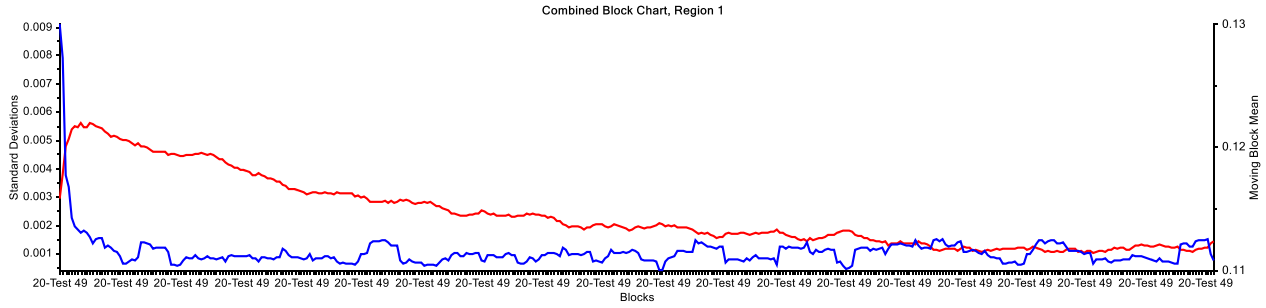
DOE 47



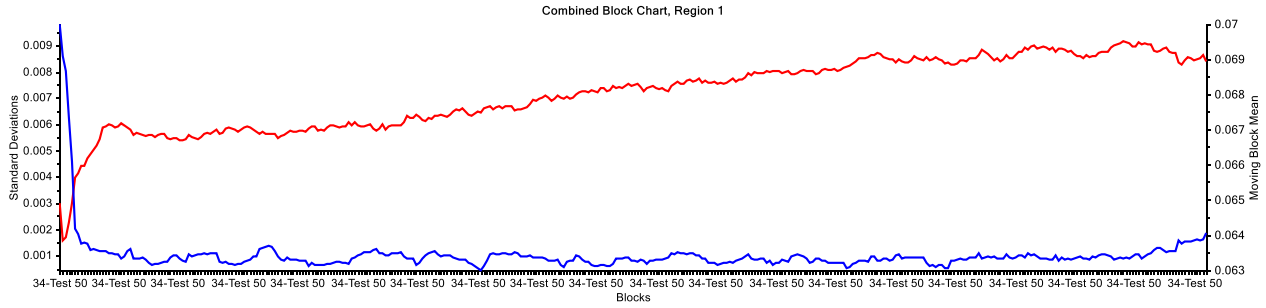
DOE 48



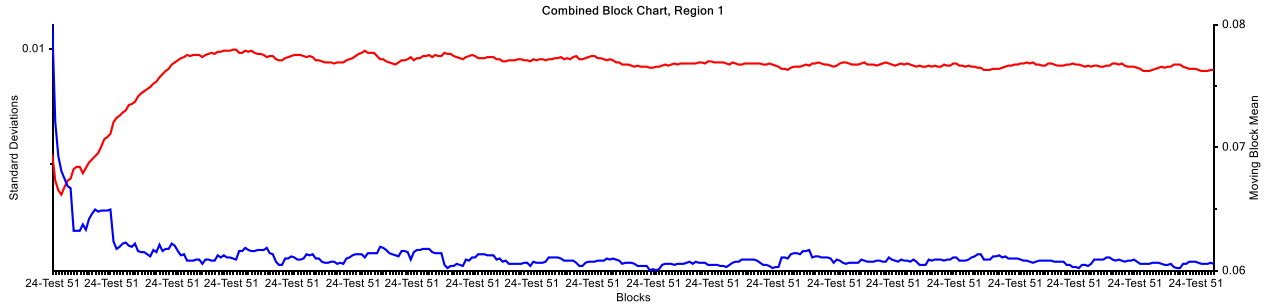
DOE 49



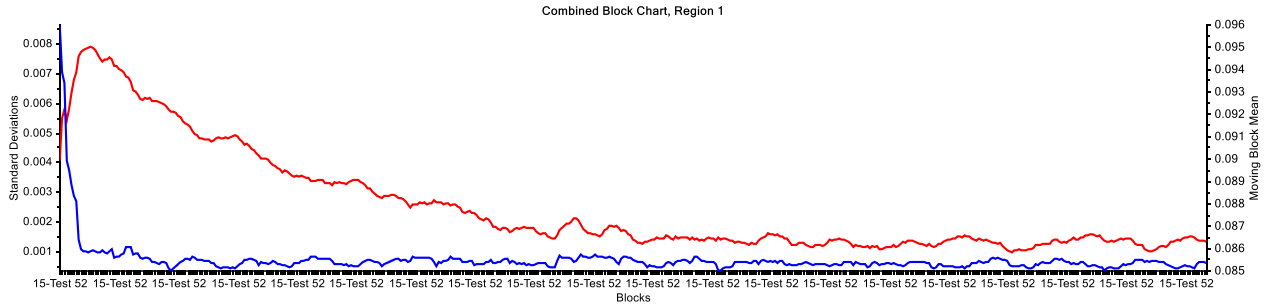
DOE 50



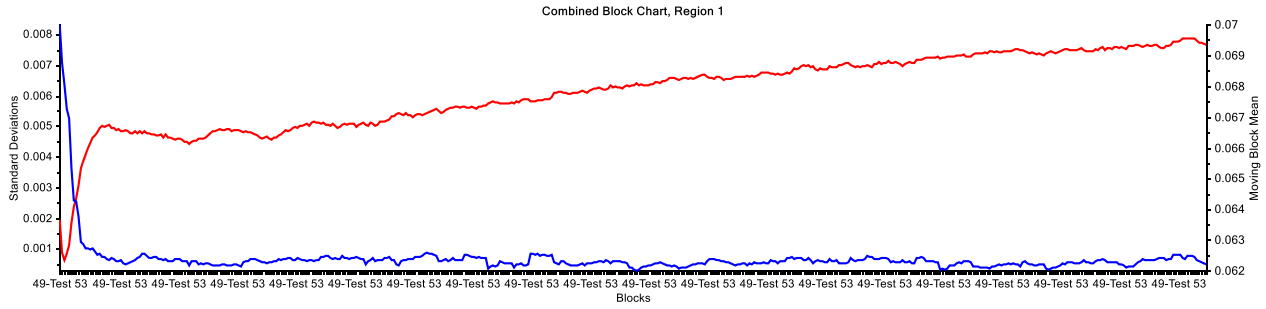
DOE 51



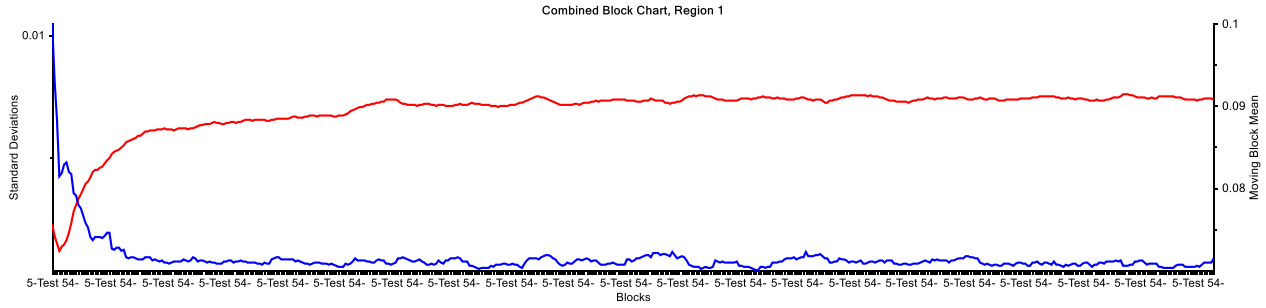
DOE 52



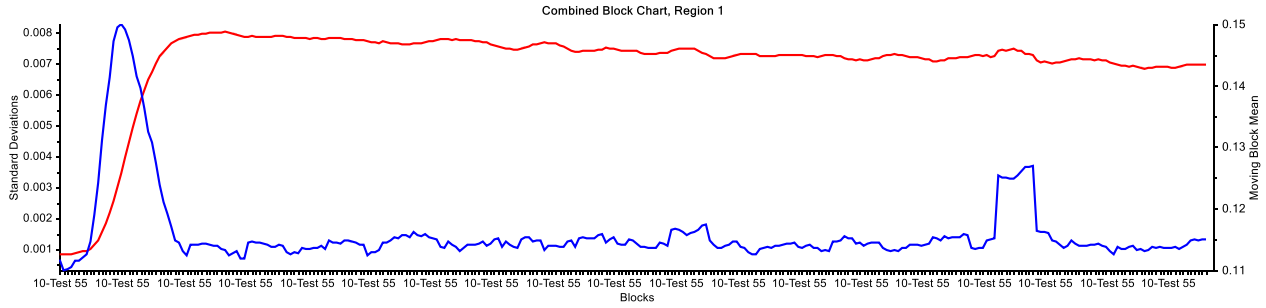
DOE 53



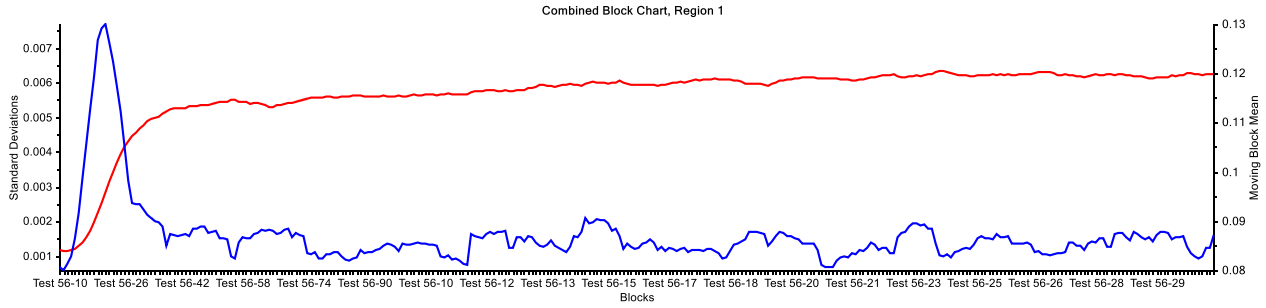
DOE 54



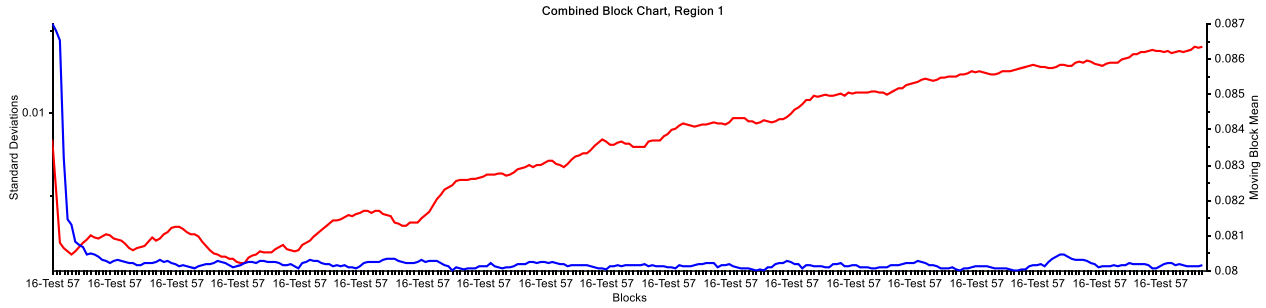
DOE 55



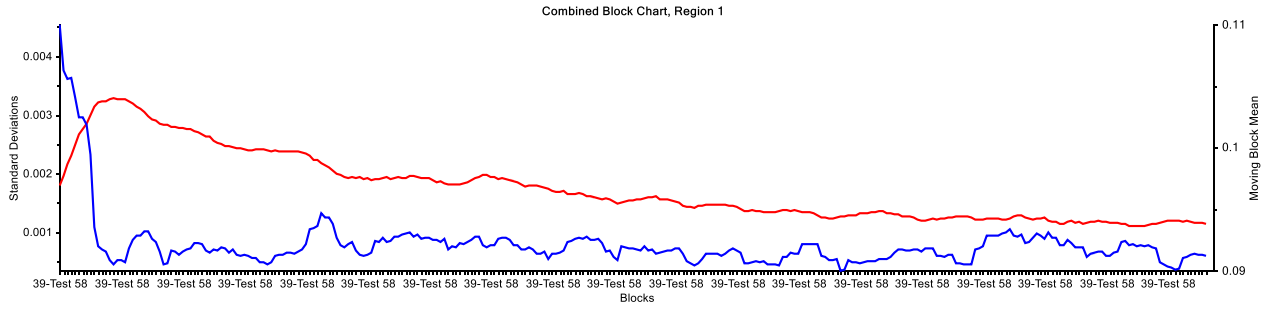
DOE 56



DOE 57



DOE 58



APPENDIX C: R CODE FOR NAS COMPUTATIONS

In this section, R codes for NAS computations are reported. The codes are generalized in order to make possible to use them for any dataset.

The first part shows the code for computing the original Lorber's NAS procedure. Hemmateenejad and Ferré variants are reported as supplementary code (at the bottom of the code) that can be implemented or not during computation. The lines were variants should be added are indicated by three hashtags (###). Then, the Bro procedure is reported.

```
## Definition of libraries and datasets, PLS computation, and creation of
## vectors to store the final results.
## This first part is in common for both Lorber and Bro procedures
library(MASS)
library(pls)
S<-#Define the signal (e.g. spectra) matrix
c.agg<-#Define the added concentrations vector
s.pure<-#Define the signal matrix of the pure analyte
s.pure.new<-apply(s.pure,2,mean)

### 1 - Add Hemmateenejad variant ###

n.sam<-nrow(S)
p.var<-ncol(S)
set.seed(1)
model.plsr<-plsr(c.agg~S,validation="CV",ncomp=n.sam-3)
rmse<-RMSEP(model.plsr)
plot(rmse,legendpos="topright")
n.comp<-model.plsr$ncomp
c0<-matrix(NA,1,n.comp)
R2<-matrix(NA,1,n.comp)
RMSE<-rmse$val[2,,1:n.comp+1]
```

Lorber NAS procedure

```
nas.i<-matrix(NA,n.sam,n.comp)
for(j in 1:n.comp){
T<-as.matrix(model.plsr$scores[,1:j])
P<-as.matrix(model.plsr$loadings[,1:j])
m<-apply(S,2,mean)
S.reb<-t(t(T%*%t(P))+m)
S.pinv<-ginv(S)
ck<-S.reb%*%S.pinv%*%c.agg
alfa<-as.numeric(1/(s.pure.new%*%S.pinv%*%ck))
K<-alfa*ck%*%s.pure.new
Sk<-S-K
Sk.pinv<-ginv(Sk)
Id<-diag(1,p.var,p.var)
H<-(Id-t(Sk)%*%t(Sk.pinv))
```

```
### 2 - Add Ferré variant ###

s.net.i<-matrix(NA,n.sam,p.var)
for(i in 1:n.sam) {
s.net.i[i,]<-H%*%S[i,]
nas.i[i,j]<-norm(as.matrix(r.net.i[i,]),"f")
}
retta<-lm(nas.i[,j]~c.agg)
R2[j]<-summary(retta)$adj.r.squared
c0[j]<-retta$coefficients[1]/retta$coefficients[2]
}
res.lorber<-t(rbind(c0,R2,RMSE))
colnames(res.lorber)<-c("c0,"R2","RMSE")
rownames(res.lorber)<-seq(1,n.comp)
res.lorber

### 1 - Hemmateenejad variant ###
s<-#Define signal matrix of zero-added samples (and remove it from S)
s.mean<-apply(s,2,mean)
S.sm<-t(t(S)-s.mean)
C<-S.sm%*%s.pure.new%*%solve(t(s.pure.new)%*%s.pure.new)
# Use C as c.add and S.sm as S
c.add<-C
S<-S.sm

### 2 - Ferré variant ###
W<-as.matrix(model.plsr$loading.weights[,1:j])
O<-t(P)%*%W
O.inv<-solve(O)
SWa<-W%*%O.inv
Q<-P%*%t(SWa)
# Use the product of Q and S as corrected signals
S<-Q%*%S
```

Bro NAS procedure

```
nas.i.bro<-matrix(NA,n.sam,n.comp)
for(j in 1:n.comp){
b<-as.matrix(model.plsr$coefficients[,,j])
H<-b%*%ginv(t(b)%*%b)%*%t(b)
s.net.i<-matrix(NA,n.sam,p.var)
for(i in 1:n.sam) {
s.net.i[i,]<-H%*%S[i,]
nas.i.bro[i,j]<-norm(as.matrix(r.net.i[i,]),"f")
}
retta<-lm(nas.i.bro[,j]~c.agg)
R2[j]<-summary(retta)$adj.r.squared
c0[j]<-retta$coefficients[1]/retta$coefficients[2]
}
```

```
res.bro<-t(rbind(c0,R2,RMSE))
colnames(res.bro)<-c("c0","R2","RMSE")
rownames(res.bro)<-seq(1,n.comp)
res.bro
```

The following lines report the computation for the standard deviation of the extrapolated value for Lorber and Bro methods. The previous matrices `res.lorber` and `res.bro` report the results obtained by the two computations (with relative R^2 and RMSE). From those matrices, it is necessary to choose the proper number of PLS-components.

```
n.comp<-#Define the proper number of PLS-components
c0<-matrix(NA,1,n.sam)
R2<-matrix(NA,1,n.sam)
```

Lorber method

```
nas.i<-matrix(NA,n.sam-1,n.sam)
for(j in 1:n.sam){
S.j<-as.matrix(S[-j,])
c.j<-as.matrix(c.agg[-j,])
n<-nrow(S.j)
model.plsr<-plsr(c.j~S.j,validation="CV",ncomp=n.sam-5)
T<-as.matrix(model.plsr$scores[,1:n.comp])
P<-as.matrix(model.plsr$loadings[,1:n.comp])
m<-apply(R,2,mean)
S.reb<-t(t(T%*%t(P))+m)
S.pinv<-ginv(S.reb)
c2<-S.reb%*%S.pinv%*%c.j
alfa<-as.numeric(1/(s.pure.new%*%S.pinv %*%c2))
K<-alfa*c2%*%s.pure.new
Sk<-S.reb-K
Sk.pinv<-ginv(Sk)
Id<-diag(1,p.var,p.var)
H<-(Id-t(Sk)%*%t(Sk.pinv))
s.net.i<-matrix(NA,n.sam,p.var)
for(i in 1:n) {
s.net.i[i,]<-H%*%S.j[i,]
nas.i[i,j]<-norm(as.matrix(r.net.i[i,]),"f")
}
retta<-lm(nas.i[,j]~c.j)
R2[j]<-summary(retta)$adj.r.squared
c0[j]<-retta$coefficients[1]/retta$coefficients[2]
}
res.sd.lorber<-t(rbind(c0,R2))
colnames(res.sd.lorber)<-c("c0","R2")
rownames(res.sd.lorber)<-seq(1,n.sam-1)
res.sd.lorber
sd.NAS.lorber<-sd(res.sd.lorber[,1])
sd.NAS.lorber
```

Bro Method

```
nas.i.bro<-matrix(NA,n.sam-1,n.sam)
for(j in 1:n.sam){
S.j<-as.matrix(S[-j,])
c.j<-as.matrix(c.agg[-j,])
n<-nrow(S.j)
model.plsr<-plsr(c.j~S.j,validation="LOO",ncomp=n.sam-3)
b<-as.matrix(model.plsr$coefficients[,,n.comp])
H<-b%*%ginv(t(b)%*%b)%*%t(b)
s.net.i<-matrix(NA,n.sam-1,p.var)
for(i in 1:n) {
s.net.i[i,]<-H%*%S.j[i,]
nas.i.bro[i,j]<-norm(as.matrix(r.net.i[i,]),"f")
}
retta<-lm(nas.i.bro[,j]~c.j)
R2[j]<-summary(retta)$adj.r.squared
c0[j]<-retta$coefficients[1]/retta$coefficients[2]
}
res.sd.bro<-t(rbind(c0,R2))
colnames(res.sd.bro)<-c("c0","R2")
rownames(res.sd.bro)<-seq(1,n.sam)
res.sd.bro
sd.NAS.bro<-sd(res.sd.bro[,1])
sd.NAS.bro
```

This last part of the code is for the computation of the further figures of merit: sensitivity (`Sn`) and limit of detection (`LoD`), the latter again divided between Lorber and Bro methods. It is again necessary to choose the proper number of PLS-components. Moreover, for LoD computations, it is necessary to define a matrix containing only the blank signal (`blank`).

Sensitivity

```
model.plsr<-plsr(c.agg~S,validation="CV",ncomp=n.sam-3)
B<-as.matrix(model.plsr$coefficients[,,n.comp])
Sn<-1/norm(as.matrix(B),"f")
Sn
```

LoD for Lorber method

```
blank.net<-matrix(NA,nrow(blank),ncol(blank))
T<-as.matrix(model.plsr$scores[,1:n.comp])
P<-as.matrix(model.plsr$loadings[,1:n.comp])
m<-apply(R,2,mean)
S.reb<-t(t(T%*%t(P))+m)
S.pinv<-ginv(S.reb)
c2<-S%*%S.pinv%*%c.agg
alfa<-as.numeric(1/(r.pure.new%*%S.pinv %*%c2))
K<-alfa*c2%*%r.pure.new
Sk<-S.reb-K
```

```
Sk.pinv<-ginv(Sk)
Id<-diag(1,p.var,p.var)
H<-(Id-t(Sk)%*%t(Sk.pinv))
for(i in 1:nrow(blank)){
blank.net[i,]<-H%*%blank[i,]
}
blank.net.mean<-apply(blank.net,2,mean)
LoD.lorber<-3*norm(as.matrix(blank.net.mean),"f")/Sn
LoD.lorber
```

LoD for Bro method

```
H<-B%*%ginv(t(B)%*%B)%*%t(B)
blank.net<-matrix(NA,nrow(blank),ncol(blank))
for(i in 1:nrow(blank)){
blank.net[i,]<-H%*%blank[i,]
}
blank.net.mean<-apply(blank.net,2,mean)
LoD.bro<-3*norm(as.matrix(blank.net.mean),"f")/Sn
LoD.bro
```